Explanation and Connectionist Models

Catherine Stinson
Rotman Institute of Philosophy

Introduction

Connectionist models are widely used in the cognitive sciences, and well beyond. This is so despite the fact that some critics have charged that we can't learn about cognition using connectionist models (Fodor and Pylyshyn, 1988). Although researchers who use connectionist models have offered a number of defenses of their methods (Smolensky, 1988; McClelland, 1988), and there is growing empirical evidence suggesting that these models have been successful in advancing cognitive science, there is no consensus on *how* they work. This chapter explores the epistemic roles played by connectionist models of cognition, and offers a formal analysis of how connectionist models explain.

The question of what sorts of explanations connectionist models offer has not received much (positive) attention. Understanding how these explanations work, however, is essential in evaluating their worth, and answering questions such as, How convincing is a given model? What makes a connectionist model successful? What kinds of errors should we look out for?

For the sake of comparison, I begin with a brief look at how other types of computational models explain. Classical AI programs explain using abductive reasoning, or inference to the best explanation; they begin with the phenomena to be explained, and devise rules that can produce the right outcome. Including too much implementation detail is thought to hinder the search for a general solution. Detailed brain simulations explain using deductive reasoning, or some approximation to it; they begin with the raw materials of the system and first principles they obey, and calculate the expected outcome. Here, inaccuracies or omissions of detail can lead to incorrect predictions. Connectionist modeling seems to combine the two methods; modelers take constraints from both the psychological phenomena to be explained, and from the neuroanatomical and neurophysiological systems that give rise to those phenomena. The challenge is to understand how these two very different methods can be a combined into a successful strategy, rather than a failure on both counts. I'll focus on the problem of why using neural constraints should be a good strategy, even if those neural constraints aren't correct in their details.

To answer this question I look at several examples of connectionist models of cognition, observing what sorts of constraints are used in their design, and how their results are evaluated. The marks of successful connectionist models include using structures roughly analogous to neural structures, accurately simulating observed behavioral data, breaking down when damaged in patterns analogous to neurological cases, and offering novel, empirically verifiable predictions.

I argue that the point of implementing networks roughly analogous to neural structures is to discover and explore the generic mechanisms at work in the brain, not to

deduce the precise activities of specific structures. As we will see, this method depends on the logic of tendencies: drawing inductive inferences from like causes to like effects. This can be combined with neuropsychological evidence, which is evaluated using graph theoretical reasoning.

How Computational Models Explain

Computational models are especially important in cases where experimenting directly on the target system is not practicable, or the system is very complex. Opening a human skull and poking around is very invasive, so this kind of intervention can only be done in exceptional cases like during treatment of Parkinson's disease or epilepsy (Engel et al., 2005). In these rare cases, single cell recordings and electrical stimulation interventions can sometimes be done on awake, behaving patients, providing important validation of models arrived at by other means. These studies are necessarily of short duration, and usually are restricted to particular brain regions, making them quite limited in terms of what can be investigated. Furthermore, these recordings are made from brains affected by pathology, and usually in patients taking medication (Mukamel and Fried, 2012). Care must be taken when drawing inferences from studies of atypical brains to neurotypical populations.

Several non-invasive means for indirect measurement from and intervention on human brains are also available. Technologies like Transcranial Magnetic Stimulation, Positron Emission Tomography, functional Magnetic Resonance Imaging, and Electroencephalography all provide valuable information about human brain functioning, but all of these methods face practical limitations like noise and limited spatial or temporal resolution.

Human experiments can be supplemented by experimenting on model species like sea slugs, mice, or macaque monkeys, but these animal models also face limitations. Most animals can't perform complex laboratory tasks, and few if any can give verbal feedback, making it very difficult to investigate higher cognitive processes. In addition, it cannot be taken for granted that the brains of non-human animals process information in the same way that human brains do.

Human brains are also extremely complex, consisting of on the order of 100 billion neurons, each with thousands of synaptic connections on average, not to mention the elaborate structures within each neuron, the chemical soup surrounding them, and all the other cells in the brain whose functions are only beginning to be understood. Computational models have the capacity to quickly analyze how complex systems evolve over time, and/or in a variety of situations, making them invaluable for investigating human brain functioning.

Explanation in Classical AI

Other chapters of this volume are dedicated to the history and explanatory uses of classical AI, but for our purposes here, a few brief notes will be helpful. Consider first the birthplace of classical AI: McCarthy et al.'s (1955) Dartmouth Proposal. In this proposal it is conjectured that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et

al., 1955). The authors optimistically suggest that one summer would be sufficient to make significant progress on the problem of machine intelligence. The idea is that we can come to understand intelligence by precisely constructing a machine that reproduces the phenomenon.

More details about how this method is meant to work are found in Newell & Simon's pioneering 1961 paper. Newell & Simon begin by analyzing behavioral phenomena into *protocols*: transcripts of subjects speaking aloud about their thought processes while they solve a problem. The AI project is then to "construct a theory of the processes causing the subject's behavior as he [sic] works on the problem, and to test the theory's explanation by comparing the behavior it predicts with the actual behavior of the subject" (Newell and Simon, 1961, 2012).

It is clear from the section of the text titled "Nonnumerical Computer Program as a Theory" that Newell & Simon intend for their programs to be scientific theories that explain the behavioural phenomena. At the time, the "Received View" of scientific theories (see Winther, 2016), supposed that theories are sets of statements cast in predicate logic, and the prevailing deductive-nomological account of scientific explanation (Hempel & Oppenheim, 1948) supposed that empirical observations, such as the subjects' problem solving behaviour, could be explained by logically deducing observation statements from statements of laws and antecedent conditions. For Newell & Simon, the antecedent conditions would correspond to the input problem, the theory would be the sequence of symbolic expressions contained in the program, and the logically deduced outputs of the program would be the observation statements. At the heart of their approach is the postulate later dubbed the "physical symbol system hypothesis" (Newell and Simon, 1976), that the processes going on inside the subject are, like their program, operations on symbols.

Established scientific theories can be used to deduce predictions, but Newell & Simon were still at the theory-building stage. The defense of their physical symbol system hypothesis "lies in its power to explain the behavior" (Newell and Simon, 1961, 2012). In other words, Newell & Simon were judging the success of their AI program as a theory of probem solving by comparing its output to human behavior. A program that counts as a good theory should produce output that matches the known behavioral data. The fact that the program gives rise to the same output is a reason for believing that the cognitive process might be the same as the program,. This is an abductive inference, or an inference to the best explanation. The inference has the form,

$$\frac{T \rightarrow O \quad O}{T} \qquad (1)$$

where T stands for the theory/program, and O for the observed behavior. If the program produces the right output, it is a candidate explanation of the observed behavior, and in the absence of any other adequate explanation, which was plausibly the case in 1961, that program is by default the best explanation.

Newell and Simon's defense of the physical symbol system hypothesis—the assumption that any explanation of cognition should take the form of symbol manipulations—is that making this assumption led to a string of successful explanations of cognitive tasks. Phenomena that previously could not be explained suddenly became

tractable with the help of that one trick. As they say, "The processes of thinking can no longer be regarded as completely mysterious" (Newell and Simon, 1961, 2016). One reasonable criterion to use when deciding between candidate explanations is unity; a single assumption that helps to explain many phenomena is preferable to multiple assumptions, all else being equal.

A more contemporary statement of this strategy can be found in Coltheart et al.'s defense of classical AI models of reading. They say, "the adequacy of the theory can be rigorously assessed by simulation. Are all the effects observed in the behavior of people when they are carrying out the cognitive activity in question also seen in the behavior of the program...?" and "if there is no other theory in the field that has been demonstrated through computational modeling to be both complete and sufficient, resting on laurels is a reasonable thing to do until the emergence of such a competitor" (Coltheart et al., 2001, 204). This clearly describes inference to the best explanation.

Explanation in Realistic Brain Simulations

"Simulation" has been used in the previous examples in a way that is common in discussions of cognitive models, but notably different than its meaning in other fields. In physics and climate science, what I'll call a *true simulation* starts from a fundamental theory, usually consisting of differential equations that describe the behavior of elementary entities like particles. A true simulation then churns through calculations based on these equations to generate a description of the state of the system at various time points (Humphreys, 1990). Often the purpose is to predict outcomes like weather forecasts, cosmological events, or the properties of a newly synthesized material. In true simulations, the inference is deductive, and has the form,

$$\frac{T \rightarrow O \quad T}{O} \qquad (2)$$

where T stands for the fundamental theory as instantiated in the program, and O for the observed outcome. In practice, true simulations are not perfect deductive tools; the starting point may not correspond exactly to the state of the world of interest, and numerical approximations are generally needed to solve the fundamental equations.

Some approaches to computational modeling in cognitive science aspire to model the brain from the bottom up, starting by modeling brain anatomy and/or physiology in detail, like true simulations. The goal of the Blue Brain Project is "to simulate the brains of mammals with a high level of biological accuracy and, ultimately, to study the steps involved in the emergence of biological intelligence." (Markram, 2006). Another large-scale, anatomically detailed simulation by Izhikevich and Edelman incorporates "multiple cortical regions, corticocortical connections, and synaptic plasticity" (Izhikevich and Edelman, 2008, 3593). Eliasmith's Spaun focuses on "explaining how complex brain activity generates complex behavior" with a simulation that generates "behaviorally relevant functions" (Eliasmith et al., 2012). In these projects, getting the anatomical and physiological details correct is a high priority.

Explanation in Connectionist Models

Connectionist models of cognition, in particular the Parallel Distributed Processing (PDP) approach, likewise incorporate details of neural anatomy and physiology. As 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). Although this statement suggests an intention to model the physiology of the brain, the physiological similarities between PDP models and real brains are quite loose, unlike true simulations, which try to get the details exactly right.

Like classical AI, connectionist models have the primary aim of reproducing cognitive phenomena. It is not immediately obvious how classical AI's top-down methods can be combined with brain simulation's bottom-up methods. Of particular concern is how incorporating neural constraints is meant to help when these constraints are taken only very loosely If we view connectionist modeling through the lens of classical AI and deductive-nomological explanation, it might look like the inference structure is only a slight variation on Inference 2, such that:

$$\frac{T^* \rightarrow O \quad O \quad t_1, \ldots, t_n \in T^*}{T} \qquad (3)$$

where T^* is a model that loosely approximates T, and $t_1, \ldots, t_n$ are statements from T (describing physiological constraints on brains) that are included in the model T^*.

This assumes that the purpose of adding physiological constraints on brains is to increase the strength of the inference to T. However, if n is small relative to the number of facts in T, the benefit of adding them to the premises would be negligible, which would undermine connectionism's claims about the importance of physiological plausibility. Another problem is that the model T^* only loosely approximates T. In order to make an inference to the best explanation, T would need to be established as a candidate explanation for O, but here it is T^* that implies O. If this were an accurate interpretation of connectionist methodology, these would be serious problems, however, inference 3 gets connectionist methodology very wrong.

During the period between 1961 and 1986, the 'Received View' of scientific theories and the corresponding deductive-nomological account of scientific explanation were largely scrapped (see Woodward, 2017). I don't think connectionists have the aim of constructing theories at all, but rather models (see Morgan and Morrison, 1999; Winsberg, 2001; Bailer-Jones, 2009 for accounts of scientific modeling). With this in mind,, we shift from interpreting T as a theory that entails all the facts about the target system, to interpreting T as the target system itself (or in propositional terms, we can think of this as the set of all facts that are true of the target system).

In the next section I analyze several examples of connectionist modeling work. I argue that connectionist models are meant to explore the mechanisms operative in the target system. On this account, explaining cognitive phenomena using connectionist models involves reasoning about mechanisms, which operates using the logic of tendencies. In Stinson (2018), I connect this argument about how connectionist models explain to the philosophical literature on idealization in modeling, and explore examples from other scientific fields where abstract, idealized models likewise offer explanatory advantages over highly detailed models.

Connectionist Explanation Examples

In this section I look in some detail at examples of connectionist models from several areas of cognitive science research. I begin by looking at the models described in De Pisapia et al's (2008) review of connectionist models of attention and cognitive control. By looking closely at how the studies are described, I discern four criteria by which the success of these models is judged. Consideration of models from several other areas of cognition confirm that connectionist models of cognition typically follow this pattern.

First, the models reviewed in De Pisapia et al. (2008) all try to capture known neurophysiological characteristics of the brain. For instance, many of the models implement feature maps corresponding to the representations computed in brain areas V1, PP and IT. Some of the models also capture more specific details about hypercolumns, patterns of inhibitory connections, neuronal dynamics, etc. This reflects the belief that "models which make strong attempts to incorporate as many core principles of neural information processing and computation as possible are the ones most likely to explain empirical data regarding attentional phenomena across the widest-range of explanatory levels" (De Pisapia et al., 2008, 423). But importantly, the neural plausibility is always limited to general or core features, not every detail.

The second criterion is that the models are expected to simulate or replicate known empirical results from psychology. For instance, "Simulations using biased competition model[s] were found to be successful in accounting for a number of empirical results in visual search" (De Pisapia et al., 2008, 431). The competing feed-forward models are evaluated in the same terms: "these models have been effective in capturing the known neurobiology of low-level visual processing, while at the same time simulating findings from the empirical visual search and natural scene viewing" (De Pisapia et al., 2008, 432).

Third, the models are judged based on their ability to explain clinical phenomena, like the cognitive effects of brain lesions and other neurological conditions. The models need both to "agree with behavioral results coming from the basic experimental paradigms and with the data from brain-damaged patients suffering from attentional impairments... the true strength of these models lies in their ability to model the qualitative pattern of impairments associated with neuropsychologically-based attentional disorders, such as the spatial neglect syndrome" (De Pisapia et al., 2008, 432).

Fourth, many of the models generate predictions about what the result of novel experimental scenarios should be, which can later be verified in the lab. One model "provided novel predictions about how patients with object-based neglect might perceive

objects when they are joined with cross-links or brought towards each other" (De Pisapia et al., 2008, 431). In another case, "reaction time slopes... obtained by model simulations were successful in predicting subsequent psychophysical investigations" (De Pisapia et al., 2008, 431).

These four criteria for successful connectionist modeling of cognition are also apparent in many other studies. For example, O'Reilly et al.'s model of working memory "is biologically plausible. Indeed, the general functions of each of its components were motivated by a large base of literature spanning multiple levels of analysis, including cellular, systems, and psychological data" (O'Reilly and Frank, 2006, 312). In addition to simulating "powerful levels of computational learning performance" (O'Reilly and Frank, 2006, 284), it also models clinical results by testing "the implications of striatal dopamine dysfunction in producing cognitive deficits in conditions such as Parkinsons disease and ADHD" (O'Reilly and Frank, 2006, 313). McClelland et al.'s model of memory likewise tries to be "broadly consistent with the neuropsychological evidence, as well as aspects of the underlying anatomy and physiology (McClelland et al., 1995, 419). Suri and Schultz (2001) model the anatomy of the basal ganglia, including only pathways that exist in the brain and through which feedback is thought to actually travel; and Billings et al. (2014) design the units in their "anatomically constrained model" to match properties like the diameters and densities of granule cells and mossy fibers in the cerebellum.

Sejnowski et al. (1988), focusing on vision, describe connectionist models as "simplifying brain models" which "abstract from the complexity of individual neurons and the patterns of connectivity in exchange for analytical tractability" (Sejnowski et al., 1988, 1301). One of the advantages they list of connectionist modeling over experimental techniques is that "New phenomena may be discovered by comparing the predictions of simulation to experimental results" and they note that "new experiments can be designed based on these predictions" (Sejnowski et al., 1988, 1300). The models they describe are not only consistent with previous experimental measures, they also make "interesting predictions for ... responses to visual stimuli" (Sejnowski et al., 1988, 1303).

Plaut et al. (1996) likewise try to simulate both experimental results, and the patterns of breakdown in clinical cases in their model of reading, as well as generating testable predictions. Some of the empirical results that the model replicates are that high-frequency and consistent words are named faster than low-frequency and inconsistent words, and that these two effects interact (Plaut et al., 1996, 7-8). In addition, "damaging the model by removing units or connections results in a pattern of errors that is somewhat similar to that of brain-injured patients with one form of surface dyslexia" (Plaut et al., 1996, 8). Finally, the assumptions of the model can be used "to derive predictions about the relative naming latencies of different types of words. In particular... why naming latency depends on the frequency of a word" (Plaut et al., 1996, 21).

Although this sample of papers has not been entirely systematic, it is representative in that it covers three decades of work, four core areas of cognition (attention, memory, language and vision), many of the main players in the field and several types of paper (experiment, theoretical paper, and review). Certainly there are connectionist models of cognition that do not meet all four criteria (and perhaps some that meet none of the four). I

do not claim that this pattern is universal, merely typical, and as we'll see later, some components can be dropped without greatly affecting the form of the inference. In the next section I offer an analysis of the kind of explanation offered by this sort of model.

How Connectionist Models Explain

Recall that classical AI's explanations of cognition employ inference to the best explanation, which involves finding a candidate explanation, then, as Coltheart put it, resting on one's laurels until a reasonable competitor comes along. Connectionist models of cognition not only provide the competition, but also make plain the methodological fragility of classical AI's dependence on inference to the best explanation. As Sejnowski puts it, "Although a working model can help generate hypotheses, and rule some out, it cannot prove that the brain necessarily solves the problem in the same way" (Sejnowski et al., 1988, 1304). In other words, simulating the behavior only shows that you have a candidate explanation; it does not show that you have the right explanation, i.e., one that produces the behavior in the "same way."

For connectionists, the "same way" means looking to the anatomy and physiology of the brain, because whatever the right explanation of cognition is, it must be at least possible to implement it with brainy stuff. Connectionists talk about taking *constraints* from both physiology and psychology, as though they are employing an inferential pincer movement, narrowing the space of possibilities from two flanks at once (although search may not be an apt metaphor for model building, because the domain is infinite, and there are no halting conditions).

A more promising way of understanding connectionist methodology is hinted at in each of the papers cited above. They all talk about the constraints they take from brains in terms of basic, or general principles. Here are some quotes to that effect: "modeling is often crucial if we are to understand the implications of certain kinds of basic principles of processing" (McClelland, 1988, 107); "connectionist modeling provides a rich set of general computational principles that can lead to new and useful ways of thinking about human performance" (Plaut et al., 1996, 2). "The study of simplifying models of the brain can provide a conceptual framework for isolating the basic computational problems and understanding the computational constraints that govern the design of the nervous system" (Sejnowski et al., 1988, 1300). The point is evidently not to model the brain in detail, but rather to model the basic processing principles used by the brain.

McClelland et al. (1995) describe this strategy in their paper about why there are two learning systems in hippocampus and neocortex. They focus on the phenomenon of memory consolidation, a gradual process that can take many years. Their goal is a model of learning and memory that goes beyond just reproducing the observed phenomena. They want to make sense of why, from a design perspective, there are two separate memory systems, and to figure out what the functional importance of gradual consolidation is. They ask, "Is the phenomenon a reflection of an arbitrary property of the nervous system, or does it reflect some crucial aspect of the mechanisms of learning and memory? Is the fact that consolidation can take quite a long time—up to 15 years or more in some cases—just an arbitrary parameter, or does it reflect an important design principle?" (McClelland et al.,

1995, 419).

In cases like this, some details, like the timing of consolidation, take on particular importance for figuring out how a phenomenon is produced. Models recreate select physiological or anatomical details of their target systems, not to strengthen the inference from model to target slightly, as in Inference 3, but in order to test the significance of those details. If that detail is changed, is there a qualitative change in overall performance? An arbitrary property can be altered without a qualitative change in performance, but a crucial aspect of the mechanism cannot. Instead of acting as piecemeal support for the theory, these details are used to probe the design of the mechanism.

Talk of mechanisms, as in the quote from McClelland et al. (1995) above, is common in discussions of connectionist methodology, but not so in classical AI, where algorithms are the main concern. For connectionists, producing the behavior in the "same way" means more than just having the right algorithm. While an algorithm provides a schematic specification of processes or activities and their coordination, a mechanism specifies both the algorithm *plus* the entities or parts involved in these activities, and their organization. (For accounts of mechanism, see Machamer et al. (2000); Glennan (2002); Bechtel and Abrahamsen (2005).) Machamer et al. (2000) stress this dualist nature of mechanisms.

The anatomical facts that are recreated in connectionist models provide a schematic specification of mechanism entities. Entities in mechanisms are not to be confused with implementation details. Rather than being specific details about the hardware or software on which an algorithm is run, mechanism entities are more like the types of structures required by an algorithm. A sorting algorithm might require a memory store and a read/write device, for example. A description of a mechanism makes explicit those entities that an algorithm takes for granted.

This focus on mechanisms rather than algorithms also helps explain why a fair bit of attention is paid to simulating neurological damage in connectionist modeling. One way of testing whether a property is an arbitrary or crucial feature of a mechanism's design is to see what happens when you remove or break it. Cognitive neuropsychology is the study of "what one can learn about the organization of the cognitive system from observing the behavior of neurological patients" (Shallice, 2001, 2128). By analyzing the kinds of cognitive deficits represented in neurological case studies, one can construct hypotheses about how cognitive mechanisms are designed. Connectionist modeling incorporates this strategy.

Traditionally neuropsychology depends on the assumption that cognitive functions are localized to specific brain regions, so that injuries affecting discrete brain regions can be correlated with deficits in specific cognitive functions. Historically, the affected brain regions would be assessed postmortem, but contemporary cognitive neuropsychology also makes use of neuroimaging data to localize lesions.

The logic involved in using data from cognitive neuropsychology to develop cognitive theories has been discussed at length elsewhere (Shallice, 1988; Bub, 1994a,b; Glymour, 1994). In this literature brain anatomy and physiology are represented abstractly as directed graphs, where nodes correspond to anatomical locations that perform particular functions, and edges

correspond to connections between these functional units, through which data is communicated. Lesions to different parts of the graph then give rise to distinct functional disturbances. For instance, in one of the field's pioneering papers, Lichtheim (1885) posits that a lesion to the "centre of motor representation of words" would give rise to symptoms like "Loss of (a) volitional speech; (b) repetition of words; (c) reading aloud..." (Lichtheim, 1885, 320). For each neurological subject, there must be a way of lesioning the graph such that the available paths through the graph correspond to the profile of that patient, i.e., their characteristic set of capacities.

The upshot of these methodological discussions in cognitive neuropsychology is as follows. Consider the possible cognitive theories as a set of possible directed graphs. Given sufficient lesion data, the correct theory should be a minimal graph whose set of path-sets contain all the profiles corresponding to dissociations seen in the neurological data (Bub, 1994a, 850). As a shorthand, I'll write this as $T^N =_{min} G^N$, where T^N stands for the cognitive model, and G^N refers to the set of graphs that can account for neurological data N. (As Glymour (1994) points out, T^N may not be unique.)

Connectionist models are more powerful than traditional cognitive neuropsychology, because they are not limited by the availability of neurological subjects with specific injuries, and they need not assume localizability of functions. Localized injuries can be simulated by damaging all the nodes in one region of the network. Other sorts of injuries can be simulated by modifying the network as a whole, such as by adding noise, changing connection weights, or adjusting the learning rule.

Connectionist modeling efforts do not use the formal approach of choosing minimal graphs, but share cognitive neuropsychology's rationale for simulating neurological data. Intuitively, the approach requires that the mechanism be such that there are distinct ways of damaging it that would result in each of the patterns of neurological deficits that have been observed, without being unnecessarily complex. (How to assess the complexity of a mechanism is a good question, but one I'll leave unanswered.) The wrong mechanism would yield a qualitatively different pattern of deficits. In cognitive neuropsychology, brain regions and their connections are treated abstractly as nodes and connections in graphs, but still can tell us a lot about cognitive architecture.

Connectionist models are likewise abstract yet informative about cognitive architecture. They are reusable, multi-purpose tools that can be reconfigured in a variety of contexts. As McClelland et al. say of their model, "These are not detailed neural models; rather, they illustrate, at an abstract level, what we take consolidation to be about (McClelland et al., 1995, 420). That connectionist models are abstract or idealized is sometimes raised as a criticism; if connectionist models were supposed to explain the way true simulations do, their lack of realistic detail would be a serious problem. However, connectionist models aim to discover only the generic properties of the mechanisms they implement, not all the details. The reasoning involved in discovering how generic mechanisms work is nothing new; in fact, it was described by Mill (1843).

According to Mill, in order to analyze causes and effects, we must first decompose each

scenario into single facts (which for Mill can be states of affairs, events, or propositions). What counts as a single fact, or how far down we have to go in the decomposition, depends on our purpose (Mill, 1843, III: 187). We then observe which facts cause which others by observing which follow from which, as circumstances vary. The advantage of experiments is that "When we have insulated the phenomenon we're investigating by placing it among known circumstances, we can vary the circumstances in any way we like, choosing the variations that we think have the best chance of bringing the laws of the phenomenon into a clear light" (Mill, 1843, III: 189).

When a regularity is discovered, such that one set of facts (such as a particular arrangement of working parts) tends to be followed by another set of facts (such as a particular pattern of behaviors), we have what I'll call a generic mechanism. According to a popular recent account, "Mechanisms are regular in that they work always or for the most part in the same way under the same conditions. The regularity is exhibited in the typical way that the mechanism runs from beginning to end" (Machamer et al., 2000, 3). Mechanisms may operate regularly only within certain ranges of parameter values, and the sameness of their results may be qualitative, or likewise specify a range of values. In general, they are like causes tending to produce like effects. Although they are loosely defined and not perfectly predictable, generic mechanisms are useful in a variety of contexts.

An illustrative example is lateral inhibition, which was first described in retinal ganglion cells (Hartline, 1940b,a), but later discovered to be "ubiquitous to all sensory areas of the brain" (Macknik and Martinez-Conde, 2009). Retinal ganglion cells have inhibitory connections to their immediate neighbors. The strength of the inhibitory signal is proportional to the activation of the cell the signal originates in. This means that when one cell is stimulated, its neighbors are inhibited. For a cell to fire strongly, most of its neighbors can't also be stimulated. Retinal ganglion cells respond to object contours or edges, which are characterized by abrupt changes in illumination. Compared to neurons in the middle of uniform patches of illumination, which are inhibited by all of their neighbors, neurons at edges receive less inhibition, so have higher relative activity. This tends to sharpen responses even further, because this activation and inhibition is ongoing. As a result, even fairly faint edges are sharpened.

The lateral inhibition mechanism has been used to explain several other biological phenomena where contrasts are detected or enhanced. One example is cell type differentiation in embryology. Cells that start to develop earliest, and are on track to specialize for a particular purpose, such as forming a particular organ, send out protein signals that act as chemical inhibitors. These inhibitory proteins prevent surrounding cells from taking on the same job, which means that the neighboring cells specialize for something different. Small initial differences in developmental schedules make for stark contrasts in developmental outcomes.

There are also economic and sociological analogues. If communities decide to focus their limited resources on their most promising students or athletes, and if the amount of investment made is in proportion to their skills, this results in a widening of the gap between the skills of the most promising and the rest. In this sort of scenario, the most promising students or athletes get more resources to the detriment of less promising ones, which makes

the best improve more quickly, further widening the skill gap between stars and non-stars.[1] Another example is the convention that ping-pong or pool tables in pubs are kept by the winner of a match. This means that the better players improve more quickly, because they get more practice, at the expense of mediocre players who get less practice in virtue of being kicked off the table after each try.

These examples vary widely in their details, but all share some very general structural properties, and have qualitatively similar effects. This is the sort of general processing principle that connectionist models are designed to discover and explore. First we discover, through a combination of mathematical demonstration and empirical observation, that a certain type of mechanism (e.g., networks with inhibitory connections among neighbors) tends to give rise to a certain type of behavior (e.g., contrast enhancement). We then make use of that knowledge to make sense of how brain structures (e.g., the retina) give rise to cognitive phenomena (e.g., edge detection).

The connection between the target system T and our model T^* is that both instantiate the same general mechanism type. We can infer that the properties of the one apply to the other based on their shared type membership. For instance, the actual retinal ganglion cell network and our connectionist model of it both belong to the general type of lateral inhibition networks. We can explore the properties of the type using the model T^*, then infer that the properties we observe, O^*, also belong to the target system T.
The first part of the strategy is making novel empirical predictions. The gap between T^* and T can be narrowed not only by showing that the model confirms the observations made in the target system, or $T^* \rightarrow O$, but also by showing that predictions work in the opposite direction. Confirming the predictions of the model in the target system, does two things: it rules out gerrymandered models that are designed to give the desired output without sharing underlying properties, and it shows that the similarities between theory and model run in both directions. The latter builds confidence that the model and theory belong to the same type. Being the same type of mechanism also involves sharing the entities that are crucial to the design of the mechanism, having qualitatively similar behavior, and having the same pattern of breakdown when damaged.

Because connectionist models are used during many stages of research, there is no single inference type that fits all cases. One important example is inferring that an established model can predict the behavior of a target system, In this case, the form of the inference might look like this:

$$\frac{T^* \rightarrow O^* \qquad T,\, T^* \in M^T}{T \rightarrow O^* \qquad .} \tag{4}$$

The first premise states that the model produces a set of predicted observations O^*. The second premise states that the model and target system instantiate the same mechanism

[1] This was rumored to be the case in a figure skating club near my childhood home, where 1988 Olympic silver medalist Elizabeth Manley trained.

type M^T. Using the logic of tendencies, we can infer that like causes (T and T^*) should have like effects, so the model's predictions, O^*, should also be true in the target system. Another example is inferring that the model is adequate, given what is known about the target system. In this case, the form of the inference might look like this:

$$\frac{T \rightarrow O \qquad T,\ T^* \in M^T}{T^* \rightarrow O} \quad . \qquad (5)$$

As with any inductive inferences, explanations using the logic of tendencies are susceptible to error. First, if the mechanism's operation is not very regular or reliable (for example, a stochastic mechanism), the model may predict different outcomes than the target in some cases, despite both being examples of the same mechanism type. Second, the output of the model is never exactly the same as the target phenomenon, so additional arguments are sometimes needed to establish that they are similar enough. The details left out will sometimes make a difference. Third, the experimental and neurological data on which assumptions about the design of the mechanism are based are of course incomplete, so a model that is adequate at one time can be ruled out by later evidence.

Given this new understanding of how connectionist models work and what the risks of error are, we can rethink the sorts of criticisms of connectionist models that are and aren't viable. For stochastic mechanisms, we should look for results that summarize probability distributions over many randomized trials rather than single runs. Because some details will always differ between model and target, results should be considered tentative until several variations with different details and assumptions all agree. Both of these suggestions are already standard practice in connectionist modeling. A more novel result is that blanket criticisms of connectionist models as either too detailed or not detailed enough are off the mark, as long as the level of detail is appropriate to the research question. Very detailed models are not only less widely applicable, but also more susceptible to being overridden by new discoveries in neurophysiology. We should expect earlier models to be more abstract, and later models to be more detailed about select parts of the mechanism. This pattern is already apparent in connectionist modeling research. Often it is the more general models rather than the more specific that receive the most critical attention, but it should be the reverse.

Despite these caveats, connectionist modeling is a powerful and nuanced set of methods that allow for the possibility of explaining cognition at many scales of generality or specificity. It can also offer explanations of how and why cognitive deficits occur as a result of particular sorts of brain lesions, which promises clinical payoffs.

Conclusion

I began by offering formal accounts of how classical AI and true simulations explain. Classical AI uses inference to the best explanation, as was clear from the methodological claims made in both older and contemporary sources. Simulation tries to deduce predictions from detailed bottom-up models. Connectionist models are puzzling in that they seem to try

to do a little of each, which should undermine both modes of explanation.

I arrived at a four-part analysis of the explanatory features of connectionist models. First, details of the neurophysiology of the brain are built into the models. Second, the output of the models reproduce known psychological data. Third, damaging the models reproduces patterns of deficits found in neurological cases. Finally, good models make novel empirical predictions that can be experimentally verified. I noted that connectionist models are intended to explore the generic mechanisms operating in the brain, and illustrated the relevant notion of mechanism with the example of lateral inhibition.

I then constructed a formal analysis of the explanations offered, which interprets connectionist models and the cognitive theories they represent as sharing membership in a type of mechanism. The inferences made from connectionist models to cognitive phenomena can be understood as involving the logic of tendencies. Models and targets that instantiate the same general mechanisms can be expected to have similar output.

One of the motivations for offering an account of how connectionist models explain is that doubts have been raised as to whether they are relevant to cognition at all. Although connectionist models have been contributing to our understanding of the mind for several decades now, there has been little understanding of how they work. I hope that this chapter will shed some light on this question.

References

Bailer-Jones, D. M. (2009). *Scientific Models in Philosophy of Science*. University of Pittsburgh Press, Pittsburgh.

Bechtel, W. and Abrahamsen, A. (2005). Explanation: A Mechanist Alternative. *Studies in the History and Philosophy of Science, Part C*, 36(2):421–441.

Billings, G., Piasini, E., Lörincz, A., Nusser, Z., and Silver, R. (2014). Network structure within the cerebellar input layer enables lossless sparse encoding. *Neuron*, 83(4):960–974.

Bub, J. (1994a). Is Cognitive Neuropsychology Possible? *Philosophy of Science*, 1:417–427.

Bub, J. (1994b). Models of Cognition Through the Analysis of Brain-Damaged Performance. *The British Journal for the Philosophy of Science*, 45(3):837– 855.

Coltheart, M., Rastle, K., Perry, C., Ziegler, J., and Langdon, R. (2001). DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud. *Psychological Review*, 108(1):204–256.

De Pisapia, N., Repovs, G., and Braver, T. S. (2008). Computational Models of Attention and Cognitive Control. In *Cambridge Handbook of Computational Psychology*, pages 422–450. Cambridge University Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., and Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, 338(6111):1202–1205.

Engel, A. K., Moll, C. K. E., Fried, I., and Ojemann, G. A. (2005). Invasive recordings from the human brain: Clinical insights and beyond. *Nature Reviews Neuroscience*, 6(January):35–47.

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

Glennan, S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science*, 69(S3):342–353.

Glymour, C. (1994). On the Methods of Cognitive Neuropsychology. *The British Journal for the Philosophy of Science*, 45(3):815–835.

Hartline, H. K. (1940a). The Effects of Spatial Stimulation in the Retina on the Excitation of the Fibers of the Optic Nerve. *American Journal of Physiology*, pages 700–711.

Hartline, H. K. (1940b). The Receptive Fields of Optic Nerve Fibers. *American Journal of Physiology*, pages 690–699.

Hempel, C. and P. Oppenheim. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, 15: 135–175.

Humphreys, P. (1990). Computer Simulations. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1990:497–506.

Izhikevich, E. M. and Edelman, G. M. (2008). Large-scale model of mammalian thalamocortical systems. *Proceedings of the National Academy of Sciences*, 105(9):3593–3598.

Lichtheim, L. (1885). On Aphasia. *Brain*, 7:433–484.

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

Macknik, S. L. and Martinez-Conde, S. (2009). Lateral Inhibition. In Goldstein, E. B., editor, *Encyclopedia of Perception*. Sage Press.

Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience*, 7(2):153–160.

McCarthy, J., Minsky, M., Rochester, N., and Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

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. (1988). Connectionist Models and Psychological Evidence. *Journal of Memory and Language*, 27:107–123.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review*, 102(3):419–457.

Mill, J. S. (1843). *A System of Logic: Ratiocinative and Inductive*. In the version presented at www.earlymoderntexts.com.

Morgan, M. S. and Morrison, M. (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press.

Mukamel, R. and Fried, I. (2012). Human Intracranial Recordings and Cognitive Neuroscience. *Annual Review of Psychology*, 63(1):511–537.

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.

O'Reilly, R. C. and Frank, M. J. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, 18:283–328.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-regular Domains. *Psychological Review*, 103(1):56–115.

Sejnowski, T., Koch, C., and Churchland, P. (1988). Computational Neuroscience. *Science*, 241(4871):1299–1306.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge University Press.

Shallice, T. (2001). Cognitive Neuropsychology, Methodology of. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 2128–2133. Elsevier Science Ltd.

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

Stinson, C. (2018). What Artificial Neurons Tell us about Real Brains. Article under review.

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

Winsberg, E. (2001). Simulations, Models, and Theories: Complex Physical Systems and Their Representations. *Philosophy of Science*, 68(3):S442—-S454.

Winther, Rasmus Grønfeldt, (2016). The Structure of Scientific Theories. *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/structure-scientific-theories/>.

Woodward, James, (2017). Scientific Explanation. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation/>.