COGNITIVE MECHANISMS AND COMPUTATIONAL MODELS: EXPLANATION IN COGNITIVE NEUROSCIENCE

by

Catherine Elizabeth Stinson

B.Sc. Cognitive Science & Artificial Intelligence, University of Toronto, 1999

M.Sc. Computer Science, University of Toronto, 2002

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Catherine Elizabeth Stinson

It was defended on

August 9, 2013

and approved by

Peter K. Machamer, PhD, Professor, History & Philosophy of Science

Kenneth F. Schaffner, MD, PhD, Distinguished University Professor, History & Philosophy of Science

James Bogen, PhD, Adjunct Professor, History & Philosophy of Science

Edda Thiels, PhD, Assistant Professor, Neurobiology

Dissertation Advisors:

Peter K. Machamer, PhD, Professor, History & Philosophy of Science,

Kenneth F. Schaffner, MD, PhD, Distinguished University Professor, History & Philosophy of Science

COGNITIVE MECHANISMS AND COMPUTATIONAL MODELS: EXPLANATION IN COGNITIVE NEUROSCIENCE

Catherine Elizabeth Stinson, PhD

University of Pittsburgh, 2013

Cognitive Neuroscience seeks to integrate cognitive psychology and neuroscience. I critique existing analyses of this integration project, and offer my own account of how it ought to be understood given the practices of researchers in these fields.

A recent proposal suggests that integration between cognitive psychology and neuroscience can be achieved 'seamlessly' via mechanistic explanation. Cognitive models are elliptical mechanism sketches, according to this proposal. This proposal glosses over several difficulties concerning the practice of cognitive psychology and the nature of cognitive models, however. Although psychology's information-processing models superficially resemble mechanism sketches, they in fact systematically include and exclude different kinds of information. I distinguish two kinds of information-processing model, neither of which specifies the entities and activities characteristic of mechanistic models, even sketchily. Furthermore, theory development in psychology does not involve the filling in of these missing details, but rather refinement of the sorts of models they start out as. I contrast the development of psychology's attention filter models with the development of neurobiology's models of sodium channel filtering.

I argue that extending the account of mechanisms to include what I define as *generic mechanisms* provides a more promising route towards integration. Generic mechanisms are the in-the-world counterparts to abstract types. They thus have causal-explanatory powers which are shared by all the tokens that instantiate that type. This not only provides a way for generalizations to factor into mechanistic explanations, which allows for the 'upward-

looking' explanations needed for integrating cognitive models, but also solves some internal problems in the mechanism literature concerning schemas and explanatory relevance.

I illustrate how generic mechanisms are discovered and used with examples from computational cognitive neuroscience. I argue that connectionist models can be understood as approximations to generic brain mechanisms, which resolves a longstanding philosophical puzzle as to their role. Furthermore, I argue that understanding scientific models in general in terms of generic mechanisms allows for a unified account of the types of inferences made in modeling and in experiment.

Keywords: neuroscience, psychology, explanation, integration, computation, models, connectionism, mechanism.

TABLE OF CONTENTS

PREFACE xii

1.0 INTRODUCTION 1

1.1 Background 1

1.1.1 Cognitive Neuroscience 3

1.2 Methods and Motivations 4

1.3 Integration 6

1.3.1 Autonomy 7

1.3.1.1 Marr's Levels 8

1.3.2 Reductionism 10

1.3.3 Emergence 13

1.3.4 Unification 14

1.3.5 Constraints on Integration 15

1.4 Integration through Mechanistic Explanation 16

1.5 Outline 19

2.0 EXPLANATION IN COGNITIVE PSYCHOLOGY AND NEUROSCIENCE 22

2.0.1 Outline 22

2.1 Explanation in Neuroscience 23

2.1.1 Integration 24

2.2 Defining Cognitive Psychology 26

2.2.1 Cognitive Psychology's Topics 27

2.2.2 Cognitive Psychology's Provenance 28

2.2.3 Cognitive Psychology's Methods . . . 29
2.3 Information Processing Models . . . 31
2.3.1 Information-Processing Diagrams . . . 37
2.3.1.1 Control Flow Diagrams . . . 40
2.3.1.2 Data Flow Diagrams . . . 42
2.3.1.3 Mechanism Schema Diagrams . . . 43
2.3.2 Elliptical Mechanism Sketches . . . 45
2.4 Theory Development in Neuroscience and Cognitive Psychology . . . 49
2.4.1 Neuroscience's filters . . . 49
2.4.2 Cognitive Psychology's filters . . . 54
2.5 Integrating Neuroscience and Cognitive Psychology . . . 59
2.5.1 Memory as a Test Case . . . 64
2.5.2 Other Approaches to Integration . . . 67
2.6 Conclusions . . . 70
3.0 COGNITIVE MECHANISMS . . . 72
3.0.1 Outline . . . 73
3.1 The MDC Account of Mechanism . . . 73
3.1.1 Hierarchies . . . 75
3.1.2 Later Elaborations by Machamer . . . 77
3.1.3 Later Elaborations by Darden . . . 78
3.1.3.1 Schema Instantiation . . . 80
3.1.4 Later Elaborations by Craver . . . 82
3.2 Beyond the MDC Account of Mechanism . . . 85
3.2.1 Ontic Explanation . . . 85
3.2.2 Schemas . . . 88
3.2.3 Regularity . . . 92
3.2.3.1 Low-Frequency Mechanisms . . . 92
3.2.3.2 One-Off Mechanisms . . . 93
3.2.3.3 Causation and Mechanism . . . 99
3.2.3.4 Generalization . . . 101

3.2.4 Types . . . 103
3.3 Cognitive Mechanisms . . . 106
3.3.1 Looking Upward, Explaining Downward . . . 106
3.3.2 Problems and Constraints . . . 108
3.3.2.1 Explanatory Relevance . . . 108
3.3.2.2 Types and Tokens . . . 111
3.3.2.3 A Role for Generalization . . . 112
3.3.3 Generic Mechanisms . . . 115
3.3.3.1 An Example: Lateral Inhibition . . . 119
3.3.4 Prospects for Integration . . . 122
4.0 COMPUTATIONAL COGNITIVE NEUROSCIENCE . . . 125
4.1 Introduction . . . 125
4.1.1 Outline . . . 127
4.2 Motivations for Computational Modeling . . . 128
4.2.1 Typical Uses of Computational Models . . . 130
4.2.2 Computational Modeling in Other Fields . . . 135
4.2.3 Simulating the Brain . . . 136
4.2.4 Simulating Behavior . . . 137
4.3 Computational Models in Artificial Intelligence . . . 139
4.3.1 Classical AI . . . 140
4.3.2 Realistic Neural Models . . . 143
4.3.3 Connectionist AI . . . 145
4.3.3.1 Defining Connectionism . . . 146
4.3.3.2 Connectionist Motivations . . . 148
4.3.3.3 The Past-tense Learner . . . 149
4.3.3.4 PDP's Impact . . . 151
4.3.4 Connectionism Attacked . . . 152
4.4 The Role of Connectionist Models . . . 154
4.4.1 Theories of Cognition vs. Implementations . . . 154
4.4.2 Neural Simulations . . . 157

4.4.3 Mathematical Demonstrations . . . 161
4.5 A Partial Resolution . . . 164
4.5.1 An Example: Modeling Rat Hippocampus . . . 165
4.5.1.1 Discovering Mechanisms . . . 169
4.6 Conclusion . . . 170
5.0 COMPUTATIONAL MODELS AS MODELS OF MECHANISMS . . . 172
5.1 Introduction . . . 172
5.1.1 Outline . . . 173
5.2 Epistemology of Modeling . . . 174
5.3 Models and Representation . . . 175
5.3.1 Representation in Explanation . . . 176
5.3.2 Beyond Representation . . . 176
5.4 Models of Mechanisms . . . 179
5.4.1 Schemas and Generic Models . . . 180
5.4.2 Non-representational Mechanistic Models . . . 182
5.4.2.1 Inferences from Non-representational Models . . . 183
5.5 Computational Models and Explanation . . . 187
5.5.1 Computational Models as Models . . . 188
5.5.1.1 Computational Models as Models of Mechanisms . . . 190
5.5.2 Connectionist Models as Generic Mechanisms . . . 191
5.5.2.1 The Explanatory Power of Hybrid Models . . . 197
5.5.2.2 Connectionist Models as Hybrid Mechanisms . . . 199
5.6 Conclusion . . . 201
6.0 CONCLUSION . . . 203
6.1 Summary . . . 203
6.1.1 Introduction . . . 203
6.1.2 Explanation in Cognitive Psychology and Neuroscience . . . 204
6.1.3 Cognitive Mechanisms . . . 206
6.1.4 Computational Cognitive Neuroscience . . . 209
6.1.5 Computational Models as Models of Mechanisms . . . 210

6.2 Prospects for Integration 211
6.3 Further Directions 214
BIBLIOGRAPHY 217

LIST OF FIGURES

2.1 Additive factors method 30
2.2 Communication theory schematic 32
2.3 Broadbent's information flow chart of the filter model of attention 33
2.4 Control flow diagram 35
2.5 Lichtheim's schematic of aphasia lesion sites 38
2.6 Early program flowchart 39
2.7 ISO/IEC/IEEE Control flow diagram 41
2.8 ISO/IEC/IEEE Data flow diagram 42
2.9 Example of a mechanism schema 44
2.10 Early schematic of the $Na+$ channel 51
2.11 Possible mechanisms for channel gating 52
2.12 Sodium channel inactivation gate 53
2.13 Comparison of Broadbent and Treisman's filter models 56
2.14 Shiffrin and Schneider's filter model 58
2.15 Schneider and Chein's attention model 60
4.1 Levels in an information processing theory of human thinking 141
4.2 Schematics of real and artificial neurons 146

PREFACE

Writing a dissertation is a little bit like taking a trip to hell and back: not just because it can be dark and frightening not just because it can make you regret every single bad decision you've ever made (as well as some of the good ones); but also because you spend a lot of time getting to know dead people.

I first realized this when I was trying to figure out what Jerome Lettvin meant by a process model. After poring over the passage for a long time, but still unable to make sense of it, I decided that maybe I could email him. It turned out he had died just a few months earlier. Procrastination had made it impossible to ever have that conversation. Then David Rumelhart, whose PDP book and past-tense learner are so central to Chapter 4 died. I started to keep a list. Ulric Neisser was next, then George Miller died shortly after.

Jim Bogen, who on many occasions during this process provided the useful service of scaring the shit out of me, joked (or threatened?) that if I didn't hurry up, my committee members would start to die before I finished writing. Peter Machamer even gave us a bit of a scare. I've been given tons (and tonnes) of support and good advice from Jim, Peter, and Ken Schaffner, most of which I've squandered and ignored to my detriment. Time and time again I've failed to follow their advice, then eventually come around to the realization that they were right. Thanks to Floh Thiels for joining the committee at a moment's notice, and asking difficult, important questions.

I've been incredibly lucky to have predoctoral fellowships from the Max Planck Institute for the History of Science, where I wrote the first draft of Chapter 2; and the Werner Reichardt Centre for Integrative Neuroscience, where I finally committed Chapters 3 and 5 to paper. Hong Yu Wong has been extremely patient.

I'm thankful to the friends and family who have generously welcomed me into their guest

rooms on numerous occasions, and stored the detritus of my existence in their basements. The supportive and collegial HPS community at Pitt made these years a pleasure, and an enlightening one. The crazies across the hall made it fun. My apologies go to the little people who got dragged on this trip with me; Mila knows nothing else, and Jesko has proven extremely resilient throughout the peregrinations of his böse Stiefmutter. A special thanks goes to my wife, Boris, who has been an enormous help with proofreading, illustrations, childcare, and (occasionally) cleaning.

I'm very much ready to reemerge into the land of the living.

1.0 INTRODUCTION

This dissertation is about the relationship between cognitive psychology and neuroscience, and between their subject matter: cognition and the brain. It mainly concerns the theoretical side of these sciences, that is, the apparatus they use for explaining. I hesitate to call these apparatus 'theories' at this point, because one of the questions I try to answer here is what kinds of explanatory apparatus are used in these fields.[1] They will be a sort of thing that gets represented in flow-charts and schematic diagrams as often as in equations. I will examine whether the explanatory apparatus of cognitive psychology and neuroscience are compatible with one another, such that they might be *integrated*, and if so, how. I am also concerned with the experimental practice of these sciences, both insofar as experiment is inextricably intertwined with theoretical work, and by way of illustration of how the explanatory apparatus I'll be discussing are discovered and used.

1.1 BACKGROUND

The relationship between mind and brain is one of philosophy's 'big questions.' Some philosophers consider the mind to be something to be examined with a different set of tools than the body, so the study of the mind is sometimes thought to be methodologically, and perhaps also ontologically, separate from the natural sciences. I will try to sidestep these issues

[1]Theory has traditionally been understood in history and philosophy of science as sets of linguistic propositions in the form of universal quantifications and observation statements. More recently accounts take theories to be other kinds of mathematical structures like set theoretic ones. It is safe to say that neither contemporary psychology nor neuroscience make much if any use of theory in any of these technical senses. For the sake of convenience I will nevertheless sometimes call the explanatory/predictive apparatus used in these fields theories, where their structure is not at issue. They might more appropriately be called models.

by taking psychology to be the science of cognition rather than the science of the mind. Whether this is a mere linguistic distinction or one of more substance I leave open. At the very least cognition has a much narrower set of potentially distracting connotations than mind. Philosophers have to a lesser, but still significant, extent discussed the relationship between psychology and neuroscience. One major question is whether and how the science of psychology might and should be *reduced* to neuroscience, and more recently, whether there might be non-reductive alternatives that still manage to integrate or *unify* these sciences in some sense.

Several branches of science try to bridge the gap between cognition and the brain by forging empirical links. *Neuropsychology* typically studies patients with brain lesions (whether hereditary, due to injury, or stroke) and correlates the cognitive deficits these patients have with the brain regions damaged. Based on these correlations, rough characterizations are made of both the subprocesses that make up normal cognitive behavior, and the brain regions involved in these processes. Brain lesions rarely affect only one neat functional region of brain, tend to happen in some regions more than others, and the patients are not numerous enough for researchers to be able to pick and choose patients with lesions in exactly their area of interest, so these methods are rather limited.

Cognitive science is also concerned with cognition and how it arises, although typically it does not involve direct investigations of the brain. Cognitive science takes an *information-processing* approach, where cognition is assumed to consist of computations operating over representational structures. It is historically tied to the field of *artificial intelligence (AI)*, as well as computational linguistics. Later developments in cognitive science brought it in closer contact with neuroscience, particularly with the rise of *connectionist* AI, which eschews logical computation over symbols in favor of computing with networks of simple neuron-like nodes. Chapter 4 will discuss this aspect of cognitive science.

More direct investigations of or interventions on brain cells and circuits are possible using electrophysiological methods like intracellular patch-clamp analyses, extracellular microelectrode recordings, or stimulation-recording experiments, where electrical stimulation or drugs are applied to an afferent cell and recording is done from the target. These cellular studies are often done in vitro, using cultured cells, harvested cells, or preserved slices of

brain, but may also be done in vivo on anaesthetized animals. These can rarely be done on humans for ethical reasons; during brain surgery to remove tumors or to treat epilepsy, some poking around is possible. These cellular studies are thus nearly always done on model organisms like the sea snail *Aplysia*, mice, rats, or macaque monkeys. Behavioral electrophysiology studies may also be done on awake, behaving monkeys with multi-unit recording devices surgically implanted inside their skulls. Inferences to human brains and behavior are obviously somewhat indirect, particularly for higher cognitive functions.

1.1.1 Cognitive Neuroscience

With the introduction of functional neuroimaging methods like Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI), more detailed investigations of and interventions on normal human brains became possible. The adaption of cognitive psychology's subtraction method for use with neuroimaging in the 1980s (Petersen et al. 1988, Posner et al. 1988) coincided (thanks to the McDonnell Foundation's sponsorship) with the rise of the field of *cognitive neuroscience*. fMRI subtraction studies remain the defining method of the field, although recent advances in neuroimaging methods have introduced much more sophisticated techniques too (see, for example, Friston (2005), Buzsáki (2006), Sporns (2011)). Behavioral studies, as well as measures of event-related-potentials with electroencephalography (EEG) and magnetoencephalography (MEG) also provide data for cognitive neuroscience.

While other fields like neuropsychology and cognitive science are also concerned with the relationship between cognition and the brain, cognitive neuroscience takes it as its main aim to forge such connections. There are two common sorts of locutions that cognitive neuroscientists use when describing their field. One is to say that they aim to *explain* psychological phenomena in terms of neural phenomena (or substrates, underpinnings, structures, events). The other is that they aim to combine or integrate cognitive psychology and neuroscience. Posner says his initial motivation for his early work in neuroimaging was to test the hypothesis that "mental operations as studied in cognitive science were localized in separate brain areas" (Posner 1994). Gazzaniga's popular cognitive neuroscience textbook is subtitled "The

Biology of the Mind." We might combine these two sorts of descriptions and say that the aim of cognitive neuroscience is to integrate cognitive psychology and neuroscience in such a way that the neurobiological details explain cognitive phenomena. The original hypothesis was that these explanations proceed by localizing cognitive operations in brain areas, although most researchers now agree that matters are more complicated than this.

1.2 METHODS AND MOTIVATIONS

My aim is an analysis of cognitive neuroscience both in terms of theoretical foundations and methodology. This involves investigating what integration is, how integration works in general, and whether integrating cognitive psychology and neuroscience is a viable project. Several strategies are required for answering these questions.

One is to look at the practices and the methodological claims of scientists in cognitive neuroscience, and the fields it seeks to integrate. From these practices and methodological claims, we can glean a picture of how integration is being approached. Scientists' claims about what they are up to are sometimes difficult to interpret in unambiguous ways. I generally assume that scientists are doing something sensible, and know what they're doing both in that sense and in that they know that what they're doing is sensible, i.e., it's not just sensible by accident. But their job is to do the science, not to describe and justify their methods in ways that philosophers can easily understand, so some interpretive work is also required. Scientists descriptions of their methods are nevertheless essential sources of information.

Another strategy is to give an analysis of the explanatory frameworks employed in the target fields, and to work out what some viable approaches to integration might be. This involves figuring out how the actual approaches used by scientists might be understood as rational ones. In other words, we can look at integrative practices, and do something like a rational reconstruction of them. The goal would be a normative picture of how to approach integration.

Finally, looking at where scientific practice does not match up with either the stated aims

of scientists or with the analysis of how integration might be achieved, reveals whether the approach is working, and the challenges that still need to be overcome. Thus three modes of investigation are intertwined in this endeavor: descriptive, evaluative, and analytical. I am neither merely describing scientific practice, nor merely critiquing it or prescribing how it should be done. Instead I am working out what a reasonable approach to integration could be, given the approach scientists are actually using, and my analysis of what should be expected to work. This bears some similarity to Darden's "compiled hindsight" which she describes as "advisory, but neither descriptive nor prescriptive" (Darden 2002), although since the historical case I'm investigating is still ongoing, foresight is involved in addition to hindsight, insofar as that is possible.

The motivations for this project are several. First, cognitive neuroscience is a large and growing field in which a huge amount of money is being invested, so it would seem wise to clearly understand what is being accomplished. It is not just the potential for wasted effort and money that is at stake. Cognitive neuroscience includes research on addiction, depression, attention deficit hyperactivity disorder, schizophrenia, Alzheimer's, and many other debilitating syndromes which deeply affect the well-being of many people. This research is also applied in the engineering of prostheses, the design of interfaces, and in education. The field claims to be well on its way towards integrating knowledge of the workings of the mind with knowledge about the biology of the brain, which has implications not only in philosophy of mind, but also in ethics, politics, and the law.

Despite its importance and prominence, there is no clear articulation of the theoretical framework organizing this field of research. Cognitive neuroscience uses an amalgam of techniques drawn from several related fields, and very little methodological analysis has been done justifying the deployment of these techniques in new contexts for novel purposes. Without a clear theoretical foundation, it is difficult to distinguish good from bad work in cognitive neuroscience, and to make suggestions for improving its methodology.

This lack of theoretical foundation is further complicated because cognitive neuroscientists often use mental and intentional language in metaphorical ways. As a result, researchers in different labs may misunderstand the theories being put forward by their colleagues, and people outside the field may misconstrue the aims and accomplishments of the field as a

whole. Bennett and Hacker (2003), for example, dismiss much work in neuroscience as nonsense fueled by conceptual confusion. Dennett (2005) argues hilariously that they have failed to understand the ways in which neuroscientists extend the use of ordinary psychological language. There is also evidence that the claims of neuroscientists are uncritically accepted by the general public. In a study by Weisberg et al. (2008), it was found that non-experts judged explanations to be better when they included irrelevant neuroscience information. That it is tricky to properly interpret the claims of neuroscientists is all the more reason to carefully examine this field.

1.3 INTEGRATION

As mentioned already, the goal of cognitive neuroscience is an integration of cognitive psychology with neuroscience. Integration is almost always left an undefined, unanalyzed term, so a few words need to be said about what it might mean.

Integration suggests intermixing, combination, coordination, linking, merging, uniting, or blending. The main connotation is a coming together, but less clear is whether the things being integrated have equal status, so that the combination shares aspects of both, or whether one is integrated *into* the other. Perhaps the popularity of the term integration is in part due to this ambiguity. By analogy to immigration policies, the integration could either be like multiculturalism, or like a melting pot. If the integration aimed at in cognitive neuroscience were like a melting pot, then clearly psychology would be the one being thrown into the cauldron. I don't think this is the goal of cognitive neuroscience, or at least it shouldn't be. The sort of integration I aim to understand and cultivate is one that encourages all parties to maintain their (scientific) cultures while finding ways of getting along.

Another ambiguity is whether the targets of the integration are the fields[2] or their theories. Insofar as cognitive neuroscience is a field, the integration is at least in part at the level of fields. There are several ways fields might be integrated, which I'll mention shortly.

[2]Neuroscience is more like a collection of closely related fields than one field, and similar questions about unity and integration within neuroscience are worth addressing (see Sullivan (2009)), but I won't deal with that set of problems here.

Psychology and neuroscience departments in universities are sometimes merged (not always for the sake of integration), or made part of the same interdisciplinary centers. Making people work in the same space is a first step towards fostering communication, but is no guarantee that any meaningful interactions will result. Psychologists and neuroscientists working together on interdisciplinary teams would be the next step towards integration.

1.3.1 Autonomy

Some psychologists are very resistant to the idea of close collaboration with neuroscience and deny that their theories should have anything to do with neuroscience. Coltheart, for example, notes that "Rather a lot of people believe that you can't learn anything about cognition from studying the brain" (Coltheart 2006), then provides extended quotes from seven such people. The last of these quotes is from Fodor, who argues for the *autonomy* of psychology. In several publications (Fodor 1968a,b, 1974) he argues that mental states are constituted by their functional role, and that these functional roles can be realized by many different physical states, thus psychology is autonomous from the study of any realizations of mental states, such as neural ones. He does not deny that neuroscience could describe how any particular token psychological state is instantiated in the brain. What he means by autonomy is a denial that descriptions of neural states can tell us anything informative about psychological states. This type of autonomy seems incompatible with the goals of cognitive neuroscience, since the point of cognitive neuroscience is for psychology and neuroscience to inform one another.

Other types of autonomy are more compatible with integration. Aizawa and Gillett (2011) argue for a kind of autonomy where psychologists take account of neuroscience, but determine how they do so based on their own theoretical needs. In particular, discoveries of differences in neural realizers only sometimes lead psychologists to make changes to their concepts and theories; other times, they treat differences in neural realizers as "individual differences" (Aizawa and Gillett 2011). Feest (2003) argues that psychology can be explanatorily but not methodologically autonomous, meaning that psychological explanations can maintain their power without this serving as a justification for psychologists to ignore find-

ings from neuroscience when formulating their theories. Both of these types of autonomy maintain a place for psychological theories, and for psychology's needs to be the determining factor in developing those theories, even while knowledge from neuroscience informs those theories.

1.3.1.1 Marr's Levels Defenses of psychology's autonomy inevitably refer to Marr for justification. Marr's (1982) book has proved extremely influential, not just in computer vision—the topic of the book—but in cognitive science generally. It is hard to find work touching on methodology that does not cite Marr. The introduction and first chapter of the book argue that an "information-processing point of view" is useful for understanding human perception. Marr describes three levels of description that are each important to consider when approaching a problem like vision. These are the computational level, which specifies the task or function being performed; the algorithmic level, which corresponds to the particular rules that are followed in order to map inputs to outputs; and the implementation level, which corresponds to the physical realization of the algorithm, or the hardware used to carry it out. He argues that the computational level of explanation was neglected in early studies of vision and psychophysics. Rather than as a correction of this neglect, which I take it to be, many take Marr's emphasis on the computational level as an endorsement of an approach where the computational level is studied in isolation, or autonomously from studies of the other levels.

Marr never claims that his levels are *entirely* independent of one another; he says that all three levels are loosely coupled (Marr 1982, 25) and that "the nature of the computations that underlie perception depends *more* upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented" (Marr 1982, 27, emphasis mine). It is often forgotten that what Marr is setting out is a method designed for particular sorts of situations. The legend under his table outlining the three levels reads, "The three levels at which any machine carrying out an information-processing task must be understood" (25). His arguments are only aimed at *machines*, and only ones carrying out *information-processing tasks*. From all indications he means both of these literally. By information processing he means something like how the visual system takes

input from the retina and processes it to produce information about the objects in the visual field. His point about the hardware being largely irrelevant to the algorithm being computed is clearly true for computers, which are usually designed to be multi-purpose platforms, but is not obviously true of biological 'machines' which are not designed to be multi-purpose platforms. In biological organisms the algorithms that can be performed are much more closely coupled to the implementation details. It seems to be an empirical question whether the computational level is independent of the other two, and one whose answer depends on the particular system of interest. His general point is that any information processing task must be studied at all three levels of analysis in order to get a full understanding of the phenomenon. If one is going to take Marr's advice, then the thing to do in most cases is to investigate all three levels of the phenomenon of interest, unless there is a compelling reason to think that the explanation is complete without considering all three.

A further confusion that a superficial reading of Marr seems to invite is to conflate his levels of analysis with what are sometimes called levels of nature, which form the subject matter of different fields of science. References to Marr in cognitive science are often accompanied by the claim that neuroscience studies the implementation level, with the implication that psychology studies either the computational or algorithmic level. In Marr's conception, all three of his levels should apply to a given phenomenon as it is studied in any field of science. This seems to be what scientists do in both psychology and neuroscience; in both fields problems are defined, and solutions to them are investigated both in abstract terms and (perhaps to a greater extent in neuroscience) in terms of specific implementations. It is an empirical question whether the phenomena studied by psychology and neuroscience are connected in some way, and if so what the nature of the connection is. In Chapter 3 I will discuss other sorts of hierarchies that might better convey the relationship between psychology and neuroscience than the idea that they occupy different levels in Marr's hierarchy.

Some manner of autonomy for both fields is desirable in integration, but not at the expense of there being no possibility for psychology and neuroscience to mutually influence each other. Fodor's version of autonomy, and the common misunderstanding of Marr's levels of analysis are both incompatible with integrationist goals, but the conceptual autonomy of Aizawa and Gillett, the explanatory autonomy of Feest, and Marr's idea that all three of his

levels should be studied are all compatible with the goal of integration.

1.3.2 Reductionism

In addition to some manner of autonomy for each field, integration requires figuring out ways in which the fields are related. For meaningful integration, there needs to be more than just mixing of departments, colleagues, projects, or methods. There needs to be integration at the level of their theories and concepts too.

An approach to relating psychological and neural theories with a long tradition in philosophy of science is reductionism. Working within the framework of deductive-nomological explanation, Nagel (1961) developed a model of theory reduction, where the theory of one science could be reduced to another theory, and thereby explained. This could be done, according to Nagel, by deriving all of the laws of the reduced theory from those of the reducing theory. This might require bridge principles that essentially translate between the vocabularies of the two sciences. Oppenheim and Putnam (1958) offered a version of reductionism where all the branches of science fit into a neat hierarchy going from social groups down to elementary particles. The goal of *unification* would have each level reduced to the one below it, ending at a fundamental level. For a science to be reduced, it must (1) contain terms not in the vocabulary of the reducing science, (2) all observations explainable in it must also be explainable in the reducing science, and (3) the reducing science must be at least as "well systematized" as the target (Oppenheim and Putnam 1958).

Schaffner (1977, 1993) extended Nagel's model to biology, where laws are rarely exceptionless, and reductions thus do not have the neat deductive character of Nagel's reductions. In later papers Schaffner emphasizes that in most cases we get "fragmentary and partial explanations of parts of a discipline, but not any type of overall sweeping reduction" (Schaffner 2006). While entire theories can only be reduced in very rare cases, particular parts and processes can be explained causal-mechanically in terms of their decomposition into sub-parts or processes and the connection of those with parts or processes in the lower-level theory (Schaffner 2006, 385-386).

Bickle (2006) claims that researchers in cellular and molecular neuroscience are achieving

direct reductions of social and cognitive behaviors down to the cellular or molecular level. Examples where researchers intervene on cells, then record behavioral changes are one sort of evidence he cites. The other sort of evidence is that neurobiologists like Kandel et al. (2000) claim to be able to link brain events with behavior. What Bickle calls "ruthless reduction" refers to researchers finding correlations between cellular or molecular events and operationalized behavioral measures, and building explanations based on these connections. Bickle is dismissive both of psychological theories and of cognitive neuroscience, although it is not clear that what he means by reduction is incompatible with integration. Finding correlations between scientific entities or concepts could be part of an integrative strategy.

Some neuroscientists describe themselves as taking a reductionist approach. Kandel, for example, describes his strategy as "radically reductionist" (Kandel 2001). The way he illustrates this is by describing how he studied learning in the sea snail *Aplysia*, despite the fact that "In the 1950s and 1960s many biologists and most psychologists believed that learning was the one area of biology in which the use of simple animal models, particularly invertebrate ones, was least likely to succeed" (Kandel 2001). He persisted, he says, because he thought there must be "conserved features in the mechanisms of learning at the cell and molecular level that can be studied effectively even in simple invertebrate animals" (Kandel 2001). What he means by radical reductionism then seems to be partly the methodological decision to study simple systems as an initial strategy for investigating complex phenomena, and partly the empirical assumption that even animals only distantly related evolutionarily will share many features at the cellular and molecular level. Neither of these is reductionism in any of the senses outlined above, and both seem compatible with the integrative goals of cognitive neuroscience. Near the end of the same paper, Kandel remarks that, "we have here only considered the molecular mechanisms of memory storage. The more difficult part of memory—especially explicit memory—is a systems problem." He goes on to say that "These systems problems will require more than the bottoms-up approach of molecular biology. They will also require the top-down approaches of cognitive psychology, neurology, and psychiatry" (Kandel 2001). This is an explicit denial of both Bickle's sweeping reductionism and deductive accounts of reduction. If anything it is an endorsement of integration.

These later reflections of Kandel's seem consistent with the approach he took in his

(1976) book *Cellular Basis of Behavior*. Although the title suggests that the book seeks to explain behavior in terms of cellular activities, which sounds like sweeping reduction, he clearly states that he means "behavior" in a restricted, operational sense, "As now used by most psychologists, the term behavior refers only to a restricted aspect of mental life, that which is observable" (Kandel 1976, 41). He talks of reduction in two ways. One is to describe how in this study, behavior is "reduced" to its observable components (Kandel 1976, 30), since he lacks experimental methods to dig any deeper. The other is the sense mentioned above of a "reductive approach" (Kandel 1976, 39) being one where simple systems are studied in the hopes that the results will also illuminate more complex cases. Although at times it sounds like Kandel is a reductionist in Bickle's sense, upon closer inspection he is at most only a reductionist in the sense of the later Schaffner papers, and perhaps not even in that sense, since it is not behavior, but just a restricted operational version of it that he is finding mechanisms to explain. Kandel's reductionism is not worrisome for cognitive neuroscience's integrative project, and the methodological decision to study simple organisms is in no way a threat to the autonomy of psychology. Kandel's practice of only dealing with restricted operationalized aspects of behaviour may nevertheless act as a practical barrier to integration, in that this approach does not lend itself easily to communication with psychologists.

One legitimate worry that Kandel's version of reduction raises is the possibility of misreading work on "behavior" as being about behavior. Neuroscientists regularly use this sort of narrow operational definition, and for good reason, but at least in some contexts, like articles in the popular press and grant proposals, it is not made clear that their work does not bear directly on the phenomena psychologists study under the same names. It's also probable that some less well-informed neuroscientists don't realize that there is a difference between their concepts and those of psychologists, or don't think much about those differences. It is not so uncommon to hear neuroscientists making light of psychological concepts. This naturally raises the ire of psychologists, and also of philosophers of mind (see Burge's (2010) tirade about "neurobabble"), which does not help the cause of integration.

Nagel, Oppenheim & Putnam, and perhaps also Bickle's versions of reductionism are incompatible with the sort of integration I'm after. Eliminating psychology by deriving all

of its claims from neural ones is neither a goal likely to succeed, nor one that lives up to the multiculturalism ideal of integration. Schaffner's later work and Kandel's radical reductionism are not incompatible with integration, but rather assume it in some sense as part of the goal of science. Schaffner's idea of causal-mechanical explanations that explain some parts and processes of (in this case) psychology in terms of their decomposition into neural parts and processes is the sort of thing cognitive neuroscientists are after. The combination of top-down and bottom-up methodologies Kandel envisions is likewise congenial to cognitive neuroscience's aims.

This is only a very brief review of some of the reductionist positions taken up. A much more detailed review can be found in Chapter 9 of Schaffner (1993).

1.3.3 Emergence

At the other end of the spectrum from reduction is emergence, which, very roughly, is the claim that novel phenomena arise at various levels of complexity. There are a number of variants of emergence; the novelty could mean irreducible, unpredictable, unexplainable, novel entities, novel activities or behaviors, or novel properties. In addition, the unpredictability or unexplainability could be in principle or in practice.

Some of these variants seem to express basic assumptions of integration, while others seem to be incompatible with it. That there are properties, entities, or activities that show up only when there is sufficient complexity is quite compatible with the separate existence of the sciences of psychology and neuroscience, and the intention of leaving these more-or-less intact in their integration. It amounts to claiming that each field has its own subject matter. On the other hand, the integration cognitive neuroscientists are after is one where neural facts explain psychological ones, and better yet, can eventually be used to predict them. This is incompatible with variants of emergence where the emergent phenomena can't be explained or predicted by appeal to the lower level. A way of making even this sort of variant of emergence compatible with integration might be to agree that many cognitive facts can be explained by neural ones, but there might also be some that are unexplainable and unpredictable at the cognitive level.

Another thing integration and emergence have in common is the idea of there being various levels at which regularities show up. Emergence is a claim about the ontology or epistemology of the world. Integration is more than that. It is a program for how to investigate a multi-level world. So while emergence is in some ways compatible with integration, it does not provide much guidance for it.

Again this is a very brief description. Bedau and Humphreys's (2008) volume includes a useful overview of emergentist positions, as well as articles by the major players.

1.3.4 Unification

Unification is a stronger form of integration involving both epistemological and metaphysical convictions. The epistemological conviction is that using fewer basic laws or phenomena to derive a result contributes explanatory power. I take this to mean not just that when it's possible to use fewer assumptions in an explanation this is preferable, which seems uncontroversial, but furthermore, that sharing assumptions with other explained phenomena is in its own right explanatory. Explanation by unification involves reducing the base of explanantia to a minimum, and was originally defended by Friedman (1974) and Kitcher (1989). Unification in the form of reducing the number of assumptions may be a goal with heuristic value, but it is not necessary for integration.

Metaphysical versions of unification can take many forms, but for our purposes here, it is the conviction that the multiple levels of phenomena from molecules to cells to neural systems to cognitive systems are all part of a unified hierarchy, where the smaller bits are parts of the bigger ones, and the behavior of all of them can in principle be explained with the same set of principles. Craver (2005, 2007) defends a version of unification in neuroscience, which he calls "mosaic unity.' According to Craver, "fields integrate their research by adding constraints on a multilevel description of a mechanism" (Craver 2005). The multilevel hierarchy he describes is one where lower levels are components of the next higher level. Craver says the mechanistic model, "describes integration as occurring in the process of building a single theory" (Craver 2005). The single theory he has in mind, is a unified, multilevel hierarchy of mechanisms. He says this in the context of cellular and molecular neuroscience, but if

we apply this to cognitive neuroscience, integration would place cognitive mechanisms and brain mechanisms in a part-whole relationship, and a unified theory would describe the whole hierarchy.

Both the epistemological and the metaphysical convictions go beyond what I'm looking for in integration. Explanations with fewer assumptions may be preferable, but they do not get their explanatory power thereby. It would be nice if research at multiple levels could converge on a common mechanistic explanation, but it is an empirical matter whether the world is organized such as to make that possible. The integration I'm after should combine results from different levels into common explanations, but perhaps the results will not all fit together into one unified explanation. Partial integrations on the level of theories may be the best we can hope for, as Schaffner suggests.

1.3.5 Constraints on Integration

The general sort of connection between theories that we're looking for in an integration is what Darden and Maull (1977) call an *interfield theory*. An interfield theory relates the subject matter of two fields, typically when the fields study different aspects of the same phenomenon. Interfield theories can take several forms: "A field may provide a *specification of the physical location* of an entity or process postulated in another field ... provide the *physical nature* of an entity or process postulated in another field ... investigate the *structure* of entities or processes, the function of which is investigated in another field ... [or they] may be linked *causally*, the entities postulated in one field providing the causes of effects investigated in the other" (Darden and Maull 1977).

These alternatives look quite compatible with the claims cognitive neuroscientists make about cognitive functions being located in the brain, performed by structures in the brain, or caused by brain processes. It might be helpful to be more specific about which of these forms (if any) the integration of cognitive psychology and neuroscience is supposed to take. Are we looking for the location in the brain of cognitive modules? Are we looking for the part of the brain that contains cognitive modules? Are we looking for the brain structures that perform cognitive functions? Are we looking for causes in the brain of cognitive effects?

These are distinct possibilities, so the goals of integration need to be made more specific.

At this point, we can gather together a wish list of features the integration of psychology and neuroscience should have.

1. The fields being integrated should both maintain autonomy over their concepts and explanations.
2. The theories of both fields should be maintained, where possible, with neither being reduced to the other.
3. Researchers in both fields should be prepared to alter their methods, concepts, and theories, when input from the other field reveals problems.
4. The theories should be connected with an interfield theory.

4b. One field should specify the physical location, physical nature, structure, or causes of phenomena from the other field.

5. Results from each field should be considered as constraints on theories in the other.
6. The current best practices of scientists in cognitive psychology, neuroscience, and cognitive neuroscience should be reflected in the integration.

It bears emphasizing that it is not just psychology that needs to accommodate input from neuroscience, but also the reverse.

1.4 INTEGRATION THROUGH MECHANISTIC EXPLANATION

A promising recent proposal for integration in cognitive neuroscience, which has the potential to meet the criteria outlined in the previous section, is integration through mechanistic explanation. It seems to meet many of the criteria on the wish list. 1. Mechanistic explanations are typically multi level, so that the fields associated with the various levels maintain some autonomy. 2. Instead of theories made up of law-like statements, mechanisms do the explaining, and again since these are multi level, each field has its corresponding mechanisms. 4. The connected hierarchy of mechanisms is an interfield theory of sorts. 4b. Lower levels in the hierarchy are components of the higher levels. According to Craver, what it means

to integrate levels is "showing that an item is a component in a mechanism and describing its role in that mechanism... [or] detailing lower-level mechanisms for a phenomenon" (Craver 2005). 5. Each level of the hierarchy mutually constrains and is constrained by its neighboring levels. As for 3. and 6., these depend on the practices of the scientists involved.

Mechanistic explanation has been variously described, but the three definitions with the most currency are the following. Bechtel and Richardson (1993) define mechanistic explanations as explanations that "account for the behavior of a system in terms of the functions performed by its parts and the interactions between these parts" (Bechtel and Richardson 1993), and introduce *decomposition* and *localization* as the main strategies for developing these explanations. Bechtel and Abrahamsen (2009) define a mechanism as "a structure performing a function in virtue of its component parts, component operations, and their organization." Glennan says, "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations" (Glennan 2002). Machamer et al. (2000) (henceforth MDC) say that "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." More detail about this last definition will be given in Chapter 3.

Several authors including Bechtel (2008), Piccinini and Craver (2011), and Kaplan (2011) have claimed that mechanistic explanation is the key to integration in cognitive neuroscience. Bechtel's proposal for integration leans heavily on the decomposition and localization heuristics he developed with Richardson in their (1993) book, and elaborated in several later publications. Functional decomposition is a top-down strategy where you "start with the overall functioning or behavior of the mechanism system and figure out what lower-level operations contribute to achieving it" (Bechtel and Abrahamsen 2005). The complementary strategy of structural decomposition is a bottom-up strategy where you seek to decompose the system into working parts that "perform the operations that figure in the functional decomposition" (Bechtel and Abrahamsen 2005). The parts and functions arrived at through these two sorts of decomposition need to fit together neatly for a successful mechanistic explanation. Localization is the linking together of working parts from a structural decomposition with the

operations from a functional decomposition by identifying the operations as those that the parts are supposed to perform (Bechtel and Richardson 1993). Bechtel argues that in cognitive neuroscience, functional decompositions stemming from cognitive science and structural decompositions stemming from neuroscience are integrated by localizing cognitive functions in neural parts. One of the case studies he describes in Bechtel (2008) will be critically examined in Section 2.5.1.

In fact typical projects in cognitive neuroscience do make use of Bechtel and Richardson's localization heuristic, by measuring correlations between performance of cognitive tasks and activation of brain regions with neuroimaging technology, then concluding that the cognitive tasks are performed in those regions. Posner and DiGirolamo (2000) provide a thorough insider view of the background, methodology, and accomplishments of this line of cognitive neuroscience work. As they note,

> A major achievement of brain imaging studies has been consistent localization of brain areas that perform particular functions. This achievement has gone a long way toward providing validation both of the decomposition of skills and of the mapping of these skills onto particular brain networks (Posner and DiGirolamo 2000).

A major limitation of this as a recipe for integration is that, used in isolation, localization results are not very informative. Subtracting the neuroimages of any two tasks is almost guaranteed to get some results appearing to show that the cognitive task being targeted is performed by the regions of highest activation, even if the tasks are poorly chosen, such that they do not represent valid constructs. Subtraction results alone do not validate task decompositions, do not reveal what operations really go on in those brain regions, and do not explain the behavior. One needs independent validation of the task decomposition from cognitive models, plus an understanding of the functional anatomy of the regions. More sophisticated neuroimaging methods do better than simple subtraction, but they do not get around the need for this further information. Neuroimaging is thus just one tool among many that can help in the discovery of mechanisms that explain cognitive phenomena. Localization is not the whole story.

Piccinini and Craver's proposal is similar to Bechtel's in some respects, but does not depend as much on decomposition and localization as the way to integration. They propose that the functional analyses characteristic of explanation in psychology are "elliptical

mechanism sketches" (Piccinini and Craver 2011). Mechanism sketches are defined in MDC as gappy, abstract representations of mechanisms that serve as guides for further research in the early stages of the development of a research project. Piccinini and Craver suggest that cognitive models are like rough drafts that can be filled in with neural details as they become available. These details that contribute to a fully-elaborated mechanism may come from many sources. For them, the distinction between a cognitive model and an integrated neural one is a matter of how fully developed the model is.

Integration through mechanisms is an interesting new proposal with a lot of apparent potential, but Piccinini and Craver's proposal is too quick and easy, so more work may be required to flesh the idea out. Furthermore, since there is considerable disagreement over what a mechanism is, further clarification may be required to make the proposal a definite one. Discussing and refining this proposal will occupy much of the chapters that follow.

1.5 OUTLINE

In Chapter 2 I discuss in detail Piccinini and Craver's proposal for how mechanistic explanation can provide a scaffolding for the integration of cognitive psychology and neuroscience. I begin by characterizing these two fields, focusing on whether cognitive psychology can be considered to be mechanistic in the way Piccinini and Craver require. This involves analyzing what an information-processing model is in the context of cognitive psychology, distinguishing process flow models from data flow models, and comparing these to the sorts of information-processing models found in computer science. I characterize types of information-processing models with the help of the ISO/IEC/IEEE standards for diagrams. Since the claim that cognitive models are mechanism sketches is a claim about the developmental stage of those models, I look at how theories are developed in cognitive psychology and neuroscience. I compare two models—one of attention, and one of sodium channels—each of which is referred to as a 'filter' and show that the ways theories are developed in psychology is not, as Piccinini and Craver describe, a gradual accumulation of mechanistic details, but rather the refinement of a model that remains abstract throughout. Finally I

draw some preliminary conclusions about the prospects for integration. The suggestion that psychology's models are elliptical sketches of neural mechanisms does not seem promising, but this does not mean that they aren't mechanistic in some other sense, nor that some other way of integrating isn't possible.

In Chapter 3 I work towards an alternative proposal for how integration in cognitive neuroscience might be achieved. The main goal of the chapter is to develop resources from within the tradition of mechanistic explanation that allow cognitive models to be seen as a species of mechanistic model. The first steps are to give an overview of the MDC account of mechanism, including more recent clarifications of, alterations to, and criticisms of the account. I then attempt to clarify and extend some aspects of the account that are relevant to integration. I discuss the issue of whether the explanatory power of generalizations can be captured in mechanistic terms, and introduce the idea of a generic mechanism as an ontic counterpart to mechanism schemas, using lateral inhibition as an illustration. A generic mechanism, like any other mechanism, is a thing in the world, but a thing in the world insofar as it belongs to a type, as opposed to a thing in the world in all its gory detail. I suggest that generic mechanisms might be a good way to understand how explanatory power can be had by the sorts of cognitive models that are and remain quite high level. I also argue that integration might take the form of multiple hierarchies of mechanisms that do not necessarily all connect together into one unified hierarchy. This allows for the possibility that some cognitive models might be neatly identified with neural components, while others do not match neatly onto the same decompositions.

In Chapter 4 I explore a poorly understood but ubiquitous practice in cognitive neuroscience: computational modeling. I pursue the joint aims of providing a clearer picture of the role of computational modeling in the cognitive sciences, and of illustrating my claims from the previous chapter about the explanatory value of generalization. Two longstanding confusions over computational modeling are what role simplification plays, and how computational models relate to their target systems. To clarify the role of computational modeling, I review the motivations researchers have for using this set of techniques, and two of the most common types of approach: classical and connectionist AI. I examine an extended debate about the relative merits of these two styles of AI from which I pick out four types of claims

made about the role of connectionist models. These are the following: theories of cognition, implementations of classical theories, neural simulations, and mathematical abstractions. Connectionist models aim at being neither realistically detailed, nor entirely abstract. This is a puzzle. One reason for the mixed claims about how connectionist models might work is that they are used for various purposes in the discovery and elaboration of mechanisms. I illustrate this with an example where connectionist models are used for several different but complementary purposes in research on rat hippocampus. This doesn't entirely resolve the puzzle though.

In Chapter 5 I take a step back to consider the role of modeling in general, focused on the question of how models that are only vaguely tied to realistic details could be helpful in investigating and explaining real-world target systems. I look beyond representationalist accounts of models, and take them instead to be like stand-ins or exemplars that are also things-in-the-world. I suggest that in many different experimental contexts, models can be understood as approximations to generic mechanisms. The inferences we draw from models, including computational ones, can then be seen to follow familiar reasoning strategies from experimental science, rather than the logic of representations. Typically explanations will require a combination of mechanisms of various specificity that work at different levels, because the causes at work in making a phenomenon occur are likewise multi-level. This analysis leads to a resolution of the puzzle of why connectionist models aim at neither realistic detail nor abstraction.

In Chapter 6 I summarize the results, and return to the problem of how cognitive psychology and neuroscience might be integrated. I propose that one promising route towards integration is to see cognitive models as generic models, instead of as models of a different Marr level. This reflects an emergentist claim of sorts, where difference-makers appear at multiple levels of nature. On this picture, integration is a matter of combining multiple mechanisms of more or less specificity. The examples of connectionist models I discuss in Chapter 4 provide an illustration of how this sort of integrative work might be approached. Finally, I offer brief discussions of some issues that remain unresolved.

2.0 EXPLANATION IN COGNITIVE PSYCHOLOGY AND NEUROSCIENCE

As discussed in Chapter 1, the aim of cognitive neuroscience is to *integrate* cognitive psychology with more basic neuroscience. I have reviewed the available scenarios for how such an integration might work, and the most promising contender is some version of mechanistic explanation. In this chapter I evaluate whether a popular picture of how this integration might be achieved is compatible with how explanation in cognitive psychology actually proceeds and the norms developed in the previous chapter for what integration should mean.

2.0.1 Outline

I first go into some detail about the nature of the integrands: neuroscientific and psychological explanations. I then critique a recent proposal by Piccinini and Craver (2011) for how mechanistic explanation might achieve this integration. I analyze the information-processing models characteristic of cognitive psychology, identifying several distinct types. I characterize the differences between these types via an analysis of the semantics of the flowcharts used to represent them. Next I compare the development of two models, one from neuroscience, the other from cognitive psychology, which are both referred to as 'filter' models. I show that the while the MDC account of mechanisms fits the example from neuroscience extremely well, the example from psychology does not at all proceed from a bare sketch to an ever more detailed mechanistic model. I then describe the sort of mapping operation that cognitive neuroscientists seem to hope will achieve the desired integration between these distinct sorts of models. I argue that what Bechtel (2008) offers as a success case of this sort of integration is in fact not so successful. Rather than acting as a sketch of the neural mechanism eventually

uncovered, psychological models of memory were very misleading guides to neural investigations, and despite being in the end a poor fit to the neural models, the psychological models have remained popular. I end by discussing some other recent approaches to integration. In Chapter 3 I then go on to describe an alternative picture of how cognitive psychology and neuroscience might be integrated, which requires an extension to the mechanistic account.

2.1 EXPLANATION IN NEUROSCIENCE

Despite its sometimes reductionist tendencies, there is fairly good agreement that neuroscience is a mechanistic science (Bechtel and Richardson 1993, Machamer et al. 2000, Craver 2007). What neuroscientists usually do when they try to explain a phenomenon is to search for and elaborate neural mechanisms. Exactly what the nature of these mechanisms is, there is less agreement about (Machamer et al. 2000, Glennan 1996, Craver 2007, Bechtel and Abrahamsen 2005, Bogen 2005).

The idea of mechanism I'll use here is based on the account in Machamer et al. (2000), and subsequent elaborations by Machamer (2004, 2011b), Darden and Craver (2002), Darden (2008), Craver (2005, 2006, 2007); I will also introduce my own extension to this account in Chapter 3. According to this account, the way a mechanism works is explained in terms of the entities making up the mechanism, and the activities they perform. On this account, there are mechanisms operating at many scales, with the entities in larger mechanisms sometimes being made up of smaller mechanisms. Mechanisms at any scale may constitute the appropriate explanation for a given phenomenon, and how far down you go in elaborating the details of the parts depends on how much you know, and how much you need to detail, given the scientific context. One of the goals of neuroscience is to unify these different scales or levels.

This general picture seems to be borne out in a wide variety of examples of neuroscientific work. In the preface to Shepherd's classic (1983) textbook, *Neurobiology*, exactly this sort of framework is described. One of the fundamental concepts of nervous system organization, he says, is that "any given region or system contains successive levels of organization, beginning with ions and molecules, and building up through cells and their circuits to behavior"

(Shepherd 1983, viii). He also points toward the need for greater unity, "Many workers in recent years have studied synaptic properties and circuits and their correlations with simple behaviors; what is still needed is an understanding of how, beginning at the single synapse, one builds up successive levels of synaptic circuits of increasing extent to mediate complex naturally occurring behaviors" (Shepherd 1983, ix).

Neuroscience on the whole seems to be open to some manner of integration with psychology, perhaps because of the historical roots of neuroscience, which was founded with explicitly interdisciplinary aims. Nevertheless, the sort of integration some neuroscientists have in mind is more like a hostile takeover than a peaceful merger. A typical attitude among neuroscientists is represented in the following quote from a chapter of Gazzaniga's (2004) textbook:

> The stark contrast between the specificity of brain function and the generality of many popular psychological constructs is a paradox. Almost every time a biologist posits a relation between two aspects of brain function, or between brain function and a psychological product, other scientists discover that the original claim was too general. The facts of nature force most neuroscientists to be splitters. By contrast, psychiatrists and psychologists tend to be lumpers, preferring the more abstract to more constrained concepts... We suggest that scientists should parse the broad constructs that dominate current psychological theory, for example, reward, fear, intelligence, and memory, into a number of more restricted concepts that are in closer accord with what is known about the brain. This suggestion is not a defense of reductionism but a plea for consistency in the level of specificity in the descriptions of brain and mind (Kagan and Baird 2004).

In short, neuroscientists typically deny being reductionists, and want to bring neural and psychological concepts into contact, but see the mismatch between psychology and neuroscience's concepts as being due to psychological concepts being too abstract or general. The resolution neuroscientists prefer is one where psychologists change their theories and concepts to better fit the neuroscience. This may not be the strongest form of reductionism out there, but it is reductionism just the same.

2.1.1 Integration

Among philosophers of neuroscience, there are widely divergent views about the prospects for integration. Craver (2007) paints a very optimistic picture of "mosaic unity" achieved through multilevel mechanistic explanation. The basic picture is that all the knowledge

we gain about the various levels from systems to circuits, down to cells and molecules will eventually all be pieced together into a cohesive whole. Sullivan (2009) argues that both reductionist aims and Craver's picture of mosaic unity are unrealistic, given the difficulty in getting results even of experiments with nearly identical protocols to fit neatly together. Revonsuo claims that "there is a clear conflict of explanatory strategies and assumptions built into the ingredients of cognitive neuroscience" (Revonsuo 2001). Bechtel, in stark contrast, claims that the integration is unproblematic, because "a common explanatory framework [is] employed in both the cognitive sciences and the neurosciences—the framework of mechanistic explanation" (Bechtel 2001). In a recent paper, Piccinini and Craver (2011) expand on Bechtel's suggestion, arguing that functional analyses are elliptical mechanism sketches, and that cognitive psychology and neuroscience can be "seamlessly integrated with multilevel mechanistic explanations."

I will argue that Piccinini & Craver's proposal for integration, while a fair description of the approach taken in a lot of work in cognitive neuroscience, is not as seamless as they suggest. Piccinini & Craver start from Cummins's (1975, 1983) characterization of explanation in psychology as functional explanation. They break that down into three types—task analysis, functional analysis by internal states, and boxology—then argue, for each type, that "properly constrained" these all amount to sketches of neural mechanisms. (MDC define mechanism sketches as gappy representations of mechanisms, where it is not yet known what all the components are and how they fit together.) Although in many cases these sketches do not contain anything that would count as a neural entity, Piccinini and Craver (2011) count these as "elliptical" mechanism sketches. One quibble I have is that the argument in all three cases relies heavily on the proviso that explanations be 'properly constrained.' What this constraint amounts to, however, is that psychologists "ought to acknowledge that psychological explanations describe aspects of the same multilevel neural mechanism that neuroscientists study. Thus, psychologists ought to let knowledge of neural mechanisms constrain their hypotheses" (Piccinini and Craver 2011). This constraint seems to me suspiciously similar to the hypothesis Piccinini & Craver set out to defend: that psychological models are compatible with and should be integrated into neural mechanisms. Certainly, if psychologists are committed to this sort of project of integration, then they

ought to work in ways that might make such an integration happen.

But are psychologists concerned with letting knowledge of neural mechanisms constrain their hypotheses? Are psychological explanations—as they really are, not as they ought to be for integration to work seamlessly—mechanism sketches in any sense less elliptical than I can claim to have a collection when all I have is a box that I could put one in? Psychologists' reactions to the increasing encroachment on was what was historically their territory by cognitive neuroscience suggest that they have different aims. For example, LeDoux and Hirst (1986) gathered together four "dialogues" between cognitive psychologists and neuroscientists, designed to identify the similarities and differences in their approaches to shared problems, and to bring them into closer contact. In the "psychologist's reply" in the chapter on attention, Hirst complains that the neuroscientists "fail to acknowledge that the processing studied by neuroscientists and cognitive psychologists are of different kinds" (LeDoux and Hirst 1986).

In the remainder of this chapter, I characterize the sorts of explanations used in cognitive psychology, compare these to explanations in neuroscience, and raise two lines of objection to the claim that psychology's models are elliptical mechanism sketches. In Chapter 3 I go on to give a positive account of how integration might occur in a way that is more fair to psychology's scientific aims.

2.2 DEFINING COGNITIVE PSYCHOLOGY

The first step in determining whether cognitive psychology can be accurately described as a mechanistic science, or if not, whether it could be turned into one, is to give a descriptive account of explanation in cognitive psychology. Cognitive psychology is typically described by philosophers as providing functional analyses. Piccinini & Craver take philosophical accounts of explanation in psychology as their starting point, and elaborate on Cummins's (1983) proposal that psychological explanation is functional analysis. They then divide functional analysis into three types which they say have been articulated in greatest detail (again by philosophers).

I take another route and start from a claim that is ubiquitous in scientists' reports of what they are up to in cognitive psychology: that cognition is information processing. I then identify several subtypes of information-processing models that are popular in cognitive psychology, and evaluate whether these can profitably be seen as elliptical sketches of neural mechanisms. As far as I know, Piccinini and Craver's is a perfectly good analysis, and may indeed map directly onto the one I give.

Cognitive Psychology is a huge field though, with researchers on multiple continents, studying a wide variety of topics, and using many different methods, so there is no simple description that accurately covers all of its branches. Further complicating matters, psychology is changing. In many domains, countries, and labs, psychology is increasingly being influenced by neuroscience. In the 2009 edition of Anderson's textbook, previous editions of which rarely mentioned neuroscience at all, Anderson states,

> cognitive psychology is in the midst of a change that may turn out to be as significant as the cognitive revolution... we are now seeing a developing synergy between cognitive neuroscience and information-processing analysis. Neuroscience data can be used to discriminate between alternative information-processing models, and information-processing models can be used to organize neuroscience data (Anderson 2009).

Anderson mentions the merger of cognitive psychology and cognitive neuroscience into one psychology department at his university (CMU) as an illustration of this trend.

Despite these difficulties, an adequate working understanding of cognitive psychology is possible. In order to construct this picture, I will draw from classic works and standard textbooks, but I will also pay attention to the changes that are discernible over time in well-cited papers from various periods and in the subsequent editions of textbooks. There are three main ways in which Cognitive Psychology is typically defined: in terms of the topics it covers, its intellectual provenance, and the characteristic methods used. I elaborate on each of these below.

2.2.1 Cognitive Psychology's Topics

A common approach for defining cognitive psychology, at least as a first pass, is to call it the investigation of 'cognition' and to list a set of topics this covers. The list of topics

typically includes perception, memory, learning, language, attention, thinking, and problem solving. This approach does not get us very far. One problem is that it is not clear why these topics and not others count as cognitive. A more serious problem is that these same topics can be approached in rather different ways some of which would not count as cognitive psychology (a molecular biology approach, for example). Nevertheless, this list of topics is a good starting point. In the sections that follow, I will often focus on just one of these main topics—attention—since it would be impractical to survey the entire field of cognitive psychology.

2.2.2 Cognitive Psychology's Provenance

Another way of defining the field is in terms of its intellectual provenance. Cognitive psychology is often said to have arisen as a reaction against behaviorism. Sources disagree about whether this was a slow-brewing turnover or a sudden revolution in response to Chomsky's 1959 review of Skinner's book *Verbal Behavior*. One element of this disagreement comes from the observation that Behaviorism's dominance was not total. Whether or not it's completely true that cognitive psychology arose in reaction to behaviorism, the contrast between these approaches is a helpful way of distinguishing what counts as cognitive.

The most important difference between the Cognitivist and the behaviorist approach is that while behaviorism views the unobservable machinations of the mind that mediate between stimulus and response as something about which it would be unscientific to speculate—a 'black box' that can't be opened, Cognitivism views these machinations as essential to explaining behavior. Cognitive psychology aims to open the 'black box' and reveal its contents. Cognition then consists in the not-directly-observable stuff that goes on in between.

There are several ways that the black box might be opened, which it will be helpful here to distinguish. One is to analyze parts into subparts. Another is to analyze capacities to perform tasks into subcapacities. Yet another is to analyze processes into subprocesses. Psychological models might do any of these, as I will describe shortly.

2.2.3 Cognitive Psychology's Methods

Cognitive Psychology can also be defined in terms of the methods it uses, which distinguishes it from other fields that study overlapping sets of topics. In a typical cognitive psychology experiment, the goal might be to test for the presence, the properties of, or the relationships between hypothesized cognitive parts, processes or tasks. Experimental tasks are designed to enable a comparison between cases where the hypothesized cognitive part, process or capacity is either present/active or absent/inactive. Performance is compared on the different sorts of trials and any differences are inferred as arising because of the presence/activity or absence/inactivity of the cognitive part, process or capacity being tested.

For example in the original Posner (1980) task, the goal is to determine whether attention can be spatially oriented without eye movement. The task is to fixate the eyes on a central cross, then push a button when a stimulus appears to the one or the other side of the display. The conditions compared are when just the cross is displayed (neutral trial), and when an arrow pointing one direction or the other is displayed, where the arrow correctly (valid trial) or incorrectly (invalid trial) indicates the location where the stimulus will appear (typically on 80% of trials the stimulus is valid). Reaction times are then compared across the neutral, valid cue, and invalid cue trials, with the result that the stimulus is detected more quickly on valid trials, and more slowly on invalid trials compared to neutral trials. This suggests that attention can be covertly oriented, and when oriented correctly, aids detection of stimuli.

Reaction time experiments like this one are arguably the defining method of cognitive psychology. Claxton (1980) claimed, "measuring reaction times is now one of the most popular activities of cognitive psychologists." Donders pioneered the reaction time experiment in 1865 (Donders 1969), demonstrating that the time taken to perform mental processes could be measured. His method involved measuring the time elapsed between the application of a stimulus, like a small shock to the foot, and a response, like pressing a button. When the tasks were gradually complicated, for instance if two stimuli (left and right foot) had to be distinguished, or if two responses (left or right button press) had to be decided between, he found that reaction times were longer. By subtracting the reaction times of the simple case (where nerve conduction was most of what was involved) from the more complex, he arrived

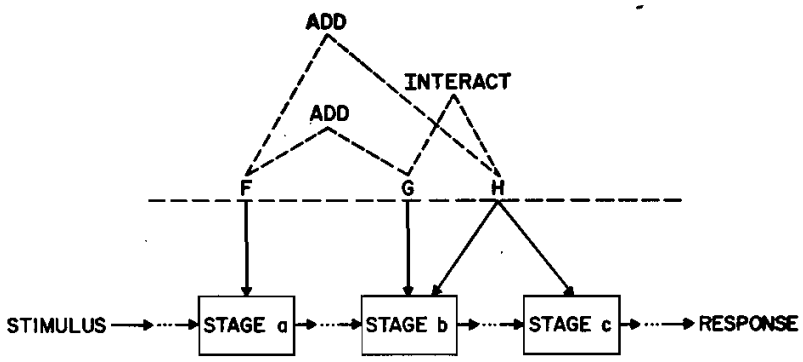

Figure 2.1: Additive factors method. Reprinted from Sternberg (1969) with permission from Elsevier.

at an estimated duration for the interposed mental steps required to distinguish or decide.

Sternberg introduced some improvements on this paradigm, which he called the additive factors method, in his 1969 paper *The Discovery of Processing Stages; Extensions of Donders' Method*. He assumed that in between stimulus and response there are a sequence of stages of processing, and that experimentally manipulated variables, or 'factors,' might influence one or more of the stages. Essentially the improvements were to add tests of the interactions between these experimental factors, in order to determine what the underlying stages might be. An example of this relationship between additive or interacting experimental factors and the underlying mental processes is shown in Figure 2.1.

One thing that is striking about Sternberg's figures is that while several complicated relationships between the factors and the stages are considered, the stages of processing are always sequential. The idea that cognition flows in an ordered sequence is continuous with 19th century psychology, but the idea of processing stages and the representation of these as stages in a flowchart seems to have been borrowed from communications theory and computer technology.

2.3 INFORMATION PROCESSING MODELS

One thing that nearly all cognitive psychologists would agree to is that they see cognition in terms of *information processing*. Although the roots of cognitive psychology go back earlier, the advent of telephone and computer technology around the same time as cognitive psychology's emergence was certainly a major influence on the now ubiquitous information-processing approach. What exactly is meant by information processing is not often made very clear, and indeed its meaning has shifted significantly through the decades. I will briefly sketch its history.

In the late 1940s, communications theory (closely related to information theory) arose as a field of study, spurred on in part by wartime efforts at developing and cracking codes, and in part by research at Bell Labs into telephone technology. Claude Shannon's 1948 technical report founded the field, which deals with the encoding and transmission of information from a source to a receiver for the purpose of communication, including the capacity of wires for carrying information, and the efficient coding of information into bits. This model of communication involves a notion of information as how much uncertainty a message eliminates. If bandwidth is limited, the more bits you transmit, the less ambiguous the message can be, but the more noise in the message, the more ambiguous. This is known as Shannon-Weaver information. A schematic of the components of Shannon's communication theory is reproduced in Figure 2.2.

By the 1950s psychologists had begun to take inspiration from communication theory, not just in models of human language processing, but as a model for cognition in general. Miller (1956) explored whether Shannon-Weaver information might be useful in psychology, and tried to characterize the capacity of humans to detect differences in stimuli of various kinds in terms of transmitted information. He treated the human observer as a communication channel, and tried to determine experimentally what our bandwidth is. This project was fairly swiftly determined to be a non-starter, although the idea that we can keep in mind about seven chunks of information at a time is still taught to psychology undergrads. As Neisser, who coined the term "Cognitive Psychology" in his 1967 book, says, "Attempts to quantify psychological processes in informational terms have usually led, after much effort,

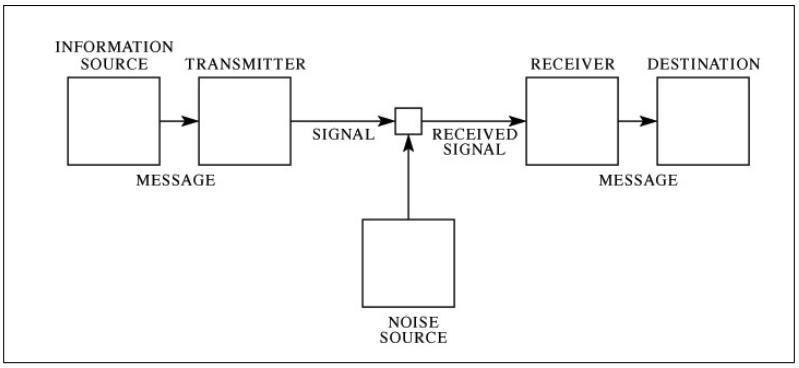

Figure 2.2: Communication theory schematic. Reprinted from Shannon (1948) with permission of Alcatel-Lucent USA Inc.

to the conclusion that the 'bit rate' is not a relevant variable after all. Such promising topics as reaction time, memory span, and language have all failed to sustain early estimates of the usefulness of information measurement" (Neisser 1967). Although the Shannon-Weaver notion of information is the most familiar—in the words of Jim Bogen, it comes to mind so fast that "if minds had knees, it would be a knee jerk reaction"—it is not terribly relevant to cognitive psychology, nor to computation more generally, which is where the information-processing metaphor eventually led. [1]

The information-processing metaphor changed its meaning slightly after this setback, but did not by any means drop out of use. Broadbent's 1958 book *Perception and Communication* presented an information-processing approach that maintained the metaphor of radio or telephone communication, with messages being sent from place to place, only without the emphasis on bit rates. For Broadbent, information processing referred to something like the journey information takes from its arrival as sensory input, its processing—perhaps in multiple steps—to its being transmitted as output in a final, processed form. His model of attention is illustrated in Figure 2.3, which he labels an "information flow chart." Like

[1] The capacity of wires for transmitting signals is of course relevant to computation in various ways including cryptography, hardware design, and network theory. It's also true that computer programs ultimately end up getting represented as binary bits and processed using binary arithmetic, but none of this is relevant to writing a program, even in a rather low-level language, unless the program happens to be an algorithm for something like encrypting messages, or checking TCP packet integrity.

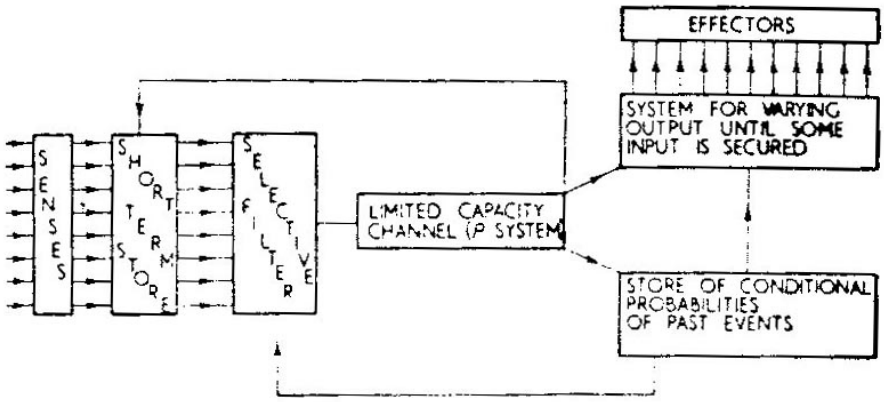

Figure 2.3: Broadbent's information flow chart of the filter model of attention. Reprinted from Broadbent (1958) with permission from Elsevier.

Shannon's diagram, Broadbent's consists of boxes representing functionally-described devices that input, manipulate or store data, and the arrows represent the route of data flow. I'll come back to Broadbent's model shortly.

Later the metaphor shifted again from being based on radio or telephone technology to being based on computer technology. Neisser's understanding of the information-processing metaphor likens cognition to a computer program. This shift in the metaphor introduces an ambiguity, however. Neisser still describes information processing in terms of information moving around from one place to the next; he likens both programs and theories of cognition to "descriptions of the vicissitudes of input information" (Neisser 1967). But at the same time, a program is not primarily about information being sent from place to place for processing. Some programs don't take any input. Others take input from several sources. Others take input and just store it somewhere without any processing. A program is more about a sequence of operations than it is about the vicissitudes of input information. Neisser recognizes this aspect of programs too. He says, "A program is not a machine; it is a series of instructions for dealing with symbols: If the input has certain characteristics... then carry out certain procedures... combine their results in various ways... store or retrieve various

items... depending on prior results... use them in certain ways... etc." (Neisser 1967). Instead of the information flow providing continuity, the journey in this case is that of the control pointer through the program, and it bears repeating, a program is not a machine. This program or operation sequence interpretation of the information-processing metaphor gained currency, but the earlier, telephone-inspired, vicissitudes of information interpretation did not disappear. They often get mixed together.

Another related version of the information-processing metaphor characteristic of cognitive science is to take it quite literally. The idea that cognition *is* computation was hinted at in Newell & Simon's early work on AI (which will be discussed further in Chapter 4), and perhaps most fully developed in Pylyshyn (1984). Newell and Simon (1961) posited an intermediate level of general "elementary information processes" that the nervous system implements in humans, and the hardware implements in a computer. The job of information-processing psychology, according to this interpretation, is to specify how these elementary information processes are combined to produce behavior, which is considered to be independent of the implementation.

Information processing remains a prominent approach to cognitive psychology. On the first page of the first chapter of several editions of his popular *Cognitive Psychology* textbook, Anderson states, "Cognitive psychology is dominated by the *information-processing approach*, which analyzes cognitive processes into a sequence of ordered *stages*. Each stage reflects an important step in the processing of cognitive information" (Anderson 1980, 1985). Figure 2.4 illustrates an example of Anderson's approach to information processing from his 1983 book, showing an analysis of how an arithmetic problem is solved. In this diagram, "The boxes correspond to goal states and the arrows to productions that can change these states" (Anderson 1983).

The method in Anderson-style cognitive psychology is to break down cognitive performances into sequences of simpler procedures, similar to the pseudocode that a computer programmer might write before writing a program. This process of writing pseudocode, then filling in the program commands to achieve each step, has very little to do with information in any technical sense, and also very little to do with its processing. In the most recent edition of his textbook, Anderson says of information processing: "It attempts to an-

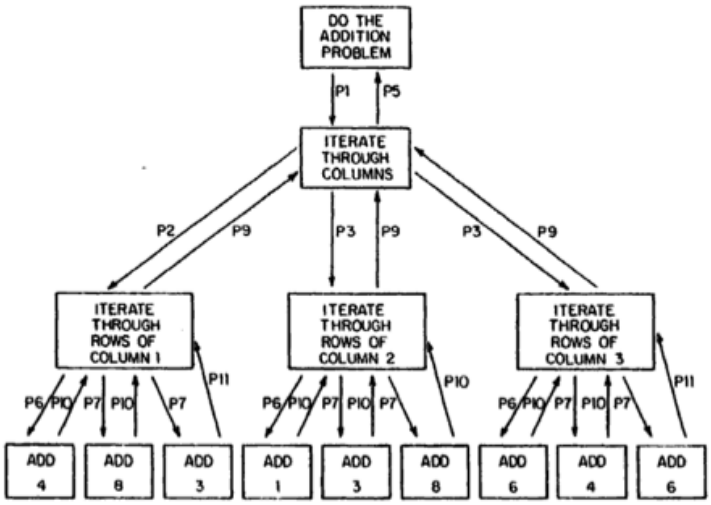

Figure 2.4: Control flow diagram. Reprinted from Anderson (1983). Copyright ©John R. Anderson, used with permission.

alyze cognition as a set of steps in which an abstract entity called 'information' is processed" (Anderson 2009). He does not say what an abstract entity is, nor how one can be processed, so I take information to be an entity only in a metaphorical sense. In a program, information is usually assumed to be available in various places and in various forms, including memory buffers, disk storage, and input from peripherals. In a program you don't get a story about the processing of information. The processing of information happens higgledy-piggledy all over the program. For programs of any complexity, the journey information takes as it is transformed into output involves a lot of numbers being copied to memory, then being copied to a register, then the number in another register being copied to memory, then input being read from a peripheral into a buffer, then the buffer's contents being written to memory, and so on, over and over and over again. Many of these numbers that are retrieved and stored aren't even data per se, but rather encoded commands comprising the program. Under this interpretation, information processing means performing a sequence of steps, like a computer executing a program.

Anderson has worked for decades in AI, and takes the somewhat idiosyncratic approach

of considering cognition to be a unified phenomenon rather than the workings of a bunch of dedicated modules, so he probably does not represent a middle-of-the-road approach. For balance, we can also look at Eysenck & Keane's 2005 textbook, which leans in the opposite direction of being very positive towards modularity and drawing connections to neuroscience. They agree with Anderson that, "Historically, most cognitive psychologists have adopted the information-processing approach" (Eysenck and Keane 2005), which they describe in very similar terms:

> Information made available by the environment is processed by a series of processing systems... These processing systems transform or alter the information in various systematic ways... The major goal of research is to specify the processes and structures (e.g., long-term memory) that underlie cognitive performance. Information processing in people resembles that in computers.

They also highlight the role of flowcharts as characterizations of cognitive theories, particularly in the 1960s and 1970s. Where Eysenck and Keane (2005) depart from Anderson (2009) is in considering contemporary cognitive psychology to include, in addition to what they call experimental cognitive psychology: cognitive neuropsychology, computational cognitive science, and cognitive neuroscience (understood narrowly to mean the use of brain imaging). However, they also admit that the integration of these four approaches is otherwise known as cognitive neuroscience (broadly understood), so I think it is fair to take what they call experimental cognitive psychology to be the pre-integration approach to cognitive psychology.

To sum up, there is general agreement that cognitive psychology takes an information-processing approach. Information-processing can be understood in several distinct ways. Two of these are: as the journey traversed by input data as it is transformed into output, and as a sequence of processing steps, similar to pseudocode.[2] Information-processing models might start with data and a series of functions to be performed on it; or they might start with a goal or task then work towards breaking it down into subgoals. What happens in a computer while it is running a program certainly could be described in either man-

[2]In addition to these two ways of unpacking the information-processing metaphor, there are a number of additional ways of describing what goes on when a computer runs a program. If the program is written in an object-oriented language, an object-centered description is more natural than a description of a sequence of processing steps, for example.

ner, depending on what you want to highlight. An additional interpretation of information processing might be as the interactions between the components in a radio, telephone, or computer system.

Although these versions of the information-processing metaphor are importantly different, the descriptions psychologists give of their models tend to look quite similar, because cognitive psychologists of all stripes typically use flowcharts to communicate the content of their models. The mechanism sketch example in MDC also looks like a flowchart, so there are at least superficial reasons for drawing comparisons. In the next section I examine the semantics of the flowcharts used by psychologists including Broadbent and Anderson, and compare them to mechanism diagrams. I argue that despite the superficial similarities, these diagrams represent, and fail to represent, quite different sorts of things.

2.3.1 Information-Processing Diagrams

One of the consequences of the influence of communications theory on psychology was that cognitive models started to be represented in flowcharts. Butler's (1993) review of the use of graphics in psychology mentions that flowcharts did not appear at all in textbooks or major journals in 1939-41(the earliest period covered in that review), but became ubiquitous later on.

For a brief period in the late 19th century, quite similar diagrams were used by the 'diagram-makers' in neuropsychology. Glymour (2001) and Shallice (1988) both discuss the diagrams from this period. One example of these are Lichtheim's 1885 diagrams from 'On Aphasia.' These are referred to as "schematic representation[s]" and he described his aim as being "to determine the connections and localizations of the paths of innervation" (Lichtheim 1885). The diagrams consist of labeled nodes, directed edges, and numbered slashes through both nodes and edges, as seen in Figure 2.5. The nodes represent what are referred to as brain "centres" or organs; the edges represent "paths of innervation" "tracts" or "commissures", the slashes represent possible lesion sites, or "interruptions of the reflex arc" (Lichtheim 1885).

It is clear that the figures are intended as schematic anatomical diagrams. The 'centres'

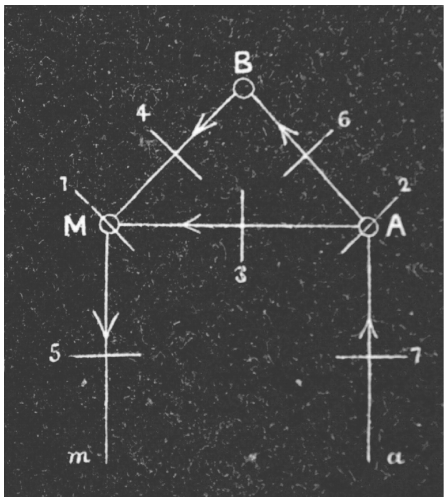

Figure 2.5: Lichtheim's schematic of aphasia lesion sites. Reprinted from Compston (2006) by permission of Oxford University Press.

are assumed to be brain areas, and the commissures are hypothesized axonal connections between them. The text discusses autopsy results for patients with various types of aphasia, relating the locations of tissue damage to the diagrams. The diagram-makers' schemas are not the immediate precursors to flowcharts in psychology, but they look quite similar.

Flowcharts re-entered into use in psychology via computer science. It is instructive to look at how flowcharts are used in that field. The flowchart was first introduced in computer science as a tool for programmers in a technical report by Goldstine and von Neumann (1947)[3]. Flowchart conventions changed quickly in those early years, but in the earliest versions, flowcharts showed the program's control flow, and specified the arithmetic or logical operations, the data substitutions and assertions, and the state of memory at each stage in a program.

Stepping through a 1947 flowchart (an example is shown in Figure 2.6), you see not only the order of processes, but also some of what is going on in terms of information transfer to and from memory. Boxes connected with dotted lines are truncated charts of the state

[3]Thanks to Stephen Morris for sending me his paper (Morris and Gotel 2011), where I found out about this report.

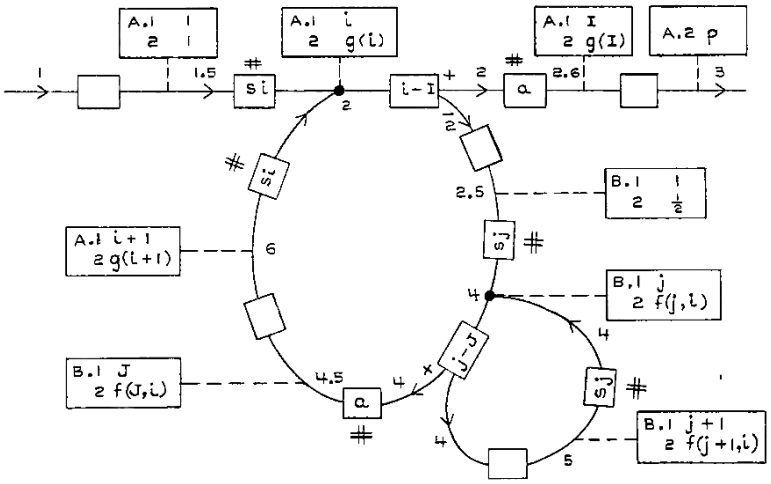

Figure 2.6: Early program flowchart. Reprinted from Goldstine and von Neumann (1947). Public domain.

of memory, and boxes marked with # are substitutions or assertions. The operations of various kinds are all mixed together in a program (and correspondingly in a flowchart of a program), so while both program flow and information transfers are represented in these older flowcharts, information flow is not in the organized form needed to tell a coherent story about information processing. The only indications of the physical entities in Goldstine & von Neumann's early flowcharts are the labels of locations in memory. In computers of that time, memory space was an important limiting factor on program complexity. Because of that, programs were often written in highly space-efficient ways. In a contemporary program one would never think to overwrite some of the program code in memory during the running of the program, but this was standard practice when memory was scarce. The most space-efficient way of writing certain parts of programs, like loops, is to overwrite completed instructions with later ones.

This dynamic character of programs is highlighted in the document:

> the relation of the coded instruction sequence to the mathematically conceived procedure of (numerical) solution is not a statical one, that of a translation, but highly dynamical: A coded order stands not simply for its present contents at its present location, but more

> fully for any succession of passages of C [program control] through it.... This entire, potentially very involved, interplay of interactions evolves successively while C runs through the operations controlled and directed by these continuously changing instructions (Goldstine and von Neumann 1947).

Specifying the memory registers used in variable assertions and substitutions is thus necessary for fully representing the flow of control, and understanding how the program works. Memory registers are no longer specified in program flowcharts, since it is no longer necessary.

Aside from memory registers, the physical entities that perform the operations are specified neither in older nor newer flowcharts. Goldstine and von Neumann (1947) state that the boxes represent "certain actions which must occur, or situations which must exist, when C passes in its actual course through the regions which they represent. We will call these the *effects* of these boxes, but it must be realized, that these are effects in a symbolic sense only." It seems that neither entities, nor their causal powers are represented in Goldstine & von Neumann's flowcharts.

As I mentioned, flowchart conventions changed rapidly after the publication of this technical report. Conveniently, computer scientists are meticulous in providing definitions, so there are international standards to consult in order to get a clear picture of contemporary flowchart use. This is helpful not just for understanding flowcharts in computer science, but also for analyzing flowchart semantics more broadly. The *ISO/IEC/IEEE International Standard for Systems and software engineering—Vocabulary* is just one of several places where the various sorts of diagrams used in computer science are defined. In this document, several different kinds of flowcharts are defined and distinguished.

2.3.1.1 Control Flow Diagrams

In the 2010 edition, the flowchart is defined as

> 1. a graphical representation of a process or the step-by-step solution of a problem, using suitably annotated geometric figures connected by flowlines for the purpose of designing or documenting a process or program... 2. graphical representation of the definition, analysis, or method of solution of a problem in which symbols are used to represent operations, data, flow, equipment, etc.... 3. a control flow diagram in which suitable annotated geometrical figures are used to represent operations, data, or equipment, and arrows are used to indicate the sequential flow from one to another (ISO/IEC/IEEE 2010).

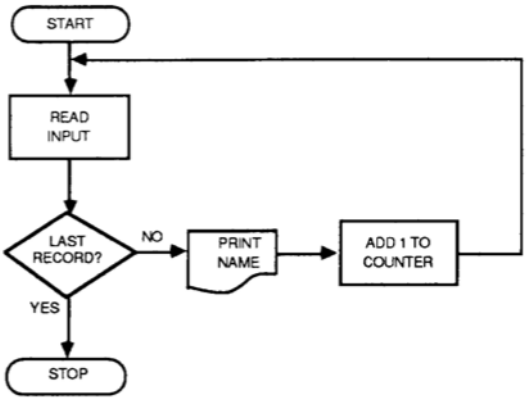

Figure 2.7: Control flow diagram. Copyright ©2010 IEEE.

The figure illustrating the definition is reproduced here in Figure 2.7. The more specific name for this kind of flowchart is a *control flow diagram*.

Models that conceive of cognition as a sequence of processes, including Anderson's, are very naturally represented in control flow diagrams, since these show a sequence of steps and the flow from one to the next. The boxes represent simple arithmetic or control operations. The arrows between the boxes indicate the flow from one process or operation to the next, in the order the central processor performs them. They do not show everything you might want to know about a model.

In a control flow diagram, arrows do *not* represent the flow of information. In fact, information flow is not indicated at all. Some processes surely involve the transfer of information, but this is not visible from the diagram. It is just assumed that there is information available in several places, including the program counter, the data register, memory, and input from peripherals like keyboards.

The arrows in control flow diagrams also do not represent causal connections, nor productive continuities. The box at the end of the arrow just happens to be what gets done next in a program. A different program might have a different next step.

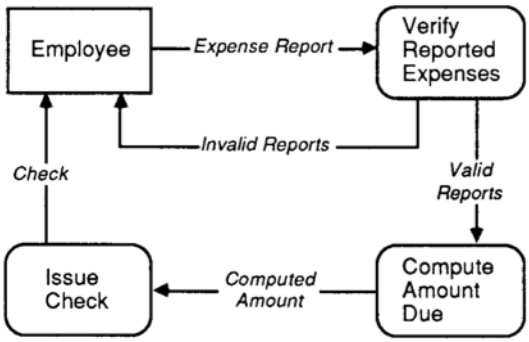

Figure 2.8: Data flow diagram. Copyright ©2010 IEEE.

From a control flow diagram you also don't get a clear picture of the entities that are doing the processing, since the geometric figures mainly represent operations or data, although obviously there are parts of the computer at work during processing. The parts involved in a particular operation depend on what that operation is and how the operations are implemented. Several operations might be done by the same part, or even, as in Anderson's conception, all the operations might be done by a unified whole. In mechanistic terms, activities and the relative order in which they occur is what a control flow diagram shows.

2.3.1.2 Data Flow Diagrams In contrast, models that conceive of cognition in terms of information flow, including Broadbent's, are very naturally represented by what are called *data flow diagrams*. The ISO/IEC/IEEE Standards define a data flow diagram as "1. a diagram that depicts data sources, data sinks, data storage, and processes performed on data as nodes, and logical flow of data as links between the nodes" (ISO/IEC/IEEE 2010). Here the arrows show the transfer of data from one source or process or device to another, and the boxes show what happens to the data during its journey, and the places it goes. The figure illustrating the definition is reproduced here in Figure 2.8.

Control flow and data flow diagrams look very similar, and this similarity further contributes to the ambiguity between these two ways of interpreting the information-processing

metaphor. What they show and what they leave out is quite different though.

Data flow diagrams show some entities as boxes, in particular those that are involved in data transfer, but other entities that are not centrally involved in information transfer may be left out. The arrows show the routes travelled by data, so do not indicate much about what the entities shown do aside from data manipulations, nor how they do it. Activities beyond those centrally concerned with data transfers are not shown, so it is difficult to figure out what the function of a system is from its data flow diagram. Many different systems might use the same arrangement of data sources, sinks, storage, and processes. The order in which the data flow operations occur is not readily discernible from a data flow diagram, so control flow cannot be recovered from this sort of diagram alone. Causal and spatial relationships between entities are also missing.

2.3.1.3 Mechanism Schema Diagrams

For diagrams of mechanisms to be informative about the mechanisms they describe, they should be able to represent the entities, the activities, and the causal relationships between these entities and activities that make the mechanism do what it does. None of the types of flowcharts described above can do the entirety of this job. The problem is not just that any diagram is only a partial representation which must leave out some details, but that for each type of diagram, there are categories of information that are systematically missing. Insofar as these types of flowcharts are representative of the kinds of models developed in various schools of cognitive psychology, the models systematically leave out some of what one would hope to find in a mechanistic model.

For comparison, consider the examples of a mechanism schema and sketch in Machamer et al. (2000). These also look like flowcharts, but they are neither data flow diagrams, nor control flow diagrams. In Figure 2.9, MDC's example of a mechanism schema, protein (entity) is the product of translation (activity) from RNA (entity), RNA is the product of transcription (activity) from DNA (entity), and more DNA is the product of duplication (activity) from DNA. The activities represented by arrows cause the production of the entities, which despite not having boxes around them, are what Darden (2005) calls "glass boxes." In this diagram, the distinct entities are represented at the nodes (without boxes), and the

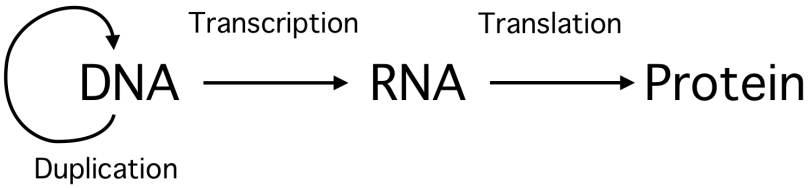

Figure 2.9: Example of a mechanism schema. Based on Figure 3 from Machamer et al. (2000, 16). Copyright ©2013, Boris Hennig, used with permission.

activities are labeled on the arrows connecting the nodes. The arrows are causal processes that help create the next entity in the chain.

There is no perfect match to this diagram in the ISO/IEC/IEEE standards, but it is similar to state transition diagrams, which use arrows to indicate "events or circumstances that cause or result from a change from one state to another" (ISO/IEC/IEEE 2010). It does not seem quite right to call DNA, RNA and protein states, but perhaps we can generalize this to a transition diagram, where the boxes might be either states that a system enters, distinct entities that are produced, or the results of some other type of causal process. We will see an example of a mechanism from Yu and Catterall (2003) shortly where the transitions are between states.

Later papers by Darden and Craver reinforce this idea that the arrows in mechanism diagrams stand for causal processes or the production of entities: "Each arrow shuttles from one region of the static diagram to another, constituting a stage of the mechanism's operation. The arrows trace the direction or flow of productivity through the mechanism" (Craver 2005).

Not all mechanism schemas are transition diagrams. Where the relative shapes and sizes or relative positions of the entities are important, a schematic diagram that represents these spatial relationships might be most appropriate. We will see examples like this from Shepherd (1983) and Hille (2001) shortly. There are surely other types too. The diagrams

of mechanism schemas that are familiar from biology and neuroscience do not look like any of the types of flowcharts described above though. They do not systematically leave out representations of entities, activities, or the causal relationships among them that are necessary for a mechanistic explanation.

The cognitive models reviewed are distinct from the sorts of mechanistic models characteristic of neuroscience and biology. Despite it being common to represent both cognitive models of various kinds as well as neurobiological mechanisms with diagrams that resemble flowcharts, a flowchart is not a flowchart is not a flowchart. Systematically different sorts of things are included and left out in these sorts of models, and yet they are not obviously incomplete. For the sort of explanation they are meant to give, they are complete.

2.3.2 Elliptical Mechanism Sketches

That the models from cognitive psychology I've discussed are not, as they currently stand, complete models of biological or neural mechanisms should come as no surprise. Piccinini and Craver (2011) claimed that these are elliptical mechanism sketches, not complete models. The question then is whether models that start out as data flow or control flow models lend themselves to being filled in with more detail to become full-fledged neural mechanisms, or whether they tend to remain distinct.

It certainly is possible in principle to transform one sort of model or diagram into another. To turn a control flow into a data flow diagram in the case of a computer program, one would have to draw a new diagram with arrows going back and forth between various types of storage devices, peripheral devices, and the CPU. Some additional knowledge about the architecture of the system would be needed in order to do this, and such a diagram would no longer tell you much about the program's dynamics or its function. Knowledge about how the program works would be lost, in exchange for knowledge about the communications channels used. To turn a data flow model into a control flow model would be very difficult, unless you already knew what the contents of the memory locations accessed were, and what sorts of operations were being done. Again, with sufficient additional knowledge, the transformation could be made. In this case, knowledge about the dynamics of data flow would

be lost in exchange for knowledge about the order of operations. A control flow diagram does not include information about data sources and recipients. Depending on the system being described, this sort of transformation could take a formerly informative, explanatory model and turn it into something uninformative and non-explanatory. For systems with different architectures than a computer, this transformation process might be more or less complicated, but in any case, additional knowledge would be needed, and some would be lost.

Turning either type of flowchart into a sketch of a neural mechanism is similarly complicated and in many cases not very useful, but could be done. The processes in a control flow model could be taken as the activities of a central processor. One could certainly identify this central processor with a part of the brain, or the brain in its entirety, but this does not make for a very informative neural mechanism. A decomposition of the processor into sub-entities, such that its activities could be understood as arising from the organization of those sub-entities and what they do would be the obvious next step in turning the black box into a glass box. But knowing what a CPU, for example, is made of and what the parts do seems irrelevant to the explanation the control flow diagram gives. The same CPU can run any number of programs, and the same program can be run on different kinds of CPUs. Making the model more detailed in this way might explain something else about the system, but it does not add to the explanation of why the program gives a particular sort of result. If knowing the order of operations explains the result, then knowing that the code for the third operation was stored in the 17th register does not make the explanation better.

In a data flow model, the boxes could be taken as entities, and the data itself could be considered another entity not explicitly represented. In Broadbent's example there is a short term store, a filter, and so on. These might be either glass boxes, if the details about them are known, or what Darden (2005) calls "gray boxes," that is, hypothetical entities where we know the function they should perform, but not what performs it. The arrows do not represent activities so much as the route taken by data. Here a similar problem arises that these are not what we might call black arrows that need to be turned into glass arrows in order to transform the model from an elliptical sketch into a full-fledged mechanism. Instead a different diagram entirely would be needed to represent the activities that the

entities perform (aside from data's activity of moving around).

In neither control flow nor data flow diagrams are the causal relationships between entities or states represented. What makes one step follow after another in a program is the invisible work of the program counter and the hardware that grabs the next command and follows it. A mechanism sketch could be constructed based on either type of flowchart, but additional knowledge—knowledge about entities and their causal relationships—would have to be added. Additionally, some of the information that is represented in these diagrams—about control flow and data flow—would get lost. In short, these diagrams, and the models they represent could be *turned into* mechanism sketches, but that isn't what they are, and isn't always what they're intended to be. You don't typically move from one type of model to the other when making progress in science, although you might have reason to make several diagrams for several purposes. What you do instead is to refine one type of model or the other type, and the diagrams you use to represent the model demonstrate the progress through the model's refinements. We'll see an example of this in the next section.

One obvious counterargument might be to point out that what a diagram shows is always partial. One can't conclude that a model does not specify causal relationships between entities based on those being absent from a diagram depicting that model. This is not the conclusion I'm drawing. The sorts of flowcharts I've been discussing are presented by their authors as representative of the models they depict, and these sorts of diagrams are routinely used for this purpose in cognitive psychology. The verbal descriptions that accompany these diagrams do not detail the missing information about causal relationships and entities either. So it is fair to take these diagrams as indications of the contents and structure of cognitive psychology's models. The models do not include the sort of information expected in a neural mechanism sketch, suggesting that the explanatory aims of cognitive psychologists may not be to describe mechanisms, or at least not the kind of mechanisms Piccinini and Craver (2011) have in mind.

Another counterargument might be to claim that these are mechanism sketches so elliptical as to consist entirely of black boxes or black arrows. One problem with this is that these diagrams are not empty. Although they're relatively empty of neural details, they aren't empty of process or data flow details. As psychological models they are not considered to be

deficient. They are complete enough to support explanations. It seems too easy a move to claim that these models are empty mechanisms, without giving a convincing set of examples where models like this have actually been elaborated from elliptical sketch into mechanism. Below I give an example where this failed over the course of several decades to happen. Piccinini and Craver (2011) do not offer any examples where elliptical mechanism sketches have been filled in with neural details.

One objection that I do think can get traction is that while the neural entities and brain activities that might be involved in a data flow model like Broadbent's or a process flow model like Anderson's are systematically left out of these models, there are nevertheless entities and activities represented, so perhaps these are the relevant ones. In a data flow model, the data's flow might be the most relevant causal process. The phenomena being explained might be the effect of this kind of data flow process alone, regardless of what the specific entities doing the data processing are and how exactly they do it. In a process flow model, the sequence of operations might be the most important explanatory factor, regardless of what kind of computing machinery performs them. These sorts of models could very well be mechanistic then, if we're content with the entities and activities operative in them being rather more abstract than the sorts of things one might reasonably identify with brain parts and the activities thereof. If this is the route we go, then these models are not elliptical mechanism sketches. Instead they are complete enough, without being specifically neural mechanisms. This is an option I will follow up on in Chapter 3.

I've argued that the sorts of models cognitive psychologists build are not very much like what we'd expect of neural mechanisms, and that turning them into neural mechanisms would require considerable effort, the addition of different kinds of knowledge, and would involve leaving out the sorts of knowledge that these models do include. In the next section I contrast cases where this sort of cognitive model underwent considerable development, and yet was not filled in with neural details, with a case where a sketch of a neural mechanism was filled in with details in just the way Machamer et al. (2000), Piccinini and Craver (2011) describe. This is not to say that there are no cognitive models that are amenable to this sort of elaboration, just that this route is not typically how cognitive models develop.

2.4 THEORY DEVELOPMENT IN NEUROSCIENCE AND COGNITIVE PSYCHOLOGY

Information-processing models in cognitive psychology are typically treated as complete, autonomous models, not as elliptical mechanism sketches that need to be filled in. In this section I compare the development of a pair of models drawn from neuroscience and cognitive psychology. The models I contrast are ion channel gating models, and Broadbent's (1958) attention filter model, which depend on similar mechanical metaphors. The neuroscientific case does proceed as expected of a mechanistic explanation, with black and gray boxes gradually filled in with details. I argue that despite Broadbent's initial hopes, the route cognitive psychology took his model in was not to develop it from a sketch into a more fleshed-out mechanistic explanation, but rather to keep refining it as a data flow model.

2.4.1 Neuroscience's filters

In neuroscience a number of structures are described as being channels, gates, or filters. One example are the channels that allow certain ions to pass through the membranes of axons, while blocking other ions from passing through. The sodium channel, for example, which is essential for the action potential, is described as having a "selectivity filter" and two "gates." The selectivity filter allows smaller $Na+$ ions through but not larger $K+$ ions. The mechanism by which it accomplishes this is first, that the pore is lined with negatively charged amino acids, which attract positively charged ions like $Na+$ and $K+$, and second, that the size of the pore is about 0.3 by $0.5nm$ wide (Hille 2001), which is too small for the $K+$ ion to fit through. During the neuron's resting state even the smaller $Na+$ ions are prevented from flowing inward because of the activation gate. When the cell is stimulated, for instance during an action potential, the activation gate opens, and $Na+$ flows in. The flow of ions is once again blocked by the subsequent closing of the inactivation gate, which happens a few milliseconds after depolarization (stimulation which makes the usually negatively charged neuron less negative). The inactivation gate remains closed until the cell is repolarized (goes back to its usual negatively charged state).

The entire channel is made of an assembly of proteins embedded in the cell membrane, and the gates are made of loops of this protein. Ion channel proteins have sections that are variously charged, like the negatively charged pore region mentioned above, and which react to various chemicals. This means that when the membrane changes its charge, or when the cell encounters certain chemicals, the protein changes shape. The inactivation gate is a loop of protein that flaps open or closed based on changes in the neuron's electrical charge. Other types of channels open and close based on the presence of particular chemicals, temperature, or physical force, and they use various mechanisms for changing the shape and size of the pore. The $nAChR$ channel opens to let $Na+$ and $K+$ ions through when it binds to acetylcholine by rotating its $M2$ helices (sections of protein forming the channel) 15°. The $K+$ channel opens by bending its inner helices on a hinge point. The $MscL$ channel opens by tilting its $TM1$ helices (Doyle 2004).

From these descriptions it should be clear that the filters and gates in ion channels are quite literally filters and gates. The filters are physical structures that let some objects but not others through an opening based on size and shape. Hille (2001) describes the channel pore as "an atomic mechanical sieve." The gates are physical structures that swing, tilt or twist open and closed over an opening allowing or preventing objects to pass through. Much is known about the constitution of the proteins forming these structures, and the ways they change shape and react to various stimuli, based on methods like X-ray crystallography. Neurophysiologists are continuing to fill in more of the details of how these mechanisms work.

But even before so much was known about these structural and dynamic details, the language of filters and gates was used with the expectation that structures corresponding to these names would be found there. Hille (2001) reproduces progressively more complete schematics of the $Na+$ channel from 1977, then from 1991. In the earlier version, the gate (it wasn't yet known that $Na+$ channels have two gates) is represented as an amorphous blob with dotted outline, and the voltage sensor is a simple box with a probe sticking into the "Channel macromolecule." Shepherd also includes a schematic diagram of an $Na+$ channel in his 1983 textbook, which looks very similar to Hille's. In Shepherd's version, the filter, gate and sensor labels are in scare quotes, since there was at the time only speculation about

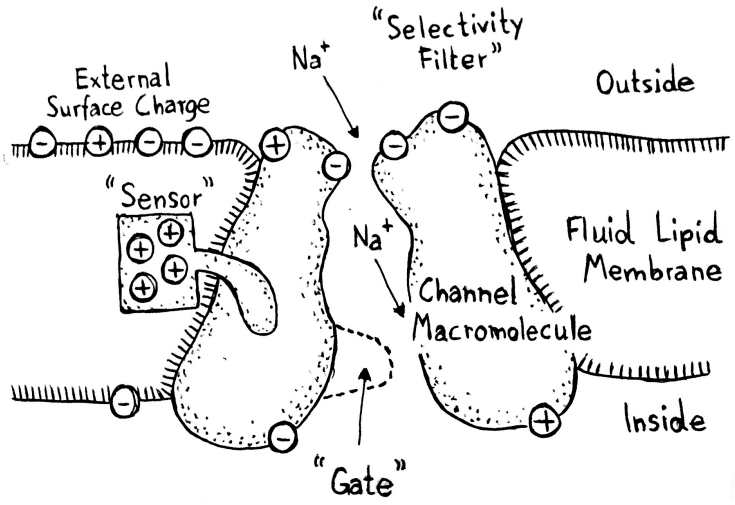

Figure 2.10: Early schematic of the $Na+$ channel, based on Shepherd (1983), Hille (2001). Copyright ©2013 Boris Hennig, used with permission.

the existence of entities performing approximately those functions, but no clear evidence about their physical natures. A composite of these two schematics is shown in Figure 2.10.

Although many of the details of how $Na+$ channels work were not yet known, the more speculative parts like the gate were treated as gray boxes only in a provisional way. Rather than resting content with the idea that the gate should be something that changes state depending on electrical and chemical conditions, thus allowing or preventing the flow of ions, a number of specific hypotheses were put forward as to how the gate worked. Hille (2001) notes, in *Ion Channels of Excitable Membranes*, "Many models have been proposed for the nature of gates." He provides illustrations of twelve such possible mechanisms (how-possibly models) that were suggested in various published articles. These include gates that swing out, assemble from subunits, pinch shut, are blocked by mobile ions, rotate, slide, twist, or are plugged by a tethered ball. His illustration is reproduced here as Figure 2.11. Hille describes hypotheses F through L as the mechanisms that remain popular and plausible (how-plausibly models), and discusses the evidence for these mechanisms in various types of ion channels. I described a few examples of these above.

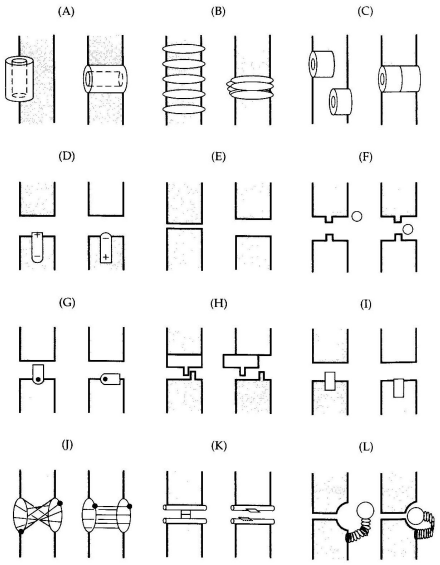

Figure 2.11: Possible mechanisms for channel gating. Reprinted from Hille (2001). Copyright ©Sinauer Associates, used with permission.

Hille's later diagram of the $Na+$ channel is superficially similar, but has a number of added details, based on results that had accrued in the intervening decade. In the 1991 version, the macromolecule is called a protein, it sticks much further out into inter- and extra-cellular space, various molecules are attached to the protein, the membrane layer and voltage sensors are much more detailed, and the gate is represented as a still schematic swinging hinge, connected to the voltage sensor. Since 1991 it has been determined that the inactivation gate of the $Na+$ channel works like a tethered ball, as in Hille's hypothesis L. It is formed by the section of protein between domains III and IV (Hille 2001). Figure 2.12 shows a state transition diagram illustrating the $Na+$ inactivation gate's opening and closing. Contemporary diagrams of the $Na+$ gate show many more details, down to the twists and turns in the channel proteins.

The development of these models from schematic drawings through to increasingly detailed models fits well with the description of mechanistic explanation MDC give. The diagrams of ion channels in Shepherd (1983) and Hille (2001) are perfect examples of mech-

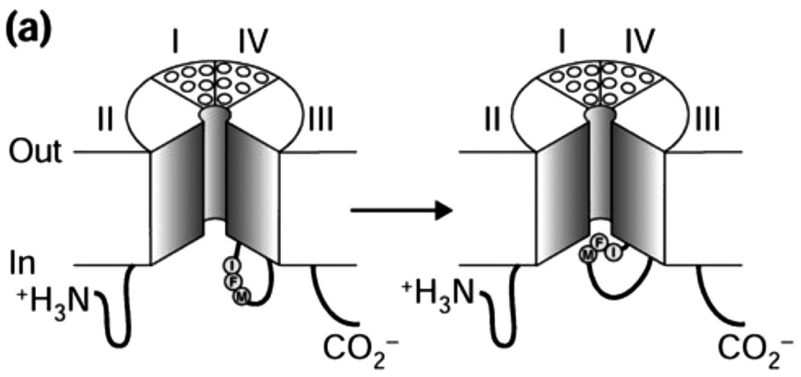

Figure 2.12: Sodium channel inactivation gate. Adapted from Yu and Catterall (2003). Copyright ©2003 BioMed Central Ltd.

anism sketches. The gate and sensors shown in dotted outline, and with scare quotes in Shepherd, began as black boxes. When it was discovered that the $Na+$ channel has both an activation and an inactivation gate, and more was known about their functioning, these boxes became gray. As the structure of the proteins making up these structures is discovered, they are gradually becoming glass boxes. Hille's (2001) collection of gating mechanisms were how-possibly models. Each option was a generic description of a possible gating mechanism, the details of which might later be filled in and evaluated.

These diagrams do differ in several respects from the example in Machamer et al. (2000), which is more like a transition diagram. These only show a static picture of the mechanism, not the transitions between states, such as open and closed gates. They are also quite a bit more detailed in terms of the entities represented than the DNA example. This is appropriate in this case, because the relative shapes and sizes of the entities are important to the channel's functioning, whereas in the DNA case, the activities that transform the entities through various stages are most important.

In any case, it is striking how the initial mechanism sketch was gradually filled in with more details, in much the way Piccinini and Craver (2011) suggest should happen with cognitive models. In this case from neuroscience, the MDC account of mechanism is very

helpful for understanding both the models themselves, and the scientific practices leading to their development. I now turn to an example from cognitive psychology to see whether the same can be said of it.

2.4.2 Cognitive Psychology's filters

Cognitive psychology also has models referred to as filters and gates. A prime example is in Broadbent's filter theory of attention. Broadbent's model is the basis for a long series of later models of attention. I will start by outlining Broadbent's model and general approach, then follow his filter model's development by later cognitive psychologists. We'll see that as these models change and develop, they do not get filled in with details of neural mechanisms.

Broadbent's interest was in exploring the limitations of auditory processing. His theory was developed to explain various results from dichotic listening experiments (where different messages are presented to the two ears simultaneously, and the subject has to "shadow" or repeat back one of the messages). The most basic result, popularly known as the cocktail party effect, is that people can usually pay attention to only one of the streams they're hearing, and effectively seem to block out other distracting sounds. The theory posits that there is a short term store where information is kept briefly, before it moves through a selective filter, which lets only some streams of information through, blocking the others. This filter was thought to be necessary, because information subsequently must move through a limited capacity channel (we can't pay attention to everything), and what makes it through the channel seems to be based on criteria fed in from higher areas; important information generally isn't lost in the shuffle.

Broadbent's filter theory is based on the technology in telephone communication systems and radios of the time, but he steers clear from making any speculations about what the attention filter might be made of in neural terms. The main concerns directing theory choice are that the functioning of the filter fit the experimental data, and that the theory be simple and general. The diagram Broadbent uses to illustrate his theory was reproduced earlier in Figure 2.3. It is made up of boxes and arrows, the boxes carrying labels indicating roughly their functions.

If we interpret Figure 2.3 as a data flow diagram, as suggested earlier, the boxes can be taken to represent entities which perform activities like storing, and filtering, and the arrows represent the transfer of information. In contrast to Shepherd's diagram, there is nothing that looks like a filter or a gate, although the limited capacity filter slightly resembles a tube. (This difference in height of boxes, which for Broadbent may have had some significance disappears in reproductions of this diagram in later texts.) That no details about the entities are offered in Broadbent's model is not surprising, since there were at the time few experimental methods available for investigating brain structures. Taken alone, it does not look very unlike a mechanism sketch consisting of gray boxes. Piccinini & Craver's account suggests that these gray boxes might gradually be replaced with how-possibly mechanisms, then how-plausibly mechanisms, and that these would be something like brain parts (possibly distributed ones) that do the storing and filtering

Broadbent even shows some interest in the question of how the theory might be implemented. He says, "we have tried to make our hypothetical constructs of such a kind that they could be recognized if it were possible to observe them directly: a filter or a short-term store might take different physiological forms, but it could be decided with reasonable ease whether any particular physiological structure was or was not describable by these terms" (Broadbent 1958). In the final chapter he explains what he took to be the advantages of describing his model in "cybernetic language." One of these is that "a description... in terms of information flow will be readily attached to physiological knowledge when the latter becomes available" (Broadbent 1958). He further comments that "readiness to connect with the neighboring sciences is highly desirable" (Broadbent 1958). He explains this point by commenting that Hebb, by wording his theory in physiological terms, leaves himself open to having his theory disproved based on irrelevant physiological findings (Broadbent 1958). So while it is clear that Broadbent meant his filter theory to be open to connections with later neurophysiological results, the way in which he did so was to keep his theory as free from physiological speculations as he could. He purposely avoided proposing even how-possibly neural mechanisms that might do the job. He claims that the relationship between physiologists and psychologists is analogous to that between auto mechanic and driver, then quips that "for many purposes a knowledge of the mechanism is not essential to the driver" (Broad-

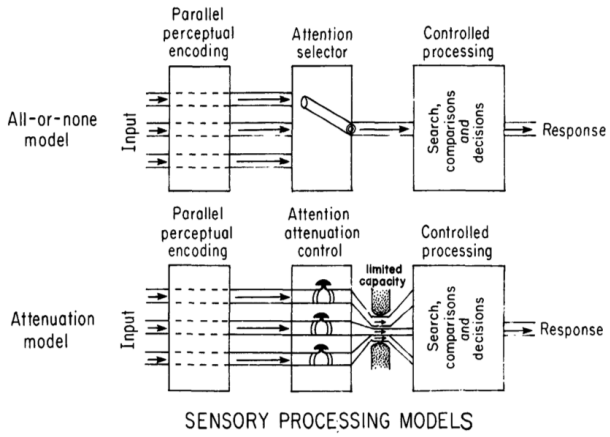

Figure 2.13: Comparison of Broadbent and Treisman's filter models. Reprinted from Schneider and Shiffrin (1977). Copyright ©1977 American Psychological Association.

bent 1958), suggesting that psychologists don't need to know about neural mechanisms[4].

Although Broadbent intended his model to be open to help from future developments in physiology, and in that sense treated as a sketch, this is not how it was treated by subsequent cognitive psychologists. Several historians have commented that although cognitive psychology was, in its early stages, open to the possibility of interactions with neuroscience, the separation between the fields became a principled one later on (Baars 1986).

If we trace the development of Broadbent's filter theory of attention forward, we see that references to finding structures in the brain corresponding to the boxes disappear, and along with them the license to call the model a mechanism sketch. The diagrams used to describe subsequent models of attention do get more complex, but the details remain of the same type as are shown in Broadbent's diagram. Figure 2.13 shows a comparison of Broadbent and Treisman's (1960) models. Treisman's main alteration was to change the functioning of the filter, so that instead of just a single stream being selected at a time, many streams might be attended to at once, but with most of them attenuated. This change was in response

[4]Of course Broadbent may not use the term 'mechanism' in the technical sense of Machamer et al. (2000)

to results suggesting that content on the unattended channels does have some effect on processing, for instance by priming certain responses, as well as the simple observation that highly salient cues, like one's name, are reliably perceived even on an unattended channel. The way the filter works changed in this updated version, but it was not made any closer to being fully fleshed-out.

The diagram above shows contraptions like swinging tubes and pinching vices, not so unlike the how-possibly mechanisms from Hille's diagrams, but in this case, these are only meant metaphorically. Treisman's own diagrams do not look so mechanical. There is also no discussion in the text of how the filter mechanism might be instantiated, and it is certainly not meant to literally be a swinging tube or a vice. Treisman (1960) discusses channels and filters at length, but the only physical entities she mentions are the ears. The goal of eventually filling the model in with physiological details had been set aside, if not forgotten, and was not to be revived for about 30 years.

The next major alteration is pictured in Figure 2.14, which shows Shiffrin and Schneider's (1977) filter model. They proposed a theory that made a functional dissociation between controlled attention and automaticity, supported by experiments where they examined the circumstances under which one task interferes with another. Tasks that aren't adversely affected when combined with other tasks are considered automatic, while controlled processing does suffer interference when combined with other tasks. In their diagram, there are multiple levels of automatic processing within short term storage, instead of Broadbent and Treisman's short-term store followed by a single filter. These multiple filters can also have feedback effects on one another. The main change to the model is the addition of a second pathway for controlled processing, which can exert effects at various stages during automatic filtering.

Once again, this model of attention has various added complications made in response to experimental data, but these complications are not details about how any of the entities might possibly or plausibly be realized in a neural mechanism. Instead the complications that are added remain, as in the older versions, essentially a data-flow diagram consisting of gray boxes connected by arrows. It seems that in this version the solid line arrows indicate direction of information flow, as before, but the thicker arrows might be meant to indicate

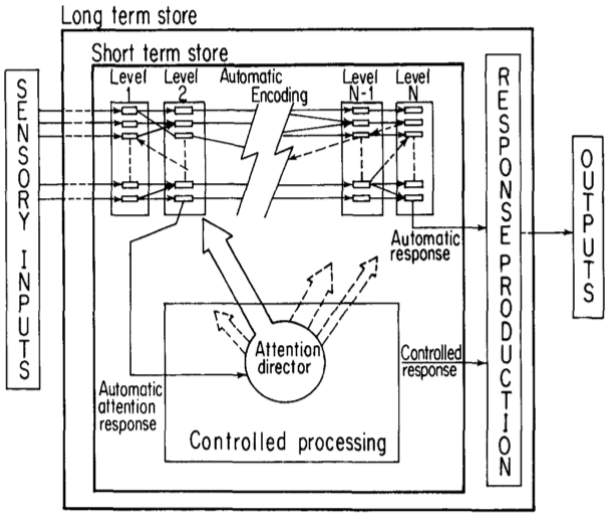

Figure 2.14: Shiffrin and Schneider's filter model. Reprinted from Shiffrin and Schneider (1977). Copyright ©1977 American Psychological Association.

that the activity of the attention director causes the filters to operate differently, so at least some of the arrows might represent causal relationships between components. Shiffrin and Schneider outline what they call a general theory in which "Memory is conceived to be a large and permanent collection of nodes" the structure of which they treat as "a very general graph with complex interrelations among nodes" (Shiffrin and Schneider 1977, 155). They suggest that the structure of long term memory is arranged in levels, which they define as "a temporal directionality of processing such that certain nodes activate other nodes but not vice versa" (Shiffrin and Schneider 1977, 155). So while they certainly are concerned with the structure of the psychological entities in their model, this structure is entirely abstract.

Another way in which models of attention developed over time is that they took into account more and more evidence from other related areas of psychology. Attention is implicated in many other psychological capacities, so a model of attention has constraints placed on it from other areas of psychology. There is a sense in which these models gradually became more integrated, but this integration was lateral, connecting with other psychological models, rather than vertical, connecting with neuroscientific models.

The differences between the filter models discussed are that cognitive psychology's models specify information-processing stages instead of entities performing activities, and, at least in some periods, cognitive psychologists were not concerned with whether their information-processing flowcharts could be filled in with neural details. They were concerned just to get the relationship between the functionally-defined stages correct. The psychologists involved show no signs of thinking that their models are mere sketches of neural mechanisms; they see them as something distinct and autonomous.

2.5 INTEGRATING NEUROSCIENCE AND COGNITIVE PSYCHOLOGY

So far I have argued for two kinds of distinctness: first, psychological models include different kinds of information than do the mechanistic models typical of neuroscience. Instead of specifying parts and their activities, these models may describe process flow or data flow. Second, psychological models do not develop from bare sketches to ever more complete descriptions of mechanisms. Instead they become ever more precise, correct, and laterally-integrated process- or data-flow models. In this section I will argue for a kind of autonomy, by showing that mapping psychological models onto brain structures does not necessarily make the model better or more explanatory.

Piccinini and Craver (2011) might counter the arguments above by saying that although cognitive psychology proceeded independently from neuroscience for three decades, the boxes in models like Shiffrin and Schneider's can be mapped onto parts of the brain, and that doing so is the key to integration. Indeed doing just this sort of mapping has become a popular move in cognitive neuroscience. Schneider does this himself with a later iteration of the model called CAP2, which is described in Schneider and Chein (2003). Figure 2.15 shows the updated cognitive model on the left, and on the right, the same model overlaid on a drawing of the brain. In the paper they talk about a 'mapping' between their flowchart boxes or functional modules, and brain regions. The mapping is supported with neuroimaging data.

This is precisely the sort of thing Piccinini and Craver (2011) suggest should be done with psychological explanations in order to integrate them with neuroscience and thus turn

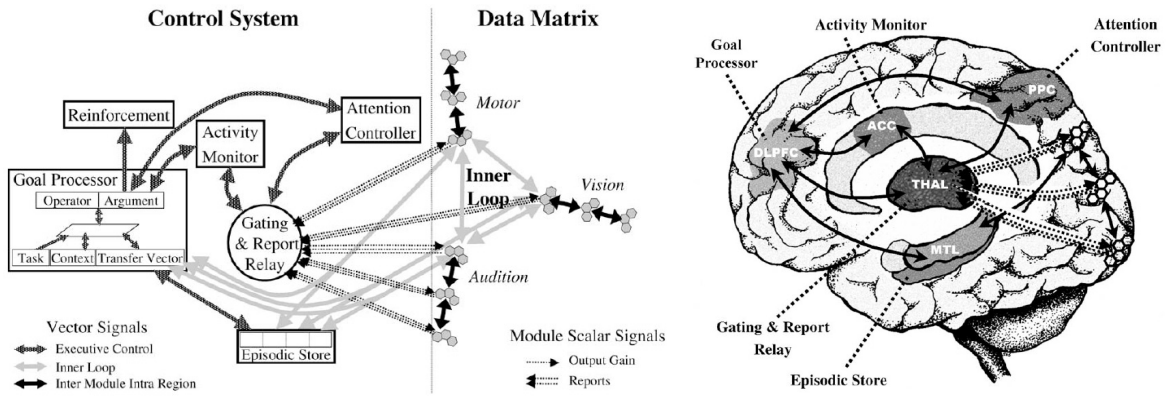

Figure 2.15: Schneider and Chein's attention model. Reprinted from Schneider and Chein (2003). Copyright ©2003 Cognitive Science Society, used with permission.

sketches into good mechanistic explanations. I'm in full agreement that this is an accurate description of the sort of integration that many projects in cognitive neuroscience try to do. What I'm not convinced of is that this sort of mapping makes for better explanations in any but the fairly rare cases where a complete reduction is possible.

For this sort of mapping to be helpful in giving a better or deeper explanation of the cognitive phenomenon modeled, several conditions must hold. The entities and activities in the cognitive model must correspond to brain entities and activities, even if fairly coarse grained ones. Those brain entities must be localizable, i.e., not extremely diffuse. The structure of interactions between components of the cognitive model must be the same as the interactions between the corresponding brain components. If all of these aren't true, a mapping will be either impossible, or useless.

Even if these conditions are met, the mapping on its own is not explanatory. In order for the mapping to provide a better explanation than the cognitive model did, we need to know how the brain regions implicated work, and furthermore how those brain regions work in that particular mechanism, since brain regions typically perform many functions and engage in many activities that aren't obviously relevant to any given function. The difficult work then is making this sort of mapping into a mechanistic explanation. This may include the

following:

- specifying in detail which parts in the brain regions indicated are to be identified with the cognitive parts,
- specifying what those neural components do and how they do it,
- characterizing the form any information being transferred takes,
- verifying that the specified information is transferred between the appropriate neural components,
- specifying which neural mechanisms make that information transfer happen, and adding them to the model,
- characterizing the relevant causal relationships between neural components,
- verifying that those causal relationships are as specified in the cognitive model,
- investigating any other activities of the neural components involved,
- investigating any other causally connected neural components,
- verifying that these other activities and entities do not interfere with the operation of the mechanism.

The mapping itself does not do much of the work involved in turning a cognitive model into an explanation in terms of neural mechanisms. Developing a cognitive model and mapping it onto brain regions may be good first steps to take in cases where the conditions for successful mapping are met, but it is difficult to tell in advance whether these conditions are met. The mapping might be fruitless, and we should expect that in most cases it will be.

As we have seen, models in cognitive psychology are typically developed entirely in isolation from neuroscience. Models that have gone through several iterations without any input from neuroscience should not be expected to be directly mappable onto the brain. This does not mean that cognitive psychology has no way of deciding which theories or models are good ones and which bad. Just as in any other science, a good theory should account for all the known data, be internally consistent and externally valid, make predictions that turn out to be true, guide us in making interventions, and so on. A good cognitive model captures relatively stable, robust regularities in how people behave. Even given all of these

constraints on good models, there are still any number of possible models one might choose, most of which we have no reason to believe would make for an uncomplicated mapping. This is not just because of underdetermination, but also because of complexity. While behavior certainly has material causes most of which come from the brain, there is little reason to believe that those neural events will be similarly organized. In complex systems, what looks stable and robust at one scale may not be so at another scale.

The question is what to do when a well-accepted cognitive model does not map cleanly onto the brain. One of the norms on a good theory is that it be consistent with theories in related fields. This means that if there are several reasonably good cognitive theories, one of which maps much more cleanly onto the brain than the others, the one with the clean mapping should be preferred. This is a perfectly legitimate way of using neuroscience to constrain cognitive theories. In cases like this, psychologists should revise their theories to better match the neuroscience. Where this happens, it would be fair to describe the cognitive model as having been a mechanism sketch. Once the sketch is altered in order to afford a better mapping, and the neural details are added, we would have an integrated cognitive-neural mechanism of the sort that should make Piccinini and Craver happy.

This sort of scenario is exceptionally rare though. Piccinini and Craver (2011) have not offered any arguments as to why we should expect these mappings to work in any but very few cases. A slightly more complicated sort of case would be where there is a cognitive theory that is clearly preferred based on the norms of explanation internal to psychology, but a rather different model maps more cleanly onto the neural structures. This is particularly likely in cases where knowledge about the relevant neural structures is not initially available, or not taken into consideration by psychologists when building their models, as was certainly the case with the filter model or attention described earlier. If later developing knowledge about the neural underpinnings of the phenomenon suggests a different decomposition into working parts than the the well-accepted cognitive theory, it is not immediately clear what the resolution should be. Piccinini and Craver (2011) would presumably argue for scrapping the cognitive theory, and I find that option extremely tempting too. This is not what psychologists in fact do, however, as I'll describe shortly. If we're to be fair to psychologists, as I have resolved to be, it is at least worth finding out why they hang on to their theories

even when cognitive and neural ontologies fail to match up.

Probably much more common are cases where there is no good mapping to be found at all, but nevertheless the cognitive theory, if judged on all the other relevant norms, is a good one. For some purposes, neat neural mappings might be judged a less important norm than some others. For example, if a model performs better than any alternatives in the clinic, there would be good reason to keep it around. This clinical superiority might very well persist even after fully fleshed-out neural mechanisms are discovered. For example, the most effective treatment for depression currently known is exercise, although the mechanism through which exercise reduces depression is unknown. Hypothetically, it may turn out that this mechanism involves the combination of vestibular system input to the thalamus (from head motion), increased gain of the hypothalamic feedback system (from sweating), and lactic acid production (from muscle strain). If this (completely fabricated) mechanism were correct, a functional analysis of exercise as say repetitive motor activity leading to increased oxygen consumption (or some plausible alternative) might have very little to do with this mechanism. Head motion, sweating and muscle strain are all just side-effects of exercise, not constitutive of it. Furthermore, knowing the neural mechanism might not turn out to be useful for clinical treatment of depression. Exercise might remain the safest and most effective treatment in all but a very few cases where depression co-occurs with a disease of one of these side-effect systems such that exercise is ineffective in bringing about vestibular input to the thalamus, increased gain of the hypothalamic feedback system, or lactic acid production.[5]

Proponents of mechanistic integration insist that models are no good as explanations unless they map onto mechanisms. Kaplan has argued in (Kaplan 2011, Kaplan and Craver 2011) for a "model-mechanism-mapping (3M) constraint." He starts from the reasonable requirement that for a model to be an explanation it must identify causally-relevant factors.[6] From this Kaplan concludes that to have explanatory power, a model must explain in terms of component parts, which in the case of cognition would be neural ones. I don't think this

[5]The pharmaceutical industry might nevertheless be motivated to make use of the discovered mechanism to devise other treatments.

[6]This leaves out some things we might want to call explanations, but if we restrict ourselves to causal-mechanistic explanation, the claim is uncontroversial.

conclusion follows.

The examples Kaplan describes involve dynamical systems. The proponents of dynamical models he argues against in Kaplan and Craver (2011) deny that dynamical/mathematical models are mechanistic, but claim that they are nevertheless explanatory. Kaplan and Bechtel (2011) concede that some dynamical models are explanatory, but only when they describe mechanisms. It is possible to identify causally-relevant factors without being mechanistic on Kaplan's account though. One difference between the MDC account of mechanism and others like Bechtel's is that where Bechtel talks about component parts, MDC talk about entities. I will argue later that there might be causally-relevant entities in cognitive models that are nevertheless not component parts. Some of the causally-relevant entities not being parts is one way in which mapping entities to brain regions might fail. [7] I elaborate on this idea in Chapter 3.

Despite these difficulties, decomposing a system into smaller parts and investigating what those parts do may be a good heuristic. Bechtel and Richardson (1993) argue that it is. They also point out that for non-decomposable systems, the heuristic will not work, but trying and failing to decompose a system is nevertheless a good strategy for figuring out when a system is non-decomposable. I have my doubts about the usefulness of this heuristic in cases where we have good reason to doubt decomposability. My worry is that functional models can prove to be quite misleading if treated as sketches of mechanisms to be filled in.

2.5.1 Memory as a Test Case

Bechtel (2008) argues much in the same spirit as Piccinini and Craver (2011), that explanations in psychology and neuroscience can be integrated by localizing cognitive functions in neural parts. This decomposition and localization heuristic is described in detail in Bechtel and Richardson (1993) then revised in Bechtel and Abrahamsen (2005), but the basic idea is that a top-down functional decomposition and a bottom-up structural decomposition are made to meet in the middle, when the decomposed functions are localized in the structural parts. The functional decomposition is roughly the same as Piccinini and Craver's idea that

[7] This may also be worth exploring with respect to dynamical models.

cognitive models can serve as elliptical sketches, and the localization step is roughly equivalent to knowledge from neuroscience being used to fill in the details of the sketch. Their ideas about the dynamics of the process are perhaps different.

Bechtel (2008) offers a case study of how work on memory developed from earlier phenomenal and functional decompositions towards a more complete mechanistic account. The main lesson Bechtel draws from this case is that working at either too high or too low a level can hinder the search for mechanisms. That seems right, and he makes several other illuminating points along the way. But what his case study also demonstrates, perhaps unwittingly, is that a cognitive model developed independently of knowledge about the neural structures that underlie the cognitive phenomenon can prove to be a very misleading starting point, or a very poor mechanism sketch in Piccinini and Craver's terms. In this case, the cognitive model not only failed to be a helpful decomposition or mechanism sketch, it seems to have *hindered* progress towards developing a more complete mechanistic model. Furthermore, although the cognitive model turned out to fit quite badly with the mechanistic model eventually developed, psychologists have not subsequently given up their model.

Psychologists' most basic divisions in terms of memory processes are between encoding, storage, and retrieval. Storage is typically divided into three types: long-term, short-term, and a very brief type of sensory storage that goes by various names, including phonological loop, iconic memory, and working memory. The very brief type is sometimes divided into separate stores for different sensory modalities, and long term memory is often divided into declarative, and procedural memory, both of which also subdivide into further types. A combination of sources of data contribute to this basic taxonomy: introspection, behavioral experiments, and studies of amnesic patients.

Bechtel relates how these basic divisions guided attempts to localize memory processes in the brain, using PET and fMRI, animal physiology, as well as lesion data from amnesic patients. H.M.'s combination of retrograde and anterograde amnesia after hippocampus resection focused localization attempts on that part of the medial temporal lobe. Much research on memory has been devoted to working out which memory functions are performed in each section of the hippocampus. However, as Bechtel (2008) relates in a footnote, animal research suggests that the main functions of the hippocampus may be in spatial information

processing, and that hippocampal lesions may affect memory by damaging axon pathways from the basal ganglia to the basal forebrain which happen to pass through the white matter of the medial temporal lobe. The suggestion that the hippocampus has little do with non-spatial memory remains controversial, but if true should count as a major failure of the localization heuristic.

Another section of Bechtel's chapter on memory is devoted to "evidence that challenges the distinctions between episodic and semantic memory, between short- and long-term memory, and between memory and other cognitive functions" (Bechtel 2008). The first piece of research reviewed also "brings into question the assumption that encoding and retrieval are really distinct operations" (Bechtel 2008), despite Tulving's earlier hypothesis that encoding and retrieval are performed, respectively, in left and right prefrontal cortex. This is followed by a summary of evidence suggesting that some semantic retrieval is done in the area supposed to be devoted to episodic retrieval, and vice versa. Next comes evidence that long- and short-term memory are not separate systems, but rather depend on shared operations in the same brain areas. Finally comes evidence that memory and language processing run together, and the suggestion that memory storage is inseparable from neural processing in general.

In summary these divisions proved a very unreliable guide to research into the neuroscientific basis of memory, and just about everything psychologists say about memory fails to map onto the neuroscience. Nevertheless, these categorizations into short- and long-term memory, encoding and storage, etc., have not by any means been abandoned by psychologists. They are still found in psychology textbooks (sometimes accompanied by a short note indicating that the categories have been called into question by work in neuroscience), they still form the foundation for experimental paradigms, and they are still considered useful for clinical purposes, since the deficits of amnesic patients do dissociate roughly into retrograde and anterograde, declarative and procedural, short term and long term. For example, people with retrograde and anterograde amnesia behave differently and require different care. Coping skills and memory aids are central to dealing with anterograde amnesia, while time, memory exercises and therapy can help with retrograde amnesia.

According to Piccinini and Craver (2011), this sort of schism between psychology and

neuroscience shouldn't occur. In discussing a task analysis of memory into encoding, storage and retrieval, they claim that, "If the study of brain mechanisms forces us to lump, split, eliminate or otherwise rethink any of these sub-capacities, the functional analysis of memory will have to change" (Piccinini and Craver 2011). The functional analysis of memory has not changed though. The situation is reminiscent (but even more challenging to deniers of psychology's autonomy, I think) of what Aizawa and Gillett (2011) describe in vision science, where discoveries about the neural underpinnings of color vision do not result in vision scientists changing their higher level categories about types of color vision.

Bechtel is perfectly right that cognitive models of memory guided later neural investigations, and that decomposition and localization was used as a heuristic. Perhaps the point was that despite the ultimate failure of localization, using this heuristic helped in the discovery of how badly the decompositions match up. I'm inclined to think that the cognitive models of memory were an extremely unhelpful guide, however, and that the heuristic failed (as heuristics sometimes do). More troubling for this view of integration is the resulting disconnect between cognitive and neural models, which does not seem to bother psychologists. This episode can hardly be taken as a paragon of integration. That this case study is presented as a prime example of the sort of integration achievable by cognitive neuroscience is not a very hopeful sign. This is not to say that integration is hopeless, nor that the decomposition and localization heuristic is not extremely important. What it shows is that something other than decomposition and localization, or treating cognitive models as elliptical mechanism sketches, is needed to make integration work in most cases.

2.5.2 Other Approaches to Integration

Finally, I briefly address two other accounts of how cognitive models and mechanistic explanation might fit together, and clarify how my position compares to them.

Weiskopf (2011) argues that functional analyses are not in general mechanistic, because they can be noncomponential (for instance if the whole rather than a part performs a task), but that they nevertheless fulfill the same norms for what makes an explanation a good one. Weiskopf makes many excellent points. I fully agree that functional analyses can be

good explanations regardless of whether they are fine-grained. I think he's right that an important category of psychological explanations are noncomponential. I also agree that functional and structural decompositions can cross-cut one another, which is a point that much of the current literature on mechanisms glosses over.

One point of disagreement I have with Weiskopf (and Craver) is over whether noncomponential functional analyses can be mechanistic. It is widely assumed that mechanistic explanations are by definition componential. My suspicion is that this is one of the subtleties underlying MDC's insistence that mechanisms be described as consisting of entities rather than components or parts. In the next chapter I describe how explanations that do not refer to decompositions into smaller entities and their activities can nevertheless be mechanistic. In short, the entities and activities which provide the explanation need not be parts or components of the mechanism, but could alternatively be larger scale structures into which the explanandum fits.

Weiskopf does much useful clarificatory work pointing out how at least part of the difference between how-possibly, how-plausibly, and how-actually models is how much evidence has been accrued in their favor; and pointing out that the accuracy of a model is a matter not just of its correctness, but also of its grain. Two more distinctions closely related to grain could also stand to be highlighted. One is the difference between what Weiskopf calls grain, which in his example is a distinction between different ways of decomposing a mechanism; and levels of mechanisms. The hippocampus as composed of CA1, CA3 and dentate gyrus versus the hippocampus as composed of cells of various types organized in several layers is a difference of grain, as in Weiskopf's example; while the difference between the entire hippocampus and just the CA1 component, or just the the pyramidal cell layer of CA1 is a difference of mechanism level.

An additional variable I would distinguish is specificity, which I describe in more detail in the next chapter. Specificity is like grain in that it points to different ways of understanding the entire mechanism, but rather than distinguishing how big the chunks are that the mechanism is decomposed into, specificity distinguishes how general the types are that the mechanism is taken as instantiating. This is the difference between hippocampus as connected network, hippocampus as connected network of neurons, and hippocampus as

connected network of neurons in my brain at this moment, for example. It is also worth noting that neither granularity nor specificity admit of just one hierarchy for any given mechanism. The hippocampus could just as well be decomposed in terms of grain into the left, right and middle sections instead of CA1, CA3, and dentate gyrus, then further into crosswise slices of cells rather than the usual cell layers. Likewise, alternative hierarchies of specificity are possible. Both are perhaps better seen as comparative variables rather than metrics.

Another disagreement with Weiskopf, is in how to understand schemas. Weiskopf places schemas on a continuum from sketches to fully-elaborated mechanisms. On my reading, schemas are abstract representations of fully-elaborated mechanisms rather than partially-elaborated ones, so I would not place them on a continuum between sparse and complete elaborations of mechanisms. Admittedly MDC's account of schemas is rather confusing, and it requires significant work to puzzle through what exactly they have in mind. In the next chapter I give a more detailed account of how to make sense of schemas in a way that is consistent with what M, D, and C say in various sources, and with the term's use in science.

Overall, our conclusions on the main points are similar: Piccinini and Craver (2011) are too quick in assimilating psychological explanation to mechanistic explanation, because some cognitive models can not be usefully decomposed further into neural parts, and functional and structural decompositions can cross-cut one another such that their parts do not neatly map onto one another. Where our views depart is over whether lack of decomposability and/or lack of neat mappings between decompositions means that cognitive models can't be mechanistic. I'm willing to let mechanisms bottom-out at any level, even if we know perfectly well how to cut the entities up into finer-grained non-working parts, and have resigned myself to the untidy possibility that hierarchies of mechanisms probably do not all fit together. [8]

An approach that differs from both mine and Weiskopf's is that of Glennan. Glennan (2005) defines mechanisms in terms of the functional relations between parts, and notes that when functional decompositions do not match up with spatially localizable parts, it is still the functional structure that is constitutive of the mechanism (Glennan 2005, 447). This means that multiple models can all be 'correct.' Glennan endorses Giere's view of models as

[8] I have in mind something like Dupré's (1995) account of the disunity of science.

being like maps, in that they represent only some aspects of the system modeled.

I agree with Glennan that models only account for some aspects of the system they model, and thus there may be many models, even apparently contradictory ones, all of which are 'correct.'[9] Where I disagree is over the role of representation. The map metaphor is a nice metaphor, and models are indeed partial, but I do not think they are partial representations the way maps are. Instead they are partial because they instantiate just some of the types to which the system belongs. A full elaboration of this point will have to wait until Chapter 5.

Weiskopf disagrees with Glennan, on the grounds that a mechanistic model "must actually be a model of a real-world mechanism—that is, there must be the right sort of mapping from model to world" (Weiskopf 2011). My position is with Glennan on the partiality of models and there being multiple correct ones, but with Weiskopf on models having to be of a real-world mechanism. Both are possible, since there are many ways in which to be of the real world. I could be modeled as a human, a mammal, a vertebrate, a native of Ottawa, a right-handed redhead, a balcony gardener, etc., each of which would capture different aspects of my person, possibly even in contradictory ways, but all of these partial models are of a real-world mechanism. This point I elaborate further in the next chapter.

2.6 CONCLUSIONS

So far in this chapter I have explored and compared the explanatory frameworks used in neurobiology and cognitive psychology. Much of cognitive psychology does not seem to be concerned with developing models that lend themselves easily to being used as sketches of neural mechanisms. Instead of increasingly detailed accounts of which parts of the brain correspond to the entities appearing in cognitive models, many of cognitive psychology's explanations take the form either of data flow models or process flow models. Piccinini and Craver (2011) insist that psychologists ought to pay attention to neural constraints when building their models, because they want these models to be seamlessly integrated as sketches

[9] See Mitchell (2000) for a detailed elaboration of this sort of pluralist view.

of neural mechanisms.

The situation seems to be like one where I want to start a collection, but need a box to put it in. If I see a friend's box which they're using for entirely another purpose, and complain to them that their box ought to have different sized compartments in it, so that my collection will fit nicely, a rational response would be for the friend to tell me to take a hike. My worry is that cognitive psychologists, like the owner of the box, might have no reason for submitting to the changes being demanded of them.

I think it's perfectly true that what some people in cognitive neuroscience are doing is trying to treat cognitive models as elliptical sketches of neural mechanisms. As a descriptive account of mainstream cognitive neuroscience, Piccinini and Craver get it just right. For this sort of project to be successful, it's also perfectly true that psychologists should let knowledge about the brain constrain their practice. Likewise, neuroscientists should let knowledge about behavior constrain their practice, and not just use restricted, operational definitions of behavior. This point could stand to be emphasized much more strongly.

My point is not to support a strong form of autonomy of psychological explanations, nor to argue that psychological and neurosceintific explanations are necessarily of distinct types. Rather my point is to mark the differences so as to understand the work that needs to be done if the goal of integration is to be achieved. Glossing over the differences and treating all psychological explanations as elliptical sketches of neural mechanisms seems like an approach likely to alienate psychologists and unlikely to provide helpful direction.

The tricky work to be done is figuring out how to go about doing psychological work while keeping in mind constraints from neuroscience, without this interfering with the main goal of forming good cognitive models; how to go about doing neuroscientific work while keeping in mind richer descriptive definitions of behavior; and most importantly, what to do when the two sets of constraints conflict.

In the next chapter, I work toward a positive account of how psychological models might be understood in mechanistic terms, despite them often making for poor sketches of neural mechanisms, and how integration might thereby proceed. It makes for a complicated story, not a seamless one, but one that might be more agreeable to psychologists.

3.0 COGNITIVE MECHANISMS

In the previous chapter I critiqued an account by Piccinini and Craver (2011) of how cognitive psychology and neuroscience might be integrated through multilevel mechanistic explanation. I argued that cognitive psychology's models are often not intended as sketches of neural mechanisms, and often do not consist of black or gray boxes standing in for components that can simply be identified with neural mechanisms. Thus, turning cognitive models into sketches of neural mechanisms is not as seamless as Piccinini & Craver make it out to be, and it remains unclear how mechanistic explanation might nevertheless be the key to integrating cognitive psychology with neuroscience.

In this chapter I work towards a less seamless account of integration through mechanistic explanation. There are good candidates for explanations in psychology that do not fit the integration story spelled out by Piccinini and Craver (2011). Rather than denying that they are good explanations on the grounds that they do not fit that account, I look for alternative ways of incorporating them into a mechanistic framework. This requires extending the MDC account of mechanism, or at least illuminating some of its darker corners.

Among the candidate explanations that do not take the form of decompositions where the components might be identified with neural structures, are explanations that appeal to general patterns or abstract principles. Several commenters on the MDC account have worried about this sort of example, and some of the more recent elaborations of the account, at least by Craver, seem to deny that generalizations and abstractions play any important explanatory role. Here I carve out some space for generalization and abstraction in the MDC account of mechanistic explanation, in the form of what I call generic mechanisms. Later in Chapter 4, I illustrate how generic mechanisms are discovered and employed in computational cognitive neuroscience, and argue against claims that the idealizations and

simplifications involved in computational models mean that they can't generate empirical evidence applicable to real cognitive or neural systems.

3.0.1 Outline

I have two main aims in this chapter. One is to review the literature on the MDC account of mechanisms, with an eye to finding resources that might be helpful for constructing an account of cognitive mechanisms. To this end I first outline the MDC account of mechanism, and review recent developments and refinements of it. Then I highlight some changes that have been suggested, and some problems and puzzles that have been raised. The second aim is to outline my proposal for how cognitive mechanisms might be understood. In Section 3.3.3 I focus on the role of generalization and abstraction in mechanistic explanation, describe what I call generic mechanisms, and argue that they solve some of the problems raised in the recent literature on mechanisms, while remaining compatible with (and possibly an intended but not well elaborated part of) the original MDC account. Craver has recently claimed that making an explanation more detailed makes it more explanatory. I argue against this claim. Finally I connect this back to the previous chapter's discussion, pointing towards a role for generic mechanisms in the integration of cognitive psychology and neuroscience.

3.1 THE MDC ACCOUNT OF MECHANISM

In order to evaluate whether mechanistic explanation is up to the job of integrating cognitive psychology with neuroscience, we need a clear elaboration of what mechanistic explanation is, and what resources it makes available. I'll focus here on the MDC account of mechanism and the literature it has spawned.

As mentioned earlier, the MDC definition is as follows: "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al. 2000, 3). Two important differences between this definition and its competitors (Bechtel and Richardson 1993, Glennan 1996) are

that it gives equal weight to entities and activities as fundamental components, rather than defining one in terms of the other, and it does not rely on laws for providing the explanatory connections between start and finish conditions (in more recent versions, Glennan has given up laws too). Activities are taken to be types of causes, and what the acceptable types are is a matter for empirical discovery.

Another difference between Bechtel's account and MDC is over whether a mechanism is a causally-active thing in the world, or a representation. Bechtel's mechanisms are representations and his mechanistic explanations are the texts, diagrams, and so on that people use in the act of explaining with mechanisms.

It is not entirely clear where MDC stand on this question at first glance. On the one hand, giving a description of a mechanism is how you explain the phenomenon the mechanism produces, according to MDC. This claim and several others like it suggest that their account of explanation is an epistemic one like Bechtel's. But on the other hand, MDC also make clear that the productive continuity in a mechanism is the key to mechanisms explaining. They say, "explanation involves revealing the productive relation. It is the unwinding, bonding, and breaking that explain protein synthesis" (Machamer et al. 2000). In this quote it seems that they intend mechanistic explanations to be ontic; unwinding, bonding and breaking are real world activities, not representations.

My interpretation is that MDC take mechanisms to be the things in the world that produce changes, but that what they usually refer to as mechanistic explanations are the linguistic acts that refer to mechanisms. There is some ambiguity here. Although for MDC, mechanisms are the causally-active things in the world, they certainly find the epistemic aspects of explanation important too. The role of diagrams and models for understanding mechanisms is highlighted, they discuss what makes phenomena intelligible, and provide an analysis of the sorts of representations of mechanisms scientists make.

MDC distinguish two kinds of descriptions of mechanisms: schemas and sketches. A schema is "a truncated abstract description of a mechanism that can be filled with descriptions of known component parts and activities" (Machamer et al. 2000, 15). A sketch is "an abstraction for which bottom out entities and activities cannot (yet) be supplied or which contains gaps in its stages. The productive continuity from one stage to the next has miss-

ing pieces, black boxes, which we do not yet know how to fill in" (Machamer et al. 2000, 18). That an epistemic distinction is made between schemas and sketches has been taken to mean that schemas must be abstractions constructed from already complete descriptions. It is unclear, however, whether for something to be considered a schema, the details of the components must already be known.

MDC use the example of the discovery of how DNA is transcribed into RNA then translated into protein to illustrate how a preliminary sketch is turned into an adequate explanation of a phenomenon, by filling in the details of the components of the mechanism. It also shows how a mechanism can be abstracted into a schema like $DNA \rightarrow RNA \rightarrow protein$. In the course of this example, proposals for how some of the stages work are described, such as Gamow's "geometrical 'holes' schema." Note that in this quote, the proposal is referred to as a schema, suggesting that schemas needn't always be descriptions of complete mechanistic explanations. I'll discuss this ambiguity further in later sections.

3.1.1 Hierarchies

Another sort of axis along which mechanisms vary are hierarchies of levels. According to MDC, "Mechanisms occur in nested hierarchies... The levels in these hierarchies should be thought of as part-whole hierarchies with the additional restriction that lower level entities, properties, and activities are components in mechanisms that produce higher level phenomena" (Machamer et al. 2000, 14). MDC point out that descriptions of mechanisms often span multiple levels, and that where they 'bottom out' depends on a given field and its interests. The bottom out activities they identify for molecular biology and neurobiology include "fitting, turning, opening, colliding, bending, and pushing... attracting, repelling, bonding, and breaking... diffusion... conduction" (Machamer et al. 2000). They note that discovering new entities and activities is an important part of scientific practice.

MDC note that the scope of generality and the level of abstraction of a mechanism schema are not the same thing. Another thing to note is that levels of abstraction and levels of mechanisms are not the same thing. Abstraction means removing detail, usually yielding a description or representation. For example, one might describe a particular sewing

machine, including all of its parts in great detail, including any rust spots, scratches, fraying of wires, the degree of bend in the needle, down to the bits of lint stuck in the grease coating each part. A more abstract description might just describe the form and position of all the parts without those extra details. Still more abstract descriptions might call it a Singer 319 Automatic Swing-Needle sewing machine from 1956, or a household object weighing over 10kg. The same object as a whole is being described in each case, just with more or less detail.

In a hierarchy of mechanism levels, on the other hand, different chunks of the machine appear at different levels. Some of the levels are, from top down, the entire sewing machine; its thread guiding system; the tension knob; the metal spring within the tension knob; and the configuration of atoms in the spring. The sewing machine could also be a component in a larger mechanism like an assembly line. In this example, the whole object is only described at one of these levels (this might not be true of non-modular mechanisms). The other levels describe either more or less of the world. Any of these levels could be described either in detail or more abstractly. Levels of mechanisms and levels of abstraction are orthogonal.

Components of various sizes can be part of the same mechanism, explanations often span several levels in a hierarchy of mechanisms, and lower-level mechanisms are generally parts of higher-level ones. However, the hierarchy, its levels, and what counts as an entity or an activity all might change when you shift to a new phenomenon. There is no guarantee that what show up as components in one mechanism are also components in others. Instead there is reason to believe that the components will lump differently for different phenomena. As a result, we should not necessarily expect the hierarchies for all phenomena to fit together into one unified hierarchy with neat part-whole relations. This is a metaphysical claim that would need defending, if someone wanted to make use of it. In MDC it is explicitly denied that there need be a fundamental level where entities and activities bottom out. If the hierarchies all matched up, then we should expect the more fundamental levels to be shared (regardless of whether some fields care to look that low).

Beyond these basics, Machamer, Darden, and Craver do not all seem to agree on every point in their paper. Each of them has elaborated their position in more detail in subsequent publications. Some points of internal disagreement are whether mechanisms must operate

regularly, the role of causation, and whether hierarchies of mechanisms are unified. In the next sections, I review their more recent statements before going on to discuss the wider commentary on the MDC account of mechanism, and my own additions to it.

3.1.2 Later Elaborations by Machamer

One thing that was not perfectly clear in the MDC paper was what productive continuity means exactly. It has been taken as a roundabout way of saying causation. Machamer seems to confirm this: "Uncovering mechanisms is a process of learning about causes. Particularly, discovering activities, the 'doing' or productive parts of mechanisms, *is* the finding of causes" (Machamer 2004). Of course calling these doings causes only clarifies the picture if we have an account of causation. Machamer thinks of cause as "a generic term," however, and says that "one does not need a theory of *cause*" (Machamer 2004). He prioritizes finding out "the possible, plausible, and actual causes at work in any given mechanism," and expresses misgivings about "philosophers who search for a general metaphysical definition of causality" (Machamer 2004). He specifically mentions counterfactual accounts in this respect, which presumably means he has misgivings about Woodward's popular account as a theory of causation (not to mention Lewis's).

One change to the definition of mechanism that Machamer (2004) endorses is Bogen's (2005) suggestion that 'regular' should be dropped, so that mechanisms just need to produce changes rather than regular changes. In MDC, a mechanism working regularly was defined as "always or for the most part in the same way under the same conditions"(Machamer et al. 2000). I'll discuss Bogen's counterargument in Section 3.2.3.

Machamer's comments about abstract activities are of particular interest. In the course of dealing with questions about the relationship between entities and activities (such as, whether activities belong to entities), and whether forces, fields, or energy are or require entities, he comments, "Activities can be abstracted and referred to and identified independently of any particular entity, and sometimes even without reference to any entity at all. So, at least, activities existing as abstract objects exist independently" (Machamer 2004). If abstract activities can exist independently of entities, this suggests that perhaps entire

mechanisms in the world (not schemas or sketches) could be abstract. It is not entirely clear what we're to make of this suggestion, but at least one possible reading is that mechanisms needn't be tied to particular instances. Later in this chapter I will elaborate on how entities too can exist as a kind of abstract object, that is, they can have details removed and still be objects in the world.

Machamer (2011a) also comments on the categorization of phenomena, suggesting that what scientists are usually interested in are types of phenomena, not particular instances, and that their categories depend on their goals. Scientific categories of phenomena that are useful for one set of purposes like psychiatric treatment may not match up well with the categories appropriate for other types of knowledge about the world (Machamer 2011a). I take him to mean that it might happen that the entities and activities that we divide phenomena up into in a field like psychiatry might not fit with the categories in a related field like neuroscience. That is, there is no guarantee that the hierarchies of mechanisms from different fields will all merge nicely.

In summary, Machamer is critical of identifying productivity too closely with a theory of causation, has been convinced to give up regularity, and endorses abstract activities like forces as existing independently of entities. He also seems to imply in both Machamer (2004) and Machamer (2011a) that phenomena in social or clinical realms might not fit well with the categorizations of components used in other fields. He seems to support a wider scope of applicability for mechanisms, but without all the fields and their hierarchies of mechanisms being unified.

3.1.3 Later Elaborations by Darden

Darden's recent papers deal with several aspects of mechanisms, including regularity, levels, schemas, and abstraction. She also elaborates on issues in mechanism discovery and how theories in different fields are related.

In response to Bogen's (2005) suggestion that 'regular' be omitted from the definition of mechanism, and Machamer's (2004) embrace of the suggestion, Darden writes, "even if it is not constitutive of what it is to be a mechanism, most molecular biological mechanisms

operate with some degree of regularity. They usually work in the same way under the same conditions... Molecular biological textbooks are filled with diagrams of mechanism schemas that have very wide scope domains of applicability" (Darden 2008). Note that here regularity is defined more loosely than in MDC. She also describes regularity as meaning working "in the same way under the same conditions" earlier in Darden (2002). This understanding of regularity, where "always or for the most part" is dropped, heads off one of Bogen's arguments, which is that some mechanisms, like neurotransmitter release in response to action potentials, work with low probability. Infrequently working mechanisms are compatible with them working the same way under the same conditions, if we take this 'same way' as incorporating their irregularity. Examples like neurotransmitter release, which we want to include despite their low probability of successful working, are regular in their irregularity, so to speak, in that they work with a consistent, stable, albeit low probability, as I'll discuss in more detail in Section 3.2.3.

Darden's second brief argument—the wide scope of applicability of many mechanisms—gets at a different issue. I take it that one of the main selling points of the mechanistic account of explanation is that mechanisms can take over the work that laws were supposed to do in deductive-nomological (DN) explanation, without succumbing to the many pitfalls of that account (see Suppe (2000) for a review of these). In particular, for a mechanism to be useful as a target of scientific study, it should be able to explain more than a single historical episode, at least in principle. I'll discuss the issue of the generality of mechanisms later in this chapter too.

A bit of Darden's terminology that I've been using already is the distinction between black, gray, and glass boxes:

> The goal in mechanism discovery is to transform black boxes (components and their functions unknown) to gray boxes (component functions specified) to glass boxes (components supported by good evidence), to use Hanson's (1963) metaphor. A schema consists of glass boxes; one can look inside and see all the relevant parts (Darden 2008).

Since mechanism diagrams often depict entities as boxes and activities as arrows, I'll extend the metaphor further to include black, gray, and glass arrows, so as to make clear that the same distinctions in status apply equally to entities and activities, regardless of how they are depicted diagrammatically.

One slight quibble with the box analogy is that there is no type of box assigned for when we have good evidence that a particular entity is involved in the working of a mechanism, perhaps because the mechanism ceases to work when the entity is removed, but we don't yet know what its function is. There might also be analogues where we know that a particular activity is involved in a mechanism, but don't know what role it plays (this often arises while watching experts like chefs, athletes, or artisans perform skilled tasks). I will not invent a color of box for these cases, for fear of stretching the metaphor too far, but do want to note the omission. It seems that the box metaphor is better suited to describing top-down research strategies than bottom-up ones, or in Bechtel's (2005) terms, scenarios where functional decomposition precedes structural decomposition.

3.1.3.1 Schema Instantiation Darden highlights an additional use for schemas aside from their glass box use of representing known mechanisms in a concise, abstract way. They can also be used as tools in scientific discovery. In a discussion of scientific discovery, Darden says,

> Mechanism schemata are abstract frameworks for mechanisms. They contain place-holders for the components of the mechanism... Often these place-holders characterize a component's role in the mechanism. Discovering a mechanism involves specifying and filling in the details of a schema, that is, instantiating it by moving to a lower degree of abstraction Darden (2002).

If a schema can be filled in with details during the process of discovery, then at least in this context, there is no clear contrast to be drawn between sketches and schemas. The glass box notion of schema seems superficially incompatible with a role in discovery, since glass boxes are representations of adequate, complete mechanistic explanations.

A simple explanation might be that Darden is using different terminology than was used in MDC, and describing here what in MDC was called a sketch. I don't think this is the case. She talks about schemata being either chosen or sketched in discovery episodes, and in historical cases. She also talks about how drawing analogies to other phenomena, or phenomena in other fields can be sources of schemata. The idea is that while in some cases, a new schema must be built up through the process of filling in the details of a sketch to yield a mechanism, only later to be represented in the truncated, abstract form of a schema;

in other cases, a schema that has been successful elsewhere can be put into action in a new context. Darden cites the example of reverse transcriptase to show how "discovery of a new type of module... expands the space of possible mechanisms" (Darden 2002). In contrast, the $DNA \rightarrow RNA \rightarrow protein$ schema, once worked out, "could be instantiated whenever a protein sysnthesis mechanism was needed" (Darden 2002). This is the difference between sketching a new schema and reusing a known one. Darden (2002) also mentions that analogous theories can be grouped and the specific details removed to construct an abstract schema, citing examples like wave phenomena and selection. In this use too, schemas are not glass boxes.

Darden calls this way of using a known schema in discovery episodes *schema instantiation*. Schema instantiation does not mean providing the details omitted from a glass box representation. It means figuring out how a schema known to work in other cases can be used as a template for discovering the components of new mechanisms that rely on the same basic principles, as described here:

> Schema instantiation begins with a highly abstract framework for a mechanism, a schema, which is then rendered less abstract during the process of instantiation. Instantiation is usually characterized as supplying values for the variables in a schema, as in Kitcher's (1989) discussion of the instantiation of deductive argument schemata. This is too restrictive. Schemata may be stated with varying degrees of abstraction; one may specify details piecemeal to make a mechanism schema less abstract before one gets all the way down to a description of a particular mechanism (Darden 2002).

As Darden mentions, schema instantiation is in some ways similar to Kitcher's idea of instantiating variables in argument schemata, but much more flexible. It also differs from Kitcher's schemas in that unification is not a specific goal. Instead, making use of schemas that have previously been successful is a way of generating promising hypotheses. For example, using the strategy of modular subassembly, "one hypothesizes that a mechanism consists of known modules or types of modules." Darden argues that "finding these recurrent motifs has been a powerful tool in discovery in biology" (Darden 2002).

It is important to note that although Darden does not endorse Kitcher's unificationist view of explanation, nor his specific means of achieving unifications through argument schemata, she does defend the importance of drawing connections between different theories and phenomena, whether it be within one field, or across different ones. Schemas are ways of

representing this general knowledge that applies to multiple examples, possibly across fields, and they are important tools in mechanism discovery.

Also important is that the relationships between phenomena and the mechanisms that explain them can take many forms, just as interfield theories can posit identities, part-whole relations, structure-function relations, or cause and effect relations (Darden and Maull 1977, Darden 2005). Decomposition into smaller parts is one way of explaining mechanistically, but "finding the mechanism that produces a phenomenon may require not further decomposition of a system, but instead going 'up' in size level" (Darden 2005). Darden repeatedly makes the point that decomposition into smaller scale components is not always the most promising route to take when looking for mechanistic explanations: "a characterization of mechanisms and their working entities shows that one does not always move down to a lower size level to find the mechanism producing a phenomenon" (Darden 2005), "the analysis of working entities in a mechanism ... serves to block the reductive move that one always gains understanding by investigating the gory details of the smallest entities present" (Darden 2005).

These points—that mechanisms can explain in multiple ways, and not always by going into gory details—will be key to developing an account of cognitive mechanisms later in this chapter. More will need to be said about how going up in size works exactly, and how these upward-looking mechanisms explain. Additionally, a line of argument taken up by Craver stands in opposition to Darden's claims, as we'll see in the next section, so this disagreement needs to be resolved.

3.1.4 Later Elaborations by Craver

Craver has expanded on the MDC account of mechanism in several directions, including its applicability to psychiatry (Kendler et al. 2010), and dynamical systems (Kaplan and Craver 2011), as well as discussions of role functions (Craver 2001), top-down causation (Craver and Bechtel 2007), and integration (Craver 2005, 2007, Piccinini and Craver 2011). I'll focus here on one aspect of Craver's work: norms of explanation.

In his 2006 paper and 2007 book, Craver argues that explanatory adequacy requires

more of a model than just phenomenal adequacy. In addition to getting the description right, the mechanisms responsible for producing a phenomenon must be identified, and, according to Craver, in detail. Craver argues that the lesson from "several decades of attack on covering-law models of explanation at the hands of advocates of causal-mechanical models of explanation" is that "merely subsuming a phenomenon under a set of generalizations or an abstract model" does not explain the phenomenon (Craver 2006). I will argue that Craver's move away from generalization and abstraction goes too far. That the covering-law approach failed is uncontroversial, but descendants of Salmon's causal-mechanical approach are perhaps not the only alternatives. One could just as well take the failure of covering-law approaches as evidence that explanation should not be thought of as a syntactic relation between sentences.

Craver contrasts mechanistic models with models that provide an equation that describes the phenomenon, but without specifying what the underlying details are that make the equation a good fit. He also criticizes how-possibly models, which he characterizes as "loosely constrained conjectures about the mechanism that produces the explanandum phenomenon," and he complains that "One can have no idea if the conjectured parts exist and, if they do, whether they can engage in the activities attributed to them in the model" (Craver 2006). Craver's main example is the Hodgkin-Huxley model of the action potential, which was constructed with the goal of being phenomenally adequate, but also suggested possible mechanisms. Craver makes a convincing argument that the Hodgkin-Huxley model was not considered either by its proponents or its critics to be an explanatory model until much later, after neuroscientists had figured out some of the (actual) molecular mechanisms responsible. What Craver is after in an explanation are the real components of the mechanism that actually produce the phenomenon.

One worry motivating Craver to reject models that do not correctly identify and describe the details of mechanisms is that they might go wrong in unusual circumstances or exceptional cases. An adequate explanation, he says, must "account for all aspects of the phenomenon, not merely part of it. One sign of a mere model (i.e., one that is not explanatory) is that it is phenomenally adequate only for a narrow range of features of the target phenomenon" (Craver 2006). A good explanation should reveal the exceptions: "an

explanation shows why the relations are as they are in the phenomenal model, and so reveals conditions under which those relations might change or fail to hold altogether" (Craver 2006). Craver's worry is that a phenomenal model that summarizes the known data perfectly could still go wrong in untested cases. By getting the underlying mechanism right, Craver hopes to avoid that possibility. This is why he thinks it's important not just to identify the mechanism, but to describe it in detail: "the model must correctly characterize the *details* of the mechanism" (Craver 2006, 7, my emphasis). If it's right that any adequate explanation must account for *all* aspects of the phenomenon, then going into the gory details of the mechanism should be a way of achieving that.

In Section 3.3.3 I'll discuss both this desideratum and Craver's strategy for achieving it. For now I'll just note that I see it as two separate questions whether what actually produces a phenomenon must be identified in a good explanation, and whether it must always be identified in gory detail. Perhaps there are ways of identifying the real difference-makers that do not depend on the details. I also worry that Craver's strategy may be unrealistic. Getting every potentially relevant detail for any context correct in a model would mean including everything, which defeats the purpose of making a model.

Craver says that "A model can be richly phenomenally adequate and non-explanatory" (Craver 2006). Presumably a richly phenomenally adequate model would account for all known aspects of the phenomenon, getting even the exceptional or unusual cases right, so it is unclear how his concern with exceptional cases justifies rejecting phenomenal models. I gather the sticking point is whether a phenomenal model would also deal with counterfactual cases. Craver argues that constitutive explanations, and by extension mechanistic explanations, must identify causally relevant components if they are to get all the details right. Four of his five "norms of explanation" from Chapter 2 of Craver (2007) mention causes. Getting the underlying causes right is connected to counterfactuals for Craver, since he endorses Woodward's interventionist account of causation (see Craver (2007), Chapter 3), which relies on reasoning about counterfactuals.

In summary, Craver gives two types of argument for detailed descriptions of causally relevant components: the need to avoid the problems that arose for covering-law accounts, and the need to account for all aspects of the phenomenon, even counterfactual ones. The

first argument supports a productive account of explanation, and counts against syntactic ones, but there should be several options available that could meet this requirement. The second argument supports the model being highly detailed, although there might be other ways of getting the underlying causes right aside from cataloguing details.

3.2 BEYOND THE MDC ACCOUNT OF MECHANISM

The MDC paper has become the most cited paper in the journal *Philosophy of Science*, and many papers have been written in response it. Thus I cannot do justice to all of the new mechanist literature, but will review several papers that deal with topics relevant to my goal of carving out space for cognitive mechanisms. These topics are ontic explanation, schemas, regularity, types, and generalization.

3.2.1 Ontic Explanation

One major disagreement in the literature on mechanistic explanation is over whether a mechanistic explanation is more like a cause that produces the explanandum, or more like a linguistic act that produces understanding of the explanandum. These are ontic and epistemic views, respectively. Both are acceptable ways of using the term 'explanation' in everyday and scientific language, although the epistemic interpretation is perhaps primary, especially in everyday language.

The job of a scientific explanation is to indicate how the production of the explanandum comes about. In the DN tradition, explanations were collections of logically-related sentences. The logical relationships between these sentences (their forming a proof of the explanandum sentence) was where the explanatory force came from. It just happened that in DN explanation, what produced, or made the explanandum so, were the sentences, or more precisely, their logical structure. Both the vehicle and the content of the explanation were in the sentences. I take it that one of the major failings of the DN model of explanation was that logical relations between sentences do not do a good job of capturing explanatory

relations. We should not expect the logic of these representations to inform us about the logic of how the world produces phenomena, even though the content of these representations should be informative about what scientists believe goes on in the world, and if they're doing a good job, about what actually goes on in the world.

In mechanistic explanation, we look to the world for the productive or making so relationships that explanations are supposed to capture; the vehicle and the content of the explanation are thus split. Since the question of primary interest to scientists when they are looking for explanations is what conditions in the world produce or make a phenomenon so—not which sentences are used—it does not seem unreasonable to use 'explanation' to refer to the content rather than the vehicle of an explanation. In mechanistic explanation that content is a mechanism in the world, not a set of sentences. I take the more fundamental question—that of what brings about the phenomenon—to be the one of greater scientific interest. Perhaps this is a matter of preference.

This is not at all to deny the importance of studying the rhetoric of science, and how scientists think about and understand science. Scientists need ways of representing knowledge both internally, and for the purpose of communication. How they do this is just a different question than what the content of those representations are. My inclination is to allow for both uses, and to recognize the importance of both. In this chapter I am more concerned with the ontic aspect of explanation, so will use the term in that way here. In Chapter 5 I am mainly concerned with models, so will have more occasion to discuss explanation in the epistemic sense. Where there is the possibility of ambiguity, I will refer to ontic explanations as either the contents of explanations or explanations^{o}, and epistemic explanations as either explanatory texts or explanations^{e}.

Two of the most prolific defenders of mechanistic explanation, Craver and Bechtel, disagree on whether mechanistic explanations are ontic or epistemic. Bechtel & Abrahamsen argue that explanation "is an epistemic activity, what figures in it are not the mechanisms in the world, but representations of them" (Bechtel and Abrahamsen 2005). They distance themselves from older epistemic views of explanation like the DN model (Hempel 1966) in which explanations proceed by logical deductions from statements of laws. They allow for explanations to take forms other than propositions—such as diagrams—and for the structure

of those explanations to take forms other than logical deduction—such as spatial, temporal or similarity relations. Bechtel argues for his epistemic view of explanation again in Bechtel (2008) in similar terms. Wright and Bechtel (2007) argue that by talking about descriptions of mechanisms, MDC are admitting that representations are necessary to explanation. They charge that MDC vacillate on the ontic view, and bring in representation through the back door: "This vacillation reflects recognition that a necessary condition on mechanistic explanation is that the structure, function, and organization of mechanisms needs to be captured and codified representationally" (Wright and Bechtel 2007, 52). As I mentioned above, it is clearly the case that in order to explain in the sense of communicating knowledge, representations are necessary. So in a trivial sense Wright & Bechtel are right. It is also clear that in order to communicate what they want to say about the structure, function, and organization of ontic explanations, MDC must use language. I don't see how the necessity of representing for communication makes the things discussed also necessarily representational. Either this is a weak argument or there is a deeper metaphysical disagreement lurking here.

Craver follows Salmon (1984) in taking an ontic view of explanation. He argues that "good explanations in neuroscience show how phenomena are situated within the causal structure of the world" (Craver 2007), and defends against the varied epistemic views of Hempel, Paul Churchland, and Kitcher. For Craver, the explanans is the mechanism itself, not a sentence describing it, a diagram representing it, nor an argument schema unifying it with other explanations. Throughout his 2007 book, Craver treats mechanistic explanation as being synonymous with causal explanation. I gather Craver sees ontic explanation as a broader category containing causal-mechanical explanation as well as other types though. One of my concerns in this chapter is to argue for shifting those boundaries a little so as to make room for mechanistic explanations still under the umbrella of ontic explanation, but less tightly knitted to causation. I'll take this up again in later sections.

Another way this issue will arise later is in coming to grips with abstraction. I will argue that abstraction plays an important role in mechanistic explanation, and will try to illuminate that role. Abstraction playing a role in explanation raises a problem for the ontic view. Abstraction is closely associated with representation, and once we're dealing with representations, we're on the epistemic side of explanation. So it is not at all obvious how

one can make use of abstraction within an ontic framework. In Section 3.3.3 I will show how this can be done.

3.2.2 Schemas

There has been much confusion over what exactly a mechanism schema is. Recall that a schema is defined as "a truncated abstract description of a mechanism that can be filled with descriptions of known component parts and activities" (Machamer et al. 2000). No elaboration is offered for what truncated is supposed to mean, but perhaps it means that only a slice of the whole mechanism is shown, and not necessarily the entire thing from startup to termination conditions. It might also mean that some of the contextual factors that affect the working of the mechanism are not considered. These might be necessary reactants that are assumed to be present, or possible inhibitors that are assumed to be absent. Since schemas are described as complete relative to sketches, more could be said about what it means for a schema to be at once complete and truncated.

The abstractness of schemas bears some explaining too. Given that sketches and schemas are representations rather than the things represented, they are in one sense already abstract. They are called abstract representations then, because they leave out some of the detail from the representation. "Abstraction is an issue of the amount of detail included in the description of one or more mechanism instances... One can describe a single instance of a mechanism more or less abstractly" (Machamer et al. 2000). Abstractness is distinguished from generality, which refers to the wideness of the scope in which a schema can be instantiated. This is a perfectly reasonable distinction. Nevertheless, describing a mechanism more abstractly gives it the potential to apply more generally. Sometimes the reason for abstracting is increasing generality. Little is said about generality in MDC, aside from the drawing of this distinction, but the definition of schema comes right after a discussion of how scientific development proceeds by the discovery of *types* of mechanisms and *types* of entities, and the statement that scientists are "typically interested in types of mechanisms" (Machamer et al. 2000). Part of the point of schemas is to summarize general phenomena, or how types of mechanisms work.

A source of much confusion about schemas is that as the term is defined by MDC, it looks like an historical distinction. Overton (2011) and others have pointed out that whether something is considered a schema or a sketch is context-dependent. It is true that sketch is usually a context-sensitive designation. A sketch has gaps in its productive continuity, and only becomes a schema when it is sufficiently filled in relative to the needs of a field, and when its correctness is sufficiently established, relative to the needs of that field.

Schema is often used in MDC as a success term for sketches at one extreme of this developmental axis. It is perhaps unfortunate that schemas were introduced in contrast with sketches, as this has misleading side-effects. As Glennan (2005) has pointed out, the progression from sketch to schema is a continuous one, not a two-stage one, as may not have been clear in MDC. Further, that schemas are contrasted with sketches gives the false impression that a schema is tied to one historical episode. However, once a schema is successful in one case, it becomes part of the toolbox of model builders, as Darden has argued, and can then be deployed in various contexts, where it may or may not end up being the basis of other successful explanations. In that sense, schema is not always a success term, and does not belong on an axis from sketch to fully-elaborated mechanism.

Overton also finds it troubling that mechanisms have a different ontic status than schemas and sketches, and he takes it that schemas and sketches a have different ontic status from one another. He writes,

> The ontic distinctions are particularly problematic since the three cannot all lie on a spectrum: a mechanism is an arrangement of entities and activities, while a mechanism schema is a description, and a mechanism sketch is an abstraction. For instance, glycolysis would be a mechanism in molecular biology, a mechanism schema in quantum chemistry, and a mechanism sketch in quantum field theory. We would like 'glycolysis' to refer to the same thing in all three cases, but surely we cannot change the ontic status of glycolysis simply by changing the context (Overton 2011).

Part of the problem is easily resolved: sketches and schemas do not have different ontic status: they are both representations of mechanisms. Sketches are just ones that are considered unfinished or gappy. So the problem is not as dire as Overton describes.

Overton is right that fields might disagree about whether a representation of a mechanism is detailed enough to be considered non-gappy. Disagreements about the representative adequacy of representations of a mechanism do not shift the ontic status of those representa-

tions, so do not raise a serious problem for the definition of a mechanism. It is true, however, that scientists in different fields have different ideas about what should count as a mechanism. Systems biologists might think of glycolysis as a mechanism, but biochemists in fact call glycolysis a pathway. What biochemists call mechanisms are the molecular processes that occur within pathways like glycolysis. In many cases, it is the processes one level lower than the main focus of investigation in a given fiels which are typically referred to as mechanisms in that field. There is a narrower use of the term 'mechanism' specific to particular fields.

This poses a bit of a puzzle for the MDC account of mechanism, which advertizes itself as being based, in part, on the practices of scientists, and what they claim to be doing. I think there's a fairly simple solution, however. The MDC account of mechanism is supposed to be a general account of how explanations tend to operate in the biological sciences. It is meant to apply across fields, such that anything called a mechanism in any branch of biological science might be covered. If the goal is to understand particular explanations in biochemistry, for example, the sketch, schema, mechanism terminology is certainly not adequate for describing all the different levels involved. That is a different project than the more general account, so we needn't be too concerned about whether the scientists who study glycolysis indeed call it a mechanism.

In the broader use of the term 'mechanism' suitable for the cross-disciplinary project, glycolysis would count as a mechanism regardless of what researchers in some fields call it, how gappy their representations of it are, or whether they are concerned with it at all. Whether it is a relevant mechanism for study in a particular field certainly varies. Some fields might take it to be at the level they investigate, while other fields might investigate its components, or the parts of its components, and still others might consider it a process occurring within its components. This makes it a lower or higher level mechanism in relation to the field, but, coming back to Overton's point, this does not make it more or less abstract. Regardless of changes in perspective, glycolysis remains a mechanism, according to the terminology used here.

As mentioned, schema is not always used as a success term in science (nor always by MDC), which may be another source of confusion. The other use of schema is for types

of mechanisms that might be instantiated in different phenomena, even in different fields. In the examples of wave formation and selection mentioned in Darden (2002), a schema is constructed based on one phenomenon then applied to another within the same field through a local analogy, or applied to a phenomenon in another field through a regional analogy. In these cases, the schema consists of glass boxes in the original case, but when being applied to a new case, schema is no longer a success term, and the boxes are no longer of glass, because it remains to be seen how well the schema fits the new case and what the details in the box might be. Correspondences between the components represented in the schema and the components in the new target would need to be found and evaluated before it could be considered a successful schema in that new context. Darden (2002) calls this "schema instantiation" and one of the standard uses of 'schema' in science is for this sort of abstract structure that might be applied to any number of phenomena, whether successfully or not.

While it is inconvenient that schema has these two meanings, the solution I favor is a conservative one: to continue to use the term in both ways. MDC's use of schema as a success term reflects the fact that these abstract structures, which might be applied to any number of phenomena, often enter the toolkit of science by having first been schemas in the sense of completed, successful, filled-in mechanisms described abstractly. It is only after something has been recognized as a schema (as a success term) in one context that it becomes a viable option for application to other contexts. A hypothesis for how a mechanism might work that is unproven in any context would properly be called a sketch. So schemas as successful sketches are the same sorts of things as schemas as part of a scientist's toolkit, just at different times or in different contexts. There is not much danger in mistaking the one for the other, but in cases of ambiguity, it should not be too difficult to make this distinction. Where it isn't relevant whether a representation is a successful one yet in a given context, one can call it either a schema or a sketch (or perhaps just a model), and make a clear ontological distinction between that and a mechanism. In most cases, it is probably safe to assume that instances of the term 'schema' refer to abstract representations of mechanisms that were deemed successful in one context and are now available to be applied in other contexts.

3.2.3 Regularity

Another source of controversy in the MDC account of mechanism is their appeal to regularity. Recall that the definition of mechanism reads, "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al. 2000), and that the characterization of regularity is given as working "always or for the most part in the same way under the same conditions"(Machamer et al. 2000). As mentioned, Bogen (2005) has convinced Machamer (2004) to drop 'regular' from the definition, but Darden disagrees. I think Bogen makes several good points that need to be taken into account, but agree with Darden that some notion of regularity should be retained in order to recognize the importance of mechanisms working more or less the same way under the same conditions, and being of general applicability.

Bogen argues that some mechanisms work unreliably or infrequently, and that there might even be mechanisms that work only once, such as if the components of the mechanism were to break down after operating the first time. His main conceptual points are that causes needn't operate regularly in order to be causes, and that generality does not contribute to the truth of causal explanations, even if it may have other practical advantages. I agree with all of this, and yet resist the conclusion that regularity should be dropped entirely from the definition of mechanism.

3.2.3.1 Low-Frequency Mechanisms Bogen offers as an illustration of an irregular mechanism, how the action potential releases neurotransmitter to initiate post-synaptic activity.[1] The neurotransmitter release stage is apparently stochastic, and it fails much more often than it succeeds. Nevertheless, it is reliable mechanism, that can be precisely modeled by a Poisson distribution. The release frequency is very stable, and high enough to support the function it performs. Bogen is perfectly correct that this is not regularity in the sense of working "always or for the most part," but its working is regular in its irregularity, so to speak. Andersen (2011) offers a taxonomy of ways in which mechanisms might be regular to various degrees, and at various organizational locations, which takes into account this sort

[1]This example is also mentioned in Schaffner (1993) and explained in detail in Kuffler et al. (1984).

of infrequent regularity. These degrees of regularity are somewhat loosely and qualitatively described (deterministic, reliable, sporadic, infrequent), and seem to be meant as descriptions of some of the varieties of regularity that might be found in mechanisms rather than as criteria for what does and does not count as a mechanism.

Incidentally, since the writing of Bogen and Andersen's articles, more has come to light about how neurotransmitter release works. Within a synaptic connection, the individual synapses are heterogeneous in terms of their release probability, and these probabilities can change over time (see Branco and Staras (2009) for a review). Changes in the release probability parameter occur due to vesicle pool replenishment, and Ca^{2+} concentration, among other factors (Branco and Staras 2009). This arguably makes them even less regular than previously thought, although the lower level mechanisms underlying that irregularity are now better understood. I agree with Bogen and Andersen that there is no good reason to think that stochasticity couldn't remain all the way down, and that whether a mechanism turns out to be irreducibly stochastic, or dependent on still unknown factors should not be the deciding factor in whether we are correct to call it a mechanism.

Bogen's (2005) example of the irregularity of neurotransmitter release justifies a revision of the MDC account. The revision Darden implicitly suggests is to drop "always or for the most part" from regularity, but to retain working "in the same way under the same conditions" (Machamer et al. 2000, Darden 2008). Earlier I suggested that this 'same way' could be interpreted as including features like low frequency parameters, but more work needs to be done to fully characterize this suggestion. Andersen's (2011) taxonomy of regularity is a more elaborate option that also avoids Bogen's counterexample.

3.2.3.2 One-Off Mechanisms

Andersen's taxonomy of regularity allows for sporadic or infrequent mechanisms if they obey either of two conditions: "Known statistical distribution of indeterminacy" or "Known interfering factors" (Andersen 2011). Neither of these conditions supports one-off causal chains. Any mechanism for which there is a known statistical distribution of indeterminacy must be one that has succeeded multiple times, so one-off mechanisms are not covered by that clause. One example she gives of a case where known interfering factors are responsible for a mechanism not bringing about its end state

reliably is the attempt to produce Higgs bosons in CERN's Large Hadron Collider. In that case, the mechanism that should produce the bosons is part of a well-supported scientific theory, but producing Higgs bosons at all, and what is more, producing them in sufficient quantity for reliable measurement, is a process hindered by multiple interfering factors of a financial, practical, and political nature. Again, this is not a one-off causal chain. In fact, Andersen explicitly excludes one-off causal chains from being mechanisms; she states, "there exist causal chains that only occur once and thus are not mechanisms" (Andersen 2011). I agree that there might be causal chains that shouldn't count as mechanisms, including some that occur only once, but I think Bogen is right that occurring only once is not what makes an instance of causation fail to be mechanistic.

DesAutels agrees that jettisoning regularity completely from the account of mechanisms is a mistake. He argues that without regularity, mechanism becomes a term so wide in scope as to cease to be very useful as a scientific concept. Getting rid of regularity means that "any singular causal chain seems to be allowed to count as a mechanism" (DesAutels 2011).

As an aside, I do not find DesAutels's rather brief arguments for why irregular mechanisms could not be used for intervening and predicting convincing. He suggests that if mechanisms can be singular causal chains, any single result of an intervention would be attributed to a unique mechanism, so could not ground predictions of further occurrences, and likewise interventions could not be used to test general claims. This seems too strong. Even with a regularist account of mechanism, doing a single intervention isn't usually a good test of a general claim, and single experimental results are only very rarely used as the foundation for predictions. Scientists repeat their experiments when possible, and gather statistics about how frequent different outcomes are. These statistics are what ground predictions and test claims. Scientists can still do this if they are not regularists. They can still group together singular causal chains that came to the same result and talk about types of mechanisms and their frequencies of occurrence when making predictions or testing general claims. Allowing singular causal chains as mechanisms does not prevent one from sometimes grouping such chains together, nor force one to consider each individual chain as something sui generis.

DesAutels is onto something in tying the regularity of mechanisms to their usefulness

as a scientific concept. I think it's worth a closer look at one-off causal chains, to decide which ones we might want to include as mechanisms and what the criteria for their inclusion should be. Andersen's two conditions that can allow infrequent occurrences to nevertheless be counted as examples of mechanisms both involve the possession of background knowledge about a type of occurrence. One is knowing statistics about how frequently (or infrequently) it happens; the other is knowing what sometimes prevents it from happening. Although we can't have this sort of knowledge about a one-off causal chain, we might have this sort of knowledge about the types of causal chains to which the one-off chain belongs or about the causal chains that are its parts. We may not be sure of having produced a Higgs boson in a high energy collision, but we know about these sorts of collisions more generally, and about the production of other sorts of particles thereby. We also have physical theories that include a role for Higgs bosons.

Part of Bogen's point was that whether something was caused mechanistically by the coordinated activities of its components does not depend on how often it happens, and he allows for the possibility of mechanisms that operate only once. He does not go into detail, but I imagine the sorts of one-off causal chains he might have had in mind as mechanisms are events like the birth of the universe, or the speciation of *Homo sapiens*, which cosmologists and paleontologists certainly seek to explain, and quite possibly in terms of the coordinated activities of the relevant components. These are also cases where there is a background of scientific theory that the examples can be fit into. The speciation of *Homo sapiens* is an example of speciation more generally, and we know what some of the general mechanisms for that are. The Big Bang theory of the birth of the universe is likewise connected to theories of star formation and singularities. What makes these examples of one-off chains acceptable mechanisms is their relationship to more general mechanisms.

Some one-off causal chains should perhaps not count as mechanisms. Luka Magnotta murdering his lover, or me falling off my bike and skinning my knee are causal chains that I'm less sure should be called mechanisms. In these less obvious cases, I tend to think it depends on the details and the context of scientific knowledge. Presumably forensic psychologists would be interested in the Luka Magnotta case, so it may well be the subject of scientific explanation, but I would only call the causal chain leading to those events a mechanism if

it relates either to other cases of sadistic murder, or to general psychological principles. If Luka Magnotta ended up murdering his lover based on a series of random events like rolls of a die (however hard that is to imagine), and not because of precipitating circumstances in his psychological life, this might still be of interest to gossips and bloggers, but perhaps not to forensic psychologists. The case might be of scientific interest if it tells us something about the effects of living with abusive grandparents, for example, or if the habit of creating fake online personas turned out to be a reliable warning sign of murderous intent. If in some way Magnotta follows patterns characteristic of other sadistic killers, or if his actions can be explained in terms of principles of abnormal psychology, the causal chain leading to his actions might make for a mechanism. I do not want to suggest that the test for what should count as a mechanism is whether it happens to be of interest to scientists, because this might change on a whim. The rough idea, which I will make more precise shortly, is that the test for what should count as a mechanism is whether the process fits with other scientific theories. The interests of scientists are just an imperfect proxy for this.

Similarly, explaining how I skinned my knee might be a subject of scientific interest to emergency medicine specialists insofar as it relates to other cases of skin abrasions. They might fairly describe skidding along gravel and asphalt as a mechanism explaining a particular pattern of skin injury. Civil engineers might also find this causal chain of scientific interest insofar as it relates to the safety of intersections, and they might, as far as I know, talk about heavy vehicle traffic creating potholes and loose gravel as a mechanism explaining bicycle accidents. But simply as a one-off causal chain, if it were unrelated to more general mechanisms of skin abrasions or traffic accidents, it seems a bit much to glorify as a mechanism this episode of skidding around a gravelly corner, falling off my bike, and skinning my knee. The intuitive idea is that a context of scientific theory matters to these decisions of what counts as a mechanism, but how exactly the connections are made between these sorts of examples and scientific theories remains to be clarified.

So far I have used examples to motivate the point that some but not all one-off causal chains should count as mechanisms, and claimed that the important distinction between mechanistic and non-mechanistic causal chains is whether they relate to more general scientific principles. Now I will break down the problem further. A mechanistic causal chain

might be unique in either of two ways: because its steps have never before been executed in that particular order, or because it involves a unique step that has never occurred before. [2]

Take the first case, where it is previously unknown that a given effect can be produced by performing the steps of a mechanism in a different order, or mixed in a novel way with steps from other mechanisms. If the new combination were nothing mysterious, in that there is productive continuity, and each step reliably produces the right type of intermediate effect, then it seems uncontroversial that this would be considered a variation on an old mechanism, or a newly discovered mechanism.

If the steps consist only of accidental, random events that just happen to combine in such a way as to produce an effect that none of these steps are usually involved in producing, it's less clear that this should be called a mechanism, although there is no problem with saying that the random series of accidents caused the effect. That it's a random collection of accidents that bears little relation to known patterns of behavior for those sorts of entities seems to me a stronger reason for denying that it is a mechanism than that it has only happened once. If the same weird constellation of accidents were to happen a second time, I'd still be inclined to deny that it is a mechanism. There may be reasons for preferring mechanism to have a wider scope that includes such random causal chains, but Bogen (2005) and Machamer (2004) haven't said what they are.

The case where a hitherto unknown step occurs is more interesting. It might happen that a step bearing some similarity to a known effect occurs, but where this version of the effect was previously unknown. Early discoveries of light's wavelike properties provide an example. Grimaldi's posthumous 1665 publication described for the first time the diffraction patterns produced when light passes through a narrow slit. He concluded light must be a fine fluid in a state of constant vibration. Within a couple of decades Hooke and Huygens proposed precursors to the wave theory of light. Hooke delivered a paper to the Royal Society in 1672 critiquing Newton's 'New Theory about Light and Colors,' and suggesting that light is propagated by waves, like the vibration of a string. Huygens developed a wave theory of light that explained reflection and refraction results, which he first presented in 1678 and

[2]To even get started with this discussion, we need to allow for steps to be identified in a somewhat abstract way, otherwise every instance of causation would be a one-off event.

published in 1690. Much later, in 1845 Faraday showed that light's wavelike properties are related to electromagnetism, then Maxwell (1873) developed a full mathematical description of the phenomenon, and finally Hertz confirmed the theory experimentally in 1887. Because waves were already a known phenomenon, and even in the 17th century there was some theoretical apparatus for explaining their behavior, this newly discovered behavior of light could be fit into the wave schema. Similarly, Eddington's observation during a solar eclipse in 1919 that light from stars was deflected as it passed near the sun was, at the time, just a single observation of this phenomenon occurring. Nevertheless, because it fit with available theory, it immediately was taken as a reliable effect.

If a hitherto unknown causal chain occurs, but there is no analog for it to be compared to, no known schema that it fits, no type of phenomenon it can be generalized under, then matters are rather different. The new phenomenon would be considered a mystery to be solved. If trying to characterize it in terms of existing theory fails, the next step would be to try to get it to happen again. If the mysterious effect were found to be repeatable, it would be considered a gap waiting to be filled with a mechanism. The Vikings may have discovered circa 700 AC that Iceland spar polarizes light, and can thereby be used as a navigational device (Alcoz 2012). The sunstone is described in several sagas as a stone that can show the position of the sun on overcast days. The Vikings presumably had no established science of waves to help them understand the phenomenon, so polarizing light with sunstones would not have been understood in a wider context of scientific knowledge in that context. My Icelandic friends assure me that the sunstone is depicted in their sagas as an everyday tool rather than a mysterious object. In this sort of case, that a causal chain is repeatable and reliable is grounds for believing that it must be produced by a mechanism, even if the mechanism is completely unknown.

If neither connecting the event to other phenomena, nor repeating it worked, people would probably be inclined to call it a miracle. Others might try to deny it happened by calling it a measurement error or a malfunction of equipment. In such a case, it would seem strange to call it a mechanism that only happened once, although barring miracles, the effect presumably was caused. Repeated occurrences can ground belief that there must be a hidden mechanism at work, but failing to repeat is not grounds for denying that there is a

mechanism. In that case, a connection to general scientific principles trumps repeatability, and one-off instances of causation are treated as mechanisms.

What is essential about regularity for mechanisms is not how often a causal chain occurs. Repeated occurrence is a good reason for believing there to be a mechanism at work, but is not constitutive of a mechanism. The deciding factor in whether we should call something a mechanism or not, is whether there is a background of theoretical knowledge into which the causal chain can be fit. Mechanisms should be repeatable in principle, or should instantiate more general schemas if mechanisms are to be a useful scientific concept.

3.2.3.3 Causation and Mechanism Although I think Bogen is right when he claims that, "It's not that causes and their effects never instance natural regularities. It's just that causality is one thing, and regularity, another" (Bogen 2005), it does not follow that any causal chain, no matter how trivial and scientifically uninteresting should be considered a mechanism. In this section I argue that there are also mechanisms that are not like causal chains, so that mechanisms neither include all causal chains, nor are included in the set of all causal chains. To paraphrase Bogen, causality is one thing, and mechanism, another. In questioning the synonymy of causality and mechanism, I am aware of breaking with established views.[3]

The problem with conflating mechanism and causation is not merely that some instances of causation are not mechanisms, but also that some instances of mechanism are not obviously causal. Take the example of a wheel and axle, a prototypical mechanism. An explanation of how it works involves entities (wheel and axle) and activities (rolling, turning) organized (axle passes through a hole in the wheel) such that they are productive of changes (locomotion) from start (point A) to finish (point B) conditions. It works well, because round things tend to roll when accelerated, and rolling friction is usually relatively low compared to sliding friction. The roundness does not cause rolling in the sense of being a preceding step in a causal chain—that they tend to roll when pushed is perhaps best described as a property of round things. Likewise rolling is not a cause of low friction in the sense of being a preceding

[3]Cartwright argues for an account of laws and causal explanation in her 1983 book that may be compatible with mine.

step in a causal chain—low friction is more like a corollary of rolling. The coefficient of rolling friction is a function of the wheel's sinkage depth and its diameter. The coefficient of friction (for sliding) is an empirical measurement that tends to be higher. Applying a force to the wheel is the only precipitating event appearing in the explanation of how a wheel and axle work. The rest of the explanatory work is done by more general facts, such as that round things roll and that rolling friction tends to be lower than sliding friction. This is imperfect terminology, but I'll call causes of the type that appear as steps in causal chains *cause*^{cc} and causes of the type that have general facts as corollaries *cause*^{gf}.

Take the more complicated example of a gyroscope, another prototypical mechanism. A gyroscope consists of a disc that can spin within an inner ring with two degrees of rotational freedom, attached to an outer ring with one degree of rotational freedom, mounted within a frame. When started spinning with the disc oriented horizontally, a gyroscope maintains this orientation despite motions of the external platform. When started spinning on an angle, the free end of the axis describes a circle in the horizontal plane such that the angle is maintained. Gyroscopes are used as mechanisms to guide navigation in helicopters, on space missions, in mining, and in wireless devices. Simpler versions are used in bicycles, tops and yo-yos. Again there are entities (discs, rings, axes) and activities (spinning) organized in particular ways so as to produce changes (responses to external torque), but the key to explaining how the mechanism works is a general principle: conservation of angular momentum. Spinning doesn't cause^{cc} angular momentum to be conserved, it just is conserved, and its conservation doesn't cause^{cc} the gyroscope to maintain the angle of its axis, this is just what it means for angular momentum to be conserved in this context.

Gravitational forces produce changes too, but it's just the case that masses in a gravitational field attract each other; this is a redescription of the phenomenon, not something caused^{cc} by it. In short, causal chains are not the only way of making things happen. General principles like friction coefficients, angular momentum, and gravitational forces also play a role in determining how mechanisms behave and in producing their results. These simple examples are from physics, where it is uncontroversial that explanations often depend on general principles or laws, but this is not exclusive to physics. Later in Section 3.3.3 I'll give an example of a general principle at work in mechanistic explanation in neuroscience.

My point is certainly not that laws should take priority in accounts of explanation, but rather that in reaction to law-centric views of explanation, mechanists have perhaps swung too far in the opposite direction. Generalizations play a role in mechanistic explanations, and this role merits further attention.

3.2.3.4 Generalization On the topic of the explanatory value of generalization, Bogen states, "Contrary to the Regularists, the goodness of the explanation of one or more instances does not depend on whether the generalizations are true (even approximately) of further instances" (Bogen 2005). I completely agree that how many instances an explanation covers does not in any way contribute to that explanation being correct of any given instance. Neither whether it counts as subsumption under a law, nor its unity with other explanations in any way contribute to an explanation being correct. As I argued earlier, frequency of occurrence is not a deciding factor in whether a causal chain should count as a mechanism. If one wanted to compare explanations in some global sense, then ones that cover more cases might be called more powerful, but this has little to do with what makes for a good explanation of a particular instance.

Nevertheless, there are two ways in which generalization is important to explanations of particulars. One is that if there is a more general fact that can adequately explain a phenomenon (judged on criteria other than its generality), then the general fact might be considered a better explanation than a more detailed account of the same phenomenon, at least in many cases. If both general and detailed explanations are correct, in that they identify the right causes and principles, and adequate, in that there is productive continuity between explanans and explanandum, then it does seem justified to prefer the more general explanation. Adding superfluous, unnecessary details to an already good explanation does not make it better. Getting at what really makes a difference to whether the phenomenon occurs is the goal of explanation. There might be other explanatory contexts where those extra details would add to the explanation, but in those contexts, the more general facts should not be judged a good explanation in the first place. They would have to miss subtleties that make a difference in that context.

In the wheel and axle example from earlier, if you want to explain why less force is

necessary to move a mass with this device than by dragging the weight along the ground, the explanation needs to refer to the wheel's roundness. If you want to explain why a precise amount of force is needed, the explanation needs to also refer to the materials the wheel is made of, its diameter, and the material it's being rolled over. If you want to know all of that plus why a blue streak is left behind, then the explanation also needs to specify that the wheel was freshly painted blue. That the wheel is blue is not at all relevant to the first explanation, and an explanation that included this unnecessary information would not be as good as one that leaves out this irrelevant detail. This is akin to the infuriating habit of journalists of adding tantalizing details about their subjects' personal lives to articles that have nothing whatsoever to do with Ronald Reagan being a former actor, Jack Layton being an avid cyclist, Luka Magnotta being bisexual, or Nadya Tolokonnikova having big lips. These tidbits may add interest, but they detract from the power of the explanation. I'll call this the *norm of explanatory relevance*.

In terms of mechanistic explanations, I argued earlier that whether we should call something a mechanism depends on whether the event can be connected to a background of scientific theory. A phenomenon that can only be explained by referring to all of its unique details would be a random collection of accidents, not a mechanism. In this sense too, generalization is important to explanation. General facts are not the opposite of rarities; general facts are the opposite of specifics. To make both this difference clear, and that I'm referring to something in the world (as opposed to ideas or representations), I will use the term *generic* to refer to these general facts that are opposites of specifics, but still things in the world, unlike abstract representations. A specific mechanism may work only once or very often; a generic mechanism likewise.

Glennan's account of mechanism refers to direct, invariant, change-relating generalizations, a term he borrows from Woodward (2000).

> These generalizations characterize the causal interactions between parts of mechanisms. They function as laws in Mitchell's (1997) pragmatic sense, though, unlike laws on the positivist conception, they depend in essential ways on particulars, are subject to breakdowns, and are not of unrestricted scope (Glennan 2005).

While I noted in Chapter 2 that I disagree with some aspects of Glennan's account, I agree that generalizations are at least one important part of what makes something a mechanism.

3.2.4 Types

Several recent commentaries express worries about whether mechanisms are tokens or types. Overton (2011) objects to the MDC characterization of a mechanism on the grounds that it does not make clear whether a mechanism is a token causal chain, or a type of causal chain. He wants the mechanism of glycolysis, for example, to be one thing, but worries that since the chain of chemical reactions can include just one or several repetitions of the phosphoglucose isomerism "submechanism," treating a causal chain as a mechanism would unnecessarily multiply glycolysis mechanisms. Overton's proposed solution to this and to the ontic mismatch he sees between sketches, schemas and full-fledged mechanisms, is to redefine mechanisms as abstractions over entity and activity types. I think this concern over types is well-founded, but such a major change to the MDC account is not necessary to deal with the problem.

Andersen (2011) takes the fact that mechanisms may be either types or tokens as a point in their favor. Like others, she identifies token mechanisms as "actual causal chains of appropriately causally connected entities," and type mechanisms as "descriptions or schemas of such mechanisms" (Andersen 2011). On her view, a token mechanism can explain, for example, why a particular neuron fires on a given occasion, and a type mechanism can explain what happens on all the occasions when neurons fire. Anderson claims that, "The way in which mechanisms comprise both types and tokens thus serves to connect the explanation of a single instance of firing with a generalization about what happens when neurons fire" Andersen (2011). She does not say how we can draw this connection though. Allowing mechanisms to be either causal chains in the world or representations of these strikes me as a *dis*connection that bears explaining. We do call both types and tokens mechanisms, but more needs to be said about the connection between them, and about what the ontic and explanatory status of each is.

I think Andersen is perfectly correct to want to connect token mechanisms with type mechanisms; the explanation for why a particular ion channel opens should be connected somehow to general facts about whether ion channels open under similar circumstances. Drawing this connection is key to understanding the role of generalization in mechanistic

explanation, as I see it. If token mechanisms explain occurrences by causing them, then representations of mechanisms cannot explain occurrences in the same way, since they do not cause the occurrences they represent (and, *pace* Hempel, logically entailing a description of the explanandum does not explain it either). If for Andersen, type mechanisms are representations, it is not clear how these can explain mechanistically. Showing how general facts can contribute to mechanistic explanations is a problem that I'll try to solve in this chapter.

Kuorikoski (2009) also worries about the ambiguity between causal chains and types of interactions. He draws a sharp line between what he sees as two concepts of mechanism:

> First there is the concept of mechanism as a componential causal system, which is accompanied with the heuristics of decomposition and localization. Second, there is the concept of mechanism as an abstract form of interaction, accompanied by the strategy of abstraction and simple models (Kuorikoski 2009).

Kuorikoski's argument is aimed at Bechtel's account of mechanism rather than the MDC account, but as we've seen from other commenters, the same tension between mechanisms as causal chains and mechanisms as abstract types is also present in MDC.

Elaborating abstract forms of interaction in the context of economics, Kuorikoski says that the "relevant or legitimate causal constituents are individuals, households or firms" (Kuorikoski 2009, 150) and

> the *relevant* properties (preferences, strategies, aggregate demand and supply) of the agents are in reality essentially relational... Decomposing the individual behavioral dispositions of constituent parts (such as individual agents) into some set of lower level component operations (such as psychological mechanisms) is also unhelpful for understanding the original macro phenomenon, since the intrinsic causal properties of the parts are not of primary interest. (Kuorikoski 2009, 151)

He makes similar arguments for natural selection. That it is relational properties between entities rather than properties of their parts that do the explaining seems quite true in these cases. The problems I pointed out with seeing information-processing models of cognition as mechanism sketches in Chapter 2 were along the same lines.

The examples Kuorikoski gives are ones where MDC would likely counsel one to look upwards for the operative mechanism, rather than looking down. That they have this strategy at their disposal, at least in principle, is perhaps a point in favor of their account of

mechanism. Bechtel leans much more heavily on the strategy of localizing causally relevant properties in component parts. Kuorikoski goes on to say that, "Since the form of interaction is not in itself dependent on the way the causally relevant properties of the component parts are constituted..., the same sample models and hence 'the same' mechanism schemata can be utilized in many different kinds of contexts or domains" (Kuorikoski 2009, 152). Here we can see that the explanatory strategy of schema instantiation, discussed by Darden, is at odds with componential mechanistic explanation. Where the form of interaction does not depend on the properties of the components, those components are not causally relevant to the mechanism. On Bechtel's account, this poses a major problem, since on his definition, "A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization" (Bechtel and Abrahamsen 2005). If the function is performed in virtue of higher order properties, rather than component parts, either these examples fail to be mechanisms, or Bechtel's account can't deal with them. The MDC account does not have this problem, since it refers only to entities and activities, and these entities and activities needn't be parts. Nevertheless, MDC have not made it at all clear how explanations that look upward work.

The problem Kuorikoski raises is not just that we seem to want to use the term 'mechanism' for both token causal chains and for abstract types of interactions, but furthermore, that in many branches of science, the factors that seem to do the work in explanations are not the details about the parts, but rather relational properties between these parts. In Kuorikoski's examples, filling in the details about components does not make for a better explanation. Arguably cognitive psychology is full of examples like this where what explains is an abstract or general form of interaction or processing among components, and not the gory details of their components. What is at stake is whether types can do explanatory work on a mechanist account. Recall that according to MDC, "Scientists... are typically interested in types of mechanisms, not all the details needed to describe a specific instance of a mechanism" (Machamer et al. 2000). If we're to include this common form of explanation within mechanistic explanation, we need to work out how abstract forms of interaction, types of causal processes, or abstractions over entities and activities can be made to fit the picture of what a mechanism is, and how these abstractions or types can produce changes.

This is the task of the next section.

3.3 COGNITIVE MECHANISMS

So far in this chapter, I have given an overview of the MDC account of mechanism, and some of its critical reception. Along the way I have highlighted aspects of the account that might be recruited for showing how cognitive systems could be explained mechanistically. At this point, the groundwork is done, and what remains is to spell out my proposal for how cognitive systems can be understood as mechanistic within an extended MDC account, thus affording the possibility of integrated cognitive-neural mechanisms.

First I will return to the problem raised in the previous chapter, where it was not clear how information-processing accounts of cognition could be made compatible with neural mechanisms. Then I will summarize three related problems that arose earlier in this chapter for the MDC account: explanatory relevance, types and tokens, and generalization. In Section 3.3.3 I will introduce the notion of a *generic mechanism*, illustrate it with several examples, and propose that it can solve all four of these problems. Finally, I discuss the implications for integration.

3.3.1 Looking Upward, Explaining Downward

In the previous chapter I argued that information-processing models are not just elliptical sketches of neural models that can be filled in with details to make mechanisms of them. Data flow models explain in virtue of how data is manipulated, not what does the manipulating. Process flow models explain in virtue of the order in which simple steps are performed, i.e., the algorithm, not in virtue of how those steps are implemented, nor which parts of the machine are involved. Intuitively, the force in a cognitive explanation does not come from the parts involved in making it operate, but rather from its functional properties. The challenge remains to see how looking upward to these system properties can still be described in mechanistic terms.

MDC do not directly address the question of whether higher-level mechanisms can explain lower-level phenomena. On the one hand, they add the following restriction to their comments about hierarchies: "lower level entities, properties, and activities are components in mechanisms that produce higher level phenomena" (Machamer et al. 2000, 13), suggesting that mechanistic explanation has to go from lower to higher levels. On the other hand, they refer to abstractions over the details of underlying mechanisms as "activities in higher-level mechanisms" (Machamer et al. 2000, 16), and claim that mechanistic explanation "involves showing or demonstrating that the phenomenon to be explained is a product of one or more of these abstract and recurring types of activity or the result of higher-level productive activities" (Machamer et al. 2000, 22), which together suggest that higher-level activities can explain particular phenomena. Their final word on the topic is that, "It is the integration of different levels into productive relations that renders the phenomenon intelligible and thereby explains it" (Machamer et al. 2000, 23). The complete picture then seems to be that productive relations involve entities and activities at both higher and lower levels, and that the explanation is an 'integration' of all of these. That mechanisms are often multi-level is not what is at stake here.

Darden clearly endorses upward-looking mechanistic explanations. She says that "finding the mechanism that produces a phenomenon may require not further decomposition of a system, but instead going 'up' in size level" (Darden 2005). One example she gives is of segregation of genes being explained, not in terms of their parts, but in terms of the larger whole, the chromosomes, on which they are found. A simpler example along the same lines might be that the daily change in altitude of the *E. coli* in my stomach can be explained by my whole body (an entity) traveling (an activity) to and from work on a hill.

Craver also claims that explanations sometimes look upwards. The examples he gives are of higher levels constraining explanations by specifying something about the context, rather than examples where higher levels actually explain phenomena at lower levels. For Craver, mechanistic explanation is a species of constituent explanation, where the parts explain the whole. Upward-looking explanations seem to be incompatible with constitutive explanation, however.

According to the MDC account, the entities and activities that explain a phenomenon

need not be parts of the whole that exhibits that phenomenon, as in Darden's example of chromosomes explaining phenomena at the lower level of the genes. This is the sort of explanation we seem to need for cognitive mechanisms. What remains is to show how exactly upward-looking explanations work. My solution to this problem also addresses several issues brought up in the commentaries on MDC.

3.3.2 Problems and Constraints

As we've seen in earlier sections, there are several open issues about the MDC definition of mechanism which need to be cleared up. Each of these problems points towards features that a solution to the problem of upward-looking mechanistic explanations should have. The first problem is that Craver's conviction that more details in an explanation are always better is at odds with the norm that explanations should only refer to explanatorily relevant factors. The second problem is that we seem to want both types and tokens to count as mechanisms, and it is unclear how we can have both in the same account. The third is that generalizations seem to play a role in even prototypical mechanistic explanations, but this has been ignored in the MDC account.[4] Below I elaborate on these problems, outlining how each of them constrains a solution that might serve all of these purposes.

3.3.2.1 Explanatory Relevance Craver's case study of the Hodgkin-Huxley model is a great example illustrating how adding more details can be essential for turning a phenomenological model into a mechanistic one, and thereby explaining rather than just summarizing empirical results. This case study was intended to show that mathematical models like the Hodgkin-Huxley equations do not provide explanations without those additional details. That what was needed for explanation in this case was more mechanistic detail does not necessarily generalize to all of neuroscience though. The examples I give in Chapter 4 will illustrate that in some cases more abstract models provide better explanations than highly detailed ones.

[4]Both Schaffner (1993) and Strevens (2004) have offered accounts of explanation that intertwined or unified causal-mechanistic explanation with generalization, but neither is a mechanistic account in the manner of MDC.

Intuitively the problem with taking it as a general rule that adding details always improves explanations, is that explanations are then no longer responsive to the explanatory context. That context includes the specific question the explanation is supposed to answer, any background information that is being assumed, and the audience it is intended to convince. There are usually, if not always, multiple explanations available for any given phenomenon, all of which are correct, in that all of them account for the phenomenon's being so. This is clearly true of explanatory texts, but I maintain it is also true of ontic explanations. These multiple alternative explanations might differ in terms of how much detail they go into, and whether they depend on particulars or on general principles, to name just two axes of difference; there may be more. I'm calling this constraint that explanations be pitched at the appropriate level given the explanatory context the norm of explanatory relevance.

An anecdote Salmon (1981, 1990, 1992) relates about a helium balloon moving forward as a plane starts moving demonstrates the second sort of difference. In Salmon (1992), he uses the example to make the point that there are often several explanations of the same phenomenon, and that these are not necessarily in competition in terms of which is correct. We choose which is more appropriate based on the context in which we're explaining. A child might better understand a story about the back of the cabin pushing the air forward, which pushes the balloon forward, because it is lighter than the air, while a university physics class might better understand an application of the equivalence principle. We might further distinguish between a generic explanation about air pushing against the balloon, which would be expected, perhaps in most airplane designs, to result in the balloon moving forward, and a specific explanation of why in this particular plane, with this particular arrangement of air molecules, and these particular acceleration forces acting, this particular ballon moved forward. Which of those is more appropriate depends partly on the question being asked. We might be asking why helium balloons on planes move forwards when the plane does, or why it happened in this particular case.

The equivalence principle provides a covering law explanation. Even if we throw out the DN apparatus because of its many faults, we can still explain some things by citing truths along the lines of 'things of type T behave as B' and 't is a thing of type T'. In some cases we might still want to ask why things of type T behave as B, but once we have an explanation

of that connection, a good explanation of the event will usually just cite the rule. We use this form of explanation even when the rule isn't a law. To apply such a rule when there are exceptions involves a further claim that this is one of the cases where the rule applies, or that there are no interfering factors in this case.

The explanation where the back of the plane's cabin pushed the air forward is described by Salmon as a causal-mechanical explanation. The specific causal story where the back of this plane's cabin pushed this air forward I will call a causal narrative, by which I mean a causal story specific to a particular instance. Bogen's insistence that mechanisms might operate only once suggests that he sees mechanistic explanation as providing causal narratives. In cases of one-off causal chains, it may be necessary to go into a lot of detail about the specific circumstances, because these cases are exceptional. This does not mean that in all mechanistic explanations one needs a detailed causal narrative specific to a particular instance. There is always a finer grain of detail one could go into, even in the case of a mechanism that operates only once, but more detail than is required to account for the explanandum is superfluous; it makes an explanation worse, not better.

It is easier to see how this norm can be upheld in explanatory texts than in ontic explanations. An explanatory text can include or leave out details as required by the context, but explanations as things in the world can't be partial in the same way, or so people seem to think. One possible response to this difficulty is to set aside explanatory relevance as being of less importance than accounting for phenomena in their entirety, including any rare or exceptional cases. In order to account for all cases, Craver claims explanations must go into the gory details. Since he takes explanations to be things in the world, perhaps he thinks that as a thing in the world, the explanation must remain the same thing in the world, regardless of the explanatory context. This is intuitively tempting. Since the various explanatory contexts might require all the weird little details, the explanation must have to include these if there is to be just one explanation. An unstated assumption, however, is that there is just one explanation for any given phenomenon, so that same explanation needs to work in any explanatory context. I think we can have multiple explanations that are all in the world, without having to have a strange metaphysics where there are multiple objects or systems in the world that supply these explanations. This is one of the constraints on

explanation that needs to be heeded if we're to have both ontic explanation and explanatory relevance.

The downside of not upholding the norm of explanatory relevance is that the explanation for something like why a wheel rolls can't just be that the wheel is round. The explanation would have to be something along the lines of the wheel being made of 3 billion iron atoms and 5 million copper atoms, with various forces acting between the atoms such that they form a solid, in a particular crystal structure such that only certain wavelengths of light are reflected making it appear blue, organized in a particular spatial arrangement such that it is round, and so on, and so on. I think it's fair to say that an explanation like this fails at its job. One might assign to the explanatory text the task of picking out which parts of the explanation are relevant to the context, but this makes the explanation^{o} nothing less than a slice of the world in all its complexity. That notion of explanation does not seem to do any useful work. It also makes one wonder why going down just to the molecular layer should be good enough for neuroscientists. If all the details that might be relevant in any explanatory context are needed, there is no bottoming-out. Another constraint on explanation then, is that it should be able to do the work of picking out which parts of the world are relevant to the phenomenon being explained.

An additional problem is that it is unclear how explanations can count as looking upwards to higher levels if they have to go into so much lower-level detail. What I'm trying to carve out space for are explanations where more abstract or higher-order features of phenomena can explain either particular phenomena, or general patterns. If higher-order features cannot do this explanatory work, then we haven't accounted for the types of information-processing explanations from the previous chapter.

3.3.2.2 Types and Tokens Earlier in Section 3.2.4 I reviewed several commentaries on MDC that pointed out an ambiguity between mechanisms as types and mechanisms as tokens. Kuorikoski (2009) pointed out that many social science explanations depend on the relational properties of the systems of interest, rather than the properties of their parts, and suggested that these abstract forms of interaction are a different notion of mechanism than what neo-mechanists describe. Overton (2011) was puzzled by the ontic differences between

mechanisms, sketches and schemas, and suggested redefining mechanisms as abstractions over entity and activity types. Andersen (2011) saw the dual nature of mechanisms as a positive feature that shows how particular instances of causation are connected to generalizations about those types of instances, although she didn't spell out what the connection is exactly.

I have argued, in agreement with Bogen (2005), that particular causal chains, including some one-off ones, should be included as mechanisms, but have also argued, in agreement with Darden (2002, 2008), that one of the most important things about mechanisms is their generality, or the possibility of instantiating them in new contexts. I think Andersen is right that there is a connection between these two notions of mechanism. Furthermore, it seems to me that connecting explanations of particular facts to explanations of general ones is one of the most important conceptual services mechanisms can provide. Being able to relate type mechanisms to token cases also helps solve the problem of explanatory relevance. When a particular causal chain instantiates a type, then a good explanation of that token might be in terms of the behavior typical of the type.

The key to drawing this connection is to figure out a way to allow type mechanisms to be things in the world just like token mechanisms, rather than being abstract representations of them. If this constraint on a solution is heeded, we can have token mechanisms and type mechanisms that explain in the same way: by producing the phenomena they explain. The connection between explanations of particulars and generalizations over them could then be that they bear to one another the relationship of a specific to a generic. Put another way, the explanation of both could be the same thing in the world, but in one case the explanation is that thing in all its detail, and in the other case, the explanation is that thing insofar as it belongs to a type.

3.3.2.3 A Role for Generalization I take it that calling something a mechanism rather than a machine implies two things. First, the term mechanism need not be restricted to artifacts; machine carries the connotation of a made object, and some object to referring to living things as machines. Second, the term machine refers to the thing in the world, qua physical object. It makes sense to say things like that the machine is made of metal, the machine weighs 40 kg, or the machine just went out for repairs. As I understand it, a

mechanism is not the thing in the world qua particular object, but rather the thing in the world insofar as it operates in a particular way to perform a particular function or task.

As noted earlier in the examples of wheels and axles and gyroscopes, even prototypical mechanisms depend on more than just causes^{cc} for their explanations, and this is also true for mechanisms in the cognitive and neurosciences. Mechanistic explanations depend also on general principles, mathematical facts, laws of nature, as well as rules of less than general applicability. It is understandable that accounts of mechanistic explanation, and in particular the MDC account, have de-emphasized the role of these sorts of generalizations, since mechanisms are a response to the DN account's failure to work in the biological sciences. But if we're to allow for type mechanisms, and if we're to heed the norm of explanatory relevance, then some role for generalization needs to be brought back into the picture.

There are ways in which the entities and activities constituting things in the world in virtue of their organization (i.e., mechanisms) make things so, that are not causal^{cc}. Roundness doesn't cause^{cc} rolling, having mass doesn't cause^{cc} gravitational forces, the square of the hypotenuse doesn't cause^{cc} the sum of the squares of the other two sides to equal it, but nevertheless, this sort of making things so does contribute to explanations of phenomena, and does so in a way that fits the definition of mechanism. Even if these ways of making so aren't causes^{cc}, they are certainly important to how mechanisms work. In Woodward's (2005) terms, they are difference-makers.

In the section on schema instantiation, we saw that Darden argues for the importance of finding "recurrent motifs" (Darden 2002), and drawing connections between different theories and phenomena, but not as with Kitcher for the sake of unification as a goal in itself. Instead unifying phenomena through mechanism schemas is presented as a way of representing general knowledge, and as a tool for generating hypotheses and discovering mechanisms. Skipper (1999) has similar ideas about expanding Kitcher's notion of argument schemata to mechanism schemata. He sees "explanatory unification as proceeding via *mechanism schemata*, in which unificatory explanations are schematized causal mechanisms" (Skipper 1999). His paper explores

> whether unification can be conceived of as the reduction of types of mechanism scientists must accept as targets of their theories and explanations, and whether it proceeds through the delineation of pervasive causal mechanisms via mechanism schemata (Skipper 1999).

In Skipper's system, mechanism schemata explain general phenomena like selection. I think this is almost correct.

As Weber and Bouwel (2009) point out, Skipper-style unification "consists in showing that the mechanisms which lead to different events contain similar causal factors." They go on to conclude that "This does not require subsumption under a law, so this kind of unification does not proceed by constructing arguments and showing that the events could be expected" (Weber and Bouwel 2009). But Skipper's unification does consist in constructing arguments, even if they do not involve laws. His general mechanism schema for selection sets out an argument structure, and instructions for instantiating the variables in that structure. I am not looking for mechanistic explanations of general phenomena that just *show* that mechanisms leading to different events share common causal factors. What I'm looking for are mechanistic explanations that *are* the common factors leading to different events. If it is things in the world that explain, then mechanism schemas can't do the job. What is needed are in-the-world analogues to schemas.

In a very recent paper, Levy (2013) argues for the value of abstraction in explanation, focusing on the Hodgkin-Huxley model. He complains that accounts of mechanistic explanation miss cases where "a model is *deliberately* 'sketchy', i.e. where gaps aren't the product of ignorance or theoretical limitations, but of an intentional strategy" (Levy 2013). I completely agree, and yet I am aiming for a still more general account than Levy's. The cases he cites as benefiting from more abstract explanations are ones where lower-level entities are treated as collections or aggregates, such that the individual details of each entity are less important to the explanation than the properties of the collective. This is just one way in which abstraction from details can be useful. In that sort of case the type to which the mechanism belongs is an aggregate of some variety. There are many other types of types beyond just aggregates.

Making space for abstraction and generalizations in mechanistic explanation is something that needs to be done. The solution I propose for making space for generalizations has several additional benefits: it upholds the norm of explanatory relevance while remaining within the framework of ontic explanation, it provides the missing connection between type and token mechanisms, and it makes space for cognitive explanations that look upwards for difference-

makers.

3.3.3 Generic Mechanisms

In the previous section I drew out several constraints on a solution to the problem of cognitive mechanisms. The first was that explanations should be things in the world that are responsive to explanatory contexts, without any weird metaphysical consequences like a multiplication of mechanisms in the world, one for each context. The second was that it should be the explanation^{o} itself, i.e., the mechanism, that does the work of picking out which parts of the world are relevant. The third constraint was that type and token mechanisms should have the same ontic status, so that they can explain in the same way. This requires a way of abstracting that does not produce a representation. The fourth constraint was that there should be a role for generalizations in mechanistic explanations. This requires that an in-the-world analogue to mechanism schemas can serve as the common factors that produce the phenomena in the world that we want to generalize over. Additionally, we have the main point, which is to find ways in which cognitive explanations of the kinds discussed in Chapter 2 can be understood as mechanistic. As promised, these problems have a common solution: *generic mechanisms*.

The idea of a generic mechanism is quite simple. Things in the world are what they are in many respects; they belong to types of varied scope, level, and sort. A small object like a lump of gold is a specific thing with a unique causal history, but it is also generic thing—a piece of gold—in addition to a still more generic thing—a homogeneous lump of solid matter. It is also a different more generic thing—an object of value. It might be a thing that belongs to Ken, as well as a thing coveted by Peter, but it remains just one thing, despite belonging to many types, both specific and generic. The lump of gold qua valuable object explains why it's kept in a safety deposit box. The lump of gold qua its molecular structure explains its being a good conductor. Aristotle calls this a difference in being that is not a difference in number. Entities of any size, and at any position in a hierarchy of mechanisms can be either specific or generic (or more correctly, both). A generic mechanism is no less a thing in the world than a specific mechanism is, it is just that same thing in

the world qua a more abstract type. This 'qua abstract type' predicate is how things in the world can be abstracted without thereby becoming representations. I call them generic mechanisms rather than abstract ones, in an attempt to make clear that unlike mechanism schemas, they are not representations of mechanisms, but rather full-fledged mechanisms (and also because some people consider 'abstract object' to be a contradiction in terms).

To take an example that looks more like a typical mechanism, take my sewing machine. As an object that weighs more than 10kg, it is still the same thing in the world. As a household appliance, it is the same thing in the world. As a Singer 319 Automatic Swing-Needle sewing machine from 1956, it is that same thing in the world. As a machine of that model that I inherited from my grandmother and have in storage in Pittsburgh, it is that same thing in the world. It is also a collection of billions of atoms, a configuration of metal, wood, and rubber parts coated in dust and grease, as well as many other things.

The object qua sewing machine is a generic mechanism, and the object qua Singer 319 from 1956 is a more specific mechanism. Which of these figures in a causal story or an explanation depends on what is to be explained. It being a particular model and year explains why I must switch to a different disc in order to do a particular embroidery stitch. It being a sewing machine explains why it sews. It being an object weighing more than 10kg explains why it crushes my toe when I drop it, and why I can't take it in my carry-on luggage. It being made of a particular configuration of parts explains why it makes a certain noise when it runs, and why it requires a certain voltage power source. For those explanations, it does not make any difference that I inherited it from my grandmother, even though that's also true of the object.

That the same object can occupy various degrees from specific to generic mechanisms allows it to uphold the norm of explanatory relevance. That the mechanisms always remain the same object despite contextual shifts forestalls weird metaphysics. Since a generic mechanism is a thing in the world, it is itself the explanation^{o}. The variety of specific and generic mechanisms an object can instantiate allows it to pick out the relevant difference-makers for a given explanatory context. Of course, which of the available types that the mechanism belongs to in fact are the difference-makers, depends on the explanatory context. Just as with explanatory texts, only the parts of the picture that make a difference to the phenomenon

being produced need to be included in a good explanation.

We're now in a position to spell out the connection between token and type mechanisms. The connection is that the token mechanism and the type mechanism are both instantiated by the same thing. Although a type is an abstract thing like a category, a type mechanism needn't be. Instead a type mechanism can be a generic mechanism, that is, a mechanism insofar as it belongs to that type.

Take the example of a neuron firing an impulse down its axon: the way the impulse travels is explained by that axon being an instance of the generic mechanism of myelinated axon. This can be further abstracted to the mechanism of insulated wire. That this mechanism is an insulated wire (in addition to everything else it is) is why neurophysiology textbooks can talk about axons in terms of general principles of electronic transmission. The token mechanism of a firing neuron's axon is connected to general principles about insulated wires, because that same thing in the world is also a (generic) insulated wire. The token mechanism is related to the type mechanism by virtue of the same thing in the world being instances of both. Type and token mechanisms have the same ontic status and explain in the same way: by causing phenomena. Multiple instances of the same generic mechanism—say an axon qua insulated wire and a stereo cable qua insulated wire—share properties and causal powers in virtue of instantiating the same generic mechanism.

Generic mechanisms point towards a role for abstraction and generalization in mechanistic explanation. That a neuron's axon is an insulated wire allows generalizations about insulated wires to play a role in mechanistic explanations involving the neuron. That the $Na+$ channel's inactivation gate is shaped like a ball and chain allows the generalization that balls and chains can plug holes to play a role in the mechanism of $Na+$ channel gating. Where general facts about the entities and activities in generic mechanisms make a difference to the production of the phenomenon to be explained, those generalizations form part of mechanistic explanations. Gravity is a factor in the working of some mechanisms, and in those cases, the fact that the mechanism is a massive object allows the ways that massive objects act to be part of the explanation.

One worry this proposal might raise is that the multiplicity of mechanisms that a thing can instantiate leads to causal overdetermination. Allowing that difference-makers operate

at multiple levels, not just in different sciences, but within a single mechanism, might seem to imply that the same event could be caused in different ways simultaneously. There is no real problem with having multiple difference-makers or causally-relevant factors in an explanation, however. As long as there is some fact of the matter as to what the real difference-makers are, and what their causal influence is, it should be no problem to combine them into a complete explanation (at least in principle). It may even be that in different explanatory contexts, the same phenomenon might have different causes. If anything, the recognition that there are difference-makers at different levels of specificity points toward a resolution of the dilemma that motivates the problem of causal overdetermination in the first place. It allows us to have cognitive-level causes even though they supervene on neural realizers.

A second worry might be that by Leibniz's law, if a thing qua type A has different properties than a thing qua type B, then they must be two different things. For instance my sewing machine qua weighing more than 10kg crushes my toes, but my sewing machine qua inherited from my grandmother does not. In these cases, it is not so much that the qua-objects have different properties. It is more that which properties apply to them differs. It is open or unknown whether my sewing machine qua inherited from my grandmother is heavy or light. The thing that instantiates both the generic and the specific has one consistent set of properties though. I do not claim that that 'thing qua type A' is the same as that 'thing qua type B' when $A \neq B$.

A third worry might be that since qua-objects are things in respects, they might be interpreted as representational, mental, or epistemic. One detail that needs to be worked out is how to justify the claim that qua-objects remain things in the world. At least the grammatical evidence suggests that the first operand of the insofar as and qua operators has the same ontic status as the result. More certainly needs to be said about the detailed metaphysics of what a generic mechanism is, and how the insofar as or qua operator works. We have now at least an intuitive idea of how the constraints developed in the previous section might be satisfied by generic mechanisms.

3.3.3.1 An Example: Lateral Inhibition One of the the most important functions of schemas (and likewise generic mechanisms) is that they reappear in several guises, but always do the same sort of job, because there's a general principle operating such that this type of mechanism behaves a certain way. We have to invent or discover these generic mechanisms, but once we have discovered them, we can reuse them in our designs and understand how similar behaviors arise when we see analogous mechanisms in other contexts.

An example of a generic mechanism from cognitive science that can and has been applied in this way to other contexts is lateral inhibition. Retinal ganglion cells (among other cell types) 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, more active cells inhibit their neighbors more, and for a cell to fire a lot, most of its neighbors can't also be stimulated. The result of this architecture of lateral inhibition is that contrasts are enhanced and detected. Laterally inhibited neurons respond best when their neighbors are doing the opposite of what they are, so where contrasts among neighbors arise, that is where activation is highest.

This phenomenon is useful in retinal ganglion cells for detecting object contours or edges, which are characterized by abrupt changes in illumination. Cells near an illumination change in the retinal image will only get inhibited by neighbors on one side of the edge, i.e., about half of them. 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 from their neighbors, so have the highest relative activity. This tends to sharpen responses even further, because this activation and inhibition is ongoing. Once some neurons are activated more than others because of being located near edges, they inhibit their neighbors more, which in turn decreases the inhibition they get back from them. Over time, areas of slight contrast gain more and more of an advantage over their neighbors. As a result, even fairly faint edges will appear sharpened.

Kuffler (1953) discovered that retinal ganglion cells have center-surround receptive fields (so do cells in the lateral geniculate nucleus, it turns out). They come in two types: 'on center' cells respond best to light in the center and dark in the periphery, and 'off center'

cells respond best to dark in the center and light in the periphery. These receptive fields come about because of lateral inhibition. Hartline (1940b,a) first described lateral inhibition in terms of retinal ganglion cells, and eventually won the Nobel Prize in 1967 for this work. The *Encyclopedia of Perception* entry on lateral inhibition describes how widespread this mechanism is in the brain:

> Lateral inhibitory circuits are currently known to be ubiquitous to all sensory areas of the brain, and they play an important role in many sensory, cognitive, motor, affective, and limbic processes. The most common mechanism by which neurons suppress their neighbors is through the inhibitory neurotransmitter gamma-aminobutyric acid (GABA) (Macknik and Martinez-Conde 2009).

Somatosensory neurons, for example, which have very large receptive fields, are still able to discriminate fine details because of lateral inhibition, which increases their spatial resolution. Lateral inhibition is a still more general phenomenon though. It works across many contexts with many sorts of units, forms of inhibition, and types of activity.

The generic lateral inhibition mechanism has been used to explain several other scientific phenomena where contrasts are detected or enhanced, and the principle has been used to engineer devices like hearing aids, and the motion sensors in computer mice. The natural phenomena for which it has been used as an explanation are widely varied. 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 differences in how far along cells are in development, make for small differences in how much protein is emitted. These small initial differences in protein signals make for stark contrasts in developmental outcomes, since these signals inhibit neighboring cells from developing in the same way, which increases the contrast in how cells are allowed to develop.

There are also economic and sociological analogues. For instance if communities decide to focus their limited resources of money, coaching hours, practice space, etc., on their most promising athletes, and if the amount of investment made in an athlete is in proportion to their skills, this will result in a widening of the gap between the skills of the most promis-

ing athletes and the rest. In this sort of scenario, the most promising athletes get more resources to the detriment of less promising athletes, which makes the best athletes improve more quickly than the others, further increasing their skills and hence the investments they get. Any athletes who live in the same community as a star will get less investment than they would otherwise, so will develop their skills less, while the star is developed as much as possible. This sort of organization increases the contrast in sports ability within the community. In a community where stimulation of some is not coupled with inhibition of others, the resources would be distributed more evenly, and the skill gap between stars and non-stars would not be further widened. These are the sorts of concerns amateur sports associations and school boards discuss in strategy and budget meetings.

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. When the better players improve more quickly, the contrast in abilities widens.

The explanation for why contrasts get enhanced in any of these cases is that the structure of the relationships between units or agents instantiate lateral inhibition, and lateral inhibition gives rise to contrast enhancement. Here maybe it's important to note that this isn't so because of a rule we might formulate or a model we might construct that describes the connection between lateral inhibition and contrast enhancement (although there do exist equations describing the phenomenon). It would also make for a very poor explanation if you ignored the mechanism of lateral inhibition in any of these cases, and explained the effect by making a huge sum of the local effects between units or agents. What makes this phenomenon occur is that a system has a particular organization of components and connections such that it counts as an example of lateral inhibition. The system's being constituted in this way means that contrast will be enhanced.

If you want to explain not only the contrast enhancement, but also the exact values the system settles into, or the timing of the change in contrast, then you need both the fact that it's an example of lateral inhibition, plus some of the gory details. But even in that sort of case, you don't throw out the generic mechanism and just go straight for

the details. A good explanation would be a balance of both. In many explanations, it's both generic aspects of the mechanism and specific aspects that make a difference in how and whether the phenomenon occurs. Those explanations are hybrids between generic and specific mechanisms. I'll give examples of these in Chapter 4.

There is explanatory power in this ability of generic mechanisms to apply the same generic principles to a variety of cases, and to do so independently of the details of the instantiations. Whenever units inhibit their neighbors from doing what they are doing, the result will be higher contrast in the result. This can mean sharpening responses, grooming the strongest members and hindering the rest, or forcing specialization. Going into details about the specific processes at work in any given case of lateral inhibition could also explain what happens in that case, but that sort of explanation misses out on what really makes a difference to whether the effect occurs. The mechanism of lateral inhibition explains the contrast enhancement. The specific causes—neural inhibition, chemical messaging, focused coaching, pub conventions—explain the details of how exactly it occurs in an individual case, such as the timing of the contrast enhancement, or the exact values of parameters. An explanation that uses only the specific facts despite there being a more generic account would not be a good explanation, because it would miss the real difference-makers.

There might also be examples where the details get so complicated that the best explanation no longer involves the generic mechanism, and only the details matter. Likewise, there should always be many types that a given system exemplifies, and not all of them might make a noticeable difference at once. For example the sewing machine being an object weighing more than 10kg might not play into the explanation of why it sews. Strevens (2004), among others, has discussed the criteria that might be used in deciding what is explanatorily relevant. I'll leave that set of problems aside.

3.3.4 Prospects for Integration

Generic mechanisms respect the four constraints discussed above, but it remains to be seen how useful they are in fostering integration between cognitive and neural systems. I do not claim that this contribution alone solves the problem of integration. This is meant as one

useful resource for making space for cognitive mechanisms, or higher-order mechanisms more generally. It almost definitely does not cover all cases. It can do things that Bechtel's localization heuristic can't, like upward-looking explanations. It is more realistic than Piccinini & Craver's elliptical mechanism sketches, which trivialize psychological models. But there still might be reason to believe that some cognitive explanations aren't compatible with even generic neural mechanisms. Weiskopf (2011) raises several objections to Piccinini & Craver's picture of integration that are perhaps not affected by generic mechanisms. We might have to give up on complete integration. We might have to embrace pluralism of the variety described by Mitchell (2003). What generic mechanisms do offer is one way in which we might make sense of psychology's explanations as being compatible with lower-level neural mechanisms.

Integration might occur in a number of ways. As Darden and Maull (1977) argue, interfield theories can posit identities, part-whole relations, structure-function relations, or cause and effect relations. Part-whole relations are only one option, and although that relationship can only go in one direction (higher levels cannot be parts of lower levels), there is no reason why wholes cannot sometimes explain properties of their parts. Structure-function relations do not have a strict direction in terms of hierarchies of mechanisms, and likewise explanations can go from structure to function or the reverse. Causes can explain effects and not the other way around, but at least according to some, downward causation is possible, so higher levels could cause phenomena at lower levels (the reverse is clearly possible). So in interfield theories, the explanations may operate either upward or downward.

A remaining issue is whether all of these types of interfield theories would qualify as mechanistic. Explanations going from function to structure would be teleological explanations, but the remaining possibilities all could fit the definition of mechanistic explanation. Generic explanations could take the form of whole to part explanations, (generic) structure to function explanations, and possibly downward causation, if such a thing exists. Wholes can explain parts mechanistically, as in the examples of gene segregation during gamete formation and *E. coli* daily altitude changes. Higher-level structures might explain the functioning of lower-level ones, as in the example of lateral inhibition. An example Machamer likes to give of top-down causation is how running causes blood pressure to rise. So there

are several ways in which interfield theories might operate downwards, linking higher-level explanantia with lower-level explananda.

Another thing to note in trying to understand the relationship between psychology and neuroscience is that there is no guarantee that all the mechanisms will appear as part of one unified hierarchy. There might be multiple hierarchies, if the entities at different levels just don't fit together in clean pieces. Even if in some cases cognitive models fail to be neatly accommodated into the hierarchy (or hierarchies) of neural mechanisms, there could still be a separate hierarchy (or hierarchies) of cognitive mechanisms. The question of whether cognitive psychology can be made mechanistic might come apart from the problem of whether they can be integrated. It also seems quite possible that even if there had to be multiple hierarchies, there might still be some other way of achieving integration. Integration need not mean exclusively combination into a unified hierarchy of mechanisms. Meaningful relationships across hierarchies that meet the desiderata for integration listed in Chapter 1 might take the form of cross-hierarchy constraints of some kind.

In the next two chapters, we will see generic mechanisms put into action in computational cognitive neuroscience. In part this is a digression into another set of more methodological problems in cognitive neuroscience. In part thinking about how computational models work was how I came to think about generic mechanisms. But this application of the tools developed here also serves to illustrate how generic mechanisms work in practice. In the conclusion I will return to the question of how generic mechanisms can help with the problem of integrating cognitive psychology with neuroscience.

4.0 COMPUTATIONAL COGNITIVE NEUROSCIENCE

4.1 INTRODUCTION

In this chapter I explore the epistemic role played by computational models in cognitive neuroscience: how they contribute to theories, and how they relate to experiment. One motivation for looking in detail at this aspect of methodology in cognitive neuroscience is that computational modeling is an essential tool in this field, for reasons I will give shortly. In order to understand how cognitive neuroscience goes about achieving its aims, this is one of the areas that needs to be examined. A second motivation is that the role played by computational models in science generally is not very well understood, and in the cognitive sciences even less so. There is disagreement over whether computational models are on par with experiments in their ability to provide empirical evidence, whether they are just tools for deducing implications, or something *sui generis*. Within cognitive science, questions have been raised as to whether certain kinds of computational models like connectionist models can tell us anything about cognition. A third reason, which was more a discovery than a motivation, is that computational modeling provides a good illustration of the points argued for in earlier chapters, namely that the aim of much work in cognitive neuroscience is elucidating multi-level mechanisms, and that these often include the sorts of upward looking explanations that might be accounted for by generic mechanisms. In this chapter, I introduce examples of computational models in cognitive neuroscience, and argue that they fit the picture of mechanism discovery methods described by MDC. In Chapter 5 I go on to argue that they explain in virtue of being hybrid generic-specific mechanisms.

It will be convenient here to broaden the field of interest to include the cognitive and neurosciences more generally, as most of the epistemic issues related to computational mod-

eling are continuous between this family of fields, and there has been almost nothing written about the epistemology of computational modeling in cognitive neuroscience specifically. I will therefore discuss examples that fall outside the project of cognitive neuroscience, and will work to resolve more general questions in the philosophy of computational modeling, along the way towards investigating explanation in cognitive neuroscience.

Before beginning, a few terminological notes are necessary. By *computational model* I mean a computer implementation, which might be either software or hardware, or some combination of the two, used as a scientific tool for investigating a target system. These can take many forms from analog models[1] like slide rules, or more complicated systems made of hydraulic parts or vacuum tubes, like the Phillips Hydraulic Computer in the London Science Museum, or the Colossus, a rebuild of which is housed at Bletchley Park's National Museum of Computing; to programs running on modern digital computers. What I mean to emphasize is that a computational model is a physical entity that operates and produces results. It is also common to use the term 'computational model' when referring to a plan for such an entity, or the abstract model that such an entity implements. For example theoretical Turing machines, which can't be implemented, are typically referred to as computational models. Important as they are, I will not be discussing that sort of theoretical computational model.

As briefly noted in Chapter 1, *theory* has several competing technical meanings, and which one is intended when the term appears in discussions of cognitive neuroscience is a question very much tied up with the issues being explored in this chapter. I will take care to specify what exactly is meant by 'theory' at various points in the discussion. The notion of a *model* is likewise contested and complicated. Some of the things that are referred to as models in science are: equations, diagrams, scale models, set theoretic structures, computer programs, animals, and collections of assumptions. Models can play the role of representations, representatives, pseudo-experiments, explanations (in both the epistemic and the ontic sense), and stand-ins for fundamental theories.

Although theories in the technical sense are rare in cognitive neuroscience, and explanatory models are much more common, I will often use the term 'theory' here when referring

[1] Interestingly, the origin of the term analog model is that they were considered analogies for the target system.

generically to explanatory/predictive apparatus. This is both in keeping with usage in the science being discussed, but more importantly, it avoids the awkwardness of talking about models in the sense of explanatory/predictive apparatus and computational models in the sense of software/hardware in the same breath. Using 'model' to denote a kind of explanatory apparatus would make it particularly difficult to clearly make the point that computational models often serve as models in the sense of explanatory apparatus. Where the distinctions between the various notions of theory and model become important, I will be careful to avoid ambiguity.

In dealing with models, one of the major puzzles is to figure out where models stand exactly with respect to representations and real world target systems, and another puzzle is to figure out what their role can be in theorizing and explaining, given where they stand. One of my aims is to clarify how it is that models can provide explanations, and in what sense of explanation they do so. Morgan and Morrison (1999) argue that the partial independence of models from both theories and the world is what allows them to be useful tools for exploring theories and the world. We can learn things from models, because they are not derived entirely from either source, but also contain outside elements. This will be the topic of Chapter 5.

4.1.1 Outline

I begin by motivating why computational modeling plays such an important role in cognitive neuroscience. I then describe how computational modeling is typically used in cognitive neuroscience. I compare this to cases like physics, economics, astronomy, and climate science, where the epistemology of computational modeling has been more widely discussed, but the motivations for computational modeling, and the modeling methods typically used, are somewhat different.

A long-running tradition of computational work on cognition falls under the umbrella of artificial intelligence (AI). In Section 4.3, I give a brief historical overview of developments in AI, emphasizing the changing relationship between theoretical convictions and modeling methods. I review comments from both modelers and philosophers relevant to the epistemo-

logical questions I'm exploring. I cover in some detail an extended debate about the relative merits of classical and connectionist AI that began with some philosophers claiming that connectionist models can't tell us anything about cognition. This debate reveals a deep ambivalence over what the role of computational models is in this field. There is an unresolved tension in that discussion between computational models as theories (in the Syntactic View) and computational models as physical implementations. There is also a confusing mixture of claims about neural plausibility being a virtue, while at the same time, simplicity and idealization are deemed essential. Despite this uncertainty about what exactly computational models are good for, if anything, cognitive scientists of various stripe continue to make liberal use of them, and apparently with great success.

One way of countering accusations that connectionist models aren't good for anything and to begin to resolve tensions about their role is to observe that many different strategies are involved in developing theories and building models. As a result, connectionist models might play various roles in this process, some of them closer to the role of mathematical deduction, and some closer to the role of experiment. As illustration, I introduce a recent example of computational cognitive neuroscience work that highlights several distinct roles that connectionist models can play in the search for mechanisms, even within a single research project. I note that if we look beyond the examples of work in AI most discussed by philosophers, this pluralism of modeling aims was always manifest in connectionist work. In Chapter 5 I go on to explain how computational models, and connectionist models in particular, can act as explanations.

4.2 MOTIVATIONS FOR COMPUTATIONAL MODELING

In many branches of science, we turn to computational modeling when experimenting directly on the target system is impractical, or the problem is very complex. In cognitive neuroscience, which studies the neural substrates of cognition, both of these motivations are in force, and few alternative methods exist.

Experimenting directly on human brains is only rarely practicable. Opening the skull

and poking around in the brain is obviously very invasive, so this kind of intervention normally wouldn't pass ethics board approval. Some rare exceptions are for the treatment of movement disorders like Parkinson's disease, or epilepsy (Engel et al. 2005). In these rare cases, single cell recordings and electrical stimulation interventions can be done, sometimes on awake, behaving patients, providing important validation of models arrived at by other means (in addition to the clinical uses which are their main purpose). It is important to note that these studies are necessarily of short duration, and usually 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). Likewise, neuropsychology provides some evidence about how the brain and behavior are related, by studying the psychology of humans with brain lesions. Care must be taken when drawing inferences from studies on brain-injured patients back to healthy populations (see Bub (1994a,b), Glymour (2001), Shallice (1988)), but this is also an important source if information.

In addition to these rare opportunities for invasive studies, several non-invasive means for measurement from and intervention on human brains are available, although all of these are indirect. Technology like Transcranial Magnetic Stimulation (TMS) allows for transcranial interventions on brains using electromagnetic induction. Neuroimaging technologies like Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI) allow for measurements of blood flow in the brain, which is related to brain activity. Electroencephalography (EEG) measures electrical activity rather than a proxy for it, but only through the skull, which adds interference, and makes it difficult to investigate deeper brain regions. All of these methods are limited in their spatial and/or temporal resolution. A number of arguments have been made suggesting that these methods may also be limited in more principled ways, in terms of the sorts of inferences that can reliably be drawn from them (Bogen 2001, Coltheart 2006, Logothetis 2008, Schall 2004, Uttal 2003, Vul et al. 2009).

These limitations on human experiments can to some extent be gotten around by experimenting instead on model species like nematode worms, sea slugs, mice, rats, or macaque monkeys, but when investigating higher cognitive processes, these animal models can't do all the work. Animals can't understand complex tasks, and can't give verbal feedback, making

it very difficult to investigate higher cognitive processes. Computational models provide an additional way of investigating human brain functioning.

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 barely understood. Computational models make understanding this complex system more tractable. Computational models thus play a very important role in cognitive neuroscience, since the other experimental methods available are limited in various ways, and the subject matter is so complex.

4.2.1 Typical Uses of Computational Models

In cognitive neuroscience, a typical computational model proposes, explores, and/or tests a hypothesized neural explanation for a psychological effect or neurological disorder. In order to do this, a computational model is built that is structured vaguely in accordance with some known neuroanatomical or physiological facts or hypotheses. In what sense a model needs to share the structure of the target system is open to several interpretations. A successful computational model should reproduce in its results the psychological effect or neurological disorder aimed at, and do a better job of it than any competing models. This is referred to in the field as simulating the data. (Simulation takes on quite another meaning in the philosophical literature on computational modeling.)

Many models compare two (or more) conditions: one where the model is 'damaged' and one where it is not. The damage might take the form of connections between parts of the system being disabled, a module being removed, or a change being made to the function that determines the activity of some parts; any of these may be rough methods of simulating neural damage. The goal is to reproduce the behavior of both patients and controls. If the model succeeds at all of this, a tentative conclusion is drawn suggesting that the neural system might work in the same way as the computational model.

If the computational model produces interesting results or suggests a neural structure that is not yet known to exist in the target system, this is considered a prediction of the

model. Confirmation of these predictions, whether behavioral or neural, would make the model more of a success, especially if the new finding is surprising or counterintuitive.

An often mentioned benefit of computational models is that compared to 'verbal' models (models of how a system is thought to work in the form of diagrams, charts, or verbal descriptions), computational models must be completely specified, so conjectures about whether the model works as intended are settled by running the program.

This picture of a typical computational model's methodology is informed by a wide survey of models selected from cognitive neuroscience textbooks, the top journals in the field, and citation lists. The particular papers that I describe in detail below to illustrate this picture were selected by searching for 'computational model cognitive neuroscience' on GoogleScholar, and selecting all the papers and chapters about attention models published in 2000 or later in the first two pages of hits. Although I focus on models of attention, computational models of other cognitive phenomena follow the same pattern (see O'Loughlin and Thagard (2000), Coltheart et al. (2001) for well-cited models of autism and reading errors, for example).

Amos (2000) presents a model of performance on the Wisconsin Card Sort Task (WCST), a measure of executive attention, that compares patients with schizophrenia, Parkinson's disease, Huntington's disease, and healthy controls. Amos criticizes two previous models for not taking into account anatomical details, and one for not implementing a simulation of patient data. He says,

> A network constrained by neuroanatomy and patient data might engender a finer level of description of processes involved in the performance of the WCST, and of the information processing performed in the frontal cortex and other areas of the cortico-basal loops (Amos 2000).

Anatomical details of frontal cortex and the cortico-basal loops are incorporated in the model, including columnar organization, excitatory projections to the striatum, and topographically organized projections from the substantia nigra and globus pallidus. Physiological research is also incorporated, for example by modeling dopamine dysfunction in schizophrenia as a reduction in the gain on affected neurons' activation functions. Although these and other biologically realistic details are built in, the model is still highly simplified in terms of the connections between brain areas implemented, the number of 'units' in each modeled brain

area, and the details included in these units.

Damage to the system is simulated in several ways for the various conditions considered, and to explore alternative hypothesized models for these conditions, for example by increasing bias against firing in frontal units, decreasing gain in the striatal module, reducing the output of striatal neurons, and adjusting noise parameters. The patterns of errors produced by the model under the various damage conditions are compared to behavioral results from the patient populations. The hypotheses corresponding to patterns of errors that better match the behavioral data are taken to be confirmed. Amos (2000) says that, "Simulation of quantitative data from patient populations may help differentiate the explanatory value of these approaches," which I take to mean that implementing competing hypotheses in computational models, and trying to reproduce the patient data in some detail is a way of deciding which of the hypotheses provides the best explanation of the data.

Amos suggests several predictions arising from his model and gives some details as to how these might be tested in single-unit studies on monkeys, and behavioral studies on schizophrenic, Parkinson's and Huntington's patients. In addition, Amos suggests a possible link between the three diseases considered; they may overlap in their symptomatology because of a common mechanism. Several times Amos mentions that he is looking for "information processing" mechanisms, although it is never made clear what exactly he means by this.

O'Reilly and Frank (2006) also describe a computational model of executive attention that is "based on the prefrontal cortex and basal ganglia." Their goal is a biologically plausible model that performs working memory tasks as well as previous, less biologically plausible models. They suggest that their model has "direct implications for understanding executive dysfunction in neurological disorders such as attention deficit-hyperactivity disorder (ADHD) and Parkinson's disease" (O'Reilly and Frank 2006), so again there is interest in accounting for both normal psychological effects and neurological disorders. Furthermore, the model generates testable predictions, and the authors claim their model can test hypotheses about the causes of neurological symptoms: "we think the model can explicitly test the implications of striatal dopamine dysfunction in producing cognitive deficits in conditions such as Parkinson's disease and ADHD" (O'Reilly and Frank 2006).

This model is rather more focused on the computational problems it needs to solve than that of Amos (2000), but still makes a point of getting both the biological details (the anatomical connections between frontal cortex and basal ganglia structures) and the behavioral data (performance of various groups on working memory tasks) right. They go into relatively more detail about certain aspects of the system, like the timing of the signals passing between the different parts of the circuit, but again the biologically plausibility is limited; the model "omits many biological details of the real system" (O'Reilly and Frank 2006, 31). Again there are comparisons to other models of the same behavioral data, and the selling points of this model are described both in terms of it matching the behavioral data more closely than other models, and in terms of it being more biologically plausible.

A telling indication of their general approach to integration is the following claim, "Because the PBWM model represents a level of modeling intermediate between detailed biological models and powerful, abstract cognitive and computational models, it has the potential to build important bridges between these disparate levels of analysis" (O'Reilly and Frank 2006). The virtue of a model being 'intermediate' in this way is a topic we will return to later.

De Pisapia et al. (2008) give an overview of computational models of attention in their textbook chapter. They highlight many of the same methodological points. Biological plausibility is valued, but only up to a point. They note that most computational models of visual attention share the same organization "which follows at least coarsely the structure and organization of the visual perceptual system" (De Pisapia et al. 2008), including modules representing brain areas V1, PP and IT, hypercolumns, refractory periods, local inhibitory connections, center-surround receptive fields, and so on. They also highlight integration of multiple explanatory levels "from single-cell neurophysiology to observable behavior" (De Pisapia et al. 2008).

The method I have described as being typical for computational models in cognitive neuroscience involves a combination of inferences. The logic of tendencies seems to be involved in the concern with roughly sharing some structural properties with brain systems, as it might be inferred that systems with like structure should behave in similar ways. Inferences to like causes from like effects seem also to be involved in concluding that a

simulation whose results match behavioral data the most closely are most likely to capture the underlying mechanisms giving rise to the behavior. What seems to be going on is inference to the best explanation; models are compared to one another to see which most closely matches both the underlying structure and the behavioral data, and the one that would make for the best explanation of the evidence is preferred, at least until a better competitor model comes along. What Amos (2000) may be getting at when he talks about constraints coming from both neuroanatomy and patient data is that there is a sort of squeezing from both sides, or an inferential pincer movement being used. The methodology depends on both sets of constraints being used together to narrow down the space of possible models.

In terms of the kind of explanation being sought in this sample of studies, there is much talk of underlying mechanisms. Uncovering the neural mechanisms that give rise to psychological phenomena or neurological conditions is what these studies are after. But the aim is very obviously not to get all the nitty-gritty details just right. The mechanisms they are looking for may be quite general. A nod is made to the value of unification in explanation in Amos's appeal to a common mechanism shared by schizophrenia, Parkinson's and Huntington's diseases. O'Reilly et al. talk about their model being intermediate between detailed biology and abstract computation. In later sections I will discuss in more detail why models with this intermediate status between biological plausibility and abstract mechanism might be preferred by scientists.

This sample is not intended to argue that this methodology is the only one employed in computational modeling in cognitive neuroscience. In fact plenty of computational models in cognitive neuroscience have other aims entirely, and correspondingly different methods. This plurality of methods and aims will become important shortly. The particular sort of model I describe here is significant, because computational explanations of cognitive phenomena do tend to take this form, and as will become clear in Section 4.3, questions have been raised as to how this sort of intermediate model can help explain cognition.

4.2.2 Computational Modeling in Other Fields

In some respects, the motivations for using computational modeling are quite similar in neuroscience as in fields like meteorology, astronomy, or economics. In these fields direct experimentation is likewise often infeasible. In astronomy distance and huge physical and time scales make direct experimentation impossible. In economics a combination of practical and ethical considerations make experiments on realistic scales rare. In meteorology, again spatial scale is a limitation. Additionally, these systems are, like the brain, too complex to understand without the help of computer models. The epistemology of computational modeling has been more widely studied in these fields.

In one important respect the motivations for computational modeling are quite different. In physics and economics, computational models often take the form of *simulations*. Simulations in this sense (I'll call them *true simulations* where there might be ambiguity) start from a fundamental theory, usually consisting of differential equations that describe the behavior of elementary entities like particles or agents. 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. Winsberg (1999, 277) explicitly focuses his discussion of simulation on "the practice of modeling very complex physical phenomena for which there already exist good, well-understood theories of the processes underlying the phenomena in question." Humphreys (1990) similarly restricts the term 'simulation' to "any computer-implemented method for exploring the properties of mathematical models where analytic methods are unavailable."

The output of a simulation can be compared to experimental results to test the fundamental theory, or to validate the simulation methods, but most often simulations are used for making predictions. Typical uses of simulation are to predict the progress of storms, predict the result of policy changes, or predict how a new galaxy will form. Humphreys (1990) lists three central uses of simulation:

> 1) To provide solution methods for mathematical models where analytical methods are presently unavailable.
> 2) To provide numerical experiments in situations where natural experimentation is inappropriate (for practical reasons) or unattainable (for physical reasons)...
> 3) To generate and explore theoretical models of natural phenomena.

Much of the work in simulation is in coming up with numerical methods for approximating solutions to the fundamental equations, because typically the mathematics becomes intractable when even a few of the basic entities interact. This is a problem cognitive neuroscientists wish they had.

4.2.3 Simulating the Brain

In cognitive neuroscience, computational models aren't typically used as a way of finding numerical solutions to intractable fundamental equations, because there are no such fundamental equations. We have no unified set of differential equations that describe the fundamental entities in neuroscience. We don't know what all of those entities are; we don't know all of the neurotransmitters at work, nor all the types of ion channels or receptor sites, and so on. Even without fundamental equations, we can do a pretty good job of predicting how cognitive agents will behave—psychology is ripe with effects. Prediction then is rarely the goal of computational work in cognitive neuroscience. More often the goal is to develop descriptions of how the underlying brain mechanisms work, which is akin to trying to discover a fundamental theory or model. The challenge in cognitive neuroscience is to figure out which mechanisms might underlie the observed effects, not to figure out what the effects might be, given the underlying mechanisms.

It is nevertheless true that in neurophysiology some computational models would count as true simulations. For example, simulations of action potentials or ion channels in neural membranes consider the micro-scale interactions between a number of parts, including Na^+, K^+, and Ca^{2+} ions, water molecules, proteins and lipids, in various concentrations, and given various electrical stimuli. Stochastic behaviors are modeled in the basic equations of neurophysiology, such as the Hodgkin-Huxley, and Nernst equations. Perhaps the closest thing to a fundamental theory is the Hodgkin-Huxley model of the action potential. This is a set of differential equations that describe the electrical properties of neurons, so it bears some formal resemblance to the laws that form the basis of physics simulations. However, predicting the behavior of larger scale systems with realistic detail is only rarely attempted, because these systems are extremely complex. There are dozens of known neurotransmitters,

many of which act in ways we don't yet understand, and there are many more neurotransmitters that have not yet been identified, so although we know some fundamental equations, we would need many more to get a full description. These sorts of simulations aren't done with the aim of overcoming the technical challenge of the mathematics being too complex to calculate the outcomes. Rather simulations in neuroscience are often done to try to figure out which equations might do a good job of describing the system.

McClelland (2009) speculates that building a real-time neural network model on the scale of rodent hippocampus might become feasible once computer power is 10,000 times what it is now. Most computational models of cognitive systems, however, involve highly idealized artificial neurons and many simplifying assumptions (and are in no way based on the Hodgkin-Huxley, nor Nernst equations), so even if the right number of them were included to be able to model, neuron by neuron, a big enough chunk of brain to produce cognitive phenomena[2], it is unclear whether such a model should be considered a true simulation. Without a fundamental theory underlying rigorous calculations, the logic of simulation does not apply, and the outcome can not be justified in the same way. If we can't assume the correctness of a fundamental theory, and don't have rigorous calculations, what we have is just a model, not a true simulation.

There are also more ambitious modeling efforts like the Blue Brain project, which has the long-term goal of modeling the entire human brain. So far they claim to have accomplished a "biologically detailed model of the neocortical column in the somatosensory cortex of young rats" (Waldrop 2012). Here too the point (or at least the more immediate one) is not to predict how a column of rat cortex or an entire human brain should behave, but rather to get the model to behave the way we already know rat cortex and human brains behave. The point is to perfect the model for the sake of understanding the brain and associated diseases.

4.2.4 Simulating Behavior

Computational models in cognitive neuroscience that claim to be simulations are simulations of behavior rather than true simulations. In a true simulation, we assume that the model

[2] I do not mean to suggest that there is or that we know of a minimum number of neurons for constituting a cognitive system.

gets the output right in virtue of having started with the right fundamental theory. The inferences work in the opposite direction when a model is said to have simulated the behavior. We infer that the model is right (and therefore explains the behavior) partly on the basis of the model successfully simulating the behavior. There is disagreement about just how this inference is supposed to work, or what justifies it. For Fodor (1974), the model explains the behavior simply in virtue of having simulated the behavior. For others (Newell and Simon 1961), simulating the behavior in a sufficient variety of cases is inductive evidence that the model got the propositional structure of the target system right. Still others (Craver 2006) think that a model explains only if it gets the underlying mechanism right, but that simulating the behavior isn't sufficient evidence of this. The interesting question is how we might justify the inference from successful simulation of behavior to getting the mechanisms right.

Insights from the literature on epistemology of simulation may not directly apply to modeling work in cognitive neuroscience, since here the modeling process does not involve developing tractable approximations to equations, computational models bear quite a different relationship to theories, and have rather different goals than do simulations. Although I've just argued for these differences, despite these differences, many of the same epistemological questions do arise. These questions surround the simplifications and idealizations that are made in building models, and how these are justified.

The simplifications inherent in connectionist models have a similar character to what is described for examples in physics and meteorology. Just as large volumes of atmosphere are treated as homogenous units in climate science (Norton and Suppe 2001), and parameters are used to approximate the behavior of the whole volume, single units in a connectionist network are sometimes taken to represent populations of neurons. The activity of a unit in a connectionist network is sometimes taken as a stand-in for the average activity in a population of neurons. This is seen as an unproblematic simplification, and connectionist modelers don't seem to get all twisted into knots over this simplification, because neurons are thought to work as populations, with the population vector or average firing rate being the measure that is causally relevant for brain areas upstream. Simplification is often seen as a virtue by connectionist modelers.

It may be that realistic detail is not usually the focus of attention in computational cognitive neuroscience, because prediction isn't the main goal. In meteorology the goal of prediction may predispose modelers to think in terms of deduction. The lack of concern with prediction among connectionists doesn't seem to hamper their equivalent aim of external validity. In fact, we test in exactly the same way to see whether a model can make predictions as we do to see whether a model can account for the data we already have. In my experience, models intended as predictors are tested with data that was previously prepared and set aside as the test data, not with data generated since the model was built. The same sorts of tests of external validity are how simplifying assumptions are justified in both cases.

In the next section I look at philosophical discussions and commentaries by scientists about computational modeling in the cognitive sciences, and begin to explore the relationships between theories, computational models, and the phenomena they target.

4.3 COMPUTATIONAL MODELS IN ARTIFICIAL INTELLIGENCE

In this section I look at examples of work in AI and philosophical discussions of such work, insofar as these bear on questions about the role computational models play in cognitive neuroscience. A number of competing and contradictory views emerge as to how computational models relate to theories of cognition. These examples and commentaries reveal that computational modeling is done in a variety of ways, with a corresponding variety of epistemic goals, but that there is no clear consensus on how to make sense of this variety, nor on how to justify the use of particular modeling methods for each type of epistemic goal. Not much progress has previously been made in making sense of this assortment of possible roles.

There are a number of approaches that can be taken to AI, including theoretical (e.g. convergence proofs), applied (e.g. algorithms for scoring bank loan applications, or controlling robot navigation), and scientific ones (i.e., investigating the nature of intelligence). AI approached as a scientific pursuit seeks to develop theories of cognition through the use of computational models. Some of these efforts seek to model cognition in terms of its neu-

ral basis, but others remain agnostic about how cognition is instantiated in brains, or even hostile towards that project. There are also various opinions within AI about how models relate to theories.

4.3.1 Classical AI

In the early days of AI, Allan Newell and Herbert Simon pioneered an approach where protocol analyses (verbal reports of thought processes) of human problem-solving were used as the basis for designing AI systems like their General Problem Solver (GPS) program (Newell and Simon 1961). This approach came to be known as symbolic or classical AI. Classical AI uses logical rules to manipulate syntactic structures, and hypothesizes that a cognitive system is like a programmed computer (or according to some *is* a computer). This approach has included many practitioners working on a variety of problems, but I'll take Newell and Simon's work as a representative early example.

Newell and Simon postulated that cognitive behavior is produced by elementary information processing over symbols, and that neurophysiological mechanisms in turn produce these information processes. This is represented in Figure 4.1. GPS is assumed to share with human cognition the level of elementary information processing, intermediate between its program output and hardware instantiation. The claim that any intelligent system shares this intermediate level is codified in their *physical symbol system hypothesis*, which states that, "A physical symbol system has the necessary and sufficient means for general intelligent action" (Newell and Simon 1976). The defense of this postulate "lies in its power to explain the behavior" (Newell and Simon 1961), they claim, and early successes with their method led them to conclude that GPS could be considered a *theory* of problem-solving behavior.

Note that in this approach, theory development follows simulation. In physics and economics, theory usually comes before simulation, and forms the basis for it. In Newell and Simon's case, a set of fundamental assumptions did go into the construction of the simulation, but the point of simulating was to discover what makes problem solvers tick. Their method is to start with an educated guess about underlying structure, see what their guess

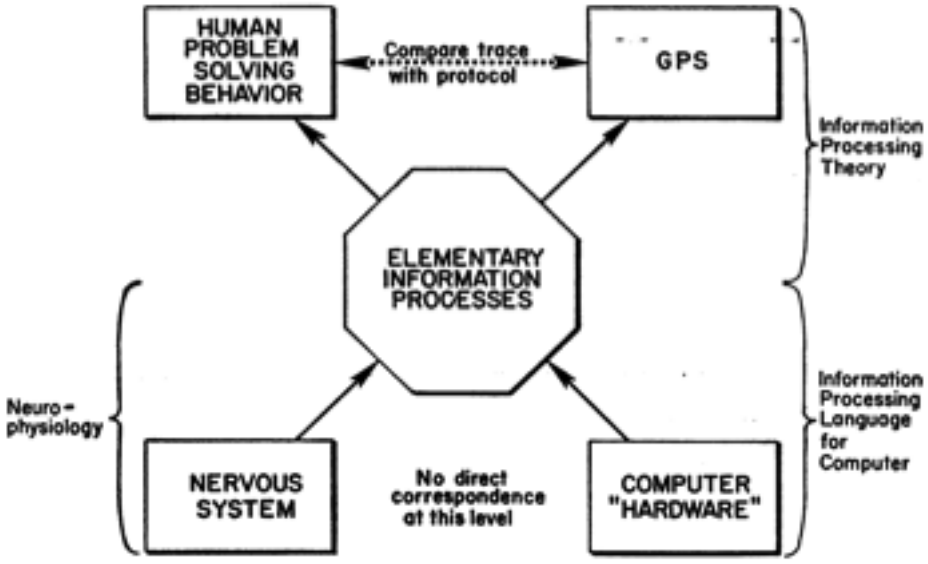

Figure 4.1: Levels in an information processing theory of human thinking. Reprinted from Newell and Simon (1961). Copyright ©1961 American Association for the Advancement of Science, used with permission.

generates when simulated, alter the model based on this feedback, and repeat.

Newell and Simon make several interesting methodological remarks. They cash out what they mean by a theory in the language used to describe theories in physics: "From a formal standpoint, a computer program used as a theory has the same epistemological status as a set of differential equations or difference equations used as a theory"(Newell and Simon 1961). They then go on to compare boundary conditions governing the applicability of differential equations in physics with environmental inputs determining the successive states of an AI program. What they mean by theory then, is the Syntactic View.[3] Newell and Simon acknowledge that a program being a theory, in this sense, does not necessarily make it a good theory: "How highly we will prize this theory depends, as with all theories, on its

[3] Since every program computes some computable function, a program like GPS certainly could be represented in the form of law-like statements. Sun (2009) goes to the trouble of spelling out some of the details of how to do so.

generality and its parsimony" (Newell and Simon 1961). Beyond these criteria they are, of course, concerned with whether the same elementary information processes show up in their participants' protocol reports as in their program traces. More detailed criteria like realistic reaction times were not a concern, at least at this early stage.

In addition, Newell and Simon acknowledge the role of neurophysiology in a complete theory of cognition; this is significant, because later defenders of classical AI reject the idea that neurophysiology plays any role in the study of cognition. Newell and Simon's work on GPS is only concerned with discovering what the elementary information processes are, based on protocol reports, and checking whether these processes can generate convincingly similar output when programmed into a computer. Nevertheless, they note the need for "a second body of theory ... to explain information processes on the basis of neurological mechanisms" (Newell and Simon 1961). They justify setting this work aside on methodological grounds. They suggest that, "Tunneling through our mountain of ignorance from both sides will prove simpler... than trying to penetrate the entire distance from one side only" (Newell and Simon 1961), presumably meaning that their psychological study could be done top down, while the physiological study would be done bottom up.

Their belief that internal psychological processes consist of symbol manipulations also plays a role in this methodological choice. If the physical symbol system hypothesis is true, then there shouldn't be any problems making the top-down and bottom-up analyses meet in the middle. Without this hypothesis, supposing that psychological systems must be supported by symbolic processes like those used in programs like GPS based on their effects being similar, would be a shaky inference from like effects to like causes. It is because of this assumption, based on introspection and common-sense psychology, that their methodology seemed justified. Without the assumption of a shared intermediate level, it would be hard to justify an exclusively top-down methodology.

Although the physical symbol system hypothesis was originally presented as an empirical claim to be tested, and the choice to work top down from behavior to information processes began as a proposal for dividing labor, these aspects of the classical AI approach became entrenched by the 1980s. A number of developments played a role in this entrenchment, including: Minsky & Papert's (1969) effective, if unfair (see Boden (2006)), quashing of

Rosenblatt's (1958) connectionist approach, plus Fodor's Language of Thought hypothesis (Fodor 1975), and his arguments for the autonomy of psychology based on multiple realizability (Fodor 1974).

In the meantime, theory had taken on a much more nebular status in philosophy of science. Particular AI programs designed to perform specific tasks were (and are) still referred to as theories, but this no longer harkened back to the Syntactic View, perhaps because this view was no longer so well received among philosophers of science. More often, by the 1980s, collections of basic assumptions or conceptual frameworks were referred to as theories, as in the computational theory of mind.

The classical AI approach has proven extremely useful for modeling some examples of intelligent behavior, like playing chess and solving logic problems. But symbolic AI didn't pick up again the other half of the problem: figuring out how physiology gives rise to elementary cognitive processing. This neglect makes it particularly unsuited for figuring out how cognitive phenomena are achieved by the brain, which is the goal of cognitive neuroscience.

4.3.2 Realistic Neural Models

At the other end of the spectrum, there are highly detailed, neurophysiology-inspired computational models. I already mentioned the Blue Brain project. Bower and Beeman's so-called GENESIS system is another example of a "realistic" approach to modeling brains, in which parts of axons and dendrites are treated as cylindrical sections of wire, and several varieties of ionic currents can act at each axon hillock. Bower and Beeman (2003) describe computational neuroscience as involving an interactive iteration of experimental work in physiology and anatomy with computational modeling. Modeling, they say, fills the need for a "quantitative approach to exploring the functional consequences of particular neuronal features" (Bower and Beeman 2003). According to this approach, models provide direction for further experiments, and aid in interpreting the data obtained experimentally. Bower and Beeman (2003) note that experiments do more than just test theories, and argue that modeling is "a mechanism for generating new ideas based on the anatomy and physiology of the circuits themselves." Computational models are embedded in the process of scientific discovery, in

this approach.

In contrast with classical AI, Bower and Beeman (2003) assume that "the structural and physiological details of the nervous system matter" because they believe that "an eventual understanding of the way nervous systems compute will be very closely dependent on understanding the full details of their structure." Some of the benefits they cite over more abstract approaches to neural modeling (it is not entirely clear whether they are referring to classical AI or connectionism, which I'll discuss next) are: that experimental data can be used as constraints, which limit the otherwise vast possibilities in terms of component types and parameter space to be explored; that biologically relevant outputs are generated, making them comparable to experimental data for the sake of prediction; and that the process of building a model brings to light what is not yet known about the system, requiring more careful consideration of what should go into the model. These seem like fair criticisms of classical AI's approach, but we will see later whether these advantages are not also available to connectionist models, which are much less realistic than GENESIS.

In defending their realistic approach to computational modeling, Bower and Beeman (2003) claim that more abstract models have usually been concerned with testing cognitive theories, or demonstrating the biological plausibility of pre-existing cognitive models, rather than exploring the computational features of neurally-realistic structures. Since classical AI is not at all concerned with biological plausibility, I take connectionism to be the target of these claims. My reading of the literature on connectionist models does not bear out this picture, as I'll argue in the next sections. Bower and Beeman (2003) then make the puzzling suggestion that Churchland and Sejnowski (1988) are critics of realistic neural modeling, when their point is actually that "Neurobiological data provide essential constraints on computational theories," and continue with the even more puzzling suggestion that Marr (1982) had proposed that "any particular biological neuron or network should be thought of as just one implementation of a more general computational algorithm" (Bower and Beeman 2003). In the end, Bower and Beeman (2003) do not settle on any clear methodological points. They admit that it is sometimes best to start with a lot of detail, sometimes less, and then adjust as required. What they describe as their most important methodological advice is "to simply get started and not worry about it" (Bower and Beeman 2003).

It is certainly true that for some problems a rather more detailed computational model would be most appropriate, and the Blue Brain and GENESIS projects, as well as others like them, have made valuable contributions to our understanding of the brain. But as we've seen here, some of the justifications for using very realistic models include misreadings of connectionists, and blind faith. The more plausible justifications are that neural realism provides useful constraints on the possible models to consider, ideas and detailed predictions for new experimental work, the impetus to fill in gaps in knowledge, and a way of finding out the functional consequences of neural structures.

In the next section we will see that all of these also serve as justifications for connectionism, which is also concerned with modeling brain structure, although in a less realistic, more schematic way. It will take some work to see how these same advantages can be had despite a more abstract sort of neural realism.

4.3.3 Connectionist AI

In contrast to classical AI's assumption that cognition consists of logic-based, information processing over symbolic representations, Rosenblatt hypothesized that symbolic representations may not be involved in cognition at all. He pointed out that sensory information may not ever be stored in the form of coded representations. He suggested that, "the images of stimuli may never really be recorded at all, and that the central nervous system simply acts as an intricate switching network, where retention takes the form of new connections, or pathways, between centers of activity" (Rosenblatt 1958). Rosenblatt experimented with networks called Perceptrons made up of simple units, with random connections between them, roughly mimicking the neural structure of the retina. Contemporary connectionists picked up this project quite a bit later.

The two-volume publication by the PDP Research Group (Rumelhart and McClelland 1986b, McClelland and Rumelhart 1986), sparked renewed interest and philosophical debate about what connectionist networks are capable of doing. Several theoretical problems for which Rosenblatt's program was criticized had been worked out in the meantime, including a convergence proof for multi-level networks, and the backpropagation training algorithm.

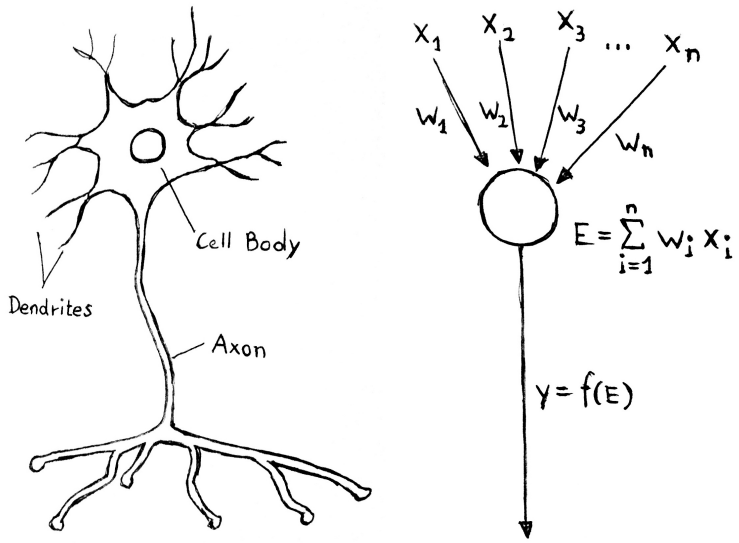

Figure 4.2: Schematics of real and artificial neurons. Copyright ©2013 Boris Hennig, used with permission.

The methodology used by the PDP group, and their underlying assumptions about what it would mean to develop a theory of cognition are quite different than those of classical AI, and more carefully worked out than those of Bower and Beeman (2003).

4.3.3.1 Defining Connectionism Connectionist models have an architecture roughly analogous to networks of neurons. They consist of a number of simple units (sometimes called artificial neurons) with multiple input connections from and output connections to other units. This mimics the structure of real neurons which typically receive input through their dendrites, then provide output to other neurons by generating action potentials down their axon. Schematics of real and artificial neurons are shown in Figure 4.2. The standard network architecture is a three-layer, feedforward neural network, where each unit sends output to every unit in the next higher layer. Any pattern of connections is possible though, including lateral, feedback, or recurrent connections. Units have activations and their connections have learned weights. The activation of a unit is typically a function of the weighted sum of

its input activations, for example a sigmoid function on $[0, 1]$, or a step function.

Here I will define connectionist models broadly to include any modeling architecture consisting of multiple simple units governed by local activation rules. This includes Hopfield nets[4] with layers of fully-connected dynamic units with random time steps; spin-glass models where units are connected to their nearest neighbors on a 2-D surface; cellular automata, which are connected much like spin-glass models, but have discrete-valued activation functions and time-stepped change points; and Bayes nets, among others. The general category of graphical models, which includes many dynamical systems and Bayesian models is a superset of what I'm calling connectionist models, and much of what I'll say applies to graphical models more generally.

Connectionist models don't fit neatly into the taxonomies in the philosophical literature on simulation. Rohrlich (1990) distinguishes simulations that churn through numerical solutions to differential equations from Cellular Automata. Hartmann (1996) casts this distinction as between continuous and discrete simulations. In a recent review of computational methods, McClelland (2010) says he was asked to write about statistical models, in contrast to symbolic models, but since he didn't think this captured the relevant distinction, he instead wrote about connectionist models in terms of emergence. Connectionist models are simultaneously like simulations that find solutions to differential equations, and include discrete models like cellular automata.[5] Connectionist models can be either continuous or discrete, as the set of possible states of the system is continuous when the activation function is continuous, which it often is, or discrete when the activation function is a step function, for example.

A taxonomy that makes more sense to me would first divide connectionist models (and graphical models generally) from the rest. These models are based on local computations between nodes in a graph, so could be called local models. A second division of the non-local models would have symbolic models on the one hand, and numerical solutions to differential

[4]See Anderson and Rosenfeld (2000) for discussions of the controversy over the provenance of so-called Hopfield nets. Grossberg invented them in 1974, Amari described them again independently in 1977, and Hopfield didn't describe them until 1982.

[5]To see how cellular automata can be a kind of connectionist network, consider arranging the units into a 2-dimensional array instead of layers or modules. Connections would go from each node to its neighbors and all have a fixed weight of 1 (that is, the weights could effectively be ignored). The activation function would then be the same as the cellular automata's update rule.

equations on the other. Both of these approaches are based on calculating a number of overall system variables, so these could be called global models. The discrete vs. continuous distinction might almost fit the division within global models. In any case, there are also any number of ways of combining all of these approaches into hybrids of various kinds, so I don't consider these distinctions particularly useful, and it turns out that many of the observations that philosophers have made about numerical solutions to differential equations apply equally to connectionist models.

4.3.3.2 Connectionist Motivations The PDP approach is characterized by an interest in psychological and neural plausibility. In the introduction to the PDP volumes, the editors state, "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). In the next few sections I'll try to spell out what exactly a 'physiological' flavor is, and why having one might be epistemically or methodologically appealing

One reason they give for using a more physiologically-plausible architecture is that the serial processing of symbols characteristic of classical AI is too slow to account for some kinds of behavior:

> the biological hardware is just too sluggish for sequential models of the microstructure to provide a plausible account... Each additional constraint requires more time in a sequential machine, and, if the constraints are imprecise, the constraints can lead to a computational explosion. Yet people get faster, not slower, when they are able to exploit additional constraints (McClelland and Rumelhart 1986).

Paying attention to the pattern of reaction times, and concluding that your model needs to be revised when its predictions don't match the behavioral data is a move straight out of the cognitive psychologist's toolbox. Hinton, Rumelhart, and McClelland were trained as psychologists, after all (unlike Newell and Simon). At the same time, the PDP approach posed a radical challenge to the methodology and theories of cognitive psychology.

The PDP group's work questioned the assumption that cognition is fundamentally composed of symbolic, information processes. Rather, they suggested that the information-processing character of deliberative thought might emerge from a connectionist microstruc-

ture. Their models showed how complex behavior could arise from simple systems with minimal assumptions, and no built-in rules. For example, Rumelhart and McClelland (1986a) showed that learning past tenses of verbs didn't necessarily require separate rules for regular verbs and exceptions. Other PDP work challenged standard methods in cognitive psychology. For example, McClelland (1979) showed that it may not be valid to infer discrete processes from additive reaction time results, challenging Sternberg's (1969) additive factors method. Plaut (1995) later showed that double-dissociations in behavioral deficits can arise even if there are not separate modules responsible for the two types of behavior, calling into question one of the key methods in neuropsychology (in addition to the dual-route model of reading). These share the basic aim of demonstrating that the sorts of models often proposed by cognitive psychology, involving distinct modules or processes for each sub-task, and operating according to explicit rules, might be radically wrong. Simpler mechanisms without modules or explicit rules are capable of achieving the same results.[6]

4.3.3.3 The Past-tense Learner Here I'll look at one of these early examples in more detail. Grammar books usually insist that there are a set of rules for forming the past tense of English verbs (add -ed if it ends in a vowel other than e or a short vowel followed by a consonant, double the consonant if preceded by a long vowel, change y to i and add -ed, and so on), but that there are also irregular verbs that have to be memorized (went, came, did, said, ...). A cognitive theory of past tense formation based on observations of behavior would suppose there to be a set of stored rules, and a memorized list of exceptions somewhere in the neural machinery of English speakers. Rumelhart and McClelland (1986a) showed that this need not be the case.

Rumelhart and McClelland (1986a) set out to show that our knowledge of the rules of past-tense formation could be implicit and inaccessible, ie., we might not really be following such rules at all, despite appearances. The past-tense model they built had word representations as inputs and outputs, a standard pattern associator architecture trained with the classic perceptron convergence procedure, and a training set the contents of which changed over time. The early training set consisted of verbs commonly heard by small children; this

[6] Boden (2006) provides a detailed history of this period.

just so happens to include a lot of irregular verbs like go, come, do, say, etc. The later training set included a wider selection of verbs commonly heard by slightly older children.

With nothing more than a basic connectionist architecture and plausible training sets of English verbs, the past-tense model produced surprisingly realistic and nuanced output. When tested early in training, the system got the most common verbs right, which meant correctly conjugating a number of irregular verbs. A little later in training, the system began to get more regular verbs right, but over-regularizing the irregular, common verbs that it had previously gotten right (comed, goed, or even camed). Finally at the end of training, the model performed well on both regular and irregular verbs, including new ones not seen during training. These developmental dynamics mimic language development stages in children; kids also get the irregular, common verbs right early on, then start over-regularizing them, only later to get them right again. The standard dual-system theory has trouble explaining this pattern.

These results were surprising and impressive. First, it showed that the underlying memory architecture need not make any distinction between regular and irregular verbs, despite the distinction seeming so obvious to grammarians and observers of language performance. More impressive and surprising though, was that the curious facts about children's language development dropped right out of the model without being built in. It was surprising that a complicated set of behaviors like over-regularizing verbs that had previously been memorized correctly could turn out to just be a result of a simple memory model and the changing vocabularies heard by young children.

Many criticisms of this model have since been raised, such as that the 'Wickelfeature' representations of words make the task easier than it seems, that the developmental dynamics that seemed so impressive are a result of switching training sets in a somewhat arbitrary way, and that the supposedly plausible training sets are not actually very much like the vocabularies children hear during language development. I do not mean to minimize these criticisms of the model. It is probably true that the deck was stacked in favor of the model in one or more of these ways, and my point here is not to stake any claims about whether the past-tense learner succeeds in the end at explaining how children learn to conjugate verbs in English. What this model definitely did succeed at showing was that just by setting up a

network with a vaguely 'physiological' flavor, you can get functionally interesting properties to fall out without any extra modeling effort. In short, neurally-plausible constraints on models can explain much more than previously believed. It also showed that the learning and unlearning dynamics could be explained by shifts in training set characteristics, even if the particular training sets used were imperfect. That is, the statistical properties of training sets can result in significant architectural differences in learned models. These two discoveries demonstrated how powerful connectionist models could be.

4.3.3.4 PDP's Impact The impact of early examples of connectionist work in the PDP tradition went well beyond just shaking up theories of language acquisition. Other examples in the book questioned how psychologists should think about major areas of research like memory, visual perception, language processing, and problem solving. As mentioned already, PDP-related work also questioned reaction time methods, and whether dissociations warrant inferences about modular structure. In this respect, connectionist modeling seems to set itself against several assumptions of experimental psychology's methodology, so represented a novel way of approaching cognitive science.

One major issue at stake in the disagreement between symbolist and connectionist AI concerns the relative merits of top-down and bottom-up approaches to the study of cognition, and perhaps even which order they should best be done in, if tunnelling through the mountain, as Newell and Simon put it, needs to be done from both sides. Newell and Simon defended starting with a top-down approach, as did Marr. Connectionist modeling takes a more bottom-up approach, although their point of departure is not as low down as the bottom from which others like Bower and Beeman begin, and it's not clear that the approach is entirely bottom up. Connectionists claim that the sort of neural architecture underlying cognition does put important constraints on how cognitive processes can work. Chapter 4 of Volume 1 of the PDP book lists the constraints they take from neuroscience (while also making clear that these do not come close to exhausting our knowledge about the brain). Some of these are,

> *Neurons are slow... There is a very large number of neurons... Neurons receive inputs from a large number of other neurons... Learning involves modifying connections... Neurons communicate by sending activation or inhibition through connections... Connections in the*

brain seem to have a clear geometric and topological structure (Rumelhart and McClelland 1986b).

The claim that the neural architecture constrains cognitive models amounts to a denial of the physical symbol system hypothesis. Since the physical symbol system hypothesis was the justification for classical AI's top-down method, these developments could be seen as shaking cognitive science at its very foundations, rather than merely introducing a novel method.

It is clear that adopting the physical symbol system hypothesis was a watershed event in AI research, whether the hypothesis is true or not. Newell and Simon's decision to approach AI in this way allowed for significant progress to be made. The PDP approach's impact did not reverse this, nor was the intention to shut down symbolic-level AI projects. Rumelhart and McClelland (1986b) freely admit that some phenomena might be best described in terms of symbol processing. The point they stress is that some phenomena are not well accounted for by macrolevel descriptions, and might be better accounted for as "emerging out of the interactions of the microstructure of distributed models" (Rumelhart and McClelland 1986b, 125).

The current consensus seems to be that classical systems do a good job of modeling some kinds of behavior like mathematical reasoning, and planning, while connectionist models do a good job of modeling other types of behavior like memory, and forming associations. It is also now clear that connectionist architectures can implement, although perhaps not always elegantly, the sorts of processing that classical AI does well, like serial processing, and structured representations (Touretzky and Hinton 1988, Hinton 1988). In the end it is not necessary to choose between them. Hybrid systems may be the most appropriate approach for larger modeling projects.

4.3.4 Connectionism Attacked

In a 1988 paper, Fodor and Pylyshyn pushed back against what many saw as connectionism's attack on classical AI. Fodor and Pylyshyn (1988) charged that connectionist models either fail to adequately capture cognition, or do so merely by implementing classical mod-

els. The specifics of the argument have been discussed in many books and papers, including Aizawa (1994), Antony (1991), Clark (1989), Cummins and Schwarz (1987), Fodor and McLaughlin (1990), Fodor (1997), Green (1998), Horgan and Tienson (1991), McClelland (1988), Narayanan (1988), Ramsey et al. (1990), Ramsey (1997), Rowlands (1994), Smolensky (1988a, 1991), among many others. Fodor and Pylyshyn's argument boils down to the claim that cognition requires representations that are structured in particular ways, by definition, so any model that lacks this structure must be inadequate. Much of the discussion cited above is about whether such structured representations are possible with non-classical architectures. If not, as Fodor and Pylyshyn argue, then any 'adequate' connectionist model must implement a classical one. Mere implementation details are, for Fodor and Pylyshyn, not very interesting, because they think cognition is multiply realizable, and what they're after is an account that is true of any realization.

Smolensky (1988a) replied to this attack by claiming that the proper description of cognition is at the subsymbolic level, as opposed to the symbolic level dealt with in classical AI. Smolensky claims that the symbolic level at best provides approximations to what is really going on at the subsymbolic level. This pair of papers spawned several rounds of replies from both sides, and about a decade of debate, mostly concerned with the issue of structured representation in one way or another. One point of disagreement is over whether cognition *by definition* involves structured representations. Another point of disagreement is over which level of description is more successful at accounting for cognition. Questions have been raised about how rigid and logical cognizers are, but questions have also been raised about whether PDP models capture the phenomena any better.

Another widespread criticism is that although connectionists try to make their models neurally plausible, in fact they are not. Backpropagation, which was a major technical breakthrough for the PDP group, is notoriously neurally implausible. The realization that connectionist models are not neurally plausible is often taken to be an unfortunate oversight, or an embarrassing mistake. I'll deal with this charge of neural implausibility in the next section. For my purposes, we'll leave aside questions about whether structured representations are essential to cognition. The resource I'm interested in mining here is that this debate brought up interesting questions about the role of computational models which so

far haven't been resolved within that debate.

4.4 THE ROLE OF CONNECTIONIST MODELS

Smolensky and other defenders of connectionism don't give a very clear picture of what role their models are intended to play in cognitive science. I've identified four different, partially contradictory stories about the role of connectionist models. These are that connectionist models provide a theory or theories of cognition, that they are implementations of such theories, that they simulate brain processes, and that they provide 'proofs of concept' or mathematical demonstrations. Several of these claims can be found within single papers (see Smolensky (1988a,b), Thomas and McClelland (2008) for examples).

4.4.1 Theories of Cognition vs. Implementations

The first prospective role for connectionist models is as an alternative to the prevailing computational theory of mind. Smolensky says that connectionism is intended as a "proper description of processing," as a "cognitive hypothesis," and as posing a challenge to the received view of cognition (Smolensky 1988a). The idea of connectionism as a theory of mind requires a bit of unpacking though.

Green (1998) directly addresses whether connectionist models are theories of cognition. He defines a scientific theory as consisting of empirical and theoretical terms, then asks what these terms might be in a connectionist model. He is perfectly correct that connectionist models sometimes treat the units as individual neurons, or particular concepts, but other times as unspecified summaries of we're not sure what. That the units need not correspond to anything in particular is touted as a positive feature by some connectionists, since it shows that the network underlying cognition need not be highly structured, and this lack of specificity allows for powerful, flexible learning. Green wonders "whether they are, indeed, TOO flexible to be good theories" (Green 1998). He complains that connectionist networks don't offer specific theories of the phenomena, since they aren't meant to literally model

brain activity. I suppose what he has in mind is something like how Newell and Simon's GPS was intended as a theory of problem solving. Connectionist models are not generally theories in this sense.

Boden describes one sense in which connectionism presented an alternative theory, "The past-tense learner was theoretical dynamite: It cast doubt on nativism and modularity... It undermined belief in the psychological reality of explicitly represented processing rules, and of explicit (symbolic) representations. And it threatened the then popular forms of philosophical functionalism" (Boden 2006). This new 'theory' on this reading is not a theory in the syntactic sense, but rather a loose set of assumptions: that the physiological microstructure is important, that complex behaviors can arise from simple systems, that learning and distributed memory can replace explicit rules and symbolic representations, and so on. Particular connectionist models are sometimes intended as theories in yet another sense: they are hypotheses about how a given behavior might arise and how it might be explained, although as Green (1998) points out, these hypotheses are quite general. Rumelhart and McClelland (1986a) is a theory of past-tense learning in this sense, and Plaut (1995) is a theory of the reading system. It remains unclear what use we can make of these sorts of theories that do not propose specifics, but just general ways in which something might work.

Connectionist models are pitched to some extent at the level of physiological details, which opens the door for Fodor and Pylyshyn's criticism that they are mere implementations (based on Marr's (1982) distinction between computational, algorithmic, and implementation levels). Smolensky says that connectionist models occupy the "sub-symbolic level," which while lower than the symbolic level, may not be quite as low as the implementation level. That Smolensky locates connectionist models at a lower level, in this sense of level, could be taken as an admission that Fodor and Pylyshyn are right about connectionist models not being at the level where they think cognitive phenomena are found, even if Smolensky thinks this level is not as low as they say. Of course the restriction to three levels may be artificial. In an earlier paper, even Marr includes four levels: computation, algorithm, mechanism, and hardware (Marr and Poggio 1976). It is thus unclear whether connectionist models are implementations in Marr's sense. Even if they are, it remains unclear whether they implement classical models or some alternative.

Rumelhart and McClelland (1986b) note that in these discussions it is often assumed that cognitive models should occupy the computational level, when actually, much of what concerns cognitive psychologists is at the algorithmic level; they want to know how cognition happens, not just which problem is being solved. This suggests that Fodor and Pylyshyn are wrong to place cognition as high as they do. In the PDP book's introductory chapter, McClelland and Rumelhart (1986) claim that: "macrostructural models of cognitive processing are seen as approximate descriptions of emergent properties of the microstructure." The idea is that a good theory of cognition will be found at a lower level than the symbolic level, because more robust generalizations can be found there. So even though connectionist models are at a lower level, they are not *mere* implementations. This is a different point than Smolensky's. Rather than denying that connectionist models are at what Fodor and Pylyshyn would call the implementation level in terms of descriptive detail, they deny that the symbolic level is able to provide good descriptions of cognitive phenomena. They claim that the robust generalizations do not lie at the symbolic level. On this view, the question of whether connectionist models are implementations is a moot point.

Connectionist modelers are, however, sometimes given to proclaiming the virtues of implementing one's theories. Unimplemented theories (i.e., ones that haven't been modeled on a computer) can be vague on points that seem innocuous but turn out to be important. They can also posit ideas that seem promising but just don't work the way we thought. Experiment is of course one way of checking that our theories aren't flights of fancy, but computational modeling can also play this role. Thomas and McClelland (2008) note that connectionist models have "forced more detailed specification of proposed cognitive models via implementation." Implemented models allow for comparison with behavioral evidence and the making of predictions that might be verified/falsified experimentally. Note that this was one of the advantages of realistic neural models that Bower and Beeman (2003) claimed can't be had by more abstract models. Connectionist models, although more abstract, also claim this advantage.

The quarrel between these first two options—connectionist models as theories of cognition, and connectionist models as mere implementations—is partly over which level provides the better theory in terms of accounting for the behavioral data more accurately and com-

pletely, which is an empirical matter, and partly over whether implementations are useful in building and testing theories (and if so, which implementations). According to connectionists, various sorts of constraints affect how cognition works, so including constraints of various kinds can be helpful, and can point in the direction of better models. Furthermore, implementations are a way of improving and verifying those models. This challenge to the physical symbol system hypothesis affects not just the prevailing computational 'theory' of mind, but also undermines the top-down methodology that Newell and Simon justified based on the assumption of a shared level of basic information processes.

Here I will note just briefly that both of these questions are instances of general issues we encountered in Chapter 3 on mechanistic explanation. There one concern was which sorts of constraints, coming from which direction, are relevant when developing a mechanistic model. Another concern was how to decide at which level the explanation of a phenomenon should be sought. These are the questions connectionists and symbolists disagree on in this struggle over whether connectionist models are theories or mere implementations. Shortly I'll argue that seeing connectionism as a method for discovering mechanisms in the manner MDC describe helps to resolve the disagreement, but before I get to that, I address two more suggestions as to the role of connectionist models.

4.4.2 Neural Simulations

Another way of looking at the attempt by connectionists to model cognition in finer detail than symbolic models do, is in terms of simulation. There is a widespread belief that connectionist models try to be realistic brain simulations. Connectionists sometimes encourage the idea that their models are intended this way; the PDP Group's rhetoric about their models being neurally plausible suggests this view. It is also true that connectionist models set up parameters defining the number and configuration of units, specify equations governing their interactions, then run the calculations to see what the dynamics of the system turn out to be. In this sense, they are not so different from simulations in physics that start with equations and parameter settings then churn through calculations to generate the physical consequences of those settings. The difference is that the equations connectionists start with

do not come from a fundamental theory.

Critics charge that if connectionist models are supposed to realistically simulate the anatomy and physiology of the brain, then they aren't doing a particularly good job of it. As mentioned already, people make almost a sport out of pointing out how these supposedly neurally plausible models fail to be plausible. Thomas and McClelland (2008) discuss how connectionist models have been criticized on the grounds that they "either include properties that are not neurally plausible and/or omit other properties that neural systems appear to have." In particular, the backpropagation algorithm, which was much touted as a technical breakthrough, is not neurally plausible (although there are now more plausible alternative learning algorithms that are also guaranteed to converge). In addition, dendrites don't just sum their inputs the way neural network units typically do; there is morphological and physiological variety among neurons in the brain, which the models often fail to reflect; and neurons are not typically connected to every other neuron within a system, the way a standard 3-layer feedforward model supposes.

One would think that if neural plausibility is supposed to be helpful, then it should be important to get the details of the target system right. In a simulation, any deviation from the details is a potential source of error that can lead to results that diverge widely from the behavior of the target system (but see Küppers and Lenhard (2004) for an example of how this fails to be true for simulations in meteorology too). At the very least it would be helpful to spell out what the relevant features of neurons are that must be accurately modeled, that is, to say what the positive analogies and negative analogies are, in Hesse's (1966) terms, and to argue for why the negative analogies can be safely ignored.

Not only are the positive and negative analogies not spelled out, but it is not always clear how much neural detail connectionists intend to model. The units in a neural network are fashioned roughly after individual neurons, which suggests that each unit is supposed to correspond to a neuron. The correspondence is obviously a loose one, since the similarity between real neurons and artificial units is not very close. Furthermore, far fewer units are used in connectionist models than would be realistic if each is supposed to correspond to a neuron. Certainly this is mainly because bigger models would take too long to run to be practical, and in addition would be very complex to analyze. In some connectionist networks

single units stand-in for populations of neurons, with the activation of the unit representing a population vector (Georgopoulos et al. 1986), so this kind of simplification may be justified. If in many cases units are intended to model large collections of neurons, one wonders how their rough structural resemblance to individual neurons justifies concluding that connectionist models are more plausible than models with no built-in neural-like constraints. Modeling populations of neurons would make more sense, if this is how the models are used. The limited biological plausibility connectionist networks do have is thus not even employed in modeling the level at which it is moderately plausible, so it is not at all clear whether the constraints provided are the relevant ones.

Indeed it was never a secret that the 'physiological' flavor stopped well short of realistic detail. This lack of detail seems to have been a feature that was included by design. Connectionists readily admit that their models are not realistic. Smolensky (1988a) notes that despite surface similarities between real neurons and the artificial neurons in connectionist models, there are significant differences. This point that connectionist models are not really very neurally plausible was also made several times in the PDP book. It is discussed in Volume 1, Chapter 4, then all of Volume 2, Chapter 20 is devoted to detailing the ways in which artificial neural networks are not like real brains. This chapter complains, among other things, about how "modelers seldom state exactly what their models are supposed to demonstrate" (Crick and Asanuma 1986). The introduction to the PDP volumes hedges on whether physiological plausibility is the point. It says,

> Though the appeal of PDP models is definitely enhanced by their physiological plausibility and neural inspiration, these are not the primary bases for their appeal to us. We are, after all, cognitive scientists, and PDP models appeal to us for psychological and computational reasons. They hold out the hope of offering computationally sufficient and psychologically accurate mechanistic accounts of the phenomena of human cognition which have eluded successful explication in conventional computational formalisms (McClelland and Rumelhart 1986).

They do not spell out how it is that a 'physiological' flavor can offer computational sufficiency and psychological accuracy.

Thomas and McClelland's (2008) response to criticisms of connectionism's lack of neural plausibility is to point out that simplification is necessary when modeling cognitive phenomena, and warranted assuming that the simplified models "share the same flavor of computa-

tion as actual neural networks." They cash this out in terms of the "information processing properties of neural systems," that is, by treating neurons as idealized electrical circuits that receive and send signals. In a class he was teaching on neural models of cognitive systems, Dave Touretzky quipped that "putting too much detail into a model is a novice mistake." In an interview, Jack Cowan says,

> the purpose of making models and doing theory is to provide insight and understanding of what's going on. It's not necessary to put in the kitchen sink to get insight... You have to abstract. If you're good at abstracting the essence of something, then you'll get real insights. That's what good theorists do. I think that just to simulate the hell out of populations of everything in the model is mindless. You've got to pick out of it the things that your taste tells you might be important and study those in some simplified context, where you can actually learn something. Otherwise, you're just mimicking what's there without necessarily understanding anything (Anderson and Rosenfeld 2000, 123).

The goal of modeling the target system accurately seems to be in tension with the modelers' need to keep the model simple so that it can produce insight rather than just a jumble of details. McClelland (2009) argues that while there is a cost to making simplifications in modeling—the inferences you draw from simplified models can always be challenged on the grounds that the properties of interest might result from the differences rather than from the similarities—it is necessary to simplify in order to achieve understanding.

There are arguments for neural plausibility along the lines of simulations needing to be as accurate as possible, and constraints of all kinds being potentially useful, and there are pragmatic arguments for keeping models simple by including approximations and idealizations. But why should it be a good idea to combine the two, yielding models that are neither neurally plausible nor simple? If connectionist models aren't really neurally plausible, then the logic of simulation where the fundamental theory and the accuracy of the representation justify the results doesn't apply, and arguments about the value of implementation don't quite connect. That connectionist models take into account constraints from neuroscience is supposed to be one of the main reasons for preferring them to symbolic accounts. If the neural plausibility turns out to be highly approximate and idealized, it's not clear why this should help.

We don't know how adding a 'physiological' flavor can help a model tell us about the target system. We don't know how much detail is the right amount, nor which details should

be included and which not. We're back to the question explored in the previous chapter of how it is that abstract models can tell us something relevant to real target systems. We'll come back to these questions later.

4.4.3 Mathematical Demonstrations

Another of the ways connectionists describe their methods is in terms of mathematical demonstrations. That simplification is seen as a virtue by modelers makes sense if the aim is to provide general mathematical demonstrations. In a mathematical model, the details of particular systems are abstracted away, and the aim is results that are true regardless of the details of a particular system.

McClelland (2009) discusses the role of connectionist models in demonstrating sufficiency and optimality results, and in "exploring the implications of ideas." Smolensky (1991) points out that symbolic models only explore the realm of discrete mathematics, and leave continuous mathematics untouched. He says that connectionism is committed to "uncovering the insights this other half of mathematics can provide us into the nature of computation and cognition." Another common claim is that computational models act as proofs of concept, which means demonstrating that you can get a particular result from a given set of assumptions, as in the past-tense learner. Thomas and McClelland call connectionist models, "a sub-class of statistical models involved in universal function approximation" (Thomas and McClelland 2008).

Some connectionist models are clearly aimed at making general points of this kind. For instance the motivation for Touretzky and Hinton (1988) is, "to demonstrate that connectionist models are capable of representing and using explicit rules." They say that another motivation was to show

> how 'coarse coding' or 'distributed representations' can be used to construct a working memory that requires far fewer units than the number of different facts that can potentially be stored. The simulation we present is intended as a detailed demonstration of the feasibility of certain ideas (Touretzky and Hinton 1988).

This sort of model that aims to prove a general point is at least one of the common types of connectionist model.

Critics of connectionism also sometimes claim that connectionist models are mathematical demonstrations as a way of discrediting them. It has been charged that because of their many parameters, there are so many degrees of freedom that networks with any characteristics at all can be trained up. If the same architecture can be used to approximate any function, one might legitimately wonder how it could be that the results of such an exercise could be informative about how cognitive models might be structured. What, in other words, can abstract mathematical results tell us about real minds and brains?

Anyone who has tried to train a network to approximate even one function knows that it is not so simple. Rosenblatt anticipated this criticism and offered counter-arguments as early as 1958. He invokes the "remark attributed to Kistiakowsky, that 'given seven parameters, I could fit an elephant.'" then argues that the remark does not apply to connectionist networks, where "the independent variables, or parameters, can be measured independently of the predicted behavior." In this type of system, he says, "it is not possible to 'force' a fit to empirical data, if the parameters in current use should lead to improper results... a failure to fit a curve in a new situation would be a clear indication that either the theory or the empirical measurements are wrong" (Rosenblatt 1958, 22).

Connectionists take themselves to be doing more than just abstract mathematical exercises though; they also build in features that roughly mimic neurons. It does seem perfectly reasonable to suppose that you might build certain assumptions into your model, run it to generate results, then conclude that in any system where your assumptions are true, you'll get the same results. This is sometimes questioned in the literature on simulation, but unless we get into deep questions in philosophy of math about how it is that formal proofs apply to objects in the world[7], it seems like an obviously valid inference. We can black box the computer's execution of the program and assume that what it's doing is equivalent to the same mathematical or logical calculations done longhand (except in rare cases of the machine malfunctioning, or if a lazy programmer hasn't taken care that floating point math isn't going to cause problems, and the user isn't taking those into account as an assumption). That the computer's operation can be black boxed in this way is exactly what it means for

[7] Batterman (2009) provides an account of how mathematics can play explanatory roles in empirical science.

something to be a computer.

Where there is room for worry is in knowing what all the assumptions are that are built into the program, and knowing whether all these assumptions are true of the case you're interested in. This is not a trivial problem. A mathematical result should hold for any system fitting the assumptions made in the model, but the more details that are built in, the harder it gets to come up with a general characterization of all the assumptions, and correspondingly, the harder it gets to know whether the system of interest fits them. Simplification per se does not pose any problems for being able to draw inferences from models to real systems. When these simplified models are nevertheless complex, it can be a problem in practice.

Connectionist models do not just simplify though, they also idealize: they make assumptions that simplify the implementation, but are not true of real brains. For example, a node's activation function is deterministic, whereas real action potentials are stochastic. By adding details to the models that are not true of the target system, their applicability as mathematical demonstrations is hindered.

Once again we have this tension between apparently contradictory roles; the aim of neurally-plausible models and the aim of general mathematical results seem to work at cross purposes. Some of the tension may be due to the interdisciplinary nature of cognitive neuroscience. Psychologists tend to value theories and not implementations. Neuroscientists tend to value neural plausibility and not abstraction.

Cartwright discusses a similar sort of puzzle in the context of physics. She notes that the best descriptions of a phenomenon, which in this case we might take to mean the most detailed descriptions, do not tend to correspond to the best mathematical models. "The descriptions that best describe are generally not the ones to which equations attach" (Cartwright 1983, 131).

What cognitive neuroscience tries to do, and what likewise the PDP approach to modeling tries to do is to navigate a middle road in between detailed description and abstract mathematical model. In order to navigate this middle road, we need a different conception of what a good theory is than the standard options available in either field. In the next sections I begin to make sense of the mixture of neural plausibility and abstractness exhibited

by connectionist models. One step towards resolving these tensions is to acknowledge that models are used for a variety of purposes, which requires them to take a variety of forms.

4.5 A PARTIAL RESOLUTION

The problem we face is that four quite different and somewhat contradictory roles have been attributed to connectionist models, and the characteristics of typical connectionist models don't seem to exactly fit any of them. Part of the solution to this problem is to observe that connectionist models are built for a variety of purposes. There are many different kinds of models that play a variety of roles in science; so too are there various kinds of connectionist models. Models intended for different epistemic roles require different characteristics.[8] That some of the roles conflict with other roles does not mean that connectionist models may not sometimes play one role, and sometimes another.

Looking back at the history of connectionism, it is apparent that there are various motivations and purposes for using connectionist models. The interviews in Anderson and Rosenfeld's (2000) oral history of connectionism suggest that among connectionist modelers there were widely differing views right from the start about what level of detail is best, how much physiological detail is desirable, and what the goals are in building these sorts of models. The modeling goals of the researchers interviewed are varied, and include engineering questions, mathematical questions, psychological questions, and neuroscientific questions. Several of the interviewees reflect on these differences as having been a source of unnecessary conflict. Additionally, connectionist models are applied to other areas of science like linguistics, epidemiology, and aeronautics. If connectionist models are used for so many purposes, then for some of these physiological plausibility may be important, while for others it may not be.

There might be roles to be played by models pitched as cognitive theories, other roles for implementations of cognitive theories, still other roles for detailed brain simulations, and

[8]This point was partly inspired by Friedrich Steinle's work (Steinle 1997, 2002), in which he argues that the sorts of experiments scientists do vary with the stages of research projects and their concommitant epistemic goals.

yet others for general explorations of mathematical space. These could act, respectively, as challenges to classical theories, tests of the feasibility of classical theories, prediction generators for neural theories, or tools for discovering neural mechanisms, for example. Given that this variety has gone unappreciated, it is likely that the apparent mismatch between the roles and the forms the models take is at least partly attributable to equivocation. It is easy to find examples of connectionist models that fail as realistic simulations of brains, for example, because this is often not their purpose.

To give an illustration of how the characteristics of models vary with their epistemic roles, I describe below a research project in which several connectionist models were built, each with somewhat different epistemic goals. At different stages in the project, the models built served different purposes, and correspondingly varied in terms of how much neural detail was included, and what kind.

4.5.1 An Example: Modeling Rat Hippocampus

The hippocampus is one of the main brain areas thought to be responsible for memory, and much of the research on it is done by studying rats finding their way around mazes. Rats are made to swim in cloudy water mazes with hidden platforms, run through plastic tubing with scented compartments, and wear microelectrode arrays to record neural activity. We have data about the circumstances under which rodents can learn and remember the location of these hidden platforms and scented tubes, and we also know a lot about what some of the neurons in the hippocampus and surrounding regions do while the rodents are performing these tasks.

Since the 1970s we've known about place cells in hippocampus, which fire when the rat is in a particular area or field in an environment. Different populations of cells code different locations, and together they form a cognitive map. In different environments the fields of particular cells and populations are unrelated; new fields spontaneously pop up. These cognitive maps have some other peculiar properties too. They rotate when the visual cues rotate, suggesting that they are tied to visual input, but they also persist in the dark, suggesting that they are independent of visual input. So although these cells respond to

visual cues, they keep tracking the rat's location as the rat moves, even if the rat can't see where it is in the environment. Using a mixture of cues, these cells keep track of where in the environment the rat is located.

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

Until recently it wasn't known where in the brain this path integration might be performed, nor how it worked. Redish and Touretzky (1997) proposed a connectionist model of how information from place cells could be combined with information about head direction and self-motion to achieve path integration, and laid out criteria that must apply to the brain area responsible for this component of the system. Essentially these were that it must be connected to the head direction, vestibular, and motor systems, as well as the place cells, and that its activity should be correlated with the animal's position. This first model was partly a mathematical demonstration, showing how a problem could possibly be solved, and partly a rough anatomical hypothesis to be tested. They proposed a how-possibly mechanism that could solve the problem, and a functional description of an unknown component in that mechanism.

Anatomists and physiologists took over from there, and subsequently, Fyhn et al. (2004) found grid cells in entorhinal cortex that had the required properties of Redish and Touretzky's model, confirming the anatomical hypothesis. The functional description provided by the connectionist model led to the discovery of a region with the right characteristics. The discovered grid cells turned out to also have the peculiar property of having firing fields that form hexagonal grids at various scales and orientations. It was assumed that these hexagonal firing fields must either be involved somehow in supporting the function these cells perform in path integration, or be a side-effect of the mechanism supporting grid cells' role in path integration, but it was not clear how exactly hexagonal firing fields and path integration

might relate. Over the course of the next couple of years the electrophysiological properties of these cells were more fully characterized, providing further clues about their exact role in achieving path integration.

The partial description of the grid cell sub-mechanism provided by these electrophysiological studies led in turn to further modeling work. Fuhs and Touretzky (2006) proposed a more detailed computational model of path integration, incorporating the physiological data about grid cells that had been gathered. The new model showed that "hexagonally spaced activity bumps can arise spontaneously on a sheet of neurons in a spin glass-type neural network model," and provided "a mechanism by which such cells could satisfy the computational requirements for path integration" (Fuhs and Touretzky 2006). Spin glass models are a type of connectionist network where each unit is connected to its closest neighbors in a multi-dimensional grid. The spin glass model described a how-plausibly mechanism for how the hexagonal grids might arise, based on what was known about the local network structure in entorhinal cortex, and the assumption that dendrites are closely packed.[9]

O'Keefe and Burgess (2005) proposed an alternative schematic model of how grid cells and place cells connect to give rise to path integration. Their model is "not inconsistent with" Redish and Touretzky's model, they say, and like the spin glass model, supposes that grid cells must be connected to their nearest neighbors. What they emphasize is the importance of another known fact about the hippocampus: place cells have cyclic firing patterns which correlate with movement (called the phase precession effect). An animal's location in a field correlates with the timing of place cell spikes relative to the EEG theta wave, while firing rate correlates with running speed, according to O'Keefe and Burgess (2005). In Burgess et al. (2007) they expanded on this model, and implemented it computationally. Essentially their model explains the hexagonal grid pattern as the effect of interference between multiple dendritic subunits tuned to different directions. Their model unifies the hexagonal grid pattern phenomenon and the phase precession effect as elements of the same system. They also note that the effects Fuhs and Touretzky (2006) describe might be added to their model

[9]Incidentally, spin glass models are another nice example of a generic mechanism. The model's formalism was first developed for describing the behavior of disordered magnets, but it has been widely applied in connectionist modeling of psychological and brain processes, as well as in financial modeling. Spin glass models are a particular kind of Ising model, which is even more general. Additionally, they got the name spin *glass* models, by analogy to the positional disorder of glass.

"to maintain the relative locations of the grids and enhance their stability and precision" (Burgess et al. 2007). Because the Burgess and O'Keefe model accounts for more data, and incorporates it into a more complete picture of the anatomy and physiology of the hippocampus and surrounding regions, it is considered more plausible than the Fuhs and Touretzky (2006) model.

My point is not to argue for or against either of these models; characterizing the hippocampal system in rats is an ongoing project. What I take from this example is that within a single research project, computational models played several roles. First a model was proposed requiring a fairly general kind of physiological mechanism. The characterization of this mechanism provided guidance to anatomists and physiologists, helping them to locate a previously unknown component of the system based on its required functional properties. Once an anatomical region with the right functional properties was found and further investigated by physiologists, its newly discovered properties required explanation. A more detailed model was built explaining both how the newly discovered properties might arise from plausible physiological conditions, and how the component might perform its function within the whole system. An alternative model was then built that could explain the same phenomenon in a different way, and at the same time unify the explanation with that of another, related phenomenon.

In short, computational models were used to aid in the discovery of a brain region of interest, to provide possible mechanisms to explain a curious effect, to show how several components might function together in a complex system, and to challenge an hypothesis by providing an alternative, more unified solution. For the purpose of discovery, very little biological detail was required. In showing that an effect could arise given plausible constraints, a more mathematical demonstration was provided. To argue for a more unified alternative theory, relatively more biological detail was needed.

In this example, models intended for different purposes leaned towards different characterizations of the role of connectionist models from among the options described earlier, demonstrating that part of the ambiguity may stem from there being a variety of reasons for building connectionist models. None of the models in this example fit just one characterization perfectly though. Redish and Touretzky (1997) was mostly theory of mind, combined

with proof of concept and implementation. Fuhs and Touretzky (2006) was predominantly proof of concept, combined with neural simulation. Burgess et al. (2007) was fairly evenly split between neural simulation and alternative theory (but not of mind), with a bit of proof of concept mixed in. This hybrid nature of typical connectionist models is a topic I'll pick up in the next chapter.

4.5.1.1 Discovering Mechanisms As may have been apparent, this episode of research on rat hippocampus happens to fit the mechanistic framework described in Chapter 3 extremely well. The path integration system began as a functional description, combined with a few of the known entities responsible for it, but mostly gaps. Redish and Touretzky (1997) proposed a mechanism sketch; they described the sorts of features they expected to find in the unknown path integrator entity, described what each of the known entities must do, and specified how they all should work together to provide productive continuity. Some of the details of the other entities were already known, but the path integrator entity remained a gray box: its function was specified, but the entity performing the function was unknown. The entity was then identified by physiologists, and some details about its activities were filled in, turning it from a gray box into a more fully elaborated entity. Fuhs and Touretzky (2006) then proposed a lower-level sketch of how that sub-mechanism might work, plus added more details to the overall sketch describing how it fit into the mechanism as a whole. In doing so, they instantiated the spin-glass schema which originated in physics, but had previously been instantiated in other neural networks. O'Keefe and Burgess (2005) provided an alternative mechanism sketch, which connected laterally to a previously proposed mechanism for another phenomenon. Burgess et al. (2007) added more details to that sketch, bringing it closer to being a fully-specified mechanism. Given all the constraints available, Burgess et al.'s model seems to be the more plausible alternative.

This interpretation of connectionist modeling as being concerned with discovering, exploring, and adjudicating between mechanisms is also consistent with what the modelers involved say they're up to. Fuhs and Touretzky (2006) describe their work as proposing and describing mechanisms. Redish and Touretzky (1997) say that their model "is constrained by both behavioral and neurophysiological data," consistent with a multi-level hierarchy of

mechanisms with constraints coming from both above and below. Burgess et al. say that their model provides a mechanism for path integration and an alternative to previous models (Burgess et al. 2007, 810). The advantages they cite are that their model integrates more diverse findings and makes testable physiological predictions. The method used in these models also matches quite well with the typical examples of computational modeling in cognitive neuroscience that I reviewed at the beginning of the chapter.

Thinking of connectionist modeling as a set of tools for discovering and exploring mechanisms may be a more apt characterization of the role they play in the cognitive and neural sciences than the various suggestions made by supporters of connectionism in response to the attack by Fodor and Pylyshyn (1988), and by critics of connectionism. I will further explore this role for connectionist models, and models more generally in the next chapter.

4.6 CONCLUSION

In this chapter I began by motivating the importance of computational modeling for cognitive neuroscience: the brain is extremely complex, and usually not available for direct experimentation. I then described how computational models are typically used: some biological constraints are built into a computational model, and the results are compared to behavioral data from patients with neurological disorders and healthy controls. Successful models are said to simulate the behavior of these populations. This is compared to typical uses of computational models in fields like physics and economics.

I then turned to the use of computational models in AI, which shares some of its aims with cognitive neuroscience, and has been much discussed by philosophers. I outlined classical AI methods, pointing out that they rely on the assumption of an intermediate information-processing stage, and that theories in the syntactic view were the original aim. I then reviewed attempts to model brains in realistic detail, which I take to be useful in some ways, but justified based on faulty reasoning about the value of abstraction. Finally I introduced connectionist AI, which was the topic of the rest of the chapter.

After briefly outlining how typical connectionist models work, I pointed out that there

is much confusion and ambiguity about the roles connectionist models play in the cognitive sciences, and how it is that models that are pitched neither at the level of cognitive theories nor at the level of neural plausibility may play those roles. I dug deeper into commentaries by connectionist modelers to find clues as to what their motivations are. I introduced the past-tense learner (Rumelhart and McClelland 1986a) as a prototypical example of the PDP approach to connectionist modeling. The discussion sparked by Fodor and Pylyshyn's (1988) attack on PDP brought out several suggestions: that connectionist models are alternative theories of mind, that they are mere implementations of classical theories, that connectionist models are intended as brain simulations, and that they explore mathematical space. I then offered a partial resolution to this mixture of contradictory suggestions: connectionist models have various uses, some of which demand more neural detail, some less.

I gave an example illustrating how connectionist models are used for various purposes even within a single research project. In work on path integration in rats, three different computational models were used for different purposes, and correspondingly they had different characteristics. I argued that seeing these models as tools for the discovery and exploration of mechanisms gave a more adequate description of them than any of the characterizations found in the AI literature. This suggestion will be elaborated in more detail in Chapter 5. There I go on to explore whether computational models, and connectionist models in particular, can explain and discover real-world phenomena, and what sorts of inferences are required.

5.0 COMPUTATIONAL MODELS AS MODELS OF MECHANISMS

5.1 INTRODUCTION

In this chapter, I take a step back to deal with broader questions about the role of models in science. The literature on models has addressed questions like: where to place models in relation to theories and the world, what the roles of abstraction and idealization are in modeling, how model explanations work, and how models relate to experiments. Building on this literature, I propose that seeing models foremost as tools for the discovery and exploration of mechanisms helps to clarify how they operate with respect to the above questions. That models are representations of their target systems is the standard orthodoxy. The representational role of models has, however, been overemphasized, at least in the cases I'm interested in. Models being representations is certainly sometimes relevant to them acting as explanations, but other explanatory models get their explanatory power elsewhere. I argue that another way for a model to act as an explanation is in virtue of its being or approximating a generic mechanism that is instantiated in the target system. I describe the inference patterns that can be used to explain or learn about target systems based on models that act as generic mechanisms.

After discussing these issues in terms of scientific models in general, I turn to computational models. Philosophical discussions of simulations have raised many of the same issues as arise with respect to models in general. There too, accurately representing a theory or a target system is assumed to be essential to the role simulations play in explanation and experiment. I argue that in many cases, computational models are not intended to operate as representations. Instead these models operate as generic mechanisms. Finally, I apply this idea back to the particular sorts of connectionist models discussed in Chapter 4.

5.1.1 Outline

In Section 5.2, I begin by introducing the notion of *model* that appears in the philosophical literature. Models have been described as mediators or intermediaries between theory and world. Contemporary wisdom has it that models take on many of the roles traditionally assigned to theories, including supporting explanation. In addition, they can take over some of the roles traditionally held by real-world physical systems, like being the material basis for experiments or quasi-experiments.

It is generally assumed that models are representations, and that this is essential to the roles they play. In Section 5.3, I briefly review some of the arguments defending the view that models are essentially representations, and that it is insofar as they are accurate representations that they are useful in explanations and experiments. I also briefly review some challenges to the assumption that representations can play these roles. The nature of representations is to be simplified and idealized, which works against the ability of models to represent accurately. I argue that simplification and idealization should instead be seen as important virtues of models, and can be if the assumption that models must be representations is relaxed. Generic mechanisms operate instead as something like stand-ins, or exemplars.

I argue further that the inferences involved in applying computational modeling results to target systems are not different in kind from inferences from more traditional sources like mathematical demonstrations, controlled experiments, experiments on model organisms, or experiments on samples. I spell out how these inferences work in a way that allows us to make sense of cognitive neuroscientists' liberal use of computational modeling.

In Section 5.4, I return to the use of mechanisms, and argue that models of mechanisms often operate as generic mechanisms, rather than as representations of mechanisms. The view of mechanisms I argued for in Chapter 3 comes in handy here. Earlier I argued that generic mechanisms, the in-the-world counterparts to mechanism schemas, get their explanatory power in virtue of being simplified and idealized. This is exactly the property we need for models.

Despite some apparent differences, computational models are models in the same sense

as in these general discussions; they can be aids to theories, experimental systems, representations, or models of mechanisms. Connectionist models in particular can be fruitfully understood as models of mechanisms. In Section 5.5, I argue that the connectionist models characteristic of the PDP approach can be made sense of by seeing them as approximations to generic mechanisms. The tensions surrounding what connectionist models are supposed to be good for can be resolved by noting that they operate as generic mechanisms in terms of how they contribute to explanation and experiment. That these models need be neither fully abstract nor fully detailed, but rather have a mixed or intermediate status, is exactly what allows connectionist models to play the roles set for them.

5.2 EPISTEMOLOGY OF MODELING

In this section I briefly review the philosophical literature on models, discussing what a scientific model is, how they relate to theories, and what roles they play. Scientific models can take many forms, including diagrams, equations, sets of assumptions, physical scale models, carefully bred fruit flies and mice, logical or set theoretic structures, or computer programs. Traditionally models have been thought of as rough representations, or pragmatic aids, but not as something that has any role to play in serious scientific work.

More recently, models have been afforded more status, as an essential tool in scientific practice. Several converging lines of argument point towards models playing a variety of important roles in science beyond the most obvious one of representing features of the world. Cartwright (1999) argues that models are not derived entirely from theory, but instead, "models are blueprints for nomological machines, which in turn give rise to laws" (Cartwright 1997). Morgan and Morrison (1999) argue that models act as autonomous mediators between theories and the world. Bailer-Jones (2009) notes that models are essential for making theories applicable to the world. In these accounts, models are placed between theories and the world. According to these authors, models have more than just a pragmatic role as helpers; models are essential for science. Theories can't be applied without them. Systems can't be understood and experiments can't be interpreted without models.

In discussions of cognitive and neurosciences too, it has been suggested that models can take over some of the jobs that theory used to be thought to do. McClelland (2009) claims that one of the roles computational models play is generating hypotheses to test. Craver notes,

> [Models] are used to make precise and accurate predictions. They are used to summarize data. They are used as heuristics for designing experiments. They are used to demonstrate surprising and counterintuitive consequences of particular forms of systematic organization (Craver 2006).

Cruse (2001) describes simulations as representing hypotheses for possible underlying mechanisms that might serve as explanations, and producing predictions that can lead to new experiments. There is consensus that both in general and in the cognitive and neurosciences, models do more than just represent.

Of the multiple roles that models can play, I'm interested foremost in their explanatory role. I begin by briefly reviewing some recent work on model explanations. As in other discussions of models, this work assumes a major role for representation. I argue that several familiar sorts of model explanations usually understood in terms of representation can be recast as non-representational model explanations.

5.3 MODELS AND REPRESENTATION

It is a common assumption about models that they are representations. This the starting point for several debates about scientific models. A number of papers discuss the issue of what it might mean for non-linguistic representations like set theoretic structures or isomorphisms to represent a phenomenon: see Frigg (2006), Giere (2004), Suárez (2003). Arguments for different views of how models represent (whether by similarity or isomorphism) also abound: see French (2003), Giere (2004), Teller (2001), van Fraassen (1980). Furthermore, it is commonly assumed that being representations is essential to models playing the roles they do in science. Bokulich (2011) evaluates three accounts of what makes a model explanatory, all of which assume that models are representations or descriptions of phenomena.

In this sense, models are still considered second class scientific tools; in order to be useful they must approximate, by accurately representing, one of the first class scientific tools like a theory or a real-world system. In this section I first review some of the arguments for models being essentially representational, then explore the possibility of non-representational models. I argue that in some cases, it is not in virtue of being representations, but rather in virtue of being appropriate stand-ins or exemplars that models make themselves useful.

5.3.1 Representation in Explanation

I'll call models that act primarily as representations, *representational models*. In some cases, representational models can be used to explain. Teachers use globes to explain geography to their students, and it is the fact that the globe represents the relevant geographical features that makes this possible. That the globe accurately represents the shapes and locations of countries and bodies of water is what makes it useful for explaining why Syria is not Iran's route to the sea, for example. Here the globe is a model of the earth, and its explanatory power comes from its being a representation of the target system.

Another teacher might use a globe and a piece of string to explain the unintuitive fact that the shortest flight path from Pittsburgh to Paris goes roughly over Gander, Newfoundland, despite it being further north than either the start or end point. In this case the globe needs to represent geometrical features rather than geography: the relative positions of the three cities, and the latitude and longitude lines are the essential features for this explanation. If these features of the system are accurately represented, then once again the globe can be used as a representational model, this time to explain why shortest flight paths between distant cities often curve significantly north or southwards.

Roughly this manner of models being useful in explanation in virtue of being representations is overwhelmingly emphasized in discussions of model explanations.

5.3.2 Beyond Representation

There is, however, a growing chorus of voices questioning the hegemony of representation in modeling, both within the general philosophy of science literature, and in cognitive science.

One problem is that there is no clear consensus about what a representation is. Some of the competing notions include similarity accounts, isomorphism, and homomorphism. There are known problems with each of these in terms of whether they include too much or too little, and whether they can do the work representations are supposed to do, of standing in for their targets and grounding inferences back to those targets.

Knuuttila (2011) argues for what she calls an "artefactual approach" motivated by the conviction that seeing models as representations is too limiting. She denies that the representational relationship is "privileged." Instead, she emphasizes how models function as "external tools for thinking, the *construction* and *manipulation* of which are crucial to their epistemic functioning" (Knuuttila 2011, 263). In Knuuttila and Boon (2011) this criticism of representational accounts of modeling is continued, and the artefactual approach explored further.

Ramsey (1997) argues that appeals to representation in discussions of connectionist models are largely confused. In Ramsey (2007) he expands on this critique to consider how representation is appealed to in computational theories of cognition more generally. He argues that the internal states and structures invoked as representations in connectionist modeling and cognitive neuroscience do not really serve as representations in those models at all. Both what he calls the receptor notion and tacit notions of representation are rejected as not being capable of doing the jobs representations are expected to do in these models, and as adding "nothing of explanatory significance" (Ramsey 2007, 186). He claims that if connectionist models provide an accurate picture of how the brain gives rise to cognitive capacities, "then those capacities are not driven by representational structures" (Ramsey 2007, 187).

If models do not operate primarily as representations, it makes sense to look beyond representations, and explore other, non-representational options for understanding models. My view of how models work is compatible with Knuuttila and Boon's idea of models as concrete objects that can be manipulated to generate knowledge. Their focus on models being constructed and on the role of agents in interacting with models is a bit different than mine, as they are interested in answering mainly epistemological questions. I'm mainly concerned with the ontic status of the models, and the methodologies used in modeling. Another source of inspiration for my non-representational view of models is Morgan (2002),

in which she distinguishes between models that are representations and models that are representatives (although Morgan defines models as representations).

The proposal I'll articulate in more detail in later sections takes the type of model I'm interested in to be physical stand-ins for their targets. These stand-ins allow for something like *surrogative explanation*, as described by Swoyer (1991). Swoyer's surrogative explanations are enabled by what he calls structural representations. The line between this notion of representation and the models I'm referring to as non-representational may be a fine one. The point I want to emphasize is that in drawing inferences from the model to the target, the fact that the model represents the target (if it does) plays no significant role, so calling these representations does no useful work (for my purposes). There are certainly other functions for models in which representation does play a role.

Returning to the example of explaining why shortest flight paths often curve north or southwards, a prop like a globe is certainly a useful pedagogical tool, but there are other ways of explaining the same phenomenon where it is less clear what explanatory role representation is playing.

Shortest flight paths can also be explained using equations for great circle distances and geodesics. These equations can be thought of as representations, certainly, if you let variables represent features of the system like the coordinates of each city. It is not obvious that the variables representing coordinates is where the explanatory power comes from, however. The equations on their own, without any variable assignments, are provably true. They guarantee that for any pair of non-antipodal points on a sphere, there is a unique great circle, and the shortest distance between those points is the shorter arc into which the points divide the great circle. The form of the equation for that arc shows whether it curves north, south, or neither.

None of this depends on the variables representing anything in particular. You can do the proof very well with the variables uninstantiated. In applying the proof to a particular case, there are two options available. In the representational option, you might substitute points within Paris and Pittsburgh for the variables, then calculate the equation for the flight route, and check the form of that equation.

Another option is to take the proof with uninstantiated variables as a generally proven

fact, applicable to anything that meets the assumptions required for the proof to work. Namely, Paris and Pittsburgh are on a surface close to a sphere, and are not antipodal, so they meet the assumptions. From this you can conclude that there is a unique great circle passing through them, the shorter arc of which curves northward. This can be proven without instantiating any variables. This second use of the proof depends on facts about the locations of the cities on an almost-sphere and the truth of the equations about great circles and geodesics, not on instantiating the variables such that they represent the locations of the cities.

In this sort of demonstration, a general fact is proven, of the form, $\forall x(P(x) \rightarrow C(x))$, where $P(x)$ means a set of premises is true of x, and $C(x)$ means the conclusion is true of x. From this general fact, if the premises are true of the target system t, then by universal instantiation and modus ponens, the conclusion is true of the target too. This does not require the target to be represented by the model (in this case the equation). The explanation just requires a target system to meet certain criteria, and for a generalization to be true of things meeting those criteria.

In this example, I would not say that the equations are physical stand-ins for the target system. The point is to show the form the inferences take in this sort of explanation. This same inferential structure is how I think experiments on models can be applied to their target systems.

5.4 MODELS OF MECHANISMS

Now I turn specifically to the use of models in mechanistic explanation and mechanism discovery. The sort of mechanistic model that has garnered the most attention are mechanism schemas, which play a major role in explanation and discovery. Schemas are defined in MDC as representations, so it is worth wondering how these relate to the non-representational models I've been discussing.

In this section, I describe how non-representational models of mechanisms, in particular generic mechanisms, the in-the-world counterparts to schemas, can be used in explanation

and discovery too. I then explore the patterns of inferences that connect models with the phenomena they explain and the evidence they provide. I describe how in several sorts of experimental contexts that are typically thought of as involving representational models, the methodology can be alternatively understood as involving generic mechanisms as models. The inferences involved in explaining with generic mechanisms do not involve the logic of representations, but are nevertheless very familiar ones. They have the advantage of connecting the model to the target more directly.

5.4.1 Schemas and Generic Models

One of the most important types of mechanistic model is the mechanism schema. Schemas are abstract, truncated representations of mechanisms. They can be built up from sketches full of gaps and guesses, by adding and correcting details about the entities and activities comprising the mechanism then abstracting (Machamer et al. 2000).

Although they are representations by definition, they may take many forms. Some examples are mathematical or chemical equations, some are linguistic descriptions, others are formal diagrams like flowcharts, and still others are more detailed pictorial diagrams. The ways in which these representations operate as explanations vary accordingly. Pictorial diagrams might explain by showing the shape and physical organization of entities in a mechanism, flowcharts might explain by showing the relative timing of activities, descriptions might explain by providing an intuitive exposition of how the mechanism works, and so on.

According to Darden, schemas also operate as something more than just representations of mechanisms, in particular during mechanism discovery. In schema instantiation, schemas are instantiated via analogy from other, better-understood phenomena. It was unclear in Darden (2002) how this use of schemas was to be reconciled with their primary use as representations of mechanisms. When a schema is instantiated in a new context, it cannot be a glass box anymore, since schema instantiation occurs in the context of discovery, where we do not yet know all the details of the mechanism. There seemed to be two distinct notions of schema in these discussions, although not necessarily incompatible ones. Darden did not go into specifics about how the analogies that allow for a schema to be instantiated in a new

context should work. Below I will clarify these two points: what the second type of schema is exactly, and how the analogies between different cases that schema instantiation affords might work.

My suggestion is that this second type of schema—ones that are instantiated in new contexts—do not operate primarily as representations. A schema starts out as a representation of the mechanism in the original context, which means that the schema is *about* that original mechanism. But the representation being about the mechanism in the first context, which is part of what it means to be a representation, doesn't inform us about a different mechanism in a different context that the representation is not about.

For a schema in the glass box sense to become useful as a schema in the schema instantiation sense, it must be more generally applicable. If it is a representation at all, it must be about both the original case and any newly instantiated case, so what it is about must be something that is shared between the two or more contexts where it is instantiated. The idea of a generic mechanism that I introduced in Chapter 3 fits the bill. A generic mechanism is a mechanism qua an abstract type. If two mechanisms belong to the same type, they are both instances of the same generic mechanism. Instead of seeing the schema in this case as a representation of the fully-instantiated mechanism, we can see it as a representation of a generic mechanism that the fully-instantiated mechanism instantiates. In cases where schema instantiation is fruitful, there should be a shared generic mechanism that the schema represents.

This second type of mechanism schema can help in explanations in a more substantial way than representations of mechanisms that aid explanation by acting as illustrations, props, or descriptions. An abstract mechanism schema that is in common to a case that is already understood and one that is not, can provide the basis for the explanation of the second, not previously understood, phenomenon. If the generic mechanism represented by the schema is instantiated in the new case, then any knowledge we have about the generic mechanism is thereby knowledge about the new case, and can be directly applied. This knowledge then goes some way towards explaining the new case, although a full explanation might also require knowledge about some specifics.

In the next section I'll discuss in more detail how non-representational models of mech-

anisms can contribute to explanations.

5.4.2 Non-representational Mechanistic Models

Mechanistic models can either serve as representations of mechanisms, or can be mechanisms themselves. As representations, they play an important role in explanations in the epistemic sense (explanation^{e}). Diagrams, descriptions, equations, etc., can all serve as teaching aids for communicating understanding about mechanisms, or as cognitive aids during episodes of mechanism discovery. These models might be quite specific or rather abstract, depending on the explanatory context and what is most conducive to understanding.

This variety of forms and of levels of abstraction does not exhaust the breadth of mechanistic model types though. In addition to diagrams, descriptions and equations, mechanistic models can also be scale models, animal models, samples, or computational models. Each of these might sometimes serve as representations of mechanisms too, but in other cases it is less clear that these types of mechanistic models act as representations.

A mouse used as an animal model is more obviously a fully-fledged mechanism than an abstract representation of one. A sample taken from a population is another kind of model that I'm tempted to call concrete rather than abstract. A scale model may or may not operate as a representation. Logical equations too might act as mechanisms themselves, since it is the properties of the vehicle or the syntax that logically entails their implications. Models of these types may play a causal (i.e., ontic) role in explanation^{o}.

As mentioned above, Morgan (2002) has described some of these cases as involving representa*tives* of or for rather than representa*tions* of their targets. She calls samples representatives of populations and animal models like mice representatives for other species like humans. I think this is right, and in the next section I expand on how the inferences from these sorts of representative models to their targets might be drawn. If we see these representatives or stand-ins as generic mechanisms, or approximations of such, it turns out that these inferences follow patterns that are very familiar from experimental reasoning. The inferences are indeed stronger and more direct than if we depended on the logic of representation.

5.4.2.1 Inferences from Non-representational Models Here I'll detail how we use and learn from models in cases where the model is intended as a representative, stand-in, or exemplar of a mechanism. In these cases, the model is treated as a mechanism itself rather than as a representation of one, and the key inferences are between generics and the specifics that instantiate that generic.

To take a simple case, consider an experimental intervention on an individual in order to find out about that individual's current state of health. A nurse might check a patient's urine for blood, in order to find out whether their kidney stone removal was successful, for example. In this case, the individual's kidney and its urine production is the mechanism of interest, and what we intervene on or observe is that mechanism itself. It would not do to deal with representations of the kidney and urine, nor to observe something highly similar; taking a pee sample from the person's twin would not do, no matter how similar their pee. Here the model, if we can call it that, is the same as the target system. The urine, a product of the kidney, indicates something about the kidney. We make an inductive inference from effect to cause. More complex cases can be built from this base.

Now consider a case where a nurse checks a person's urine with the aim of finding out about human kidneys and urine in general. For example, alien invaders might try to find out about humans based on a person they abduct. The abducted human's pee would be taken by the alien doctors as being representative of human pee, in that it is an example of such. It is that individual's pee, but it is also generic human pee. In the case of mechanisms like animals, there is reason to believe that there is a robust generic category to which the animal belongs, namely their species. There is then reason to believe that from a single example, we can find out about that generic. The abductee is not a representation of humans, but they are an instance of the generic human, with the addition of many details specific to that individual.

The alien nurses and doctors might investigate the pee sample, and draw various conclusions not just about that one abductee's kidneys, but about how human kidneys in general work. They might of course draw some false conclusions if they are not able to distinguish which characteristics of the pee are due to their abductee being human, and which are due to the abductee being the particular human they abducted. If there happened to be blood

in the pee, the aliens might wrongly conclude this to be a normal state of affairs, just as medical science assumed for centuries that human fetuses usually grow in the fallopian tube based on a dissection of a pregnant woman who happened to have an ectopic pregnancy (which may have been the cause of her death).[1]

Inferences from single instances of a generic are not entirely reliable. Specific instances are usually (perhaps always) also many other things beyond being instances of any given generic, and it is difficult to pick apart which of the types an instance belongs to are the causally relevant ones for a given phenomenon. This means that we need to use methods like random sampling, replication, and Mill's methods in order to draw reliable inferences. But inferences from experiments and observations are always unreliable in this way. Models do not have a different status than experiments in this regard.

Experiments on samples that are meant to investigate the characteristics of a population in general operate in a similar way, with the sample acting as a representative model of the population. In this case, the mechanism of interest might be something like human beings, or those falling into a category, such as children with Autism.[2] Each individual in the sample is an instance of the generic mechanism of interest, with some potentially-confounding individual factors added. For each individual, their being an instance of the target generic, plus all the other things they are, explains^{o} their behavior.

If the goal is to find out what the causal contribution of Autism in particular is, one could either investigate this generic directly, by finding or constructing an instance of Autism that is not also a bunch of additional things (which is unrealistic in this case), or investigate many different instances of Autism and figure out what they all have in common. Experiments on samples use the second strategy. With most human experiments, we cannot, for ethical reasons, construct an instance that approximates the generic closely enough to be able to investigate that generic more-or-less directly. When using a sample to investigate a generic, we draw inferences about the causal contribution of the generic target from all the cases in the sample, by taking random samples, and analyzing the statistics of the sample data.

[1] This example may be creative remembering on my part. I can't find a source.

[2] A different type of experiment using a sample might look for the distribution of characteristics over a population instead of what they all have in common, and the inference pattern would be correspondingly different.

In the case of animal models, a mouse might act as a representative of the generic mammal, or a macaque monkey might act as a representative of the generic primate. In some cases, animal models do act as representations; when a mouse is dissected during a teaching exercise, it serves as a representation, for example. In typical experiments with animal models, the mouse or monkey could still be taken as a representation of a type of animal, but this is not how it operates as an animal model. When monkeys are used in psychophysical experiments, they act as stand-ins for generic primate perceivers. Any monkey-specific aspects of the results are controlled for, so far as possible, through breeding, training, and avoiding species-specific behaviors in the task design. What remains is applied as a hypothesis about how primates in general perceive stimuli. Macaques are often chosen as animal models for a complicated set of reasons including practical and political ones, but in principle, because they are as close as we can reasonably get to a generic primate in terms of behavior and genetics. From the animal model we draw conclusions about generic primates, based on having investigated one. What we learn about that generic is then applied to humans via syllogism: primates are X, and humans are primates, therefore humans are X.

The use of mice in drug trials works in a similar way, with mice standing in for generic mammals. Knowledge about generic mammals is then applied to humans, as instances of the generic. Gene-knockout experiments use the same basic inferential structure, but involve additional technological manipulations to make the model animal as close as possible to the desired generic. These animals are specially bred to have a particular genetic make-up, with selected genes turned off. This is accomplished by adding artificial pieces of DNA to blastocysts, with markers attached that can be chemically inactivated, then breeding for two generations to create homozygotes. In this way mice can be bred that lack particular genes of interest, and they can be compared to mice that are otherwise genetically identical. The knockout mice are used as disease models, for investigating the role of particular genes in humans, and the phenotypic consequences of alterations to or variants of those genes. From these mice as representative mammals, we can infer that other mammals' genes likely work in the same way. The inference goes from mouse to generic mammal, and from generic mammal to humans specifically. Once again, the first inference, from specific to generic, is susceptible to error if the model animal's specific characteristics, beyond being an instance of

a generic mammal, are the real cause of the behavior. This inference can be strengthened by testing other mammals too, or by amassing other kinds of evidence, such as from molecular genetics. These confounding factors that may make the model behave differently than a 'pure' generic are analogous to uncontrolled variables in an experiment.

Scale models can also serve either as mechanisms themselves or as representations. Architectural models, for example, typically serve as representations of planned building projects. But in some cases it is unclear whether scale models operate as representations or not. For example, my former neighbor, Richard Manicom, needed to figure out whether his baby grand piano was going to fit in the door and around the corner into the living room before committing to buy his house. Since he had math and engineering degrees, he first tried to solve the problem algebraically. He tried for several days without success. Finally he decided to draw a simple floorplan using his son's school supplies, and cut a miniature model of the piano out of paper. He then pushed the piano cut-out over the floorplan to see whether the piano would fit. He determined that it would, so bought the house (and it did fit).

In this case, the floorplan and piano cutout are clearly representations of the house and piano, and it might seem that the model being useful for solving the problem depended on these accurately representing the house and piano in some respects. The paper cutout and floorplan don't share many physical properties with the piano and the doorway of the house, but the properties that are important are instantiated by the model. These properties—two-dimensional shape, relative size, and sufficient rigidity—are shared between the house and piano, and the paper model (assuming it is not folded or bent). The piano cutout actually fitting through the paper aperture, and these actually having the appropriate shape, relative size and sufficient rigidity are what allow the model to serve its purpose. We can call instantiation a kind of representation, but I think this is misleading.

An alternative to seeing this model as a representation of the doorway and piano, is seing it as something approximating a generic that they instantiate. Under this interpretation, Richard's quasi-experiment showed that rigid objects of the piano's shape fit through apertures of the doorway's shape and relative size. We know from Euclidian geometry that such a demonstration should also apply to similar shapes (ones with the same internal angles, but different lengths). We first infer from the paper instance to the generic model, based

on a reasonable expectation that it being made of paper is not relevant to the result, then we infer from the generic result to another instance of it, in this case the real house and piano. The result is more general than just that house and that piano though. If another Costain Redwood model house were for sale, and someone with the same kind of piano were considering buying that house, Richard's result would suffice to show that the other piano would fit into the other house. This may never happen, but the result does generalize.

A generic has the potential to be instantiated in many cases. When we know what is true of the generic, we thereby know what is true of any instance of the generic, as long as other types to which the instance also belongs do not have interfering effects. If the piano were made of iron and the doorway were strongly magnetic, for example, it might be that although in terms of its shape, the piano should fit through the aperture, the force required to get it through would be too strong to be able to manage it.

It remains a problem to know when generalizations about a generic will fail to hold of an instance because other factors interfere. My claim is merely that the pattern of inferences that we use in such cases more plausibly follows the logic of generics than of representations (for which there is no agreed-upon account). The inferences used in the above examples are familiar ones like Mill's method, Euclidean geometry, syllogism, and statistical inference, so reasoning with generics is continuous with familiar modes of scientific reasoning.

In this section I have argued that several familiar scientific uses for models, which might usually be thought of as essentially involving representations, can alternatively be understood as involving non-representational models, where the model acts as a stand-in for a generic mechanism. In the next section I will build on this suggestion, applying the same kind of analysis to the somewhat more complicated case of computational models.

5.5 COMPUTATIONAL MODELS AND EXPLANATION

In this section I argue that computational models sometimes operate as generic mechanisms. I begin by arguing that computational models are models in the technical sense from philosophy of science. I then argue that computational models are often models of mechanisms,

which was already suggested in Section 4.5.1. I next argue that some computational models, including some connectionist networks, are best understood as approximations to generic mechanisms. Understanding connectionist models in this way finally resolves the puzzle discussed earlier about what role these models play in cognitive science. Finally, I discuss in more detail the explanatory role of computational models as generic mechanisms, and explain why the odd mixture between abstraction and detail characteristic of connectionist models might have explanatory benefits.

5.5.1 Computational Models as Models

Computational models, which I defined earlier as programs running on or hardware implemented in computers, seem like quite different sorts of things than the models that are most often discussed in philosophy of science. They are superficially unlike the Bohr model of the atom, or the Lotka-Volterra model of population dynamics, for example, and also unlike diagrams, descriptions, gene-knockout mice, scale models, etc. The view of models inspired by Morgan and Morrison (1999), fits computational models quite naturally, however. It is perhaps even easier to see in the case of computational models than for other types of models, that they are simultaneously like theoretical objects, as well as things in the world. A computer program is formally very much like a mathematical proof, and can be made to fit the criteria for a theory on either the syntactic or the semantic account. A computer running a program is also a real-world event, very much like an experimental trial. The role of models as intermediaries between theory and world, as mediators, as autonomous from both theory and world, or as tools that aid in the application of theory all come quite naturally to computational models. My claim here is that computational models are models in the same rough sense as discussed by Morgan, Morrison, Bailer-Jones, Cartwright, etc., and I will show how they too might take over some of the roles traditionally held by theories, such as providing explanations and acting as experimental systems.

Given that they are called computational *models*, it may seem so painfully obvious that they are models as to not be worth arguing for, but there is resistance to the idea. There is very little agreement about what the epistemic status (also the metaphysical status) of

computational models is. Humphreys (2004), Norton and Suppe (2001), Winsberg (2003) and Barberousse et al. (2009) discuss whether simulations can generate empirical knowledge. Parker (2009), Winsberg (2009), and Arnold (2011) discuss whether the materiality of simulations matters and how so. Frigg and Reiss (2009) collect together a number of claims along the lines of simulations (the type of computational model that is the subject of most discussions in philosophy of science) being something *sui generis*. According to Frigg and Reiss (2009), a number of people—Rohrlich (1990), Galison (1996), Winsberg (1999, 2001), Humphreys (2004)—have argued in one way or another that simulations create novel metaphysical conditions, demand a new epistemology, disrupt our views of theories and models, or are a novel kind of method that is neither theoretical nor experimental. Frigg and Reiss disagree with all of these points.

My view, and one that is compatible with most of the papers cited above, is that while there is methodological novelty in computational modeling, it does not require an upheaval of our metaphysics, epistemology, or ideas about theories and experiments. There is philosophical work to be done in clarifying the role of computational models, but this work is continuous with work on models of other kinds, and with work on the epistemology of experiment (even if the current consensus on simulations is that they should not count as experiments).

Computational models can act as models in two ways. First, they are the sort of models that connect empirical findings with theory. How exactly is it that the things we believe to be true about how the world works, like the laws of physics, lead to the phenomena we observe? It is not always clear how to make the connection, and models, including computational ones, fill this role. A typical simulation begins with equations derived from theory, and works out what the phenomenological outcomes should be, given various environmental conditions, codified in the simulation's parameter settings. When these outcomes fit well with empirical knowledge, we gain understanding of how or why those real world phenomena are produced from the theory. The simulation allows us to connect the two.

This is not particular to simulations. The models of rat hippocampus discussed in Section 4.5.1 also worked this way. Fuhs and Touretzky (2006) and Burgess et al. (2007) provided ways of connecting the functional characterization of grid cells and the mysterious

phenomenon of their firing fields forming hexagonal grids, to basic anatomical facts about neurons and dendrites. Connectionist models intended as proofs of concept do something similar: they show how, for example, the developmental dynamics of verb conjugation errors might be connected to and arise from the statistical character of environmental inputs combined with a distributed memory system (Rumelhart and McClelland 1986a).

Another way in which computational models can serve as models is to be a platform on which to perform quasi-experiments. Scale models of buildings can help us decide whether a room will be sunny or a piano will fit through a doorway, and macromodels of molecules can help in the discovery, for example, of the double helix structure of DNA. In these examples complex systems are simplified, re-scaled, and modeled in different materials that are easier to manipulate than the target system. Connectionist models, similarly, simplify brain systems and model them in a different medium that is easier to manipulate. All of these models help us to figure out how things in the world of a particular type might be expected to behave, whether that type be buildings of a particular design, molecules with particular internal forces, or networks with particular topologies.

5.5.1.1 Computational Models as Models of Mechanisms

So far I have argued that computational models are models. To be more specific, the connectionist models I'm interested in are, like the models of rat hippocampus we saw in Section 4.5.1, models of mechanisms.

That episode of modeling work began with unexplained behavioral data: the fact that rats are able to do path integration. The first model by Redish and Touretzky (1997) proposed a sketch of a mechanism that could explain the behavior, and which predicted the existence of grid cells. Grid cells were then found and characterized by physiologists (Fyhn et al. 2004), who filled in some of the details of the mechanism sketch. The later models by Fuhs and Touretzky (2006), Burgess et al. (2007) then proposed more detailed mechanistic models to account for the behavioral and physiological data.

The remaining piece of the story is to argue that these models of mechanisms are approximations to generic mechanisms, and to show how models of this type can explain their target phenomena.

5.5.2 Connectionist Models as Generic Mechanisms

I have already described how several other types of models such as animal models and samples from populations can be understood as approximations to generic mechanisms, and have pointed out that the connectionist models that have so far eluded understanding are models of mechanisms. Here I argue that these can be understood as non-representational models of mechanisms that work in much the same way as the previous examples: by approximating generic mechanisms. Seen this way, the inferences we draw from connectionist models reveal themselves to involve the familiar inferential moves used in other modes of scientific discovery, modeling, and experimentation. Thus they are not *sui generis* insofar as various other types of models can be understood in the same way. Nevertheless, making sense of them does involve some rethinking of how models in general explain.

Earlier I asked what McClelland and Rumelhart (1986) might have meant when they said that "the appeal of PDP models is their obvious 'physiological' flavor." My interpretation is that PDP models approximate the generic physiology of brain networks. This interpretation is consistent with more recent comments by McClelland in the context of a debate about the relative merits of connectionist and structured probabilistic approaches to computational models of cognition. Griffiths et al. (2010), who defend Bayesian models, describe connectionist models as taking "a bottom-up approach, beginning with a characterization of neural mechanisms" (Griffiths et al. 2010, 357). McClelland et al. (2010) deny that they are focused exclusively on working up from neural models, claiming that they "emphasize function, algorithm and implementation equally and seek accounts that span levels" (McClelland et al. 2010, 354). In describing why they consider it an advantage that connectionist models are constrained in what they can do easily, they say, "The constraints arise from a commitment to mechanisms similar to those that implement real minds, thus they provide useful clues as to how real minds solve important cognitive problems" (McClelland et al. 2010, 354). I take mechanisms similar to those that implement real minds to mean approximations to generic mechanisms implemented by real brains.

Brain networks can be structured in various ways in terms of the exact patterns of connections between neurons, the chemical properties of their environments, and so on, but

what is generally true of them is that they consist of numerous units that affect one another through relatively dense axonal connections, and do some form of input integration and connection weight adjustment. Whatever else is true of the brain, these things are also true. To call brains densely interconnected networks of simple units certainly leaves out many details, and these details that are left out certainly make a difference to the behavior of real brains, but it is not a leaving out of details that yields an abstraction or a fiction. It is a leaving out of details that yields real things in the world that are true of brains. Networks of real neurons are instances of graphical models of particular kinds.

Insofar as brains are such networks or graphs, whatever is true of such networks should also be true of brains. If PDP models are investigations of the properties of networks of densely interconnected simple units, then the results of these investigations should tell us about brains, because brains belong to this type. PDP models can act as a sort of mathematical demonstration, where a result is demonstrated for a class of graphical models, and this warrants the conclusion that the result applies to members of this class of graphical models.

It is not quite so simple in practice, of course, because PDP models are not exactly the same as the generic graphical models that brains instantiate. PDP models also involve some idealizations for the sake of making them easier to implement and run. Actual PDP models are only approximations to the generic mechanisms that are instantiated in brains. The approximations involved are chosen such that they should not significantly interfere with the drawing of inferences from PDP models to brains when the work is done well.

Idealizations like the backpropagation algorithm and the standard, fully-connected, 3-layer architecture are perhaps not true of any real networks of neurons. Backpropagation was the first algorithm developed for updating weights in a multi-layer network that was guaranteed to converge. Because no other algorithms were available, it was used in many early connectionist models, despite its neural implausibility.[3] For the purposes to which models employing idealizations like backpropagation are put, their falsehood should be irrelevant. Backpropagation is not used in models of how reinforcement learning is done, for example,

[3]It is still used, despite the existence of more plausible alternatives, because of its wide availability in software packages, and its relative mathematical simplicity.

but rather in models of how networks might learn to do some other task, assuming some form of reinforcement is present. An example like this is NETtalk (Sejnowski and Rosenberg 1986), where the particular algorithm used for feedback learning does not make much difference, and the point is to show that a system capable of pronouncing English words needn't encode a complicated set of rules.

In contrast, in models focused on investigating reinforcement learning itself, such as Suri and Schultz (2001), the anatomy of the basal ganglia is modeled in detail, and only pathways that exist in the brain and through which feedback is thought to actually travel are included in the model (i.e., idealizations like backpropagation are not used). When reinforcement learning is the target of investigation, how realistic the reinforcement methods are is relevant, so grossly implausible methods of reinforcement would not be appropriate. Other aspects of that model might, however, be simplified in order to make it easier to run. What counts as an acceptable idealization depends, of course, on the details of the model and its intended purpose,.

The alternative to making appropriate idealizations would be to try to make one's models as realistic as possible in every way. This, however, is what Touretzky referred to as a novice mistake and Cowan called mindless. Computational modeling is a tricky endeavor. If the details of one part aren't quite right, that can make the entire model fail, and the cause of the failure is harder to track down the more details there are. Setting all of the details that are not under direct investigation to the most neutral and harmless values makes the job not just easier, but feasible. This is similar to the challenge faced in running a controlled experiment. One always has to trade off external validity (how well the experimental condition matches what happens in the world) with control. Too many realistic details make an experiment impossible to control and therefore interpret. Holding as many variables as possible constant allows the independent variable's influence to show clearly. Likewise, idealizations are beneficial in computational modeling.

It is, nevertheless, quite common to think of connectionist models in terms of representation. The representational features of connectionist networks have been exhaustively discussed. Just a few of the sub-topics within that discussion are: whether nodes represent individual neurons or populations of neurons; whether assigning concepts to nodes in ad-

vance makes the network an implementation of a classical architecture; what can be achieved with distributed representations; whether connectionist networks can realize structured representations; and whether post-hoc analyses of networks show that they have merely learned to implement classical theories.

Despite this obsession with representation in discussions of connectionism, I agree with Ramsey (1997) that the representational features of connectionist models have been exaggerated. He says,

> in an important area of connectionist theory where internal representations are constantly invoked, the notion is actually doing no explanatory work whatsoever, and that proponents of connectionism have failed to provide adequate motivation for treating features of their networks as representations (Ramsey 1997, 35).

It is understandable that connectionists would go to such lengths to justify their models as being capable of various kinds of representation, even if this turns out to be mostly false and unnecessary, given the attack by Fodor and Pylyshyn (1988). It is also true that some models were built with the sole purpose of demonstrating the representational capacities of connectionist networks, so this is not to say that connectionist networks cannot represent. Typical connectionist models just have a different aim than representation: that of contributing to explanations of psychological phenomena; and the way they explain psychological phenomena has little to do with representation.

The past-tense learner, for example, aims to show that the way children learn to conjugate verbs could be explained in a different way than previously believed. The standard account of having two separate systems for regular and irregular verbs is one possible explanation of how children learn both the rules and the exceptions, although some of the dynamics of the learning process are difficult to square with that account. The past-tense learner provided an alternative explanation by showing that a network that learns associations, paired with a training set that changes over time, could also bring about the result to be explained, and even does a better job with the dynamics. The past-tense learner's ability to act as an explanation of how children learn to conjugate verbs is not due to the network representing the parts of the brain involved. Rather, the past-tense learner is a candidate explanation of that behavior because it causes the same pattern of behavior, because it does so in virtue of being a network that approximates a generic network that is also instantiated in the brain,

and because the brain instantiation of that generic actually causes the behavior in the cases to be explained.

Dror and Gallogly (1999) have argued for a similar point: that biologically implausible models are useful and necessary in cognitive neuroscience. Some of the ways such models can be useful are in characterizing the problem space, analyzing the complexity of computational problems, identifying the benefits of particular constraints and modeling choices, and identifying tradeoffs between competing constraints. All of these can be done with models that are either biologically implausible, or that, as Dror and Gallogly (1999) say, transcend the biology. One of the examples they cite is Tsotsos's (1990) work on the complexity of visual search problems, which established that biological vision systems could not be solving the computationally intractable problem of exhaustive search, and pointed the way to more realistic solutions. Another of the examples is McClelland et al.'s (1995) analysis of memory and learning, which "revealed a computational tradeoff between rapid learning and slow consolidation" (Dror and Gallogly 1999), showing that the hippocampus and neocortex are complementary rather than redundant memory systems.

My point is a slightly different but compatible one. Dror and Gallogly (1999) argue for the value of implausibility or 'aplausibility' in computational models. I argue for there being a kind of generic plausibility to models that may on the surface seem implausible. McClelland et al. (1995) describe their model not in terms of implausibility but in terms of detail and abstraction. They say, "These are not detailed neural models; rather, they illustrate, at an abstract level, what we take [memory] consolidation to be about" (McClelland et al. 1995, 420). A model being less detailed needn't mean that it is implausible, nor even 'aplausible' as Dror and Gallogly suggest. What I think McClelland, here and elsewhere, means by connectionist models of this type being abstract is not that they are just ideas without physical existence, but that they implement the sort of generic mechanisms I introduced in Chapter 3. They are real, i.e., physical, and plausible despite not being highly detailed.

Such models can tell us about real brain systems because they are networks of the same type, with the same architectural constraints as real brain systems. When connectionist models explain, it's in virtue of the model being the same sort of thing as what causes the phenomenon to be explained in the target system. A physical model can tell us something

about how physical things work, because it is one. Likewise a network model can tell us about how networks work, because it is one. As generic mechanisms, connectionist models demonstrate what is true of things belonging to an abstract type. The types are chosen to include the neural structures of interest, so that inferences can be drawn back to those neural structures. There is a difference between a how-possibly model that puts forward a concrete suggestion that may turn out to be the wrong suggestion, and a how-generically model that puts forward a suggestion of what type of thing must be involved, where this isn't mere speculation. It may be that you can explain something more or less general with a model that has some generic parts, whereas you'd need several more specific models for each of the cases if it were required that explanations be fully specified. An explanation being more general is a good thing, so when we can get them, we shouldn't throw them out.

There is a sense in which it is true that connectionist modeling uses a bottom-up strategy, but not in the sense of starting from the details of brain anatomy and physiology, as sometimes assumed. Instead, the goal seems to be to find out how little detail can be included in a nevertheless biologically plausible model while still accurately simulating the behavior. For that sort of project, it makes sense to begin with very general models with few constraints, and to add more details only as needed. It also makes sense to do exploratory work where the anatomical constraints that we know to be true of brains are built into models, to see what sorts of phenomena naturally result from models structured in those ways.

Another strategy is to try out generic mechanisms that are known to produce similar results, then to fit them to the details of the model. This is essentially Darden's (2002) method of schema instantiation. In the rat hippocampus example, close packing of neurons was one plausible explanation for hexagonally spaced grids; it is a general geometric fact that hexagonal spacing has the highest packing density for circles and spheres, so this holds for a wide variety of materials, plausibly including axons and dendrites. Another plausible explanation for the hexagonal spacing was as an interference pattern of oscillators at three different angles. That hexagonal patterns result from the intersection of three sets of lines at 60° angles is also a very general phenomenon that can be seen in applications as far removed from rat hippocampus as basket-weaving. These models also instantiated the generic mechanisms of spin glasses and cognitive maps. These generic mechanisms are used in the

rat hippocampus models, not just for the sake of aiding understanding by referring to similar known effects, or in an attempt to be clever, but because these same generic mechanisms might actually be responsible for bringing about the result in this case. Fuhs and Touretzky (2006) and Burgess et al. (2007) are doing neuroscience, not engineering.

Little has been said about how to actually go about instantiating schemas. Clearly when a known mechanisms is applied in a novel way, many of the details will have to be different, and these differences may well change the behavior of the mechanism in important ways. This makes it more difficult to know how the mechanism will work, but is essential to generic mechanisms having multiple applications.

Whether using the bottom-up strategy or the schema instantiation strategy, what is important for making a model suitable for its purpose is not how much detail it includes overall, but whether it includes the right amount of detail about the right aspects of the model. Suri and Schultz (2001) for example, include realistic details about the basal ganglia, because they are modeling reinforcement learning, but they do not model the rest of the brain in as much detail. There is nothing special about Suri and Schultz (2001) in this regard. It is extremely common for computational models to go into much more detail about some features than others. A connectionist model is typically neither purely generic nor purely detailed, but rather a hybrid of the two. In the remaining section I argue that this hybrid nature is often essential to their explanatory power.

5.5.2.1 The Explanatory Power of Hybrid Models

That typical connectionist models are partly generic and partly detailed is neither a failure on the part of modelers to get all the details right, nor evidence of indecision. Including selected details in an overall generic model serves a purpose. Although this purpose is rarely if ever explicitly articulated by modelers, my reconstruction of what connectionist modelers are up to is that they build hybrid generic-specific models in order to include all and only the details that they have reason to believe may be essential for producing the phenomenon of interest. In other words, the point is to find the real difference-makers, whether these be generic or specific properties of the system. The hunch that motivates the connectionist project is that a lot of what goes on in the brain can be explained by fairly simple, general constraints. By making a model

that is purely an instance of a generic, one can find out how that generic contributes to the system's behavior. By adding some specific details too, one can find out how that generic behaves given some additional constraints.

Craver (2006) defended explanations that are as detailed as possible. Much of 20th century philosophy of science defended explanations that are as general as possible. What I want to defend here is that the best explanations, at least in cognitive neuroscience, but likely in much of biology and the human sciences more generally, are often neither as detailed nor as general as possible, but instead a mixture of both. One of the main virtues that I see in the new mechanist account of explanation is that it can accommodate this sort of hybrid explanation (perhaps in spite of some of its proponents' views).

I argued in Chapter 3 that the sort of causal explanation mechanists are after need not pick out highly detailed causes in order to pick out causes that are the ones actually operative. Generic mechanisms are every bit as real and causally relevant as detailed mechanisms. There I argued for generic mechanistic explanations being just as good as more detailed mechanistic explanations. Given the tools developed in that chapter, hybrid generic-specific mechanisms can also be defended.

It has already been noted that mechanism schemas rather than fully-elaborated mechanisms are typically what scientists are after. Mechanism schemas, furthermore, may be, and perhaps usually are, unevenly elaborated. That is, they give more details about some entities and activities than others. Which details are included depends on what needs to be explained; all and only (or most and mostly) those deemed necessary for explaining the target phenomenon or the system's behavior in a particular context are included. So much should be uncontroversial. If an explanation^{e} is doing its job, the details represented in the schema should be the ones that make a difference in bringing about the phenomenon. An ontic analogue to this sort of schema can also be constructed.

The mechanism in the world can be a hybrid of more generic and more specific components. There is no reason why some of the generic entities in a generic mechanism could not be substituted with more specific ones to make a hybrid. Likewise, generic activities can be substituted with more specific ones, assuming the entities in the mechanism are specific enough for the more specific activities to apply. For example, some people can stop

(generic) on roller skates, by doing any of a hockey stop, a derby stop, or a plow (specifics). Other mechanisms might be able to perform the generic activity of stopping, but not some of the more specific ones, like a skater whose ankle injury precludes them doing hockey stops. In that case, one of the specific activities can't be substituted for the generic activity. Although there are these cases where one level of activity corresponds to the entities and another doesn't, there is little reason to suppose that there is any general fact of the matter about how the levels of specificity across various entities and activities correspond. There are just facts of the matter about what each kind of thing can and can't do (or does and doesn't do).

In the examples covered, we have seen several mechanistic explanations that combine generic and specific elements. Retinal ganglion cells use the generic lateral inhibition mechanism, which we know enhances contrasts. To know the resolution at which edge detection works best, and how quickly it operates, some specifics of the mechanism are needed, like the ratio of inhibitory to excitatory connections, and their relative synaptic strengths. Piano shaped objects generally fit through apertures of a given relative size, but if they are magnetized, the specific details might change that behavior. Bicycle wheels stay upright because of the generic mechanism of preservation of angular momentum, but to know how fast a rider needs to be going and how far they can lean into a turn in order not to fall over, specifics like the wheel's diameter, the rider's weight, and whether the road is full of gravel play a role.

The way generic explanations are adapted to the specific contexts in which they operate is by the addition of some specific details to the mechanism. That this move of specifying some details of generic mechanisms sometimes helps does not mean that more detailed mechanisms are better explanations in general. Enough detail is needed to not get the phenomenon wrong, but too much detail obscures the contribution of generics, which are among the real difference-makers.

5.5.2.2 Connectionist Models as Hybrid Mechanisms In Chapter 4 I raised the problem of why it should be that connectionist models are loosely, but only loosely, based on brain anatomy and physiology, and how it can be that this not quite realistic detail is

helpful for modeling cognitive systems. My claim, in short, is that generic aspects of brain anatomy and physiology partially explain some cognitive phenomena, but specific aspects of the anatomy and physiology explain other aspects. For a complete explanation, there will usually be causes acting at multiple levels which need to be accounted for. Things in the world can serve as explanations in that they are the causes of specific events or outcomes. Causal explanation works naturally for particular instances of phenomena, but less well for general patterns. A kind of explanation that works nicely for general patterns but less well for particular instances are laws or proofs. A proof being valid explains why any circumstances that satisfy the premises are also accompanied by the conclusion. But often what's taken advantage of in creating some effect is not just a general rule, and also not just a particular pushing or pulling, but some combination of the two.

If a model is to stand in for the target system, the model needs to embody the same mixture of generic and specific causes that bring about the phenomenon in the target system. This is what connectionist models do, and this is why they partially, but only partially, mimic neural structures. These models seek to give mechanistic explanations of cognition by embodying both the generic properties that partially explain the phenomena, and the more specific properties that do the rest of the explanatory work. Highly detailed models without the generic aspects would not provide explanations that are as good, because they would fail to pick up on the generic mechanisms that are among the real difference-makers.

One of the overlooked virtues of mechanistic explanations is that they provide a way of combining generic and specific aspects of explanations. Explanations can have an intermediate status between laws and causal chains. Computational models, as generic mechanisms, have an intermediate status between abstractions and things in the world. They are both general and in the world. This helps to make sense of how connectionist models explain, not despite, but in virtue of their intermediate status.

Although it is intuitively plausible that rough similarity to neurons should be a virtue when modeling a cognitive system, defenders of connectionism haven't quite articulated why this should be, or why more detail wouldn't be better. So why then does rough similarity to neural systems matter? Critics complain about the backpropagation algorithm not having neural correlates, but they don't complain about the models being made of silicon, plastic,

wire, etc., instead of nervous tissue. In order to serve as an explanation, the structure has to be the same, so that general results that depend only on that structure apply equally to target systems with that structure (unless other physical details have an opposing effect). A number of mechanists claim that constraints should come from both above and below, without giving any account of how constraints from above are to be incorporated. Recognizing the hybrid nature of typical connectionist models, as well as the need for mechanistic explanations to combine generic and specific difference-makers, points the way to how constraints from above and below can be combined, and provides an answer to the puzzle of connectionist models' 'physiological' flavor. Constraints from above and below can be combined in a mechanistic model as generic and specific components of the mechanism.

Equipping computational models with the same constraints as apply to real brains makes these models better at explaining cognition because it ensures that the model belongs to the same abstract types as do brains, which means the same causes will be at play. This warrants the use of connectionist models for making discoveries about cognition. Generalizations can be made from connectionist models to systems with the same constraints, i.e., systems that belong to the same abstract types. Connectionist models can thus produce knowledge about the generic mechanisms they instantiate, and how systems with those structures behave. Rather than providing how-possibly theories, connectionist models provide theories of how-possibly-given-neuroscientific-constraints.

5.6 CONCLUSION

In this chapter I continued my examination of the role of connectionist models in cognitive neuroscience, with the aim of finally resolving the puzzle of why models that are only vaguely neurally plausible should be helpful in explaining cognition. I began by taking a wider view, considering models in general, and analyzing the roles they may play in explanation. I provided an alternative to representational accounts. While it is usually possible to consider any given model in terms of representations, they need not be seen in this way. Instead they can often be considered as physically realized stand-ins for their target systems. I showed

how several kinds of models, including model organisms, samples from populations, and scale models can be re-understood as instantiating approximations to generic mechanisms. I outlined how inferences from such models to their target systems might work, and showed that these inferences follow familiar reasoning patterns.

Models thus can serve as explanations in the sense of being instances of the same generic mechanisms that cause the phenomenon of interest in the target system. Models can not only explain, they can also be used to discover how real-world mechanisms behave. Computational models, being models, can likewise serve as explanations for their target systems and as tools for discovery. Connectionist models in particular serve as instances of the generic mechanisms that brains also instantiate, and thereby can inform us about brains and the phenomena they produce, namely cognition.

I argued further that in typical cases, combinations of mechanisms at multiple levels of specificity are necessary for bringing about the phenomena of interest. Hybrid generic-specific mechanistic models can be used in such cases to account for these combinations of causes. The puzzling mixture of some biologically plausible details with general structural constraints characteristic of connectionist models can be understood as reflecting the hybrid nature of the mechanistic explanations being sought. Connectionist models are hybrids of generic and specific elements, because the causes at work in giving rise to cognition are likewise mixed. The explanatory power of connectionist models is not compromised by their only having a 'physiological' flavor; rather this is where their explanatory power comes from.

6.0 CONCLUSION

I began with the problem of how the subject matter and the theoretical apparatus of cognitive psychology and neuroscience are related. Cognitive neuroscience attempts to integrate these two fields, and takes itself to be gradually succeeding at this task. My main goals have been to evaluate what form this integration might take given the resources available, and to develop additional theoretical resources where these are lacking. The main development was the notion of a generic mechanism, which is a thing in the world insofar as it belongs to an abstract type. This means that generic mechanisms can simultaneously have causal powers, and be less specific and detailed than the mechanisms neuroscientists usually look for. Working out how these more generic mechanisms can work together with more specific ones is a promising direction for integration, because it would solve the problem of how higher level causes can co-exist with lower level ones, how generalization can be accommodated within mechanistic explanation, and how the norm of explanatory relevance can be maintained even on the ontic side of explanation. What remains to be done is to apply this back to the question of integration in cognitive neuroscience. Before tackling that task, I'll first summarize what has come so far.

6.1 SUMMARY

6.1.1 Introduction

In Chapter 1 I began by discussing the available approaches to studying the relationship between cognition and the brain. These include neuropsychology's lesion studies, behav-

ioral studies, classical and connectionist approaches to AI, electrophysiology, functional neuroimaging, and ERP approaches. I then discussed possible ways of integrating fields. The kind of integration I'm after is one where both psychology and neuroscience maintain some form of autonomy, but both fields take constraints from the other into account. The autonomy could take the form of each field incorporating constraints from the other based on their own theoretical needs, and maintaining explanatory authority over the phenomena that they traditionally cover. I noted that this sort of autonomy is in no way contradicted by Marr's oft quoted comments in his 1982 book.

I then briefly reviewed several versions of reduction, emergence and unity. Emergentism as the metaphysical claim that entities, properties, or regularities show up at various scales in terms of size or complexity, is one of my assumptions. The claim that emergents cannot be explained in terms of lower levels would contradict the goals of many cognitive neuroscientists. If at least some emergents on the first definition can be explained in terms of lower levels, then there is room for both the second kind of emergence and integration. It was decided that both Kitcher and Craver's notions of unity make assumptions that go beyond the goal of integration.

One recent proposal for integration that has the potential to meet the constraints I set out on the kind of integration we should be looking for is that psychology and neuroscience both employ mechanistic explanations, and that the mechanisms referred to by each might be located in a joint hierarchy. Proposals along these lines have been made by Bechtel (2008), Piccinini and Craver (2011), and Kaplan (2011).

6.1.2 Explanation in Cognitive Psychology and Neuroscience

In Chapter 2 I discussed in detail Piccinini and Craver's proposal for how mechanistic explanation can provide a scaffolding for the integration of cognitive psychology and neuroscience. I began by characterizing the two fields. Neuroscientists seem to want to explain psychological phenomena, but are not necessarily very willing to make changes to their own concepts or practices to make it happen. Piccinini and Craver argue that psychology's explanations are elliptical mechanism sketches.

My analysis of the types of explanations used in cognitive psychology turned up two distinct approaches, both of which are commonly referred to in terms of information processing. I gave a brief historical review of how information-processing metaphors in psychology developed. The point where information processing went from being about the route taken by data through wires and other transmission devices to being about data processing by computer programs was in Neisser (1967). I distinguished data-flow models from process-flow models, and compared these to the sorts of information-processing models found in computer science. This involved a slight detour into the history of flowcharts and their analysis. I characterized types of information-processing models with the help of the ISO/IEC/IEEE standards for diagrams. I then compared these types of information-processing models and the diagrams that represent them to mechanism sketches, and the diagrams that represent them. I conclude that both types of information-processing models are significantly different in terms of the sorts of things they include than mechanism sketches, and that it is not entirely clear how one could turn one kind of model into the other without adding and subtracting a lot of information. This mismatch calls into question the claim that psychology's models are elliptical mechanism sketches.

In a second line of argument, I looked at how explanatory models are developed in psychology and neuroscience. I compared two models—one of attention in psychology, and one of sodium channel gating in neuroscience—each of which is referred to as a filter. The neuroscience example illustrates what Machamer et al. (2000) described as a mechanism sketch, and its gradual filling-in to become more like a full-fledged mechanistic explanation. The psychology example shows that the ways theories are developed in psychology is not, as Piccinini and Craver suggest, a gradual accumulation of mechanistic details, but rather the refinement of a model that begins as an information-processing model and remains one throughout, never getting filled in with mechanistic details. The concern of psychologists in refining these theories was to get the relationships between the data-processing stages correct, not to figure out what entities and activities might possibly, plausibly, or actually implement those stages.

What cognitive neuroscientist sometimes try to do with this sort of model is to press it into service as a mechanism sketch, even though this is not what it was intended to

be. I gave the example of Schneider and Chein (2003) who associate a model of attention descended from Broadbent's with brain regions using fMRI. I then argued that this sort of crude mapping from cognitive model to brain map would only constitute a mechanistic explanation if a number of other intermediate steps were first accomplished. These essentially amount to the work of building a mechanistic model in the first place.

I then discussed another test case of this sort of attempted transformation of a cognitive model into a mechanistic one, which is discussed in Bechtel (2008). I argued that the example of memory research shows how ill suited functional models of cognitive phenomena are for being pressed into service as hypothesized mechanism sketches. In this case, just about all of the functional divisions and hypothesized processes turned out not to fit well with the neural picture, and even hindered progress in the neuroscience of memory, rather than serving as a guiding mechanism sketch. The current state of affairs seems to be that psychologists have hung on to their taxonomy of memory types and processes despite these being demonstrably false divisions from the point of view of neuroscience.

From my analysis of psychology's information-processing models, I concluded that Piccinini and Craver's proposal for integration looks too simple. Neither the structure of cognitive models nor the patterns of their development fit the picture of mechanism sketches to be filled with details about entities and activities. Furthermore, the case study of memory research which Bechtel holds up as a prime example of integration fostered by the decomposition and localization heuristic looked instead like a case where a cognitive model not only hindered neuroscientific work, but retained its status as the standard cognitive model later on, despite its lack of fit with the neuroscientific evidence.

6.1.3 Cognitive Mechanisms

In Chapter 3 I worked towards an alternative proposal for how integration in cognitive neuroscience might be achieved. The main goal of the chapter was to develop resources from within neo-mechanistic explanation that allow cognitive models to be seen as a species of mechanistic model, rather than sketches of such. That sort of scenario would afford the possibility of integration while taking cognitive models seriously.

I began by giving an overview of the MDC account of mechanism, with a focus on ways in which higher level mechanisms might be included. I then went on to discuss more recent extensions of the MDC account by its authors. Machamer eventually endorsed Bogen's suggestion that regularity be dropped from the definition of mechanism, suggested that there might be activities that can be abstracted from entities, and seemed to argue that the hierarchies of mechanisms relevant to different sciences might not all match up neatly. Darden disagreed with the proposal to drop regularity, and emphasized the importance of mechanisms being of general applicability. She expanded the use of schema beyond just being a success term for sketches, to include also general templates for mechanisms that can be instantiated in many contexts by analogy. Darden also talks about mechanistic explanations that require looking up in size rather than decomposition into parts, and claims that investigating the gory details does not always help to understand phenomena. Craver seems to agree that explanations sometimes look up in size, but nevertheless argues that explanations must identify mechanisms in detail. Abstract models and generalizations are not explanatory, according to Craver. His historical analysis of the Hodgkin-Huxley model shows how having an equation describing a phenomenon adequately might not be explanatory unless accompanied by a description of the mechanism involved.

Bechtel argues that mechanisms and mechanistic explanations are representations, in opposition to the ontic view expressed by Craver. I diagnosed this disagreement as stemming from the double use of 'explanation' in everyday language, and from the vehicle and content of explanations having been one and the same sets of sentences in the DN account of explanation. Now that the vehicle and content are split, it is difficult to decide which was the most important part in the first place. It seems to depend on which sorts of questions one is interested in.

Next I clarified the role of schemas in mechanistic explanation. One use for the term is as a success term for sketches. Another use for schemas are as templates of types of mechanisms. Darden's notion of schema instantiation, which she describes as a strategy for mechanism discovery, invokes schemas as types of mechanisms. Several recent papers have taken up this confusion over what a schema is and offered suggestions as to how to resolve it. I argued that it can be resolved without resorting to major changes to the MDC definition

of mechanism, such as Overton's suggestion that mechanisms be considered representations rather than things acting in the world.

The problem of regularity has inspired much commentary. I reviewed Bogen's arguments for dropping regularity from the definition of a mechanism, and the replies by Andersen and DesAutels. I agree with Andersen that infrequently working mechanisms pose no problem to maintaining some notion of regularity. However, I agree with Bogen that a causal chain occurring only once is no good reason for denying that it is a mechanism. My suggestion is that what makes the difference between one-off causal chains that are mechanisms and ones that aren't, is whether they are mere series of accidents, or whether the interactions in the causal chain follow patterns that connect them to scientific theories. Given later developments in Chapter 3, this can be elaborated in terms of whether they instantiate generic mechanisms of wider applicability. This disagreement seems to hang on whether mechanisms are something more than or distinct from episodes of causation. I maintain that they are not synonymous.

I argued further that mechanisms are not always or exclusively causal in the sense of causal chains. Causation in the sense of corollaries of general facts (which might be called structural causes) are also involved in mechanisms operating the way they do. Examples are mathematical facts or laws of physics that make things so without providing any impetus that brings them about. I argued that invoking generalizations, like these mathematical facts and laws of physics, can make for better explanations, even of particulars, in cases where they adequately account for the phenomenon. That we should prefer the most general cause that adequately accounts for a phenomenon, and not include details that are more specific than necessary I called the norm of explanatory relevance.

On the problem of whether mechanisms are types or tokens, I endorsed Kuorikoski's claim that in addition to the sort of causal chain mechanisms that at least Craver and Bogen seem to focus on, there is a notion of a mechanism as abstract type that is important in the social sciences. The idea is familiar from Darden's strategy of schema instantiation, but it is one that requires more elaboration. How the MDC account can include both type mechanisms and token mechanisms required clarification.

I then offered an outline of a solution to the problems of how to integrate information-

processing accounts with neural mechanisms, how to combine type and token mechanisms into one account, how to allow for a limited kind of regularity that takes into account Bogen's concerns, and how to include generalizations in mechanistic explanations. I introduced the idea of a generic mechanism as an ontic counterpart to mechanism schemas, using lateral inhibition as an illustration. A generic mechanism, like any other mechanism, is a thing in the world, but a thing in the world insofar as it belongs to a type, as opposed to a thing in the world in all its gory detail. I suggested that generic mechanisms might be a good way to understand how causal-explanatory power can be had by the sorts of cognitive models that are and remain quite abstract. I also argue that integration might take the form of multiple hierarchies of mechanisms that do not necessarily all connect together into one unified hierarchy. This allows for the possibility that some cognitive models might be neatly identified with neural components, while others might not match neatly onto the same decompositions.

6.1.4 Computational Cognitive Neuroscience

In Chapter 4 I looked in more detail at one of cognitive neuroscience's most important methodological tools, computational models. The motivations for this were several: these methods are poorly understood, and can benefit from further analysis independently of the problem of integration; computational models are one of the key methods for achieving integrated models, since they help tie together data about behavioral and neural phenomena; and they provide a good illustration of how generic mechanisms are discovered and used in explanations.

I first described how computational models are typically used in cognitive neuroscience. A very common method is to vaguely base the model's structure on selected biological details, then attempt to simulate the behavioral results from tests on humans. I compared this to computational modeling techniques in other fields like physics, where there has been much more philosophical analysis. I then moved on to describe several approaches to AI, which also attempts to model cognition with computers, but from the vantage point of cognitive science.

I pulled out from discussions of AI four distinct suggestions about the role of computational models in cognitive science, which raise several puzzles. It was not clear why selectively including biological details can help in modeling when these details are not entirely realistic. It was also not clear how highly simplified and idealized models can tell us about real-world systems. Connectionist approaches to AI in particular have been attacked on both grounds: Fodor and Pylyshyn (1988) charged that connectionist models either fail to be relevant to cognitive theories, or else merely implement them without adding anything of interest. In contrast, the PDP Group claimed that their models having a 'physiological' flavor was one of their selling points, and suggested that physiologically plausible models could help generate theories of cognition that better account for detailed behavioral data. At the same time, connectionists regularly tout the virtues of simplified models.

I argued that concomitant with the shift from classical to connectionist AI was a shift from the semantic view of theories to a yet unstated mechanistic view of explanation. Seeing connectionist models as tools that can be employed for several purposes in the discovery and exploration of mechanisms, helps to make better sense of the PDP approach. I illustrated this with examples from research on memory and navigation in rats, where a series of connectionist models were used for multiple epistemic purposes. Some of these purposes demand more neural detail, some less.

6.1.5 Computational Models as Models of Mechanisms

In Chapter 5 I brought the resources I developed in Chapter 3 to bear on the problem of how partly realistic, partly simplified models like connectionist ones can explain cognitive phenomena. I began by stepping back to consider models in general. The contemporary view of models is that they play many roles in science, ranging from those previously reserved for theories to those previously reserved for experiments. I argued that seeing models as things in the world that stand in for their target systems is more fruitful than seeing models as acting as representations. To back up this claim, I ran through a number of examples of models that might typically be thought of as representing their targets, and showed how these can be alternatively understood as approximations to generic mechanisms, which

act as physical representa*tives* of their targets. I showed that the inferences involved in working with this sort of model are all standard inferential strategies that are ubiquitous in experimental science.

I then argued that computational models, and in particular connectionist models operate in the same way: as approximations to the generic mechanisms that are instantiated by their target systems. With this claim established, it becomes easy to see why these models might typically include mixtures of generic and specific components. Part of what makes a difference in the operation of a given mechanism are its generic properties; but usually in complex sciences like biology and neuroscience, some of the specific properties of the mechanism also make a difference to its operation. Constructing mechanistic models that are mixtures of generic and specific components poses no serious problems. I argued that this provides the final resolution to the puzzle about why connectionist models are typically hybrids involving some realistic neural details, combined with some generic structural basics.

6.2 PROSPECTS FOR INTEGRATION

I injected into the summary a number of comments about how my arguments in the previous chapters relate back to the problem of how to integrate the cognitive theories of psychology with the biological details of neuroscience. The previous chapters developed the tools necessary for at least partially resolving this problem. The main thing missing from previous accounts of integration was a way for cognitive level explanations to play a more substantive role than that of being reduced and replaced. The notion of a generic mechanism that I introduced in Chapter 3 provides a way for mechanistic explanations to look upwards to higher level organizational features of systems, and to include the causal contributions of these in explanations. This is at least one of the requirements for an integration that treats psychology as being as valid a source of constraints or models as neuroscience.

I argued that typical mechanistic explanations in cognitive neuroscience would be hybrid models that include components of various levels of genericness/specificity. How specific the components must be (and which ones) depends on the explanatory context. Mechanistic

models are flexible enough that more or less generic/specific components can usually be substituted depending on what is to be explained. Connectionist models like the examples taken from work on navigation in rats illustrate that quite generic mechanisms like spin glasses or interference patterns may be combined with specific mechanisms like neural firing oscillations in hippocampal place cells in order to construct mechanistic models that are multilevel in quite a different way than that described by Craver (2007). These mechanisms incorporate causally-relevant difference-makers (i.e., entities and activities) that operate at multiple scales into single mechanistic explanations. Craver's picture of multi-level integration, in contrast, features multiple levels of mechanistic explanations, where each mechanism is located at just one level. (Although there is no fact of the matter as to which level a given component belongs to in general. It depends on the mechanism.) Moving upwards or downwards for Craver means looking at wholes in which the mechanism is a part, or parts which are themselves mechanisms.

The picture of integration that I envision is one where a given system might instantiate any number of more or less generic mechanisms of varied specificity. Each of these, if they make a difference to the phenomenon to be explained (for different explanatory questions, different ones will), has associated with it certain causal powers. All of these causal powers apply in the system, possibly with interferences arising between some of them. There may not be any easy formula for figuring out what the sum of these causes will be, but that is not a problem particular to this picture. How causes that are all at the same level interact is just as difficult a problem. Part of the work of elaborating a model is resolving this problem.

For example, the roundness of a bunch of axons might make them pack closely together into a hexagonal pattern. This hexagonal pattern might coordinate oscillations in the firing pattern of the whole collection. The myelin sheath covering particular axons might make them act as insulated wires. The concentration of the collective of ions in extracellular space might make ion channels in a particular axon open and close in a particular fashion, and so on. A mechanistic explanation of that axon's firing behaviour would then depend not only on the gory details, like it having a myelin sheath, but also on the activities of collectives to which it belongs, like the bunch of axons whose oscillatory firing patterns are the result of their higher level properties, and also on collective properties of other entities,

like the ions in extracellular space. The concentration of ions (a higher level property) is the relevant difference-maker, not which particular ions happen to be nearest to a given ion channel, although that might make some very small difference too. This example is a purely neuroscientific one, but it illustrates how even there, mechanisms of different specificity may all work together to bring about a phenomenon.

In an example that reaches higher up, a child who lives on a farm might spend a lot of time outside as a result. Being outside involves seeing objects at all distances, from very close, right out to the horizon. Compared to a child who spends all day indoors, this is a much greater variation in distances. The greater variation in distances seen, would cause the farm child's eyes to focus at a wider variety of distances, including a greater proportion of far distances (since the indoor child mainly sees things only as far away as the walls). Years of exercising their eye muscles to focus on a greater variety of more distant objects might cause the farm child's eyes to develop micro-scale differences relevant to muscle elasticity and strength that cause them to have better distance vision as an adult than the indoor child.[1] This phenomenon could be explained just in terms of the massive sum of details about the child's eye muscles over the years, as Craver would have to insist, but surely their having exercised their eyes looking at more distant objects more of the time is a better explanation. Once the connection between spending more time outdoors and more eye exercise focusing on distant objects is established, one might even prefer the explanation to just be that the child spent a lot of time outdoors, or even that they lived on a farm. To explain why any particular child has the strength of vision they do would require some specific details about their particular eye muscles, but this does not warrant throwing out their having spent a lot of time outdoors as a relevant difference-maker. Just as in the lateral inhibition case described earlier, both the higher level cause and some of the specific details need to be combined to yield the best explanation. This is roughly how integrating cognitive and neural models might work.

[1]The correlation between time spent outdoors and adult distance vision is reported in Aamodt and Wang (2011). The authors couldn't think of any explanation for it. That indoor objects can only be as distant as the walls seems like an obvious explanation to me, but this is just speculation.

6.3 FURTHER DIRECTIONS

There are a lot of things that I did not have the space, the time, or the knowledge to do here. I will mention a few of the gaps that I'm aware of. There are surely many others.

The brief treatment of methodology that I gave in the introduction bears expanding on. As I argued in Chapter 2, the simplest neuroimaging-based approaches to integration do not provide adequate evidence in isolation. There is much more that could be said about what can and can't be inferred from fMRI subtraction studies. Posner and colleagues started using this method in the early of days of cognitive neuroscience in the hopes that it might prove a fruitful method. Although there have been many successful cases of this type, the growing consensus is that this method is quite limited. Neuroimaging technology and methodology have matured significantly in recent years. Reviewing these recent developments, such as Friston (2005), Buzsáki (2006), Sporns (2011) and seeing how integration fares with more complex neuroimaging methods would be well worth exploring. There is also more to be said about electrophysiology, the use of model organisms, and genetic techniques in cognitive neuroscience.

As was probably evident at several points, there are connections to be developed between my account of mechanistic explanation and interventionist accounts of causation like Woodward's (2005). I referred to generic mechanisms as difference-makers at several points, which is Woodward's terminology, and the decision to count causes^{gf} as causes at all was in recognition of their being so on that account. The exact relationship between generic mechanisms and causes bears further elaboration.

I treated the disagreement between ontic and epistemic accounts of explanation only very briefly, and helped myself to a bit of both. There is ongoing discussion between Craver, Bechtel, and others on this topic, which would also be very relevant to this project. There is also a literature on ontic versus epistemic structural realism which might be relevant. Structural realism is another elephant in the metaphorical room. There may be connections between that topic and generic mechanisms.

My review of arguments for and against representational views of models was also very brief. There is a large literature on this topic, as well as on representation in general that

I have barely begun to explore. Commenters on a recent talk I gave about Chapter 5 suggested that Goodman on denotation and Manders on responsiveness suppression might be good places to look.

I had also meant to write more about the relationship between models and experiment. My arguments in Chapter 5 could be extended into an argument that models are on equal ground with experiments in terms of their ability to discover novel empirical facts. I made a few suggestions along these lines, but did not argue for it properly.

There are also a number of philosophers of science who have written on closely-related topics who I have not read carefully enough to do justice to here. There are certainly stronger and more interesting connections to be drawn to the work of Cartwright, Dupré, Mitchell, Schaffner, Strevens, and Wimsatt.

Perhaps the largest gap to be filled is giving a proper metaphysical account of generic mechanisms. I have helped myself to a collection of desirable properties that I'd like them to have, and although I think there are probably multiple metaphysical options available that could support those properties, I have not narrowed it down to any particular option (nor shown definitively that any exist). Part of that account would need to justify my liberalness about how many generics there are out there. I do not think they are restricted to natural kinds, but some people have argued that only natural kinds have causal powers associated with them. Another major question related to generic mechanisms is how we can tell when a system instantiates one. This is essential for making the proposal workable. And a final question that I mentioned briefly above, is how we can tell when and how different causes will interact.

My aim here was to make room within the MDC definition of mechanism for cognitive mechanisms. This has meant elaborating a broader notion of mechanism than that found in MDC on most readings. It may be that this broader notion of mechanism is not appropriate for other purposes. In some branches of biology, mechanism seems to have a much narrower meaning, with mechanisms covering only one of many types of explanatory structure. So while I have tried to make mechanism cover a wider variety of explanations, there may also be good reasons for narrowing its scope, and instead broadening the selection of explanation types that we recognize as legitimate. It may be more appropriate to divorce generic

mechanisms from mechanistic explanation, and instead elaborate a more general account of structural explanation. That is another question I will think about moving forward.

BIBLIOGRAPHY

Aamodt, S. and Wang, S. (2011). *Welcome to Your Child's Brain: How the Mind Grows from Birth to University*. Oneworld Publications, Oxford.

Aizawa, K. (1994). Representations without Rules, Connectionism and the Syntactic Argument. *Synthese*, 101(3):465–492.

Aizawa, K. and Gillett, C. (2011). The Autonomy of Psychology in the Age of Neuroscience. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the Sciences*. Oxford University Press.

Alcoz, J. (2012). The Viking Sunstone: Is the Legend of the Sun-Stone True? *http://www.polarization.com/viking/viking.html*, retrieved on October 29, 2012.

Amos, A. (2000). A Computational Model of Information Processing in the Frontal Cortex and Basal Ganglia. *Journal of Cognitive Neuroscience*, 12(3):505–519.

Andersen, H. (2011). The Case for Regularity in Mechanistic Causal Explanation. *Synthese*, 189(3):415–432.

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

Anderson, J. R. (1980). *Cognitive Psychology and Its Implications*. W. H. Freeman & Co., 1st edition.

Anderson, J. R. (1983). *The Architecture of Cognition*. Harvard University Press.

Anderson, J. R. (1985). *Cognitive Psychology and Its Implications*. W. H. Freeman & Co., 2nd edition.

Anderson, J. R. (2009). *Cognitive Psychology and Its Implications*. W. H. Freeman & Co., 7th edition.

Antony, M. V. (1991). Fodor and Pylyshyn on Connectionism. *Minds and Machines*, 1(3):321–341.

Arnold, E. (2011). Tools for Evaluating the Consequences of Prior Knowledge, but No Experiments. On the Role of Computer Simulations in Science. Unpublished manuscript.

Baars, B. J. (1986). *The Cognitive Revolution in Psychology*. Guilford Press.

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

Barberousse, A., Franceschelli, S., and Imbert, C. (2009). Computer Simulations as Experiments. *Synthese*, 169(3):557–574.

Batterman, R. W. (2009). Idealization and Modeling. *Synthese*, 169:427–446.

Bechtel, W. (2001). Cognitive Neuroscience: Relating Neural Mechanisms and Cognition. In Machamer, P. K., Grush, R., and McLaughlin, P., editors, *Theory and Method in the Neurosciences*. University of Pittsburgh Press.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. Routledge.

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

Bechtel, W. and Abrahamsen, A. (2009). Decomposing, Recomposing, and Situating Circadian Mechanisms: Three Tasks in Developing Mechanistic Explanations. In Leitgeb, H. and A, H., editors, *Reduction and Elimination in Philosophy of Mind and Philosophy of Neuroscience*. Ontos Verlag, Frankfurt.

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

Bedau, M. and Humphreys, P. (2008). *Emergence: Contemporary Readings in Philosophy and Science*. MIT Press, Cambridge, MA.

Bennett, M. R. and Hacker, P. M. S. (2003). *Philosophical Foundations of Neuroscience*. Wiley-Blackwell.

Bickle, J. (2006). Reducing Mind to Molecular Pathways: Explicating the Reductionism Implicit in Current Cellular and Molecular Neuroscience. *Synthese*, 151(3):411–434.

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

Bogen, J. (2001). 'Two as Good as One Hundred'–Poorly Replicated Evidence is Some 19th Century Neuroscientific Research. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 32(3):491–533.

Bogen, J. (2005). Regularities and Causality: Generalizations and Causal Explanations. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):397–420.

Bokulich, A. (2011). How Scientific Models Explain. *Synthese*, 180:33–45.

Bower, J. M. and Beeman, D. (2003). *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System*. Internet Edition.

Branco, T. and Staras, K. (2009). The Probability of Neurotransmitter Release: Variability and Feedback Control at Single Synapses. *Nature Reviews Neuroscience*, 10:373–383.

Broadbent, D. E. (1958). *Perception and Communication*. Pergamon Press.

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.

Burge, T. (2010). A Real Science of Mind. *The New York Times*, December 19.

Burgess, N., Barry, C., and O'Keefe, J. (2007). An oscillatory interference model of grid cell firing. *Hippocampus*, 17:801–812.

Butler, D. L. (1993). Graphics in Psychology: Pictures, Data, and Especially Concepts. *Behavior Research Methods, Instruments, & Computers*, 25(2):81–92.

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford University Press.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press, New York.

Cartwright, N. (1997). Models: The Blueprints for Laws. *Philosophy of Science*, 64(Supplement):S292–S303.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press, Cambridge.

Churchland, P. S. and Sejnowski, T. J. (1988). Perspectives on Cognitive Neuroscience. *Science*, 242(1968):741–745.

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. MIT Press, Cambridge, MA.

Claxton, G., editor (1980). *Cognitive Psychology: New Directions*. Routledge & Kegan Paul.

Coltheart, M. (2006). What has Functional Neuroimaging Told us about the Mind (so far)? *Cortex*, 42(3):422–427.

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.

Compston, A. (2006). From the Archives. *Brain*, 129(6):1347–1350.

Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science*, pages 53–74.

Craver, C. F. (2005). Beyond Reduction: Mechanisms, Multifield Integration and the Unity of Neuroscience. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):373–395.

Craver, C. F. (2006). When Mechanistic Models Explain. *Synthese*, 153(3):355–376.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Craver, C. F. and Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, 22(4):547–563.

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

Cruse, H. (2001). The Explanatory Power and Limits of Simulation Models in the Neurosciences. In P. Machamer, R. Grush, P. M., editor, *Theory and Method in the Neurosciences*, pages 138—154. University of Pittsburgh Press of Pittsburgh Press, Pittsburgh.

Cummins, R. (1983). *The Nature of Psychological Explanation*. MIT Press, Cambridge, MA.

Cummins, R. and Schwarz, G. (1987). Radical Connectionism. *Southern Journal of Philosophy*, 26(S1):43–61.

Cummins, R. E. (1975). Functional analysis. *Journal of Philosophy*, 72(November):741–764.

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

Darden, L. (2005). Relations among Fields: Mendelian, Cytological and Molecular Mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):349–371.

Darden, L. (2008). Thinking Again about Biological Mechanisms. *Philosophy of Science*, 75(5):958–969.

Darden, L. and Craver, C. F. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 33(1):1–28.

Darden, L. and Maull, N. (1977). Interfield Theories. *Philosophy of Science*, 44(1):43–64.

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

Dennett, D. (2005). Philosophy as Naive Anthropology: Comment on Bennett and Hacker. *Philosophical Investigations*, 28(2):193–196.

DesAutels, L. (2011). Against Regular and Irregular Characterizations of Mechanisms. *Philosophy of Science*, 78(5):914–925.

Donders, F. C. (1969). On the Speed of Mental Processes. *Acta Psychologica*, pages 412–431.

Doyle, D. A. (2004). Structural Changes During Ion Channel Gating. *Trends in Neurosciences*, 27(6):298–302.

Dror, I. E. and Gallogly, D. P. (1999). Computational Analyses in Cognitive Neuroscience: In Defense of Biological Implausibility. *Psychonomic Bulletin & Review*, 6(2):173–182.

Dupré, J. (1995). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Harvard 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.

Eysenck, M. W. and Keane, M. T. (2005). *Cognitive Psychology: A Student's Handbook*. Psychology Press.

Feest, U. (2003). Functional Analysis and the Autonomy of Psychology. *Philosophy of Science*, 70(5):937–948.

Fodor, J. (1997). Connectionism and the Problem of Systematicity (Continued): Why Smolensky's Solution Still Doesn't Work. *Cognition*, 62(1):109–19.

Fodor, J. and McLaughlin, B. P. (1990). Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work. *Cognition*, 35:183–204.

Fodor, J. A. (1968a). *Psychological Explanation: An Introduction to the Philosophy of Psychology*. Random House.

Fodor, J. A. (1968b). The Appeal to Tacit Knowledge in Psychological Explanation. *The Journal of Philosophy*, pages 627–640.

Fodor, J. A. (1974). Special Sciences (or: The Disunity of Science as a Working Hypothesis). *Synthese*, 28(2):97–115.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

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

French, S. (2003). A Model-Theoretic Account of Representation (Or, I Don't Know Much about Art... but I Know It Involves Isomorphism). *Philosophy of Science*, 70(5):1472–1483.

Friedman, M. (1974). Explanation and Scientific understanding. *The Journal of Philosophy*, 71(1):5–19.

Frigg, R. (2006). Scientific Representation and the Semantic View of Theories. *Theoria*, 55:49–65.

Frigg, R. and Reiss, J. (2009). The Philosophy of Simulation: Hot New Issues or Same Old Stew? *Synthese*, 169(3):593–613.

Friston, K. J. (2005). A Theory of Cortical Responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815.

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

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

Galison, P. (1996). Computer Simulations and the Trading Zone. In Galison, P. and Stump, D. J., editors, *The Disunity of Science: Boundaries, Contexts, and Power*, pages 119–157. Stanford University Press.

Gazzaniga, M. S. (2004). *The Cognitive Neurosciences III*. MIT Press.

Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419.

Giere, R. N. (2004). How Models Are Used to Represent Reality. *Philosophy of Science*, 71(5):742–752.

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

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

Glennan, S. (2005). Modeling Mechanisms. *Studies in History and Philosophy of Biological and Biomedical Science*, 36:443–464.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT Press.

Goldstine, H. H. and von Neumann, J. (1947). Planning and Coding of Problems for an Electronic Computing Instrument. Report on the Mathematical and Logical aspects of an Electronic Computing Instrument Part II, Volume 1-3, Institute for Advanced Study, Princeton University, Princeton, New Jersey.

Green, C. D. (1998). The Thoroughly Modern Aristotle: Was he Really a Functionalist ? *History of Psychology*, 1:8–20.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic Models of Cognition: Exploring Representations and Inductive Biases. *Trends in Cognitive Sciences*, 14(8):357–364.

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.

Hartmann, S. (1996). The World as a Process: Simulations in the Natural and Social Sciences. In Hegselmann, R., Müller, U., and Troitzsch, K., editors, *Simulation and Modeling in the Social Sciences from the Philosophy of Science Point of View*, pages 77–100. Kluwer Academic Publishers, Dordrecht.

Hempel, C. G. (1966). Laws and Their Role in Scientific Explanation. In *Philosophy of Natural Science*. Prentice-Hall.

Hesse, M. B. (1966). *Models and Analogies in Science*. University of Notre Dame Press, Notre Dame, Indiana.

Hille, B. (2001). *Ion Channels of Excitable Membranes*. Sinauer Associates, 3rd edition.

Hinton, G. E. (1988). Representing Part-whole Hierarchies in Connectionist Networks. In *Proceeding of the Tenth Annual Conference of the Cognitive Science Society*. Erlbaum.

Hooke, R. (1757). Critique of Newton's Theory of Light and Colors. In Birch, T., editor, *The History of the Royal Society*, volume 3, pages 10–15. Royal Society of London.

Horgan, T. and Tienson, J., editors (1991). *Connectionism and the Philosophy of Mind*. Kluwer Academic Publishers.

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

Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press, New York.

Huygens, C. (1912[1690]). *Treatise on Light*. MacMillan and Co., London.

ISO/IEC/IEEE (2010). *International Standard: Systems and Software Engineering—Vocabulary*.

Kagan, J. and Baird, A. (2004). Brain and Behavioral Development During Childhood. In *The Cognitive Neurosciences*, pages 93–103. MIT Press.

Kandel, E. (1976). *Cellular Basis of Behavior*. W. H. Freeman & Co.

Kandel, E. R. (2001). The Molecular Biology of Memory Storage: A Dialogue between Genes and Synapses. *Science*, 294(5544):1030–8.

Kandel, E. R., Schwartz, J. R., and Jessell, T., editors (2000). *Principles of Neural Science*. McGraw Hill, 4th edition.

Kaplan, D. M. (2011). Explanation and Description in Computational Neuroscience. *Synthese*, 183:339–373.

Kaplan, D. M. and Bechtel, W. (2011). Dynamical Models: An Alternative or Complement to Mechanistic Explanations? *Topics in Cognitive Science*, 3(2):438–444.

Kaplan, D. M. and Craver, C. F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science*, 78(4):601–627.

Kendler, K. S., Zachar, P., and Craver, C. (2010). What Kinds of Things are Psychiatric Disorders? *Psychological medicine*, pages 1–8.

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In Kitcher, P. and Salmon, W. C., editors, *Scientific Explanation*, pages 410–505. University of Minnesota Press.

Knuuttila, T. (2011). Modelling and Representing: An Artefactual Approach to Model-based Representation. *Studies in History and Philosophy of Science*, 42:262–271.

Knuuttila, T. and Boon, M. (2011). How do Models Give us Knowledge? The Case of Carnot's Ideal Heat Engine. *European Journal for Philosophy of Science*, 1:309–334.

Kuffler, S. W. (1953). Discharge Patterns and Functional Organization of Mammalian Retina. *Journal of Neurophysiology*, pages 37–68.

Kuffler, S. W., Nicholls, J., and Martin, A. R. (1984). From Neuron to Brain: A Cellular Approach to the Function of the Nervous System. *Sunderland, MA: Sinaur Associates*.

Kuorikoski, J. (2009). Two Concepts of Mechanism: Componential Causal System and Abstract Form of Interaction. *International Studies in the Philosophy of Science*, 23(2):143–160.

Küppers, G. and Lenhard, J. (2004). The Controversial Status of Simulations. *Proceedings of the 18th European Simulation Multiconference*.

LeDoux, J. E. and Hirst, W., editors (1986). *Mind and Brain: Dialogues in Cognitive Neuroscience*. Cambridge University Press.

Levy, A. (2013). What was Hodgkin and Huxley's Achievement? *British Journal for the Philosophy of Science*, 64(2):1–24.

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

Logothetis, N. K. (2008). What we Can do and What we Cannot do with fMRI. *Nature*, 453(7197):869–78.

Machamer, P. (2011a). Phenomena, Data and Theories: A Special Issue of *Synthese*. *Synthese*, 182:1–5.

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

Machamer, P. K. (2004). Activities and Causation: The Metaphysics and Epistemology of Mechanisms. *International Studies in the Philosophy of Science*, 18(1):27–39.

Machamer, P. K. (2011b). Mechanisms: Ontology, Representation, and Psychology. In *Mechanisms: Les Mécaniciens: Salon des Refusés (Pittsburgh; 9 April 2011)*.

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

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

Marr, D. and Poggio, T. (1976). From Understanding Computation to Understanding Neural Circuitry.

Maxwell, J. C. (1873). *A Treatise on Electricity and Magnetism*. Clarendon Press, Oxford.

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. (1979). On the Time Relations of Mental Processes: An Examination of Systems of Processes in Cascade. *Psychological Review*, 86(4).

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

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

McClelland, J. L. (2010). Emergence in Cognitive Science. *Topics in Cognitive Science*, 2(4):751–770.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., and Smith, L. B. (2010). Letting Structure Emerge: Connectionist and Dynamical Systems Approaches to Cognition. *Trends in Cognitive Sciences*, 14(8):348–356.

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.

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, 63(2):81–97.

Minsky, M. and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

Mitchell, S. D. (2000). Dimensions of Scientific Law. *Philosophy of Science*, 67(2):242–265.

Mitchell, S. D. (2003). *Biological Complexity and Integrative Pluralism*. Cambridge University Press, Cambridge.

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

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

Morris, S. and Gotel, O. (2011). The Role of Flow Charts in the Early Automation of Applied Mathematics. *BSHM Bulletin: Journal of the British Society for the History of Mathematics*, 26(1):44–52.

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

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. Harcourt Press.

Narayanan, A. (1988). Fodor and Pylyshyn on Connectionism: An Extended Review and Brief Critique. *Artificial Intelligence Review*, 2(3):195–213.

Neisser, U. (1967). *Cognitive Psychology*. Appleton-Century-Crofts.

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.

Norton, S. D. and Suppe, F. (2001). Why Atmospheric Modeling Is Good Science. In Miller, C. A. and Edwards, P. N., editors, *Changing the Atmosphere*, pages 67–105. MIT Press.

O'Keefe, J. and Burgess, N. (2005). Dual Phase and Rate Coding in Hippocampal Place Cells: Theoretical Significance and Relationship to Entorhinal Grid Cells. *Hippocampus*, 15:853–866.

O'Loughlin, C. and Thagard, P. (2000). Autism and Coherence: A Computational Model. *Mind & Language*, 15(4):375–392.

Oppenheim, P. and Putnam, H. (1958). The Unity of Science as a Working Hypothesis. *Minnesota Studies in the Philosophy of Science*, 2:3–36.

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.

Overton, J. A. (2011). Mechanisms, Types, and Abstractions. *Philosophy of Science*, 78(5):941–954.

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

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., and Raichle, M. E. (1988). Positron Emission Tomographic Studies of the Cortical Anatomy of Single-word Processing. *Nature*, 331:585–589.

Piccinini, G. and Craver, C. (2011). Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches. *Synthese*, pages 1–58.

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

Posner, M. I. (1980). Orienting of Attention. *The Quarterly Journal of Experimental Psychology*, 32(1):3–25.

Posner, M. I. (1994). Attention: The Mechanisms of Consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, 91(16):7398.

Posner, M. I. and DiGirolamo, G. J. (2000). Cognitive Neuroscience: Origins and Promise. *Psychological Bulletin*, 126(6):873–889.

Posner, M. I., Petersen, S. E., Fox, P. T., and Raichle, M. E. (1988). Localization of Cognitive Operations in the Human Brain. *Science*, 240(4859):1627–1631.

Pylyshyn, Z. W. (1984). The Relevance of Computation. In *Computation and Cognition*. MIT Press.

Ramsey, W. M. (1997). Do Connectionist Representations Earn their Explanatory Keep? *Mind & Language*, 12(1):34–66.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge University Press.

Ramsey, W. M., Stich, S. P., and Garon, J. (1990). Connectionism, Eliminativism and the Future of Folk Psychology. *Philosophical Perspectives*, pages 499–533.

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

Revonsuo, A. (2001). On the Nature of Explanation in the Neurosciences. In Machamer, P. K., Grush, R., and McLaughlin, P., editors, *Theory and Method in the Neurosciences*, pages 45–69. University of Pittsburgh Press.

Rohrlich, F. (1990). Computer Simulation in the Physical Sciences. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2:507–518.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6):386–408.

Rowlands, M. (1994). Connectionism and the Language of Thought. *The British Journal for the Philosophy of Science*, 45(2):485–503.

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

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

Salmon, W. (1990). Scientific Explanation: Causation and Unification. *Critica: Revista Hispanoamericana de Filosofia*, pages 3–23.

Salmon, W. (1992). Scientific explanation. In Salmon, M., Earman, J., Glymour, C., Lennox, J. G., Machamer, P., McGuire, J., Norton, J., and Schaffner, K., editors, *Introduction to the Philosophy of Science*, pages 7–41. Prentice Hall, Inc.

Salmon, W. C. (1981). Rational Prediction. *The British Journal for the Philosophy of Science*, 32(2):pp. 115–125.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Schaffner, K. F. (1977). Reduction, Reductionism, Values, and Progress in the Biomedical Sciences. In Colodny, R., editor, *Logic, laws, and life*, pages 143–171. University of Pittsburgh Press.

Schaffner, K. F. (1993). *Discovery and Explanation in Biology and Medicine*. University of Chicago Press.

Schaffner, K. F. (2006). Reduction: The Cheshire Cat Problem and a Return to Roots. *Synthese*, 151(3):377–402.

Schall, J. D. (2004). On Building a Bridge Between Brain and Behavior. *Annual Review of Psychology*, 55(1):23–50.

Schneider, W. and Chein, J. M. (2003). Controlled & Automatic Processing: Behavior, Theory, and Biological Mechanisms. *Cognitive Science*, 27:525–559.

Schneider, W. and Shiffrin, R. M. (1977). Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention. *Psychological Review*, 84(1):1–66.

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

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

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(July, October):379–423, 623–656.

Shepherd, G. M. (1983). *Neurobiology*. Oxford University Press.

Shiffrin, R. M. and Schneider, W. (1977). Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending, and a General Theory. *Psychological Review*, 84(2):127–190.

Skipper, R. A. J. (1999). Selection and the Extent of Explanatory Unification. *Philosophy of Science*, 66(Supplement):S196–S209.

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

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

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

Sporns, O. (2011). *Networks of the Brain*. MIT Press, Cambridge, MA.

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

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

Sternberg, S. (1969). The Discovery of Processing Stages: Extensions of Donders' Method. *Acta Psychologica*, 30:276–315.

Strevens, M. (2004). The Causal and Unification Approaches to Explanation Unified—Causally. *Noûs*, 38(1):154–176.

Suárez, M. (2003). Scientific Representation: Against Similarity and Isomorphism. *International Studies in the Philosophy of Science*, 17(3):225–244.

Sullivan, J. A. (2009). The Multiplicity of Experimental Protocols: A Challenge to Reductionist and Non-reductionist Models of the Unity of Neuroscience. *Synthese*, 167:511–539.

Sun, R. (2009). Theoretical Status of Computational Cognitive Modeling. *Cognitive Systems Research*, 10(2):124–140.

Suppe, F. (2000). Understanding Scientific Theories: An Assessment of Developments, 1969-1998. *Philosophy of Science*, 67(Supplement):S102–S115.

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

Swoyer, C. (1991). Structural Representation and Surrogative Reasoning. *Synthese*, 87:449–508.

Teller, P. (2001). Twilight of the Perfect Model Model. *Erkenntnis*, 55:393–415.

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

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

Treisman, A. (1960). Contextual Cues in Selective Listening. *The Quarterly Journal of Experimental Psychology*, 12(4):242–248.

Tsotsos, J. K. (1990). Analyzing Vision at the Complexity Level. *Behavioral and brain sciences*, 13(3):423–469.

Uttal, W. R. (2003). *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. MIT Press.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford University Press, Oxford.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Psychological Science*, 4(3):274–290.

Waldrop, M. M. (2012). Brain in a Box. *Nature*, 482:456–458.

Weber, E. and Bouwel, J. V. (2009). Causation, Unification, and the Adequacy of Explanations of Facts. *Theoria*, 66:301–320.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., and Gray, J. R. (2008). The Seductive Allure of Neuroscience Explanations. *Journal of Cognitive Neuroscience*, 2:470–477.

Weiskopf, D. A. (2011). Models and Mechanisms in Psychological Explanation. *Synthese*, 183(3):313–338.

Winsberg, E. (1999). Sanctioning Models: The Epistemology of Simulation. *Science in Context*, 12(02):275–292.

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

Winsberg, E. (2003). Simulated Experiments: Methodology for a Virtual World. *Philosophy of Science*, 70(1):105–125.

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

Woodward, J. (2000). Explanation and Invariance in the Special Sciences. *British Journal for the Philosophy of Science*, 51:197–254.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation: A Theory of Causal Explanation*. Oxford University Press, USA.

Wright, C. and Bechtel, W. (2007). Mechanisms and Psychological Explanation. In Thagard, P. R., editor, *Handbook of the Philosophy of Science: Philosophy of Psychology and Cognitive Science*, volume Volume 4 o. Elsevier B.V.

Yu, F. H. and Catterall, W. A. (2003). Overview of the Voltage-gated Sodium Channel Family. *Genome Biology*, 4(3):207.