Explanation and Connectionist Models

Catherine Stinson

Rotman Institute of Philosophy

University of Western Ontario

December 24, 2016

Abstract

How do connectionist models explain? Connectionist models replace experiments that for ethical and pragmatic reasons we can't do, and explore the abstract properties of complex brain-like networks, with the aim of discovering the mechanisms that give rise to cognition. Their mode of explanation is distinct from those of both classical AI, which uses inference to the best explanation; and from simulation, which approximates deductive reasoning. Connectionist models explain using the logic of tendencies combined with inference to the best explanation. This is illustrated by considering connectionist models of attention, language, vision, and memory.

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 (for example 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, exactly. This chapter explores the epistemic roles played by connectionist models of cognition, and offers a novel 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 like, How convincing is a given model? What makes a connectionist model successful? What kinds of errors should 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 use abductive reasoning, or inference to the best explanation. Detailed brain simulations use deductive reasoning, or some approximation to it. Connectionist modeling seems to combine the two methods; modelers talk in terms of taking constraints both from psychological data, and from neuroanatomy and neurophysiology. The challenge is to understand how combining these two very different methods could be a

worthwhile strategy, instead of a failure on both counts. Specifically, why should using neural constraints be a good strategy if the details of those neural constraints aren't correct?

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 general mechanisms at work in the brain. This method depends on the logic of tendencies: drawing inductive inferences from like causes to like effects. This is combined with inference to the best explanation, and neuropsychological reasoning about directed graphs, in order to constrain the possible explanations to consider.

How Computational Models Explain

I begin by exploring how computational models in general are used to explain aspects of cognition. Computational models are especially important in cases where experimenting directly on the target system is not practicable, or the problem 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 as a means of experimenting. Furthermore, these recordings are made from brains affected by pathology, and usually in patients taking medication (Mukamel and Fried, 2012), so care must be taken when drawing inferences from studies on brain-injured patients back to healthy populations.

Several non-invasive means for indirect measurement from and intervention on healthy 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. Animals usually can't always understand complex tasks or 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 as human brains do.

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 make understanding this complex system more tractable, making them an essential tool 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 Dartmouth Proposal. They conjecture 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 idea is 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). The

program consists of a sequence of symbolic expressions, and they postulate that the processes going on inside the subject are likewise operations on symbols. The defense of their postulate (later dubbed the "physical symbol system hypothesis" (Newell and Simon, 1976)), "lies in its power to explain the behavior" (Newell and Simon, 1961, 2012). A complete theory would also explain how neurophysiology gives rise to symbol manipulations, but they set that half of the project aside.

In other words, the way Newell & Simon judge the success of an AI program is by comparing its output to human behavior. A program that counts as a good theory should match the behavioral data. The fact that the program gives rise to the same output is a reason for thinking that the cognitive process can be explained by the program, i.e., they might work the same way. This is an abductive inference, or an inference to the best explanation. This inference has the form,

$$\frac{T \rightarrow O \quad O}{T} \tag{1}$$

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

I take it that their defense of the physical symbol system hypothesis—the assumption that the intermediate layer of a complete 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, all suddenly became tractable with the 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.

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 discussion of cognitive models, but notably different than its usual meaning. In fields like physics and climate science, what I'll call a true simulation begins with the fundamental equations of a well-established theory, and deduces how a system changes over time by making calculations using those equations. 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,

$$\frac{T \rightarrow O \quad T}{O} \quad . \tag{2}$$

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 are like true simulations. The Blue Brain project, and the Human Connectome Project are examples where models of human cognition are being built from the bottom up, starting by modeling brain anatomy and physiology in great detail, with the aim of being able to simulate how a brain works. There are also many models of this type that focus on particular brain areas, or even small structures like synapses. In these projects, getting the anatomical and physiological details correct is a priority. Incorrect details or imprecise calculations can lead to predictions that are wildly incorrect.

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 aim to reproduce cognitive phenomena. Another motivation for the connectionist approach mentioned in the PDP 'bible', is to improve on the performance of classical AI models, which McClelland et al. see as only approximations to psychological phenomena: "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... Yet people get faster, not slower, when they are able to exploit additional constraints" (McClelland and Rumelhart, 1986, 12).

It is not immediately obvious what kind of explanations are being offered by this combination of methods. Of particular concern is how incorporating neural constraints is meant to help when these constraints are taken only very loosely. The form of the inference looks like the following, at first pass:

$$\frac{T^* \rightarrow O \quad O \quad t_1, ..., t_n}{T} \tag{3}$$

where T^* loosely resembles T, and $t_1, ..., t_n \in T$ are physiological constraints on brains that are included in the model (perhaps loosely interpreted). Consider this in comparison to Equation 1, inference to best explanation. That some of the statements $t_1, ..., t_n$ in T are true strengthens the inference to T, but the fact that the model/program T^* only loosely resembles T undercuts the inference. 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^*, not T, that implies O.

In Stinson (2016), I argue that PDP models are abstract, idealized models of multi-level mind/brain mechanisms. Below I analyze the inferences we make using these models, by exploring several examples. I argue that the missing connection between T and $T^* \rightarrow O$ is provided by the logic of tendencies, and that this is how abstract mechanistic explanations work.

Connectionist Explanation Examples

In this section I look in some detail at examples of connectionist models from several areas of 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. The success of these models seems to be judged based on four criteria. Consideration of models from several other areas of cognition confirms that connectionist models of cognition follow this general 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 brain areas V1, PP and IT. Some of the models also go into 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 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).

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 measurements, 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 point out the obvious weakness of inference to the best explanation. As Sejnowski puts it, "The problem of inferring function from response properties applies to neurons throughout the nervous system... 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 one, i.e., one that produces the behavior in the "same way."

For connectionists, the "same way" means looking to the architecture of the brain, because whatever the right explanation of cognition is, it must be made of brainy stuff. Connectionists talk about taking *constraints* from both physiology and psychology, which I think is intended as an inferential pincer movement, narrowing the search space from two flanks. Search, however, is an inappropriate method for theory building, because the domain is infinite, and there are no halting conditions.

Another, 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 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 some detail 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). Note that they are interested in the *mechanisms* of learning and memory.

Talk of mechanisms is common in discussions of connectionist methodology, but not so in classical AI. Producing the behavior in the "same way" means getting the

mechanisms right. This helps explain why so much attention is paid to simulating neurological conditions in connectionist modeling. 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 that are represented in neurological case studies, one can construct hypotheses about how the cognitive apparatus is designed to work. 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. Originally, the affected brain regions could only be assessed postmortem, but contemporary cognitive neuropsychology also makes use of neuroimaging data to localize lesions. Connectionist models are a powerful addition to cognitive neuropsychology, because they are not limited by the availability of neurological patients 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; however, other sorts of injuries can also be simulated by modifying the network as a whole, such as by adding noise, changing connection weights, or adjusting the learning rule.

The logic involved in using data from cognitive neuropsychology to develop cognitive theories has been discussed at length elsewhere (see Shallice (1988); Bub (1994a,b); Glymour (1994)). The upshot, for our purposes, is that if we consider the possible cognitive theories as directed graphs, 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 $T =_{min} G^N$. Intuitively, this means that the mechanism has to 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 in the clinic, but that it should be no more complex than needed to account for that data. Note that this way of understanding brain anatomy, as directed graphs, is very abstract.

In many examples of connectionist modeling, the mechanisms that are supposed to explain how the brain gives rise to cognition are abstract (or basic or general) processing principles. These mechanisms are not the nitty-gritty details of brain anatomy and physiology, but something closer to the design principles or algorithms used by the brain. These are reusable, multi-purpose tools that can be reconfigured to perform a variety of functions. 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 often raised as a criticism. If connectionist models were supposed to offer deductive explanations, the way true simulations do, their lack of realism would be a serious problem. However, connectionist models are not meant to be deductive explanations; and they aim to discover the tendencies, not certainties, of the general mechanisms they implement. The reasoning involved in discovering how mechanisms work is nothing new; in fact, it was described in detail by Mill (1843).

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

each scenario into single facts, although 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 tends to be followed by another set of facts, we have what I'll call a *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 very loosely defined and not perfectly predictable, mechanisms fitting this description 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 detect object contours or edges, which are characterized by abrupt changes in illumination. Cells near an illumination change in the retinal image only get inhibited by neighbors on one side of the edge. 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 will appear sharpened.

The lateral inhibition mechanism has been used to explain several other scientific 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 protein signals 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 will result 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 are quite varied in their details, but all share some very general characteristics, and have qualitatively similar effects. This is the sort of general processing principle that connectionist models are designed to discover and explore. The point, as has been repeated several times, is not to model brains in glorious detail, but to model the general mechanisms they employ. The form of explanation is not deduction. Instead, it follows the logic of tendencies. Through some combination of mathematical demonstration and empirical observation, we establish that a certain type of structure (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).

This is the connection we were missing between our theory T and our model T^*. Both belong to the same general type. For instance, the actual retinal ganglion cell network belongs to the general type of lateral inhibition networks, which our connectionist model implements, and which we know tend to exhibit the phenomenon of contrast enhancement.

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

The final part of the strategy is making novel empirical predictions. These empirical predictions enlarge the set of observations O entailed by the model to $O^p \supset O$, and when these predictions are validated experimentally, that means we can claim the larger set O^p as a premise. Increasing the number of true observations implied by the model means that it explains more data.

In general, the form of the inference is the following,

$$\frac{T^* \rightarrow O^p \quad T, T^* \in T^M \quad O^p \quad t_1, ..., t_n \quad T =_{min} G^N}{T} \tag{4}$$

where T^M is the type of mechanism implemented. This still shows some similarities to inference to the best explanation, but with several additional sources of support. Although we don't know that $T \rightarrow O^p$, we do know that the observations are entailed by a model that implements the same mechanism as our theory. We also know that some details of the theory are true, as we have taken these from our knowledge of brain physiology, and we know that our theory is a minimal model consistent with the neurological data. These last two premises constrain inference to the best explanation to a smaller set of candidates by specifying further details about the brain mechanisms in question.

As with any inductive inference, this type of explanation is susceptible to error. First, if the mechanism's operation is not very regular, the model and the theory may lead to different output, despite both being examples of the same mechanism. Second, a detail I've glossed over is that the output of the model is never exactly the same as the behavioral observations, although this is a problem for the other approached to AI too. Third, we may not have a complete set of neurological cases,

or some of these may violate the assumptions under which the methods of cognitive neuropsychology work. Fourth, the correct theory may not be the minimal graph (evolution is not always perfectly efficient).

Despite these risks of error, connectionist modeling is powerful and nuanced. It allows 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.

Conclusions

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 general 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 theories can be understood as involving the logic of tendencies, to connect the behavior of general mechanisms to their instantiations. Inference to the best explanation is also involved, with added constraints on the pool of candidate explanations coming from neurophysiology and cognitive neuropsychology.

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 is little understanding of why they are effective. Attempts by connectionist modelers to explicate their methodology have not made it clear to doubters why we should believe that this is a valid way of learning about cognition. I hope that this chapter will shed some light on this problem.

References

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.

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.

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.

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.

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.

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. (2016). What Artificial Neurons Tell us about Real Brains. Article under review.