ON THE MATTER OF MEMORY: NEURAL COMPUTATION AND THE MECHANISMS OF INTENTIONAL AGENCY

by

ALEXANDER D. MORGAN

A dissertation submitted to the
Graduate School-New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Philosophy
written under the direction of
Frances Egan and Robert Matthews
and approved by

________________________

________________________

________________________

New Brunswick, New Jersey
May 2014
Humans and other animals are intentional agents; they are capable of acting in ways that are caused and explained by their reasons. Reasons are widely held to be mediated by mental representations, but it is notoriously difficult to understand how the intentional content of mental representations could causally explain action. Thus there is a puzzle about how to ‘naturalize’ intentional agency. The present work is motivated by the conviction that this puzzle will be solved by elucidating the neural mechanisms that mediate the cognitive capacities that are distinctive of intentional agency.

Two main obstacles stand in the way of developing such a project, which are both manifestations of a widespread sentiment that, as Jerry Fodor once put it, “notions like computational state and representation aren’t accessible in the language of neuroscience”. First, C. Randy Gallistel has argued extensively that the mechanisms posited by neuroscientists cannot function as representations in an engineering sense, since they allegedly cannot be manipulated by the computational operations required to generate structurally complex representations. Second, William Ramsey has argued that neural mechanisms in fact do not function as representations in an explanatory sense, since the explanatory role they play is more like that of a causal relay rather than an internal ‘stand-in’ for external entities.
I argue that the criticisms developed by Gallistel and Ramsey rest on a misapplication of relevant theoretical notions from computer science, and an impoverished appreciation of the rich variety of mechanisms posited in contemporary neuroscience. My central argument, though, is that the conception of representation presupposed by those researchers, according to which representations have the same abstract structure as the external systems they represent, encompasses states in all sorts of systems, including mindless systems such as plants. Thus these ‘structural representations’ are not distinctively cognitive representations, and although they play an important explanatory role in neuroscience and other disciplines, they cannot on their own elucidate the nature of intentional agency. I conclude by sketching some ways in which neuroscientists are beginning to elucidate the representational mechanisms that do mediate distinctively cognitive capacities. Thus, by removing the obstacles put in place by Gallistel and Ramsey, this work clears a path toward the ‘naturalization’ of agency in terms of neural mechanisms.
Acknowledgements

‘But is it philosophy?’ The interdisciplinary hothouse of the Philosophy Department at Rutgers has provided an intellectual climate that is inhospitable to that question. For that I’m grateful, as it has allowed me to work away on questions that I’m interested in without much regard for disciplinary boundaries. But now that I step out into the wider field of academic philosophy, I should perhaps ready myself to answer that question. I don’t think that philosophy has a distinctive content, or even an especially distinctive methodology; after all, scientists also use thought experiments and seek to clarify concepts. One way of thinking about philosophy that I find appealing is in terms of the intellectual habits or dispositions of its practitioners — a philosophical temperament. I have in mind a certain driving curiosity about the general principles that underly the patterns we find in nature, a willingness to question cherished verities, and perhaps most importantly, an eagerness to engage with criticisms of one’s own views. If this is what we mean by ‘philosophy’, then I can think of no-one who is more deserving of the label ‘philosopher’ than Randy Gallistel. Although I think that Randy is dead wrong about many things, my tentative attempts to explain why to him are never met with dismissiveness or defensiveness, but rather with good-humored explanations of why I am dead wrong. Randy’s philosophical temperament, together with his keen wit and deep erudition, are an inspiration. I am grateful for his example.

This dissertation grew out of a graduate seminar that I took with Randy, undoubtedly the highlight of my many years at university, but the interests that led me to take
that seminar, and the preparedness that allowed me to benefit from it, were fostered and developed by several outstanding undergraduate teachers. My exposure to the force of nature that is Kim Sterelny in some of my first philosophy courses at Victoria University propelled me along the path of attempting to understand the mind as a biological phenomenon from the perspective of empirically-informed philosophy. That path led me to the University of Auckland, where I tried to find my own way through the thickets of the various fields of cognitive science. I would have become lost were it not for the expert guidance of Professors Michael Corballis, Russell Gray, Justine Kingsbury, Sue Pockett, and Denis Robinson at crucial junctures, so to them I owe a deep debt of gratitude. I am especially indebted to Justine, whose encouragement and support was pivotal in allowing me to go on to pursue graduate studies at Rutgers.

Upon arriving in an environment that was deceptively familiar from so many sitcoms and college movies, this itinerant Kiwi was immediately made to feel at home by the wonderful community of graduate students in the philosophy department at Rutgers — a community sizzling with intellectual energy, but in a way that manages to warm with companionship, rather than burn with competition. I’d like to offer my sincerest thanks to my many brilliant peers at Rutgers who kept this community alive: Luvell Anderson, Marcello Antosh, Josh Armstrong, Kate Devitt, Tom Donaldson, Richard Dub, Erik Hoversten, Alex Jackson, Michael Johnson, Karen Lewis, Kelby Mason, Sarah Murray, Meghan Sullivan, Jenny Nado, Iris Oved, Karen Shanton, Jenn Wang, and many others. Special thanks to E. J. Green, Gabe Greenberg, Lisa Miracchi, Ron Planer, and Will Starr, not only for their friendship, but also for discussions and comments that directly led to improvements in many sections of this work.

The community at Rutgers is kept alive not only by the students, but also by the tireless efforts of the administrative staff. I would thus like to thank Pauline Mitchell, Susan Viola, and especially Mercedes Diaz, who has provided not only expert guidance along my somewhat circuitous path through the PhD program, but also invaluable
friendship along the way. Of course, Rutgers wouldn’t be Rutgers were it not for the inspiring faculty there. Seminars with Martha Bolton, Andy Egan, Jerry Fodor, Lila Gleitman, Alvin Goldman, Peter Klein, Ernie Lepore, Jason Stanley, Steve Stich, and many others not only enriched my understanding of the content of philosophy and cognitive science, but provided exemplary models of how philosophy ought to be done.

Of the members of the philosophy faculty at Rutgers, two deserve special mention. From the beginning, Frankie Egan and Bob Matthews have provided unflagging intellectual and moral support, even, and indeed especially, during some of my most difficult periods during the PhD program. As dissertation chairs, they have been exemplary, allowing me free rein to explore ideas independently, but always ready at hand to help me discipline those ideas when needed with their keen critical feedback. This dissertation would be a far shoddier specimen were it not for their guidance and inspiration. I also could not have asked for a better external committee member. Ever since I first contacted Gualtiero Piccinini out of the blue as a second-year graduate student, he has been incredibly supportive and generous with his time. His unparalleled knowledge of the history of computer science and theoretical neuroscience is reflected throughout this dissertation. Together with Randy Gallistel’s inspiritional influence, I feel privileged to have had such an outstanding dissertation committee.

Since leaving Rutgers, I have had the wonderful good fortune to find myself amongst a group of brilliant philosophers in the Center for Integrative Neuroscience at the University of Tübingen, under the inspiring guidance of Hong Yu Wong. I’d like to thank my colleagues Chiara Brozzo, Eva Engels, Gregor Hochstetter, Elizabeth Irvine, Krisztina Orbán, Ferdinand Pöhlmann, Katia Samoilova, and Catherine Stinson for helping me feel at home in Germany, for contributing to such a pleasantly stimulating environment in which to do philosophy, and for feedback on several sections of this dissertation. I would especially like to thank Hong Yu for his exceptional warmth and generosity, which among other things have allowed me to complete the final stages of
this dissertation in an environment largely free from distractions. Hong Yu’s personal qualities are matched by his keen philosophical acuity, and I’m thankful that I’m able to count him both as a mentor and a friend.

This dissertation has benefitted considerably from exchanges with audiences at the Rutgers Center for Cognitive Science, the University of Tübingen, the University of Porto, as well as from extremely helpful email correspondence with Paco Calvo, Chris Eliasmith, Bruce McNaughton, Ed Stabler, and especially Bill Ramsey, for which I’m extremely grateful. Chapter 5 of this dissertation was published, with slight modifications, in *Synthese* under the title ‘Representations Gone Mental’.

Finally, I’d like to offer my heartfelt thanks to Susan and Kat for their love and generosity through different parts of this journey, and to Max, Tahi and Niko for emotional support at critical moments along the way. But most of all, I’d like to express my deepest love, admiration, and gratitude to my mother, Patricia Morgan. There’s an obvious sense in which, without her, this dissertation would never have been possible. But that doesn’t do justice to the innumerable ways in which her emotional and material support, often through times of extreme hardship, have helped me to complete this journey. By instilling in me an abiding appreciation of the complexities of nature, and the courage to follow paths less trodden, Patricia has had a subtle but ubiquitous influence on this dissertation. She and I pursued our doctorates simultaneously, following very different but parallel paths. She finished first, inspiring me to catch up. We did it, mum.
# Contents

Abstract ................................................................. ii  
Acknowledgements ....................................................... iv  
Contents ................................................................... viii  

## Chapter 1  Toward the Mechanisms of Intentional Agency  
1.1 Intentional Agency: A View from the Middle .......................... 2  
1.2 Why Bother? ............................................................... 8  
1.3 Obstacles, and the Path Forward ....................................... 20  

## Chapter 2  Theoretical & Experimental Foundations  
2.1 Introduction ................................................................ 28  
2.2 Information Processing from an Engineer’s Perspective ............ 31  
2.2.1 Information ........................................................... 32  
2.2.2 Representation ....................................................... 40  
2.2.3 Computation .......................................................... 49  
2.3 The Computational Basis of Animal Learning .......................... 64  
2.3.1 Path Integration in Insects ........................................... 65  
2.3.2 Food Caching in Birds .............................................. 72  
2.4 Computing with Neurons ............................................... 77  
2.4.1 A Potted History of Experimental Neuroscience ............... 78
Chapter 3  Exorcising the Connectoplasm  111

3.1  Introduction ............................................. 111
3.2  The Scope of the Problem .............................. 120
3.3  Problems with Connectionist Representation .......... 124
  3.3.1 Symbols as Reverberating Circuits? ............. 130
  3.3.2 Symbols as Synaptic Weights? .................... 134
  3.3.3 The Inefficiency of Neural Codes ................. 140
3.4  Problems with Connectionist Computation ............ 146
  3.4.1 Constructability .................................... 150
  3.4.2 Transportability .................................... 153
  3.4.3 Compositionality .................................. 157
3.5  Problems with Connectionist Learning ................. 161
  3.5.1 Anti-Representationalism ......................... 163
  3.5.2 Empiricism ........................................ 171
3.6  A Case-Study of Connectionist Inefficiency .......... 178
  3.6.1 The Samsonovich-McNaughton Model .......... 179
  3.6.2 Problems .......................................... 183
3.7  Memory Amongst the Molecules? .................... 192

Chapter 4  Beyond Good and Evil  198

4.1  Introduction ............................................. 198
4.2  Dissent in the Computationalist Ranks? ............. 204
4.3  Manufactured Consent Amongst the ‘Associationists’ .. 213
## Chapter 5  Deconstructing Structural Representations  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1  Introduction</td>
<td>304</td>
</tr>
<tr>
<td>5.2  The Job Description Challenge</td>
<td>309</td>
</tr>
<tr>
<td>5.2.1 Structural Representations</td>
<td>313</td>
</tr>
<tr>
<td>5.2.2 Receptor Representations</td>
<td>325</td>
</tr>
<tr>
<td>5.3  Problems in the Workplace</td>
<td>329</td>
</tr>
<tr>
<td>5.3.1 Structural Receptors?</td>
<td>330</td>
</tr>
<tr>
<td>5.3.2 Mental Representations</td>
<td>340</td>
</tr>
</tbody>
</table>

## Chapter 6  Conclusion: Steps Toward a New Job Description  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliography</td>
<td>353</td>
</tr>
</tbody>
</table>
Chapter 1

Toward the Mechanisms of

Intentional Agency

Here’s a just-so story. At some point in the infancy of humanity our ancestors cracked the mindreading trick and rather went wild with it. They projected minds all over the world, not just to non-human animals — finding spirits in trees and stones, gods in stars and clouds. Presumably, for all the confusion this involved, our ancestors reaped some net benefits from the ability to identify with others, and some truths about other minds were mingled with the many falsehoods. Having started out doing intentional explanation promiscuously, we eventually grew up and out of our animism; we worked our way pure to causal explanation. We then started to retrench on seeing intentionality everywhere: anthropomorphizing the world was to be avoided as an embarrassing sign of our former collective intellectual immaturity. But our hands have been relatively clean for some time; we can now afford to think again. There may be some overlooked truths mingled with the discarded falsehoods. There may be something after all to make sense of in the strange nonhuman minds around us, without anthropomorphizing them: continuities as well as discontinuities to be described between human and nonhuman minds.

— Susan Hurley, Animal Action in the Space of Reasons
1.1 Intentional Agency: A View from the Middle

Why did the chicken cross the road? I don’t know, but I suppose she had her reasons. Perhaps she was hungry and saw a tasty morsel over yonder. Perhaps she was trying to escape from a predator. Whatever the answer, what’s notable is the nature of the question; it seems perfectly natural to understand the ‘why’ question as a request for the chicken's reasons for acting.¹ The question, understood as a request for the chicken’s reasons, seems to have application, as Anscombe (1957) might have put it. The point of course isn’t that we might reasonably ask the chicken why she crossed the road and expect her to provide a justification for her action. The point is that it’s commonsense to suppose that the chicken had a reason for acting, a reason that, by her lights, made it somehow appropriate for her to cross. Moreover, it’s commonsense to suppose that her reason for acting also features in a causal explanation of why she acted. If she crossed the road because she was trying to escape a predator, it’s natural to suppose that the ‘because’ expresses a causal relation. All of this is just a roundabout way of saying that according to our ordinary, commonsensical conception of things, human and non-human animals alike are intentional agents that can act for reasons, reasons that both rationalize and causally explain their behavior.

Now, it’s true that we often explain the activities of apparently mindless systems by citing what superficially appear to be reasons. Why did the plant reorient its leaves? To maximize its intake of sunlight. But in asking such ‘why’ questions, are we really inquiring into the reasons for which an agent acts? I take it that Granny, the philosophical avatar of commonsense, would shake her cane at the suggestion. Although there might be reasons, on some reasonable interpretation of that term, why the plant reoriented its leaves, they aren’t the plant’s reasons, reasons in the light of

¹That’s of course the conceit behind the schoolyard joke; framed as a joke, the question generates an expectation for something other than a run-of-the-mill statement of the chicken’s reasons.
which, from the plant’s point of view, it was appropriate to act. The plant reoriented its leaves because of the operation of mechanisms that were naturally selected to maximize sunlight intake, not because it intended to maximize sunlight intake. The point isn’t intended to cast aspersions on the noble kingdom Plantae. The literature on ‘plant intelligence’ reveals that plants exhibit rich and surprising forms of adaptive plasticity that are mediated by complex information-processing mechanisms (Garzón & Keijzer, 2011). Granny would find this evidence fascinating, and might even agree that it shows that plants are “cognitive in a minimal, embodied sense” (ibid., p.166). But she’d nevertheless insist that, unlike animals, plants aren’t intentional agents; they’re just not the kind of thing that can act in the light of reasons.

Of course, many think that Granny is an old-fashioned fuddy-duddy. For example, some radical upstarts have argued that agency is best understood in terms of thermodynamic autonomy, the capacity of a dynamical system to sustain a distinction between itself and the external environment in the face of external perturbations through the operation of internal, self-maintaining processes (Moreno & Etxeberria, 2005; Skewes & Hooker, 2009). This notion of agency-as-autonomy encompasses living organisms as such, including plants, protists, and all the rest. But Granny needn’t get the cane out yet, since it’s not clear that when proponents of autonomy talk of ‘agency’, they’re talking about intentional agency, the capacity for a system to act in light of a reason. Perhaps Granny and the autonomists can agree that “minimal agents are complex dissipative systems that can... dynamically self-maintain” (Moreno & Etxeberria, 2005, p.163, my emphasis), but that genuine intentional agents require somewhat more in the way of internal and behavioral complexity.

Nevertheless, some radicals really do seem to hold that the plants and protists are intentional agents. For example, Barham (2012) claims that “all organisms are normative agents, and that only organisms are normative agents in a literal, original, and underived sense” (p.94), where it’s clear that he understands normative agency
as “the power to ‘act for a reason’” (p.93). Biologists impressed by plant intelligence endorse similar views. For example, in describing the capacities of the walking palm *Socratea durissima*, Trewavas (2003) writes that “[w]hen competitive neighbours approach, avoidance action is taken by moving the whole plant back into full sunlight. Such obvious ‘walking’ is accomplished by growing new prop roots in the direction of movement while those behind die off. That this is intentional behaviour is very clear” (p.15). I don’t want to dispute these claim (yet), but I do want to get in clear view why Granny would find them problematic. We’ve already seen the answer in broad outline: the reasons for which an agent acts are somehow based on an agent’s subjective *perspective* on the world. To make the answer fully explicit: when an agent acts for a reason, she acts on the basis of what she takes to be appropriate given the way she takes the world to be — that is, on the basis of the content of her mental states. But, Granny insists, plants just don’t *have* mental states.

The commonsense view that intentional actions are explained by an agent’s mental states receives canonical philosophical expression in Davidson’s (1963) *causal theory of action*. The essential structure of Davidson’s theory is this: in order for an agent’s action to be intentional, it must have been caused by the agent’s desire to achieve a certain outcome, together with her belief that performing that action would eventuate in that outcome. This belief-desire pair constitutes the agent’s reason for acting. One virtue of this kind of causal theory is that it allows us to identify the respect in which an action counts as intentional. Suppose, to use Davidson’s famous example, that I flip the kitchen light switch, and inadvertently alert an intruder to my presence. There are many different ways of describing my action: as a flipping of a switch, an illumination of a room, an alerting of an intruder, and so forth. But I surely didn’t *intend* to do all of these things. A causal theory of action like Davidson’s allows us to pick out the description under which my action counts as intentional; it’s the description that reflects the content of the mental states that causally explained my action. To put the
point slightly differently, a causal theory allows us to make sense of the fact that my alerting the intruder was *inadvertent*.

Another virtue of a broadly Davidsonian causal theory of action is that it allows us to distinguish cases in which an agent performed an action because she intended to do so, from cases in which an agent performed an action that she intended to perform, but not *because* of that intention. Upon alerting the intruder and seeing her pull a gun on me, I might intend to fall to the ground and play dead; yet I might be so overwhelmed with fright that I faint, and fall to the ground unintentionally. It’s hard to make sense of the distinction at issue unless intentions at least sometimes play a role in the etiology of action.

There’s an additional aspect of Davidson’s thinking about intentional action that is especially noteworthy in the present context, but to appreciate it we must first step back to see how the direction of Davidson’s thinking was shaped by the intellectual currents of the tradition that he was working within. Since its inception, philosophy in the analytic tradition has been characterized by a careful attentiveness to the subtleties language. For many analytic philosophers in the early- to mid-20th century, this methodological focus on language — or at least on *English* — manifested as an attempt to reconstruct questions about certain philosophically perplexing phenomena, such as mental states, as questions about the peculiarities of the sentences used to talk about such phenomena, in order, as Quine (1960) put it, to “carry the discussion into a domain where both parties are better agreed on the objects (viz., words)” (p.272). But even among later analytic philosophers who regarded ‘analytic philosophy’ more as a label for a certain socio-historical affiliation than a description of a specific philosophical methodology, the traditional linguistic preoccupations of their discipline exerted a tacit but tenacious influence on the kinds of questions they asked, and the kinds of answers they considered.

An example of this is the received view in the analytic tradition that intentional
mental representations are (predominantly or exclusively) *propositional attitudes*. This view emerged with Russell (1940), who claimed that English mental-state verbs like ‘believe’, ‘desire’, ‘hope’ and so forth “must be followed by a subordinate sentence telling what it is that is believed or desired or [hoped]” (p.65, my emphasis). This is false, as I hope to have shown with the subordinate clause you just read. But it nevertheless seeped into the collective subconscious of analytic philosophers, and became almost universally accepted in the latter part of the 20th century.

Davidson follows the mainstream in assuming that intentional mental states are propositional attitudes. However, when it comes to characterizing the *nature* of propositional attitudes, he follows a much smaller tributary. On his view, the conditions for correctly ascribing propositional attitudes to an agent are determined by the constraints involved in interpreting the agent’s utterances. These constraints are massively holistic in that they require arriving at an interpretation of an agent’s language as a whole, an interpretation that maximizes the truth and consistency of the agent’s utterances. So, for Davidson, ascriptions of propositional attitudes don’t refer to discrete entities in an agent’s mind; rather, they capture certain patterns in an agent’s linguistic dispositions. A corollary of this view is that only creatures that have linguistic dispositions can have propositional attitudes (Davidson, 1975, 1982). Since Davidson’s causal theory of action identifies reasons for acting with propositional attitudes, i.e. belief-desire pairs, it follows that on his view, nonlinguistic creatures are incapable of having reasons for acting. They don’t count as intentional agents.

This kind of *conservatism* according to which only language users are capable of

---

2This is of course a bit quick. Ben-Yami (1997) develops the point in more detail. As he points out, it is common in English to ascribe mental states using sentences that contain a mental-state verb that is not followed by a sentential complement. For example, ‘I want coffee’. Sometimes it is possible to paraphrase these sentences into sentences that do take a sentential complement, but the result is often unnatural: ‘I desire that I have coffee’. Moreover, not all such sentences can be paraphrased in this way. Consider ‘I trust you’. These generalizations seem to be true of most languages, not just of English.
genuine intentional agency has been endorsed by other philosophers, for similar reasons. For example, McDowell (1994) argues that in order for an agent to genuinely act for a reason, it must have the capacity to reflect on whether it ought to act for that reason, a capacity acquired through immersion in a linguistic community (p.125). However, I’d like to keep Davidson’s particular brand of conservatism in the foreground, since I think it is instructive to see how strange it looks when held up against the background of his causal theory of action. The causal theory seemed promising since it captured some of our core intuitions about the nature of intentional agency. But these intuitions seem just as compelling in the case of nonlinguistic animals as in the case of normal adult humans. If the chicken crossed the road in order to avoid a predator, and in so doing caused a conscientious driver to swerve into a ditch, we’d like to be able to say that she caused the action unintentionally. Or take Davidson’s (1963) classic objection to a non-causal theory of action: “a person can have a reason for an action, and perform the action, and yet this reason not be the reason why he did it” (p.691). Just replace ‘person’ with ‘chicken’ and the point seems just as potent. While it is not transparently inconsistent to endorse both conservatism about agency and a causal theory of action, the former seems to sap the latter of much of its motivation. If we’re to discredit the intuitive reasons for thinking that reasons are causes in the case of nonlinguistic animals, why credit them in the case of normal adult humans?

So on one way of understanding the motivation for this dissertation, it is an attempt to clear a middle ground between radicalism and conservatism about the scope of intentional agency. I want to hack away at the philosophical tangles on both sides to make room for Granny’s view that there’s something special about animals, something that sets them apart from plants and protists, namely that they’re able to act in the world in virtue of the way they take the world to be. To put it another way, I want to find a place in nature for real teleology. Contrary to the radicals, I think that a genuinely goal-directed system must actually have goals; but contrary to the
conservatives, I don’t think that such a system must be capable of reflecting on those goals, or expressing them in language.

1.2 Why Bother?

At this point, one might reasonably wonder: why bother? Neither radicalism nor conservatism commands wide assent, and several philosophers have already argued quite persuasively that animals can act for reasons without having the capacity to reflect on their reasons (e.g. Carruthers, 2004; Glock, 2009; Hurley, 2003). A more pointed worry threatens to deflate any account of the nature of intentional agency: why think that intentional agents have anything theoretically interesting in common? Why not think that the concept of intentional agency merely reflects our practical, parochial interests in predicting the behavior of systems that might harm us or do us good? Sterelny (2003) expresses the worry in the form of a rhetorical question: “An Alpha Centauri biologist is unlikely to have any interest in identifying humanoids. Would an Alpha Centauri... scientist be interested in identifying agents?” (p.264).

One way to address this latter worry is to simply answer the rhetorical question directly: yes, insofar as the Alpha Centauri scientist is herself an agent, she would plausibly be interested in the nature of intentional agency. This response is superficial, but I think it is just the exposed end of a deeper point, a point that will help us respond to the general worry about the motivation for this dissertation. Excavating the point will take a bit of work. To begin, let’s survey the conceptual terrain and turn our attention back to Davidson’s views about intention and action. Davidson’s conservatism about agency and his causal theory of action make uneasy bedfellows for an obvious reason that I haven’t yet made explicit: the view that beliefs and desires a merely an interpretative gloss on linguistic dispositions, and the view that they nevertheless causally explain actions, seem deeply incompatible. Davidson is of course
aware of the apparent incompatibility between these views, and argues that it’s merely apparent. His argument is complex and controversial, but the central idea is that although an intentional action involves a genuine causal law that relates some event occurring within the agent and the event that constitutes the agent’s action, such laws don’t relate those events as described using the psychological vocabulary of reasons and intentions; genuine causal laws only relate events as described within the vocabulary of physics (Davidson, 1970). We can now begin to discern the point we were seeking: even if Davidson’s argument were sound, the most it would show is that mental events, under some description, can cause actions. But insofar as the descriptions under which mental events cause actions are expressed in an inscrutable physicalistic vocabulary, it remains entirely unclear how the fact that a mental event has a specific content contributes to a causal explanation of the action that it rationalizes.3

The promise of the causal theory of action was that it would help shed light on the intuitively attractive idea that if an agent acts for a reason, her action is caused by a mental event with a specific content, a content in virtue of which the action is somehow appropriate. It was the chicken’s perceiving a threat as a threat that caused her to cross the road. We don’t just want a theory that is consistent with that idea, we want a theory that explains it. That is, we want a theory according to which mental events have causal powers in virtue of their contents. And it seems that no matter how you fudge the metaphysics, an Interpretivist theory just isn’t going to give us that. How might we get such a theory? Granny has a suggestion that she thinks she probably got from the Early Moderns: think of mental events as tokenings of concrete particulars with ‘semantic’ properties — that is, as mental representations.

3To get at the point from another direction, Davidson’s view entails that there are causal relations between mental events and their intended consequences, as well as between mental events and their unintended consequences; but his view doesn’t explicate what the difference between those relations consists in.
In accordance with the linguistic preoccupations of their discipline, analytic philosophers have interpreted this suggestion in a specific way, a way that is arguably far more specific than Granny ever intended. In addition to their background assumption that the mental states that constitute reasons for acting are propositional attitudes, they have assumed two further theses about (i) the nature of propositional attitudes, and (ii) the conditions under which they count as reasons. First, they’ve assumed that for an agent to have an occurrent mental attitude, such as a belief or desire, toward a proposition $P$ is for her to stand in a distinctive kind of cognitive relation with a mental representation that means $P$. A subsidiary assumption is that mental representations contain component concepts whose meanings contribute to the propositional content of the mental representations of which they’re a part. So to believe that the cat is on the mat is to stand in the ‘belief’ relation with a mental representation that contains, among other things, the concept *cat*. The second assumption is that in order for propositional attitudes to exhibit the kind of rational coherence in virtue of which they’re recognizable as reasons, they must be causally related to one another, and to action, in ways that respect the implicit generalizations of folk psychology. If propositional attitudes entered into causal liaisons with other attitudes and actions that were *crazy* from the point of view of common sense, they wouldn’t, according the present view, count as reasons (Fodor, 1987, p.15). Let’s call this view *Anthropic Realism*.

Anthropic Realism gets us closer to what we wanted out of a causal theory of action. Insofar as it holds that the mental states that count as reasons are concrete particulars with semantic properties, it at least makes room for the possibility that mental states cause actions *in virtue* of their semantic properties. But we wanted to

---

4 The most influential source of the view I am about to articulate is Fodor (1987).
5 There are of course other, more familiar names for the view under consideration — such as Intentional Realism, the Representational Theory of Mind, and so on — but for reasons that I hope will soon become apparent, I think these names are descriptively inaccurate. One can be a realist about intentions and mental representations without being an *Anthropic Realist*. 
know how that possibility is made actual. Within the purview of Anthropic Realism, we want to know how the two tenets of that doctrine are connected. Fodor (1975; 1987) has an ingenious suggestion: the Computational Theory of Mind. He proposes that mental representations are structured symbolic expressions made up out of atomic symbols with semantic and syntactic properties, much like the symbol strings in a digital computer. These expressions can be manipulated by formal rules, instantiated by computational operations, that apply to an expression solely in virtue of the syntactic properties of its component symbols. Nevertheless, the formal rules are set up in such a way that they systematically preserve semantic relations between expressions, much like the derivation rules of a sound formal language. In this way, mental representations can have causal powers that ‘respect’ their semantic properties.

That’s the hope, at least. I’m skeptical that this project is workable, for reasons I won’t go into here. What I want to focus on instead is why Fodor is so invested in the project. Ultimately, he wants to help out in Granny’s garden and make room for the view that animals are capable of acting intentionally, in virtue of the way they take the world to be. But like most of his contemporaries in latter 20th century analytic philosophy, he takes this to require the ‘vindication’ of Anthropic Realism. Even opponents of Anthropic Realism have thought that its denial entails the outright

---

6A word about terminology: the term ‘Computational Theory of Mind’ is sometimes used interchangeably with ‘computationalism’. I would like to minimize the possibility of confusion by insisting on a stricter usage. I will use ‘computationalism’ to refer to the view that the explananda of cognitive processes are computational in some substantive sense; precisely what that sense is is a matter of some debate, as I will discuss in Chapter 2.2.3. I will use ‘Computational Theory of Mind’ to refer to the species of computationalism that attempts to vindicate Anthropic Realism by appealing to a ‘classical’ conception computation in the way that I am about to describe in the text. The view that I defend in this dissertation is a version of computationalism, but not of the Computational Theory of Mind.

7Fodor (1987): “Greycat is strikingly intelligent in comparison with, for example, rocks, trees, worms, and spiral nebulae. Rocks, trees, worms, and spiral nebulae are, each in its own way, quite complicated objects... Yet none of their behaviors seems remotely as clever as Greycat’s. In fact, they don’t [literally] behave at all” (p.ix). Greycat’s movements count as genuine behavior because they’re explained “by adverting to the way that Greycat takes the world to be” (p.ix).
rejection of intentional phenomena. However, I think that there are reasons to reject Anthropic Realism as a reconstruction of Granny’s view, reasons that pertain to each of the two main tenets of that view.

The problem with the first tenet is in fact the central reason Davidson’s (1975) was so reluctant to ascribe propositional attitudes to nonlinguistic creatures. The problem is that, according to Anthropic Realism, in order for a propositional attitude ascription to be true of an organism, the organism must have mastery over the concepts contained within the ascription; yet this often seems implausible, especially in the case of nonlinguistic animals. A passage from Stich (1983) illustrates this implausibility nicely:

Does Fido really believe it is a squirrel up in the oak tree? Are there not indefinitely many logically possible creatures which are not squirrels but which Fido would treat indistinguishably from the way he treats real squirrels? Indeed, does he believe, or even care, that the thing up the tree is an animal? Would it not be quite the same to Fido if he had been chasing some bit of squirrel-shaped and squirrel-smelling machinery, like the mechanical rabbits used at dog-racing tracks? The concept of animal is tied to the distinction between living and nonliving, as well as to the distinction between animals and plants. But Fido has little grasp of these distinctions. How can you say that he believes it is a squirrel if he doesn’t even know that squirrels are animals?

(Stich, 1983, p.1045)

The problem with the second tenet of Anthropic Realism is related to Sterelny’s (2003) concern about the motivation for developing an account of intentional agency that will encompass nonhuman animals. Sterelny worries that the notion of intentional agency is shaped by the role it plays in folk psychology, that folk psychology is merely a

---

8Stich expresses the point in ways that would rile Fodorians. For example, his claim that “[t]he concept of animal is tied to the distinction between living and nonliving” suggests that he is assuming a holistic theory of meaning of the kind Fodor repudiates. However, his point is independent of any such assumption. Regardless of how you think meanings are determined, the point is that, even granting that Fido has a concept of squirrelish things, it’s highly implausible that his concept is coextensive with our concept of a squirrel, i.e. the class rodents of the family Sciuridae.
species- or even culture-specific tool for interpreting the behavior of other creatures like us, and hence that attempting to apply the notion of intentional agency to nonhuman animals is “likely to be a misleading form of anthropomorphism” Sterelny (2003, p.258). Echoing Sterelny’s concern, Godfrey-Smith (2003) writes that “[i]f we think of folk psychology as a socially-evolved interpretive tool that functions to help us deal with a specific set of social tasks, then when it is used to describe non-human animals it is far from its domain of normal use” (p.267). In particular, he continues, the notion of intentional agency “is a categorization used precisely in the part of folk psychology’s domain where any quirks deriving from the social role of interpretation are most likely to be operating” (p.271).

I find this latter claim highly implausible. The capacity to distinguish between intentional agents and mindless systems is not bound to contexts of human social interaction, but is required to satisfy some of the most basic biological imperatives, such as avoiding predators and finding mates, so it is surely of much deeper phylogenetic provenance than folk psychology. However, I think that the claim about the parochialism of folk psychology is plausible, and has some troubling implications for the second tenet of Anthropic Realism. Given the parochialism of folk psychology, it is problematic to suppose that an animal’s actions only count as reason-driven to the extent that they respect folk psychological generalizations. We should surely allow, at least in principle, that animals might have psychologies utterly alien to our own, yet nevertheless be capable of acting for reasons; their own reasons.

Fodor (1987) does allow that not all of the generalizations of folk psychology must be true of an organism for it to count as an intentional agent: “one rather hopes that there will prove to be many more — and much odder — things in the mind than common sense had dreamed of” (p.15). But this doesn’t get to the heart of the worry here, which is that taking folk psychology to constrain what counts as intentional action would simply seem to get the metaphysical grounding relations around the wrong way.
What rationalizes an animal’s action isn’t determined from the outside, through the application of some idealized folk theory; it is determined from the inside, by what the animal takes to be appropriate given how it takes the world to be. Even with its proper domain, human social intercourse, folk psychology is true, to the extent that it is true, by virtue of the ways in which the contents of our mental states make certain actions seem appropriate or compelling — not the other way ‘round. Now, this talk of appropriateness could clearly do with some sharpening, but even as a blunt instrument it serves its present purpose: to carve out an explanatory role that Anthropic Realism doesn’t fill. We want to know why the contents of certain mental states make certain actions appropriate from the perspective of the agent.

The two problems with Anthropic Realism that I’ve just canvassed are really just different aspects of the same underlying problem: Anthropic Realism attempts to explicate the inner, intentional causes of action from the outside, using conceptual tools shaped by, and best suited for, the analysis of public language. This is a problem even in the case of normal humans, but focusing on agents who can’t express their intentions in language makes it vivid. The proper reaction, I think, is not to embrace some form of Eliminativism or Interpretivism about the causes of intentional action, but rather to find a new set of conceptual tools. Contrary to the assumptions of the analytic tradition, I doubt that Granny cares very much about the tools we use to theorize about intentional agency, so long as they can help us tame the thickets of radicalism and conservatism. I doubt that she cares, for example, whether intentional actions are caused by conceptually articulated propositional attitudes; she just cares that agents act — somehow — on the basis of the way they take the world to be.

In any case, the positive reasons that analytic philosophers have given for identifying the mental causes of action with propositional attitudes are remarkably thin on the ground. They seem to consist largely in the observation that ordinary explanations of action employ sentences that contain a mental state verb followed by a that-clause.
Putting aside worries about whether observations about ordinary language use can license strong metaphysical conclusions, the specific observation at issue doesn’t always hold true, either cross-linguistically or even in English (Ben-Yami, 1997). Ask Granny why the chicken crossed the road, and she might very well say something like ‘because she was hungry and saw some food over yonder’. Philosophers don’t seem to have adopted the view that the mental causes of action are propositional attitudes for strong positive reasons, but rather for negative ones; it’s just not clear what the alternative would be. To paraphrase Davidson (1999): we have many vocabularies for describing nature when we regard it as mindless, and we have a mentalistic vocabulary for describing human thought and intentional action, but what we lack is a way of describing what’s in-between.

So this dissertation is motivated in large part by the conviction that in order to articulate a robust notion of intentional agency that encompasses human and non-human animals, we must develop a vocabulary for describing what’s in-between, a vocabulary that eschews familiar terms like ‘belief’, ‘concept’, or ‘proposition’, and instead appeals to the terms employed by our best contemporary explanations of cognition. This is not because of any metaphysical qualms about beliefs, concepts, and the like; unlike Eliminativists, I do not claim to have any grounds for denying that beliefs, say, have identity conditions. Rather, my aversion is methodological; I have no idea what the identity conditions for beliefs might be. Although it is often said that consciousness is one of the few remaining scientific mysteries, and perhaps in principle eludes our understanding, it strikes me that our understanding of belief is in far worse shape. Philosophers have serious debates about whether, for example, thermostats have beliefs (Dennett, 1987), or whether only humans are so lucky (Davidson, 1975). But we can be pretty certain that thermostats aren’t conscious. Although we routinely

---

9Should we suppose that there is some thing that rains because we say ‘it is raining’?
use instruments to measure the presence or absence of consciousness — surgery would be a far scarier prospect without them — we simply have no analogous methods for empirically anchoring notions like belief, concept, or proposition.10

I’m happy to allow that once we have a detailed account of intentional agency in the terms of our best science, there might be some reasonable precisification of, say, ‘belief’, according to which the systems encompassed by the account qualify as genuine believers. As it happens, that would be my preferred way of using the term. However, I realize that there may be other reasonable precisifications of the term, according to which the systems that are encompassed by the envisioned account do not ipso facto qualify as believers. And it seems to me that there are no strong theoretical reasons to choose one interpretative scheme over the other; what matters from a theoretical perspective is that the envisioned account elucidates the phenomenon that we’re interested in: how a system can act appropriately given the way it takes the world to be. Thus, given the variable and flexible ways in which terms like ‘belief’, ‘concept’, and ‘proposition’ are commonly used, insisting on using these as theoretical terms, to elucidate the nature of intentional agency, would seem to run the risk of entering into merely terminological disputes.11

10I hope it is clear that my claim here is not that these terms are strictly meaningless. I’m not presupposing some form of verificationism or operationalism. In line with a broadly causal theory of reference, I think that we often intend to use ‘belief’, for example, to refer to some dimly-understood kind of mental state. My point is just the conditions for applying ‘belief’ are so labile and heterogeneous that it seems prudent not too rest too much theoretical weight on the notion.

11An example of the kind of terminological dispute that I think a reliance on commonsense terms like these can engender can be seen in an exchange between Hurley (1998) and Noë (2002). Hurley (1998) argues that the cognitive abilities of nonlinguistic animals are generally situation-dependent and context-bound — that they fail to satisfy the Generality Constraint — and hence that such animals lack the “richly normative conceptual and inferential capacities” possessed by normal adult humans (p.136). Nevertheless, she argues, nonlinguistic animals exhibit “islands of practical rationality” (p.136), and hence count as genuine intentional agents. She takes this to show that traditional ways of understanding intentional agency are mistaken: the capacity to act for a reason doesn’t require conceptual abilities. Noë (2002), by contrast, argues that acting for a reason does require conceptual abilities, but that our traditional ‘Socratic or Fregean’ way of understanding concepts is mistaken: even in the case of normal adult humans, concept possession needn’t require sophisticated inferential capacities, and concept application needn’t involve reflective judgment. However, once you take a step...
This doesn’t yet speak directly to the worry about motivation we opened this section with. Regardless of the vocabulary we use, why bother to describe the boundaries of intentional agency? At this juncture it’s worth pointing out that Fodor’s proposal to vindicate Anthropic Realism by way of the Computational Theory of Mind at least promised to do something quite remarkable: integrate reasons, purposes, and goals into the natural order. Since the emergence of the mechanical philosophy in the Early Modern period, adopting a broadly mechanistic conception of nature has paid remarkable dividends, enriching our understanding of otherwise puzzling phenomena and revealing them to be deeply interconnected. Nowhere is this exemplified better than in our contemporary understanding of life. Until remarkably recently, many thought that the properties of living organisms are so utterly different from those of inorganic matter that they must be the expression of an immaterial élan vital. But we now understand how the distinctive capacities of living organisms are produced by biochemical mechanisms that are ultimately made up out of inorganic materials. Descartes was an ardent proponent of the mechanical philosophy, and would have had no truck with 19th century vitalism; however, he did draw the line at the human mind. He argued that no mere machine could exhibit the creative flexibility and purposiveness that is distinctive of human mentality, and hence that the human mind is an immaterial substance, floating above the rest of nature. There are familiar reasons to think that this view is false, but this still leaves us with a puzzle: how could its contrary be true? How could a mechanistic system exhibit flexible and purposive behavior? How do we naturalize the mind?

Fodor’s answer to this question is framed in terms of Anthropic Realism, which we’ve seen is problematic. But even if we reframe the question in a way that doesn’t back and let the smoke settle, it seems clear that both authors agree about the substantive issues; they just disagree about how to use the word ‘concept’.
presuppose Anthropic Realism, the question is no less puzzling. The question, in essence, is how mental states could play a causal role that is somehow relevant to their intentional content. How is it that the chicken’s perception of food being over there causes her to cross the road? This has question has seemed especially puzzling given the widespread assumption in analytic philosophy that intentional content is essentially relational. Dretske (1989) gives voice to this assumption when he confidently asserts that “[m]eaning certainly isn’t an intrinsic property of meaningful things, something you could discover by looking in the head” (p.4). How could the fact that something internal, a mental state, stands in a relation to something external, its content, be a causally relevant fact about the internal state? In seeking to solve the puzzle about how intentional content could be relevant to the causal powers of mental states, many philosophers have focussed their attention on the question of how to explain intentional content in terms that are broadly consistent with the natural sciences: how to ‘naturalize semantics’. The conditions for answering this question have been understood in a distinctive way, expressed by Fodor (1987) thusly: “I want a naturalized theory of meaning; a theory that articulates, in nonsemantic and nonintentional terms, sufficient conditions for one bit of the world to be about... another bit” (p.97).

I am now in a position to address the motivational issue directly. The imagined scientist from Alpha Centauri, insofar as she is sufficiently curious and reflective, would be struck by a curious fact. Lots of the entities around her do things as a result of external events. Sometimes she merely does things as a result of external events, as when she startles in response to a loud sound. But some of the things she does, she does in light of a reason, her reason, “with an idea of herself as its source” (Nagel, 1979). She might wonder how that could be the case; that is, she might be struck by the aforementioned puzzle. So on yet another way of understanding this dissertation, it is the beginning of an attempt to engage with that puzzle — not an attempt to solve it, so much as a suggestion about the most promising way to approach the
solution. For I think that the traditional project of ‘naturalizing’ semantics is simply the wrong approach. For one thing, sufficient conditions for something’s having intentional content are ubiquitous; simply specify the spatiotemporal location of the state that mediates my perception of these letters and you presumably have a sufficient condition for something’s having content. But that wouldn’t elucidate how my perceptual state has content, or how that content could be relevant to the actions that my perceptual state occasions. More generally, the kinds of conditions that philosophers have provided bear little resemblance to paradigm cases of successful naturalizations in the history of science. I propose instead that if we are to naturalize intentional agency, we ought to take our lead from successful naturalization projects such as the naturalization of life. Specifically, I propose that just as life was naturalized by elucidating the biological mechanisms that explain the various capacities that are distinctive of living organisms, intentional agency ought to be naturalized in much the same way: by elucidating the neural mechanisms that explain the various capacities that are distinctive of intentional agents.

Importantly, the naturalization of life didn’t proceed by providing a definition of life consisting of a precise set of necessary and sufficient conditions; rather, our intuitive conception of life recognizes a range of different capacities that are distinctive of living organisms — such as metabolism, reproduction, and so forth — which might be exhibited in different ways and to varying degrees, and naturalization involved clarifying those capacities and elucidating their biological mechanisms. Conceptualizing the project of naturalizing intentional agency in much the same way — not as defining intentional agency, but as elucidating the neural mechanisms mediating the cognitive capacities that are distinctive of intentional agents — has a couple of notable features. First, it leaves it open that there might very well be artificial intentional agents, or in any case agents with internal information-processing mechanisms that are not nervous systems. Although the naturalization of life proceeded by elucidating the mechanisms
underlying existing earthbound organisms and biological processes, it did not rule out
that there might be artificial or extra-terrestrial forms of life; indeed, the naturaliza-
tion of life yielded an understanding of biological processes that is now being applied
in the field of synthetic biology. There’s every reason to suppose that much the same
will be true in the case of the naturalization of agency.\textsuperscript{12} The second feature of the
present conception of the naturalization of agency is that it allows that although there
is an important distinction between intentional agents and mindless systems, there is
no bright line demarcating them. As in the case of life, the capacities we associate
with being an agent might be multifarious and gradable. Are viruses alive? Are ants
intentional agents? Naturalization sheds light on these questions, but needn’t answer
them decisively.

\section*{1.3 Obstacles, and the Path Forward}

This dissertation is intended to contribute to the project of naturalizing intentional
agency by elucidating the neural mechanisms of agentive capacities. However, it would
be imprudent to immediately start work upon that project, for there at least two
major obstacles that stand in its way. These obstacles are both manifestations of a
common underlying sentiment in many circles that, to paraphrase Jerry Fodor (1998a),
‘notions like \textit{computational state} and \textit{representation} aren’t accessible in the language
of neuroscience’ (p.96). It is the primary task of this dissertation to overcome these
obstacles, and to thereby lay the groundwork for an account of the neural mechanisms
of intentional agency.

\textsuperscript{12}The underlying point here is often expressed by holding that life and mind are \textit{multiply realizable}. I
think this is probably right insofar as \textit{ontological} issues are concerned, but when it comes to natural-
ization it seems that \textit{explanatory} issues are paramount. What drives naturalization is the project of
revealing the specific and possibly idiosyncratic mechanisms mediating the capacities of the entities
around us; general ontological taxonomies arguably come after this project. To put it another way,
naturalization primarily involves revealing \textit{how things work}, not \textit{what things are}. 
The first obstacle is a system of interlocking arguments developed by the distinguished psychologist C. Randy Gallistel in a series of provocative publications over last two decades. Gallistel’s arguments are complex and multifaceted, but at their core is the conviction that the ‘connectionist’ mechanisms that contemporary neuroscientists appeal to, namely networks of interconnected neurons joined by plastic synapses, are simply incapable of explaining the capacities that are distinctive of intentional agency; they cannot explain how an animal is able to receive information from the world, flexibly integrate that information with other information into a stored mental representation, and retrieve that information when needed so as to act appropriately in the world. Gallistel sees connectionist neuroscience as the heir to a long-standing ‘anti-representational’ tradition in psychology — running from the empiricism of the Early Moderns, through the associationist psychology of the 19th century, to the behaviorism of the early 20th century — that is unified by a common assumption that learning is fundamentally associative: learning involves rewiring a plastic brain so that it can better interact with the environment, rather than recording explicit representations of the environment. Gallistel marshals a range of ‘engineering’ considerations to argue that connectionist neuroscience isn’t merely historically wedded to an anti-representational associative theory of learning, but that it simply doesn’t have the explanatory resources to avoid being committed to such a theory. He argues that well-confirmed, mathematically rigorous theories about the nature of information and computation impose ‘ineluctable’ constraints on the kind of mechanisms that are capable of flexibly recording, integrating and manipulating information under real-world time and energy constraints. Specifically, such mechanisms must be equipped with a read-write memory that is functionally equivalent to the registers of a digital computer. However, Gallistel claims, contemporary neuroscientists do not recognize anything that could play the role of such a mechanism; they hold that information is stored implicitly in the connections of a neural network via ‘associative’ plasticity mechanisms, rather
than explicitly in a functionally distinct read-write memory. Even though the plasticity mechanisms recognized by contemporary neuroscientists are clothed in shiny new bells & whistles such as STDP, underneath they’re “much the same rewiring rules that Locke suggested” (Gallistel & King, 2010, p.253). Thus, by virtue of its commitment to associationism, contemporary neuroscience “represents a resurgence of the behaviorist view that the brain does not really represent the external world” (Gallistel, 1998, p.11).

The second major obstacle to the kind of neurocomputational account of intentional agency I wish to develop is provided by a recent argument developed by the philosopher William Ramsey. Although Ramsey develops the argument for purposes that are ultimately antithetical to Gallistel’s, his arguments about the status of representations in contemporary neuroscientific theorizing nicely complement Gallistel’s. Many neuroscientists of course do in fact describe the mechanisms that they posit as representations, and locate themselves within the post-behaviorist cognitive tradition; but while Gallistel argues that these mechanisms cannot function as representations in a mechanistic sense, Ramsey argues that they in fact do not function as representations in an explanatory sense. Although neuroscientists often describe the mechanisms they appeal to as representations, Ramsey holds that when we look closely at the explanatory role that these mechanisms play within the theories in which they are posited, we see that they don’t serve as stand-ins or internal surrogates for external entities; rather, they merely serve as triggers or causal relays. For example, neurons that selectively respond to a certain class of entities, such as the famous ‘bug detectors’ in the frog’s retina, are often said to represent the entities they’re responsive to. But according to Ramsey, this representational gloss is explanatorily superfluous; all these neurons do is trigger a subsequent response. The tendency amongst neuroscientists to call their theoretical postulates ‘representations’ is, in Ramsey’s view, simply due to conceptual inertia left over from the cognitive revolution; just as cosmologists still
talked about Ptolemaic celestial spheres even after the Copernican revolution rendered such spheres explanatorily redundant, connectionists still talk about representations despite the fact that representations play no explanatory role within their theories. Now, unlike Gallistel, Ramsey is sanguine about the explanatory prospects for contemporary neuroscience, and suggests that the extant successes of neuroscience show that cognitive science is ‘moving in a non-representational direction’ — even hinting that this might eventually lead to a kind of eliminativism about intentional agency. But aside from their differences about whether contemporary neuroscience ultimately has the resources to explain cognition, both Ramsey and Gallistel agree that neuroscience is a kind of neo-behaviorism, in which, as Ramsey puts it, “cognitive representations play no real explanatory role” (2007, p.226).

Ramsey and Gallistel are in agreement about this because they have much the same conception of the nature of mental representation. They think of mental representations as internal mental models of the world. This idea is intuitively compelling because it seems to be supported by our introspective experience of ‘offline’, stimulus-independent cognitive processes such as episodic memory and mental imagery. However, like many other cognitive scientists and philosophers, they attempt to make this intuitive idea precise by explicating it in terms of abstract structural resemblance, or, in mathematical terms, homomorphism. The idea here, which I think has some claim to being the most prevalent and well-developed view about the nature of mental representations in cognitive science, is that internal neural or mental systems embody the same abstract, formal structure as the external systems they represent. This proposal has an air of mathematical rigor because it is effectively an application of the mathematical notion of representation employed in, for example, measurement theory (Krantz et al., 1971). However, one of the central tasks I undertake in this dissertation is to argue that when the ‘structural’ or homomorphism-based view of representation is properly explicated, and divested of the intuitive metaphor of a mental model, it turns
out to encompass all sorts of states, including the purportedly non-representational states posited by neuroscientists — and even some of the states posited by behaviorists such as Hull (1930) and Osgood (1956). For example, Hull (1930) seems to endorse something very much like a structural view of representation when he writes that “the world in a very important sense has stamped the pattern of its action upon [the organism]... in such a way that a functional parallel of this action segment of the physical world has become a part of the organism” (p.106). Indeed, I argue that the structural notion of representation encompasses states in mindless systems such as plants. So while the structural notion might pick out a certain kind of theoretical posit that plays an important role in various explanatory endeavors, and which deserves to be called a ‘representation’, it doesn’t pick out a distinctively mental or cognitive kind of representation, and hence — to reflect the charge back at Gallistel and Ramsey — structural representations per se are incapable of explaining intentional agency.

These arguments rest upon certain foundational theoretical concepts and experimental results. I’ll thus begin the dissertation by exploring these foundations in Chapter 2. I will first consider certain widespread ways of thinking about information, representation, and computation, which in Gallistel’s view limn the very boundaries of what it is to be a computational mechanism. In doing so, I will also identify the ‘structural’ notion of representation that both Gallistel and Ramsey take to be the most, or even the only, legitimate notion of representation in cognitive science. I then move on to discuss some of the ethological evidence that Gallistel marshals in support of his argument, considering, specifically, the ability of the Saharan desert ant to find its way home by keeping track of the distance and direction it has travelled, and the ability of the Western scrub jay to recall where and when it cached a particular type of food. Gallistel argues that the ‘connectionist’ mechanisms invoked by contemporary neuroscientists are incapable of explaining evidence like this, so I close this chapter by considering the nature of these mechanisms and the sense in which they qualify as
‘connectionist’.

Once I’ve laid the foundations, I will then begin, in **Chapter 3**, to assemble the various components of Gallistel’s argument. The keystone of the argument is the claim that connectionist mechanisms have the functional architecture of a finite-state machine, since they lack a mechanism that can store computationally manipulable vehicles of information — they lack a read-write memory. I consider why, in Gallistel’s view, none of the obvious candidates for a neurobiological read-write memory mechanism, such as synaptic weights or reverberating patterns of activity, could function as an effective read-write memory. Briefly, weights are not computationally efficacious, and reverberating patterns are not thermodynamically stable. I then discuss the various crippling computational limitations that connectionist networks allegedly suffer from as a result of not having access to an effective read-write memory. Most notably, since they lack a mechanism by which to flexibly manipulate information and thereby generate novel information, connectionist networks must, in Gallistel’s view, embody implausibly vast amounts of information within their computational architecture. I then show how these concerns culminate in Gallistel’s objection to contemporary neuroscientific conceptions of learning. Since the connectionist mechanisms assumed by neuroscientists cannot flexibly manipulate information, they cannot operate on experience to generate structurally complex representations of the world; instead they are simply molded by experience so as to interact better with the world. Thus, according to Gallistel, contemporary neuroscience cannot explain distinctively rational forms of learning, and must characterize learning as a simple reflex-modification phenomenon that is unmediated by cognitive processes.

In **Chapter 4**, I reply to Gallistel’s arguments, first by pointing out that they largely *presuppose* that contemporary neuroscience is part of an anti-representational, associationist tradition in psychology. I argue that this presumption is false: neuroscience is guided by explanatory norms that are very different from those that mo-
tivated the anti-representational behaviorism of an earlier era — neuroscientists are expressly attempting to develop mechanistic explanations of distinctively cognitive capacities, and in doing so freely posit representational structures. I then argue that Gallistel’s arguments have an appearance of theoretical rigor and legitimacy that they don’t deserve, since they misappropriate theoretical concepts from computability theory. I argue that once these presuppositions and misinterpretations have been peeled away, the arguments that lie beneath are remarkably insubstantial. In particular, Gallistel’s foundational argument that neither synaptic weights nor reverberating patterns of activity in a recurrent network could function as a read-write memory ignores the possibility, which has been the received view in neuroscience since Hebb’s (1949) seminal work, that synaptic weights and reverberating activity might be importantly complementary; computationally efficacious reverberating patterns might be stored dispositionally by a thermodynamically stable pattern of synaptic weights. I then draw from a broad palette of theoretical and experimental work in contemporary neuroscience to provide a sketch of how currently recognized neural mechanisms might not only function as vehicles of explicit information, but how those vehicles might be coordinated across different regions of the brain over time, so as to implement the computations that mediate distinctively cognitive forms of learning. This sketch is highly speculative, but it draws from an apparently productive and promising body of research that Gallistel simply doesn’t discuss. Thus, contra Gallistel, it seems reasonable to be optimistic that neuroscience is on the path toward explaining cognitive capacities.

Finally, in Chapter 5, I turn to Ramsey’s argument that, even if it turns out that neuroscience can explain cognitive capacities, it posits mechanisms that simply don’t play the explanatory role of representations, and hence might reveal that those capacities are not cognitive or intentional after all. I begin by discussing the kind of explanatory role that, in Ramsey’s view, a given theoretical posit must play in
order to be properly characterized as a genuine representation: it must function as a ‘stand-in’ for certain distal states or entities. Then, following Ramsey, I identify a type of theoretical posit that is allegedly ubiquitous in neuroscience, and which allegedly doesn’t play a representational role; these so-called ‘receptor representations’ simply function as triggers or causal relays rather than as genuine *stand-ins*, hence aren’t representations properly so-called. I then turn from exposition to criticism, and argue that when the ‘structural’ and ‘receptor’ notions of representation are properly explicated, there turns out to be no distinction between them; anything that satisfies the conditions of being a structural representation satisfies the conditions of being a receptor representation, and vice-versa. There only *appears* to be a distinction between structural and receptor representations because the former are tacitly conflated with the *mental models* that are widely thought to underly ‘offline’ cognitive processes like episodic memory or mental imagery. However, I argue that, contrary to widespread assumptions, there’s nothing distinctively *mental* about structural representations, for they are to be found in all sorts of non-intentional systems such as plants.

I conclude that these considerations motivate the need to articulate conditions for a new kind of explanatory role: what is it for something to be a distinctively *mental* representation? I embellish the sketch of computational mechanisms I that I provided earlier in response to Gallistel with complementary work on the mechanisms of offline cognition to argue that contemporary neuroscience is in fact on its way to filling that role. It thereby promises to elucidate the mechanisms of intentional agency.
Chapter 2

Theoretical & Experimental Foundations

It is certain that there may be extraordinary activity with an extremely small absolute mass of nervous matter; thus the wonderfully diversified instincts, mental powers, and affections of ants are notorious, yet their cerebral ganglia are not so large as the quarter of a small pin’s head. Under this point of view, the brain of an ant is one of the most marvellous atoms of matter in the world, perhaps more so than the brain of man.

— Charles Darwin, The Descent of Man

2.1 Introduction

Gallistel constructs an elaborate and interlocking system of arguments upon a foundation of theoretical assumptions and empirical results. So understand Gallistel’s arguments, and evaluate how well they hold up under pressure, we should first investigate their theoretical and empirical underpinnings. That is the task of the present chapter. The general observation upon which Gallistel’s argumentative structure is built is that animals are exquisitely sensitive to certain behaviorally-relevant aspects of their environment. Behaviorist methodologies artificially constrained an animal’s behavioral
repertoire, and thereby led behaviorists to radically underestimate the complexity and sophistication of the cognitive capacities of animals. But when we study animals ethologically, in the ecologies to which they are adapted, we find that the structure of their behavior often reflects quite subtle aspects of the structure of those ecologies. It is unclear how to explain this without supposing that animals in some sense receive information from the environment, store it in memory, and combine it with other information, such that it eventually comes to inform behavior that is ecologically appropriate.

The notions of information, representation, and computation that emerged from developments in theoretical computer science and electrical engineering in the 1940s and ‘50s provided a way for psychologists to understand how this might be possible, and helped dislodge behaviorist qualms about the explanatory power of ‘mentalistic’ inner states, thereby setting the cognitive revolution in motion. For many cognitive scientists, these notions played a largely illustrative or heuristic role by suggesting, in broad outline, what a mechanistic explanation of cognition might eventually look like; on this view, the information-processing concepts that best describe yet-to-be-discovered biological mechanisms of cognition in animals might be a relative of, but nevertheless distinct from, the information-processing concepts derived from computability theory and electrical engineering. However, Gallistel argues that the only definitions of information, representation and computation we have that are sufficiently precise and mathematically rigorous to play a serious explanatory role in psychology are those that derive from the aforementioned developments in computer science and engineering, and that these definitions apply just as well to the biological mechanisms of cognition as to human-engineered artifacts. Indeed, he argues that these definitions impose inviolable constraints on the nature of any physical information-processing mechanism that is capable of encoding, storing and flexibly manipulating information.

The purpose of section 2.2 is to provide a brief introduction to the notions of
information, representation, and computation as Gallistel understands them, and to describe the constraints that Gallistel thinks they impose. The upshot is that, in Gallistel’s view, if a computational mechanism is to be both powerful — in the sense that it can solve an interesting range of computational problems — and practical — in the sense that it can solve those problems in a range of novel circumstances given ‘real-world’ physical constraints — then it must be capable of encoding novel information using an efficient, ‘compact’ coding scheme, such as the binary encoding of number; it must be capable of storing encoded information in a thermodynamically stable form for indefinitely long durations; and it must store information in a ‘readable’ form such that it can be flexibly fetched from memory and transported to the mechanisms that perform basic computational operations when needed. In short, a powerful and practical computational mechanism must have access to a read-write memory that is functionally distinct from the mechanisms that perform computational operations. Equivalently, in Gallistel’s view, such a mechanism must have the flexible architecture of a Turing machine, rather than the fixed architecture of a finite-state machine or a look-up table.

Gallistel draws from a wealth of ethological evidence to argue that animals routinely solve complex computational problems of the kind that can only be solved by powerful and practical computing mechanisms, and hence that animal brains have the functional architecture of Turing machines — that is, they contain a functionally distinct read-write memory mechanism. In section 2.3, I briefly survey some of this ethological evidence. In particular, I consider the ability of the Saharan desert ant to find its way home by keeping track of the distance and direction it has travelled, and the ability of the Western scrub jay to recall where and when it cached a particular type of food. As Gallistel rightly emphasizes, it is unclear how to explain abilities like these without supposing that animals are somehow capable of recording the values of specific environmental variables, such as distances, durations, and locations; that they can store
these values for considerable periods of time; and that they can flexibly integrate these values with other stored values so as to generate new, behaviorally-relevant values, such as the direction of home or the location of a recently stored cricket, that were merely implicit in the disparate input values.

According to Gallistel, the kinds of mechanisms recognized by contemporary neuroscientists are incapable of satisfying the ‘inviolable’ constraints on powerful and practical information-processing mechanisms, primarily because, in his view, they lack a read-write memory. Thus, he argues, contemporary neuroscience cannot explain the aforementioned ethological evidence; such evidence can only be explained by the hypothesis “that the nervous system possesses a read-write memory mechanism that performs the same essential function performed by the memory registers in a computer” (Gallistel, 2008, p.229). Gallistel presents a series of interlocking arguments for this conclusion, which I’ll discuss in Chapter 3, but before addressing these arguments, I’ll first try to identify their target. I’ll thus devote section 2.4 to a discussion of the contemporary connectionist orthodoxy in neuroscience: its historical origins, its experimental underpinnings, and its theoretical content.

2.2 Information Processing from an Engineer’s Perspective

The notions of information, representation, and computation have clearly played a tremendously important role in the history and sociology of cognitive science. Ralllying around these notions helped psychologists break free from the methodological fetters imposed by behaviorists. However, these notions are often said to play a foundational explanatory role in cognitive science. Yet these notions are highly contested, and used in different ways by different researchers. Moreover, while they’re intended to apply to states or processes inside the minds or brains of cognitive agents, they’re
continuous with loose commonsense notions that apply in the first instance to agents themselves, as in the case of a person interpreting a representation, or performing a computation.\footnote{Indeed, it’s this latter aspect that has made these notions both attractive to cognitive scientists, by promising a way to overcome behaviorism, and problematic, by raising a question about whether that promise can be fulfilled.} So one might wonder whether these notions can be made sufficiently rigorous and precise to play the kind of substantive explanatory role in cognitive science that they’re often taken to play. Gallistel argues that they can, by drawing from various theoretical developments in engineering, computer science, and mathematics that emerged around the middle of the 20th century; most notably, the mathematical theory of communication (e.g. Shannon, 1948), computability theory (e.g. Turing, 1936), and the representational theory of measurement (e.g. Krantz et al., 1971). Indeed, Gallistel claims that these theoretical developments aren’t just useful tools for clarifying the notions of information, representation, and computation; they’re the only precise ways of explicating those notions we have. He claims, in effect, that they limn the conceptual boundaries of those notions, imposing ‘ineluctable’ constraints on what it is to be a system that can receive, encode and process information. In this sub-section, I’ll review Gallistel’s conception of these central theoretical notions, and the explanatory constraints that he thinks they impose.

### 2.2.1 Information

The idea that the nervous system receives information from the external environment and encodes it in the form of neural signals seems to have first appeared in the pioneering work of the British electrophysiologist Edgar Adrian in the 1920s (Garson, 2003). However, Adrian didn’t employ a precise, quantitative notion of information, and didn’t employ the notion to explain distinctively cognitive capacities. The notion of information that came to be influential in the confluence of disciplines we now
know as cognitive science derived primarily from the mathematical theory of communication developed by Claude Shannon (1948).\(^2\) Shannon developed his theory in the wake of World War II, and was motivated by practical engineering concerns, made especially salient by the War, about how to reliably communicate messages across an unreliable medium of communication, or ‘channel’. As Shannon put it, “the fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point” (1948, p.379). It’s worth noting that these ‘points’ might be spatial or temporal: messages might be communicated across space, as in the transmission of a signal across a telephone line, or they might be ‘communicated’ across time, as in the data stored on a computer’s hard disk. In confronting the general problem of how to communicate messages across an unreliable channel, engineers prior to Shannon had realized that this problem might be ameliorated by locating the channel within a communication system, a system in which a ‘transmitter’ encodes messages into a signal, which is communicated across the noisy channel, and subsequently decoded by a ‘receiver’ that inverts the initial encoding function and thereby reconstitutes the initial message. A communication system like this enables errors introduced by the noisy channel to be detected and corrected.

Shannon’s central contribution was to provide a set of quantitative mathematical tools for analyzing the reliability of channels and the efficiency of the codes employed by a communication system. One of the most important of these tools was his measure of the amount of information carried by a signal. The intuitive idea behind this measure is that information is equivalent to the reduction in a receiver’s uncertainty about the particular message that was communicated across a channel. Shannon proposed that to measure such uncertainty, we can ignore any specific content that the

\(^2\)Shannon’s theory of course has many important precursors. Perhaps most importantly, Shannon’s measure of uncertainty derives from pioneering work by Hartley (1928). I’ll just stick with the potted history here.
message might have, and simply assume that a message is drawn from a set of other possible messages, each of each has a certain specific probability of occurring. Thus in mathematical terms, we can regard an unknown message as a discrete random variable \( X \), with \( n \) possible values \( \{x_1, x_2, ..., x_n\} \), and an associated probability mass function \( p(X) \).³ Moreover, the receiver is assumed to know, in some sense, the range of possible messages that might be communicated, i.e. the support of the probability mass function.

Now, suppose we have a communication system in which there is only one possible message. Intuitively, the receiver in such a system would have no uncertainty about the identity of the message that a given signal communicates, so upon receiving a particular signal, her uncertainty about the message it communicates would not be reduced. To reflect this intuitive idea, we’d like a measure of the reduction of uncertainty — that is, of the amount of information a signal communicates — to return a value of zero in this case. But what if we have a communication system in which there are multiple possible messages, with different probabilities of occurring? Shannon showed that any measure of uncertainty that satisfies certain intuitive constraints of the kind I just mentioned has the following general form:

\[
H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)
\]

This equation is Shannon’s measure of the amount of uncertainty associated with the random variable \( X \) — or, to use the jargon, the entropy of \( X \). Because it is expressed using the binary logarithm, the units of this measure are called binary digits, or bits for short. Gallistel & King (2010) provide a nice illustration of how this equation works, by applying it to the story of how Paul Revere warned the patriots in the

³For expository purposes, I’ll focus on Shannon’s measure of uncertainty as it applies to discrete random variables, but it should be noted that his measure applies equally well to systems of continuous random variables.
American Revolution of the impending British invasion. Revere had arranged with his
friend Robert Newman, who was situated in a belfry tower across the Charles river and
could see the movement of the British troops, to communicate by lantern flashes the
direction from which the British were approaching: one flash if by land, two if by sea.
Now, suppose Revere truly believed that the British were much more likely to come
by land than by sea. To make things more concrete, suppose that the probability that
the British were coming by land was \( p(x_{\text{land}}) = 0.9 \), and hence that the probability
they were coming by sea was \( p(x_{\text{sea}}) = 0.1 \). Then Revere’s \textit{prior} uncertainty, before
receiving the lantern signal, is given by the above formula as follows:
\[
-(0.9 \log_2 0.9) + (0.1 \log_2 0.1) \approx 0.47 \text{ bits.}
\]
Thus Revere is slightly less uncertain about the message
he will receive than if he had believed the messages were equiprobable, in which case
his uncertainty would have been 0.5 bits. Now, assuming that Revere trusts his friend
in the belfry tower, that he is able to perfectly distinguish the two signals from one
another and so forth, then his uncertainty after receiving one signal or the other —
his \textit{posterior} uncertainty — will be 0 bits. That is, he will be certain of the direction
from which the British are invading. The amount of information communicated by a
given signal from the belfry tower will thus be Revere’s prior uncertainty minus his
posterior uncertainty, namely \( 0.47 - 0 = 0.47 \) bits.

This example also serves to illustrate the importance of the encoding and decoding
procedures in communication systems. Robert Newman, \textit{qua} transmitter, encodes a
given message, either ‘land’ or ‘sea’, into a signal consisting of either one or two lantern
flashes. This signal is only useful to Revere — it can only communicate information to
him — insofar as he can invert the encoding effected by the transmitter. \textit{Qua} receiver,
Revere must ‘know’ two things: the probabilities of the various messages he might
receive, and the method by which to decode signals back into messages. Perhaps
Shannon’s most important insight was that there are intimate connections between
the efficiency of information communication and the coding scheme employed within
a given communication system. We can begin to see the nature of that connection by considering the following, highly contrived, modification of the Revere example. Suppose there were ten possible directions from which a single British soldier might come, numbered 1 to 10, and that we wish to communicate the direction from which a given soldier is coming using a coding scheme in which the number associated with a particular direction of origin corresponds to the number of lamp flashes — so, for example, ten flashes would communicate that a soldier is coming from the tenth direction. Now suppose that the prior probability of a soldier coming from the tenth direction is very high, say \( p(x_{10}) = 0.9 \). In this case, the majority of messages the transmitter sends will employ signals that use up the most amount of time and energy. Clearly it would be much more efficient to use a code that reflects the probabilities associated with the various possible messages, such that the most likely message would be encoded by one flash, the second most likely by two, and so forth.

Gallistel holds that Shannon’s theory of communication provides only notion of information we have that is sufficiently precise and mathematically rigorous to play a significant explanatory role in science. Thus if we are to explain cognition in terms of the brain’s capacity to receive, store, and process information — as Gallistel thinks we must — the aforementioned points about the nature and efficiency of communication systems impose significant constraints on theories about neural structure and function. Gallistel thus applies the schema of a communication system directly to the brain, arriving at the following information-theoretic conception of world-brain communication, depicted in Figure 2.1 below: ‘messages’ are distal stimuli in the external world, which emit energy that impinges upon the receptors of an animal’s sensory organs. The receptors function as transmitters, encoding distal messages into a sensory signal, physically realized as a neural spike train, which is degraded by various sources of biophysical noise and potentially confounded with other signals carrying distinct messages. The function of the ‘cognitive system’ of the brain is to decode this signal
by inverting the encoding process, so as to reconstitute the message and disambiguate it from other messages it might have been confounded with. These reconstituted messages are physically realized as explicit symbols, which can be combined into structured symbolic representations and manipulated by computational operations so as to eventually generate ecologically appropriate behavior (Gallistel & King, 2010, p.23).

Figure 2.1: Gallistel’s information-theoretic view of world-brain communication. Distal stimuli emit various forms of energy, which are encoded into signals by receptors in sensory organs. These signals are degraded by various forms of noise, and potentially confounded with other signals. The ‘cognitive system’ of the brain decodes these sensory signals into symbolic representations of the distal stimuli. From Fig. 1.4 of Gallistel & King (2010, p.22), itself adapted from Fig. 1 of Shannon (1948, p.7).

I’ll discuss Gallistel’s conceptions of representation and computation shortly, but first I should say more about the issue of coding, since it looms so large in Gallistel’s objections to the mechanisms of information processing recognized by contemporary neuroscientists. As I mentioned, Shannon’s work highlighted the intimate connection between coding schemes and the efficiency of information communication. Let’s explore that connection by comparing a few different schemes for encoding the natural numbers using the symbols 0 and 1. One possible scheme, called a unary code, simply maps a given number to the same number of 1s. So, for example, the unary encoding of 5 is 11111. The length of a string generated by a unary code is linearly proportional to the value of the number that is encoded. Now consider an alternative system, a binary code, where a number is encoded as a binary numeral. In this system, 5 is encoded
by 110. Note that the length of a string generated by a binary code is proportional to the binary logarithm of the number that is encoded. For example, the number 5 is encoded by \( \log_2(5) \propto 3 \) symbols. Taking the number of symbol tokens required to encode a number as the criterion of efficiency, the binary code is vastly more efficient than the unary code. This is perhaps not obvious when dealing with small numbers like 5, but consider what happens when we try to encode a much larger number — say, 1 million. The unary code would require 1 million symbol tokens, whereas the binary code would require only twenty. The binary code is thus considerably more efficient, or, as it’s sometimes put, compact.

One difference between these coding schemes is that the unary scheme obeys what Gallistel calls the analog principle: “there is a natural ordering on the symbols that corresponds to the ordering of [what] they encode” (Gallistel & King, 2010, p.76). There are murky metaphysical questions what exactly a ‘natural’ ordering amounts to, but we can put such questions aside since I think the idea is clear enough for present purposes: symbol strings generated by the unary code are ordered by an intrinsic physical property, their length, which reflects the order of the natural numbers. However, the strings generated by the binary code are not ordered by their intrinsic properties in a way that reflects the order of the natural numbers; there’s no sense in which, say, the binary encoding of 5 is intrinsically smaller than the binary encoding of 6. However, that’s not to say that the mapping from natural numbers to binary bit strings is arbitrary, for the mapping is determined by a well-defined generative procedure.\(^4\) By contrast, consider what Gallistel calls a ‘nominal’ coding scheme, a scheme in which the mapping from messages to signals is entirely arbitrary. A fragment of a nominal coding scheme that maps numbers to bit strings is as follows:

\(^4\)Namely, divide the number by two, take the ceiling of the remainder to be the least significant bit of the bit string, and repeat this process on the integer result until the result of further division is zero.
<table>
<thead>
<tr>
<th>$n$</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>001</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>010</td>
</tr>
<tr>
<td>3</td>
<td>101</td>
</tr>
</tbody>
</table>

Table 2.1: Example of a nominal scheme for encoding the integers 0 to 3 into a binary string.

In a nominal coding scheme like this, the bit strings are effectively labels for the numbers they encode, and there is no generative procedure by which numbers are labeled, i.e. mapped to strings; the only way to effect the mapping is to exhaustively list each number-string pair. In Gallistel’s terminology, the procedure for effecting a nominal encoding is not compact. One disadvantage of a nominal coding scheme like this is that the non-compactness of the procedure for effecting the encoding typically infects procedures that subsequently employ the code. For example, suppose we’d like to devise a procedure for performing addition on the integers using the nominal coding scheme mentioned above. We’d like our procedure to map two bit strings, which encode integers $a$ and $b$ according to our nominal coding scheme, to a single bit string, which encodes the integer $c$ according to our scheme, such that $a + b = c$. To take one instance of such a mapping, we’d like our procedure to map 100 and 010 to 101, thereby computing $1 + 2 = 3$. Now, there’s clearly a simple rule or procedure that would effect this particular mapping; but there’s no single rule or procedure that would effect this mapping along with all of the indefinitely many other required mappings. A procedure for implementing addition using our nominal coding scheme would have to list each input-output pair individually in a look-up table. That is, the procedure would not be compact. The general point here is that, as Gallistel & King (2010) put it, “[t]here is a tight bond between the encoding procedure that generates the symbols

---

5Note the distinction between a compact coding scheme, such as the binary coding of numbers, and a compact procedure for effecting a coding.
and the procedures that act on them” (p.99).

Taken together, these basic principles about efficient information coding and transmission suggest that the method by which information is coded in an information-processing mechanism has tremendous practical implications, both in terms of the physical resources that the mechanism needs to operate, and the kinds of computational procedures it can practically perform. They suggest that insofar as such a mechanism can encode and flexibly manipulate a large number of messages, it will employ compact coding schemes that support compact computational procedures.

2.2.2 Representation

According to the picture of world-brain information transmission we’ve just sketched, ‘messages’ from distal stimuli are encoded into neural signals, which are then decoded into explicit symbolic representations, thereby communicating messages to the brain in such a way that they can be computationally manipulated. But what is the nature of these symbolic representational vehicles of distal messages? Shannon’s theory of communication provides tools for quantifying the amount of information a given message communicates, and the efficiency of the channels over which information is communicated, but it is entirely neutral with respect to the question of what a given message ‘means’, what it is ‘about’, or what it ‘represents’. As Shannon (1948) put it, “[t]hese semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages” (p.379). Understood solely in terms of its theoretical content, the mathematical theory of communication is simply a branch of probability theory.

How do we get representations into the picture? What is a representation in the relevant sense? The term ‘representation’ is used in a multitude of ways in ordinary language, perhaps canonically to refer to public phenomena that are interpreted by agents, such as words or pictures. The notion of mental representations — private en-
entities, internal to the mind, that are in some sense about the external world — probably emerged in a recognizably contemporary form in the Early Modern period under the guise of the ‘way of ideas’. Ideas were taken to be essentially conscious, introspectable entities, and something like this conception of mental representation was employed in psychological theorizing throughout subsequent centuries. This broad notion of mental representation had always had its critics, but skepticism about the explanatory utility of the notion mounted in the early 20th century, and it was rejected as being inadequate for, or irrelevant to, rigorous scientific explanations of behavior. A descendant of the notion re-emerged during the cognitive revolution under the guise of ‘cognitive representation’. These entities were no longer regarded as essentially conscious or introspectable, but they were widely regarded as meaningful — as representations of the external world. Partly because of this metaphysically puzzling semantic aspect, and partly because of a general lack of clarity about the diffuse notion of representation, cognitive representations have not escaped the controversies surrounding their mentalistic predecessors, and there is no general consensus about what exactly cognitive representations are, or what explanatory role they play.

However, there is a general conception of cognitive representation that commands wide assent in philosophy and psychology, and Gallistel has done more than anyone to defend and explicate it. Gallistel holds that the core of a concept of representation that is sufficiently precise and rigorous to play an explanatory role in psychology can be found in the mathematical notion of representation: a homomorphism, or structure-preserving mapping from one mathematical structure to another. Perhaps the most historically and theoretically important example of a homomorphism is that between geometry and algebra. As Descartes and Fermat independently discovered in the 17th century, geometry

---

6Much the same view has also been developed and endorsed by Cummins (1996), Palmer (1978), Shepard & Chipman (1970), and Swoyer (1991), among many others.
century, geometric entities like lines and curves can be perfectly mirrored, within a coordinate system, by algebraic expressions like vectors and equations. Algebra thus ‘represents’ geometry, enabling one to reason about geometric entities by manipulating algebraic expressions. In fact, the relation goes in the other direction, too; one can reason about algebraic structures by manipulating geometric entities. The relation between geometry and algebra is thus a specific type of homomorphism called an isomorphism: intuitively, a homomorphism that ‘goes both ways’. This is a very strict requirement: for two mathematical structures \( A \) and \( B \) to be isomorphic, they must instantiate precisely the same abstract structure. Homomorphisms are more permissive; to say that there is a homomorphism from structure \( A \) to structure \( B \) is just to say that all the structure in \( A \) is preserved in \( B \); \( B \) might nevertheless have additional structure.\(^7\)

Gallistel suggests that we ought to understand cognitive representation in essentially the same way we understand representation in mathematics; namely in terms of homomorphisms. His suggestion, in broad outline, is that cognitive representations are systems in the mind or brain that are homomorphic with — i.e. embody the same abstract structure as — certain systems in the external world. Now, ‘homomorphism’ is a technical term in mathematics that is arguably only well-defined in the context of discussing formally-specified abstract entities, so one might wonder how it might apply to cognitive representations, which are presumably concrete, physical systems. Indeed, one might worry that such an application would be meaningless, or at best metaphorical.\(^8\) However, the standard theory of measurement (e.g. Krantz et al., 1971) seems

\(^7\)To put this all slightly more precisely, a homomorphism from one set-theoretic structure, \( A \), to another, \( B \), is a mapping \( \phi : A \to B \) from the objects in \( A = \{a_1, ..., a_n\} \) to the objects in \( B = \{b_1, ..., b_n\} \), such that for any relation \( R \in A \), if \( R \) obtains for a subset of the objects in \( A \), \( A' = \{a_i, ..., a_j\} \), there is a relation \( S \in B \), that obtains for a subset of the objects in \( B \), \( B' = \{\phi(a_i), ..., \phi(a_j)\} \). An isomorphism is a homomorphism that is bijective, i.e. one-to-one and onto.

\(^8\)See van Fraassen (2008) for an expression of this sort of worry.
to apply the notion of homomorphism to ‘empirical’ systems in a precise, literal, and theoretically fruitful way. Very roughly, this theory explains measurement in terms of homomorphisms from real-world physical magnitudes and relations to numbers and numerical operations. Indeed, Gallistel often emphasizes the analogy between his conception of representation and measurement theory (e.g. 1990b, p.2). So perhaps the worry about whether there could be homomorphisms between physical systems isn’t so problematic. But there is a related question in the vicinity concerning how a mere homomorphism from one physical system to another could explain cognitive representation. For surely such homomorphisms are ubiquitous. This is where the distinctive features of Gallistel’s account of cognitive representation emerge. Gallistel doesn’t explicate cognitive representation merely in terms of physical homomorphisms; he explicates them in terms of functioning homomorphisms.

According to Gallistel, a functioning homomorphism from a representing system $A$ and a represented system $B$ obtains when the following three conditions are satisfied: first, there is a homomorphism from $A$ to $B$; second, this homomorphism is established and sustained by causal relations between the two systems, such that variations in $A$ are responsive to variations in $B$; and third, $A$ interfaces with an agent’s motor control systems such that it can guide the agent’s behavior with respect to $B$ in ways that reflect the relevance of $B$ for the agent. In other words, $A$ tracks changes in $B$ and is used to guide behavior with respect to $B$; in this way the homomorphism from $A$ to $B$ is exploited. It is a functioning homomorphism. As Gallistel & King (2010) put it, ‘functioning homomorphism’ is “the two-word definition of a representation” (p.55).

---

9As do other proponents of similar views. See, for example, Palmer (1978): “A representational system is an ordered triple consisting of the two relational systems and a homomorphic function that maps the represented objects into the representing objects. The basic approach is similar to that used in measurement theory for numerical representation” (p.267).

10Note that this way of putting things is somewhat misleading. A functioning homomorphism, qua homomorphism, is an abstract relation between a representing and a represented system, whereas a representation is presumably a concrete particular, i.e. the representing system. The conflation here
The general conception of representation here is pithily expressed by Gallistel (1990a) as follows: “To fit behavior to the environment, the brain creates models of the behavior-relevant aspects of the environment. The formal properties of the processes that realize these models reflect the formal properties of the corresponding external reality” (p.27). As I’ll discuss in much more detail in Chapter 5, this general conception of representation is widely endorsed throughout philosophy and psychology. For example, Palmer (1978) holds that a “representation is, first and foremost, ... some sort of model of the thing (or things) it represents”, where a model involves a “correspondence (mapping) from objects in the represented world to objects in the representing world such that at least some relations in the represented world are structurally preserved in the representing world” (p.266-7). But how exactly does Gallistel understand these internal models or ‘functioning homomorphs’, and how do they fit into his broader information-theoretic view of world-brain communication?

For Gallistel, functioning homomorphs are comprised of symbols and computational operations defined over them.\(^{11}\) Representing systems are thus *symbolic* systems. Symbols as Gallistel understands them are physical vehicles of explicit information, the embodiment of Shannon’s ‘messages’ about distal stimuli in the world. They *refer* to distal stimuli by virtue of the fact that they’re causally connected to those stimuli via channels of communication. Two additional facts about symbols are especially important to highlight given the role they’ll play in subsequent discussions. First, symbols are stable and enduring physical states that serve to carry information forward in time, unlike signals, which transiently carry information forward through space. Symbols are thus elements of *memory*. Second, symbols are physical tokens that can be

---

\(^{11}\)For the clearest and most explicit expression of Gallistel’s views about these matters, see Gallistel (2008, pp.229-39), Gallistel & King (2010, pp.55-58, 78).
manipulated by computational operations and combined with other symbols to generate new symbols and more complex symbolic structures. Symbols are thus elements of *computation*. How does all pertain to the relation between symbols and functioning homomorphisms? It’s the fact that symbols can be manipulated by computational operations that enables them to play a role within structurally complex symbolic systems that are homomorphic with systems in the external world. Moreover, it’s the fact that symbols are in causal contact with the external world via channels of communication that enables the homomorphisms they participate in to be established and sustained. As Gallistel (2008) puts it, “Symbols and the processes that operate on them create functioning homomorphisms between the symbol system and the aspects of the world to which the symbols refer” (p.230).

To illustrate Gallistel’s notion of a functioning homomorphism, and to forestall some possible misunderstandings about his conception of a symbol, let’s consider one of his central examples of a functioning homomorph, the hydraulic path integration mechanism depicted in Figure 2.2. As I’ll discuss in much more detail in section 2.3, path integration is a method of navigation by which the direction and distance that one travels is successively integrated over time into a representation of one’s location with respect to a point of origin. Any mechanism capable of performing path integration must be capable of measuring direction and distance. Without going into unnecessary details, I’ll briefly describe how the mechanism depicted in Figure 2.2 is capable of achieving at least one aspect of path integration. The mechanism is imagined to be located within a motorboat. The driveshaft of the boat is connected to a pump that drives water into a reservoir at a rate that is proportional to the rotational speed of the driveshaft. The water pressure in the reservoir thus serves to measure the speed of the boat. Water flows out of the reservoir via four tubes, each connected to a bucket that corresponds to one of the four compass directions. The outward flow is regulated by a set of valves connected via a movable plate to a magnetic compass, in such a way
that the degree to which the boat is oriented in a given compass direction corresponds
to the rate of water flow into the bucket associated with that direction. The rate of water flow into a bucket thus serves to measure the speed that the boat is traveling in a given direction, and the accumulation of water in a bucket over time effectively integrates speed, to measure the total distance traveled in a given direction.

**Figure 2.2:** Gallistel’s example of a representational mechanism that mediates a functioning homomorphism: a hydraulic mechanism that performs path integration by accumulating water into four buckets, each corresponding to a given compass direction (collectively labeled ‘Integrator’). The mechanism, imagined to be located within a motorboat, is constructed in such a way that the amount of water that accumulates in a given bucket corresponds to the total distance the boat has traveled in the compass direction associated with that bucket. Collectively, the water levels in the buckets reflect the location of the boat within a Cartesian plane with the compass directions as its axes. From Fig. 11.5 of Gallistel & King (2010, p.204).

Gallistel holds that the water levels in the buckets are symbols that participate in a functioning homomorphism. They are stable but manipulable physical states that embody explicit information about the total distance the boat has travelled in a given compass direction. Viewed abstractly, the possible states of the various buckets, and
the possible manipulations of those states by the path integration mechanism, constitute a representing system that is homomorphic to a real-world represented system, namely the possible locations of the boat with respect to its origin within a given environment, as represented by a Cartesian coordinate system. Specific states of the representing system are causally responsive to specific states of the represented system by virtue of the channels of communication from stimuli in the represented system to symbols in the representing system. Moreover, though I haven’t described the mechanism by which it could do this, the representing system could in principle be used to determine the direction and distance back to the boat’s point of origin, thereby serving to guide behavior with respect to the represented system. So it seems that the representing system — the path integration mechanism — doesn’t merely participate in a homomorphism, it participates in a functioning homomorphism. It counts as a representation in Gallistel’s sense.

This example serves to illustrate that symbols as Gallistel understands them lack many of the features that are commonly associated with symbols in cognitive science, by proponents and detractors of symbols alike. It’s important to highlight these differences now, to avoid any merely terminological disputes later. One feature that’s commonly associated with symbols is that they are themselves representations. This idea figures prominently in the work of Jerry Fodor, who identifies certain conceptual representations with unstructured, atomic symbols in a ‘language of thought’. A closely related feature associated with symbols is that they have a certain kind of content, namely conceptual content. It’s not clear what exactly this amounts to, but the intuitive idea is that symbols express the kind of ‘high level’ meaning expressed by lexicalized concepts, a meaning that might be articulated by someone with mastery over such concepts. It’s this feature of symbols that certain neo-connectionists seem to have in mind when they hold that the nodes of their networks operate at the ‘sub-symbolic’ level: they “correspond semantically to fine-grained features below the level
of the concepts consciously used to describe the task domain” (Smolensky, 1988, p.1). A third feature associated symbols derives from Charles Peirce’s influential tripartite theory of signs, according to which symbols are signs that bear a purely conventional, interpretation-dependent relation to their objects. A fourth and final feature, closely related to the one just mentioned, is that symbols are discrete tokens rather than analog magnitudes.

Symbols as Gallistel understands them have none of these features, at least none of them essentially. First, symbols for Gallistel are not themselves representations, they are *constituents* of representations; it’s only when symbols play a role in a larger symbolic system that is homomorphic with some eternal system that they play a representational role. Second, Gallistel’s symbols needn’t express conceptual contents; the examples he gives suggest that symbols typically ‘refer’ to physical magnitudes. Third, symbols in Gallistel’s sense aren’t symbols in Peirce’s sense; on the contrary, they are what Peirce called *indexes*: roughly, signs that are causally related to their object. Finally, symbols according to Gallistel *can* be discrete, but as the example of the hydraulic path integrator shows, they might also be analog; whether they are discrete or analog depends on the scheme by which they were encoded, in particular, on whether or not the coding scheme employs the ‘analog principle’ described in section 2.2.1.

What’s really crucial for whether something counts as a symbol in Gallistel’s sense is that it can play a role in a functioning homomorphism in the way I’ve described. In order for a physical mechanism to represent anything, according to Gallistel — in order for it to be a functioning homomorph — it must contain physical states that encode explicit information about distal stimuli; these states must be capable of persisting in a thermodynamically stable form, enabling the information they encode to be ‘written’ to memory; and these states must be manipulable and causally efficacious so the information they encode can be ‘read’ by computational operations. As we’ll soon see,
Gallistel holds that these requirements impose significant constraints on theories about the mechanisms of representation in animal brains. But to see the significance of those constraints, we must first get a clearer view of what Gallistel means by ‘computation’.

### 2.2.3 Computation

The notions of computation that Gallistel employs arose within a specific historical and theoretical context, and to fully understand those notions, we need to understand the context in which they developed. The term ‘computer’ today is typically used to refer to a certain kind of technological artifact, such as the laptop you might be using to read this dissertation. But for centuries before the development of computing technology, ‘computer’ was used to refer to a person performing an effective procedure, a step-by-step process that can be carried out in a finite number of steps without requiring any special insight, for the purpose of obtaining the value of some mathematical function. The rote, step-by-step process of long-division is an effective procedure for obtaining the value of the division function, and a person carrying out is procedure is, in this earlier sense, a computer. A human computer in this sense is sometimes called a *computor*.

At the beginning of the 20th century, David Hilbert posed a question about the scope of effective procedures that has come to be known as the *Entscheidungsproblem*: is there an effective procedure that can, in principle, determine whether a statement in an arbitrary formal language is true or false? Since the notion of an effective procedure is an imprecise and intuitive one, the *Entscheidungsproblem* could only be addressed in a mathematically rigorous way by providing a formal, operational definition of ‘effective procedure’. Alan Turing provided such a definition in his pioneering (1936) paper, in the form of what has come to be known as the ‘Turing machine’ — effectively, a precise, mathematical description of the essential features of a computor carrying out an effective procedure that abstracts away from inessential features such as physical
resource limitations. For example, the paper upon which a computer might write when carrying out, say, long division, is idealized as an infinite tape of paper divided into squares upon which discrete symbols might be written, which Turing formally defined as an indefinitely extensible tuple that can contain elements drawn from a particular finite set.

Turing used his formal definition to prove that there are functions that cannot be computed by any Turing machine, and — assuming that the Turing machine is indeed an adequate formalization of the intuitive notion of an effective procedure — thereby established that the Entscheidungsproblem must be answered in the negative. The claim that Turing machines are an adequate formalization of effective procedures, in the sense that they can compute all and only the functions that could in principle be carried out by a computer without regard to physical resource limitations such as the computer becoming tired or running out of paper, has come to be known as the Church-Turing thesis, and is almost universally accepted by computer scientists. It is evidenced by, among other things, the fact that various alternative formalizations of effective procedures — such as the lambda calculus, recursive functions, Post machines, and so on — have all proven to be computationally equivalent to Turing machines. In the language of computability theory, the branch of discrete mathematics to which these theoretical notions belong, these various formalizations of effective procedures all have the same computational power: they all compute precisely the same set of

\[12\] That is, 'inessential' relative to the specific purposes for which Turing developed his definition of an effective procedure, namely to limn the boundaries of what can be effectively computed in principle.

\[13\] Note that the Church-Turing thesis is a thesis rather than a mathematical theorem, since it concerns the informal notion of an effective procedure, and hence cannot be formally proven. Moreover, the Church-Turing thesis properly-so-called should be distinguished from the 'physical' Church-Turing thesis (Gandy, 1980), which entails that some Turing machine can compute any function that can be computed by a physical mechanism. This latter thesis is conceptually distinct from the Church-Turing thesis, and is arguably far less plausible, yet there is a widespread tendency throughout cognitive science to conflate the two theses (Copeland, 2000). Gallistel is guilty of this common error when he writes that the Church-Turing thesis is the view that "a Turing machine can compute anything that can be computed by any physically realizable device" (Gallistel & King, 2010, p.124).
functions, the *effectively computable* functions.

Now, there are formal specifications of effective procedures that are provably less powerful than Turing machines. Most notable is the finite-state machine (FSM), which is essentially a Turing machine that, informally speaking, doesn’t have access to a tape upon which it can write symbols. A FSM simply transitions from one state to another on the basis of the current input it receives, but it cannot, as it were, record its ‘working’ on a sheet of paper. This drastically limits the computational power of FSMs; the class of functions that FSMs can compute is provably a proper subset of the effectively computable functions. A textbook example of a simple function that no FSM is capable of computing is the following. Consider a ‘mirror language’: a set containing strings that consist of \(n\) symbols of a given type (say, \(A\)), followed by \(n\) symbols of a different type (say, \(B\)), for all positive integers \(n\), i.e. \{\(AB\), \(AABB\), \(AAABBB\), \(\ldots\)\}. The question is whether a FSM could ‘decide’ whether any arbitrary string is in this set or not — in other words, whether the FSM could compute the function defined as the mapping from all strings in this set to the value ‘True’, and all strings not in this set to the value ‘False’. As proven in most introductory textbooks on computability theory, the answer is *no*. Although one can define a FSM that is capable of deciding whether a string is of the form \(A^nB^n\) for some particular \(n\), one cannot define a FSM that is capable of deciding this question for any arbitrary \(n\).\(^{14}\) One could, on the other hand, define a Turing machine that is capable of deciding this question in full generality. The difference, intuitively speaking, is that a Turing machine has access to a tape that enables it to record the number of \(A\)s it sees, and to subsequently compare that value to the number of \(B\)s it sees. A Turing machine has *memory*.

It’s important to note at this point that, despite their names, Turing machines and

\(^{14}\)The superscript, when applied to a symbol, is the standard convention in computability theory for denoting a string consisting of a certain number of tokens of that symbol type. Thus, for example, \(A^3 = \text{AAA}\).
finite-state machines are not ‘machines’ in any physical, mechanistic sense. They’re machines only in the sense that they’re mathematical models of a person computing a function using a rote, ‘mechanistic’, step-by-step procedure. Like the other constructs in the field of computability theory that I mentioned earlier, such as the lambda calculus and recursive functions, Turing machines and finite-state machines are essentially just ways of specifying a mathematical function in intension — i.e. ways of compactly describing a function without having to exhaustively list all of the input-output pairs that constitute the function. A ‘computation’, as understood in computability theory, is not a concrete, physical process performed by a certain kind of mechanism, but rather an abstract entity: a function specified intensionally.

Nevertheless, more than any other notion in computability theory, Turing’s abstract ‘machine’ came to be highly influential in thinking about how computations might be physically realized by computational mechanisms — first, via its influence on the design of engineered computing mechanisms, and later via its influence on theorizing about the mechanisms of cognition in psychology. There are perhaps two main factors that account for this influence. The first is that Turing’s characterization of his purely mathematical formalization is suffused with causal-mechanical notions. As we’ve seen, Turing explicitly modeled his machine on the ‘mechanical’ operations that people perform when carrying out effective procedures — writing symbols, erasing them, and so-on — and he described these operations using vivid mechanistic metaphors, perhaps reflecting the fact that he had a lifelong obsession with mechanical devices (Hodges, 1983). It’s thus very easy to visualize Turing machines as physical mechanisms. The second factor is Turing’s (1936) definition of a universal Turing machine, a Turing machine that can interpret a coded description of any other Turing machine $M$ written on its tape, and thereby simulate the input-output behavior of $M$. This purely notional device is widely regarded as the inspiration for John von Neumann’s (1945) design for a general-purpose, digital computer that can store its own instructions in memory,
the basic architecture that underlies virtually all commercially available computers today. In any case, for whatever reason, it is now commonplace in cognitive science and related fields to characterize Turing machines in mechanistic terms; to think of them, effectively, as a *blueprint* for a certain kind of physical mechanism. An example of this tendency, drawn more or less randomly from the literature, is Rescorla’s (2012) recent comment that “a Turing machine... is just a device for mechanically manipulating formal syntactic items” (p.6).

As we’ll see, this tendency is pervasive in Gallistel’s work. Gallistel holds that constructs from computability theory, and associated theoretical results, impose ‘ineluctable’ constraints on the kind of physical mechanisms that can efficiently perform complex computations. To understand the nature of those constraints, let’s first consider what it is, in Gallistel’s view, for a mechanism to perform a computation. Gallistel holds that computation in a mechanistic sense is “the process of putting into action a [physically realized function]” (Gallistel & King, 2010, p.87). Now, there are many different ways of putting a function into action. Here’s one: construct a ‘look-up table’ that physically instantiates the value of the function for every possible input to the function. One problem with this approach, of course, is that most functions of interest have an infinite domain and range, so it would be impossible to physically instantiate a look-up table that would compute such functions. But in a sense this is true of any physical computing mechanism; strictly speaking, even the most powerful supercomputer cannot compute a function as simple as addition for *every possible* input to

---

15It’s important, for both theoretical and historical reasons, not to *identify* the notion of a UTM with the notion of stored-program digital computer, as some authors have a tendency to do. Although the former helped to inspire the latter, the development of stored-program computers was a significant conceptual and engineering advance. Moreover, it’s important not to *identify* stored-program digital computers with computers that have a von Neumann architecture. Although von Neumann (1945) provided one of the first detailed designs for a stored-program digital computer, there are other possible (and actual) architectures for stored-program digital computers. For further discussion of these issues, see Copeland (2008) and Piccinini (2008a).
that function, since for sufficiently large numbers the computer would simply run out of physical storage. The real problem with look-up tables as a means of putting functions into action has less to do with the physical limitations on what they can compute, and more to do with the nature of how they compute. A look-up table explicitly stores each and every value of the function that it computes. By contrast, a computational mechanism analogous to a Turing machine generates the values of the function it computes on the fly: the values of the function aren’t explicitly stored at specific locations within the mechanism and obtained by searching for the location that corresponds to a given input, rather they’re generated by the mechanism by computationally manipulating the input.

On the basis of this contrast, Gallistel identifies two specific problems that arise when we take an engineer’s perspective and ask how we might use look-up tables to physically implement computations. The first is that a look-up table can only return those values that are explicitly stored within the table, so a look-up table must be explicitly ‘pre-wired’ to produce an appropriate output value for each and every input value it might receive. Following Gallistel, let’s call this the problem of pre-specification. The second problem is that the number of entries that must be stored within a look-up table scales exponentially with the number of variables that the table can take as input. If we want to construct a look-up table that computes a function of \( n \) variables, where the number of possible values that the \( i \)th variable can take is \( I_i \) (for \( i = 1, \ldots, n \)), then we’d have to construct a table with \( I_1 \times I_2 \times \ldots \times I_n \) entries. Since physical resources must be allocated for each entry within a look-up table, the amount of physical resources required to instantiate a look-up table does not scale well with the complexity of the computation that a look-up table is required perform. Call this the problem of combinatorial explosion.

Consider how a mechanism analogous to a Turing machine — as Gallistel puts it, a mechanism with the ‘functional architecture’ of a Turing machine — deals with
these problems. As mentioned earlier, such a mechanism generates its output on the fly by processing its input. In constructing such a mechanism, one needn’t pre-wire outputs for all of the possible inputs that the mechanism might receive into the structure of the mechanism; rather, one constructs the processing unit of the mechanism in such a way that it can generate the appropriate output for whatever input it might receive. Correspondingly, the physical resources needed to instantiate a given output are only needed when the output is actually generated. The mechanism doesn’t ‘contain’ its outputs in a physical sense, as distinct elements of machinery, but rather in a dispositional sense, as capacities of the processing unit. Thus a mechanism with the functional architecture of a Turing machine appears to avoid the problems of pre-specification and combinatorial explosion because it contains a processing element that is capable of generating its output.

So from the engineering perspective we adopted earlier, it seems that look-up tables don’t provide a very promising way of putting functions into action, especially when those functions are complex functions of multiple variables. Instead, it seems, we should turn our attention to generative mechanisms. Now, just as we might think of the computability-theoretic notion of a Turing machine as blueprint for a certain kind of physical mechanism, Gallistel suggests that we might also think of finite-state machines as a blueprint for a certain kind of physical mechanism. The essential difference between such mechanisms, in Gallistel’s view, is that a mechanism with the ‘functional architecture’ of a finite-state machine lacks the mechanistic analog of a Turing machine’s tape — it lacks a read-write memory mechanism that can store symbols in a thermodynamically stable but computationally accessible form. Nevertheless,  

As Piccinini (2008a) and Wells (2004) point out, it’s anachronistic to interpret the tape of a Turing machine as analogous to the internal, read-write memory of a computing mechanism — as Gallistel does when, for example, he writes that animals “must possess a memory like the memory that Alan Turing (1936) placed at the heart of his mathematical abstraction of a computing device” (2001, p.9694). Recall that Turing (1936) thinks of the tape as “the analog of paper” upon which a human
a mechanism with the architecture of a FSM can process its inputs and generate its outputs; indeed, such a mechanism is, as it were, *all processor*. Unlike a look-up table, such a mechanism doesn’t merely *search* for an output upon receiving a particular input, it transitions through a series of computational states as it reads its input, generating its output along the way.\(^{17}\)

However, insofar as it makes sense to apply results from computability theory to physical mechanisms of a particular type,\(^ {18}\) it would seem that mechanisms with the architecture of a finite-state machine are severely limited in their computational power. As we saw earlier, finite-state machines are incapable of computing even simple functions such as the \(A^nB^n\) problem. We saw that the intuitive reason for this is that FSMs lack a memory tape: a FSM has no way to flexibly record information about incoming input, such as how many As it has seen, and to compare that with information about subsequent input, such as how many Bs it sees. Now, a FSM does have a kind of path-dependent state-based memory, in the sense that the state that a FSM enters into upon receiving some input depends on the history of states the machine has already transitioned through. But this way of dealing with problems that

---

17In my discussion of FSMs, I’ve elided over some distinctions that aren’t directly relevant to the main points I wish to make. It’s worth noting here, though, that there are really two different kinds of FSMs. In one kind, the output of the machine is determined by the state that the machine ends up in after processing its input. These machines contain privileged ‘accept’ states, such that if the machine ends up in one of those states, it is said to ‘accept’ its input, otherwise it ‘rejects’ its input. These machines are thus called ‘acceptors’, and can be understood as implementing binary functions, functions that map inputs to one of two outputs. In the second kind of FSM, called a ‘transducer’, the machine can produce an output symbol at each step of its processing, thereby generating a string of symbols when its processing is complete. Thus, unlike acceptors, transducers can map inputs to any number of outputs. Canonical examples of transducers are Mealy and Moore machines. It’s (the mechanistic analog of) transducers that I have in mind at the present point in the text.

18At the risk of ruining the suspense, I should note that I will go on to argue in Chapter 4 that this *doesn’t* make sense.
require memory requires there to be a distinct processing path for each contingency that one wants the machine to be responsive to. So, returning to the $A^nB^n$ problem, one could construct a FSM that solves the problem for several different values of $n$ by creating a distinct processing path for each value — in effect, by combining distinct FSMs, each of which solves the problem for strings of some particular length. But this would require specifying in advance precisely which contingencies, in this case, which string lengths, one wishes the machine to be responsive to. Moreover, interpreting FSMs as a distinctive kind of mechanism, this kind of path-dependent memory would involve a profligate waste of physical resources, since similar states and transitions would have to be repeated in order to enable the machine to respond appropriately to each contingency. Thus even though FSMs are generative, they seem to face precisely the same problems that beset look-up tables, namely the problems of pre-specification and combinatorial explosion.

What enables mechanisms with the architecture of a Turing machine to avoid these problems isn’t merely that they contain a generative processing unit, it’s also that they incorporate a functionally distinct mechanism that can dynamically store and update information — a read-write symbolic memory. In Gallistel’s view, it’s this latter component that enables mechanisms with the architecture of a Turing machine to solve problems, like the $A^nB^n$ problem, that cannot in principle be solved by mechanisms with the architecture of a finite-state machine: “the absence of a readable memory radically limits computational power” (Gallistel & King, 2010, p.177). So it seems that if we’re looking for computational mechanisms that can implement non-trivial functions, we ought to restrict our attention to mechanisms with the architecture of a Turing machine, viz. mechanisms with a functional distinction between a processor and a read-write memory. As Gallistel & King (2010) put it, “if we are going to physically realize computationally important functions... we are going to need architectures that permit us to make mappings from essentially infinite domains to essentially infinite
ranges using modest physical resources and without having to build every possible output into the machine in advance” (p.51), and for that, “a specific, minimal functional architecture is demanded, an architecture that includes a read-write memory” (p.128).

However, Gallistel & King hold that this is a minimal architectural requirement, enforced by theoretical considerations from computability theory. As anyone who has worked through the operation of a Turing machine using pen and paper will know, the specific way a read-write memory is implemented in a Turing machine is highly inefficient, requiring the traversal, back & forth, over large blocks of tape in ways that seem clearly unnecessary. After all, the Turing machine was developed for a specific theoretical purpose, to provide a formal definition of an effective procedure, not for any practical purposes. In the course of designing practical and efficient computing mechanisms, computer engineers have discovered various ways of improving upon the basic idea of a read-write memory tape. Most importantly, they have developed indirectly addressable read-write memory. An addressable read-write memory is one in which each unit of memory, or register, is assigned a unique numerical address, so that the processing unit can access that register directly, and ‘fetch’ the symbol contained therein, without having to sequentially traverse large blocks of memory.¹⁹ An indirectly addressable read-write memory is one in which the addresses themselves can be stored as symbols within memory, and thereby made accessible to computation. An indirectly addressable memory is one of the defining features of von Neumann’s design for a general-purpose, stored-program digital computer; it is effectively an engineering implementation of the central idea behind Turing’s universal machine.²⁰

¹⁹Addressability enables so-called random access rather than sequential access to memory. Anyone who has been frustrated by the process of searching for a desired song on a cassette tape after the convenience of being able to directly select a song on a CD will have an intuitive sense of the basic principle underlying this distinction.

²⁰I’m riding roughshod over some historical and theoretical subtleties here. See note 15 for a slightly
Gallistel identifies various practical advantages that a mechanism equipped with an indirectly addressable read-write memory has over a mechanism equipped with a serial read-write memory or no memory at all. First, as suggested earlier, addressability enables symbols to be directly fetched from memory, thus making memory access much faster. Perhaps more interestingly, though, addressability underwrites a distinction between variables and their instances. A particular address in memory can be thought of as a variable, where the symbols that might be stored at address are the values of that variable. This distinction is important for various reasons, but perhaps most notably because it enables a mechanism equipped with an addressable memory to avoid the pre-specification problem. Addressability enables the values of variables to be flexibly updated as the need arises, thus enabling the computational machinery to operate on novel information that wasn’t pre-specified in the architecture of the machinery, in a way that doesn’t require the machinery itself to be entirely restructured. Although a serial read-write memory also has the capacity to record novel information, when information is bound to a variable as in an addressable memory, it can be manipulated in a much more flexible and efficient manner.

An additional layer of computational flexibility is provided by an indirectly addressable read-write memory. Indirection allows symbols stored in memory to be interpreted not just as data, but as instructions to the processor, and addresses that ‘point’ to other locations in memory. By being physically instantiated as symbols, instructions and addresses can thus be flexibly manipulated by computational operations. This has various advantageous properties that I’ll discuss in more detail in Chapter 3.4, but the two main advantages are as follows. First, since indirection allows for locations in memory, i.e. variables, to point to other locations in memory, it effectively allows for the possibility of a hierarchy of variables of variables. Moreover, it allows for re-
lations between variables, rather than just relations between the values of variables, to be recorded in memory and computationally manipulated. This allows the processor to generate indefinitely complex hierarchical data structures, thereby allowing the mechanism as a whole to stand in functioning homomorphisms with external systems of indefinite complexity. The second main advantage of indirect addressing is that it enables an important kind of learning. Learning, as Gallistel understands it, primarily consists in receiving information from experience and decoding it into symbols that can be computationally manipulated. But indirection allows computational instructions themselves to be stored and manipulated, thereby, according to Gallistel, allowing for the possibility of a second kind of learning, in which “experience modifies the program itself” Gallistel & King (2010, p.123).

The following toy example of indirect addressing might help to illustrate how the two advantageous properties I just sketched could in principle be implemented. Imagine a simplified von Neumann machine consisting of a processing unit that can perform various operations on the basis of instructions it receives, and an indirectly addressable memory that can contain data as well as instructions. The instructions consist of the name of an operation, followed a number indicating a register address, followed by an optional ‘%’ command indicating that the value stored at the address is to be interpreted indirectly, as an address itself. Some examples of possible instructions are listed in Table 2.2 below.

Now consider the block of memory depicted in Table 2.3. The addresses of the memory registers are listed in the left-hand column, and the contents of those addresses in the right-hand column. A series of instructions is stored in registers 10-18, which

---

21 See Gallistel (2008, p.238) and Gallistel & King (2010, p.159-160).
22 Note that another important similarity between the instructions and data stored in the memory of a von Neumann machine that is not captured by the present example is that the instructions would be encoded as binary strings, and hence could be manipulated by arithmetic operations, just like the data.
define certain operations to be performed on the contents of registers 80-82 and 90-91. Let’s suppose that processing begins at register 10, and continues sequentially through subsequent registers. The instruction stored at register 10 tells the machine to clear the processing unit and add the value stored not at register 80, but — as indicated by the ‘%’ — at the register that the value stored at register 80 points to, namely register 90. That value, namely 32, is in the next instruction transferred to register 82. The following series of instructions, in registers 13-15, increment the value stored at register 80 by 1. The instruction at register 16 then clears the processing unit and adds the value stored at the register that the new value of register 80 points to, namely register 91. This value, 54, is then added to the value stored at register 82, to obtain the sum 86. Thus, after this series of instructions, the values stored at registers 80, 81, and 82 will be 91, 1, and 86, respectively. It can be seen in this example that register 80 is effectively functioning as a variable of variables; by being sequentially incremented, it can point to a series of other registers that might contain a range of different values.

Time to take stock. The various theoretical considerations surveyed in the last few

---

The choice of numbers for the register addresses here is entirely arbitrary.
Table 2.3: A block of the indirectly addressable read-write memory registers contained within the toy von Neumann machine described in the text. In a machine like this, the memory registers can contain instructions as well as data, and the data can be treated as pointers to other memory registers.

sections suggest that there are certain functional requirements that any information-processing mechanism must satisfy if it is to be both powerful — in the sense of being able to solve non-trivial computational problems and hence represent relatively complex external systems — and practical — in the sense of being able to solve those problems under a wide range of novel real-world circumstances without requiring vast amounts of physical resources. First, it must be capable of encoding novel information using an efficient, ‘compact’ coding scheme, such as the binary encoding of number; second, it must be capable of storing encoded information in a thermodynamically stable form for an arbitrarily long duration; third, it must store that information in a computationally accessible form such that it can be fetched from memory and transported to the computational machinery when needed; and fourth, it must store that information
in an indirectly addressable form so that it can be flexibly combined with other information to generate novel, hierarchically complex data structures. Underlying all of this, though, is the requirement that a powerful and practical information-processing mechanism have the computational architecture of a Turing machine rather than that of a finite-state machine: it must have a functionally distinct read-write memory.

At least, that’s what Gallistel thinks. And he’s serious about the ‘must’s; he thinks that the aforementioned functional requirements have the character of physical or even logical necessity. He holds that they are “dictated by the logic of computation itself” and that therefore “they will be observed in any powerful computational device, no matter what stuff it is made of” (Gallistel & King, 2010, p.167). In particular, he holds that these requirements impose ‘ineluctable’ constraints on theorizing about the computational architecture of the brain: “basic insights taught in computer science courses on, for example, automata theory, are a more secure basis for considering what the functional architecture of a computational brain must be than are the speculations in neuroscience about how brains compute” (ibid., p.ix).

This set of claims provides the conceptual foundation for Gallistel’s argument that the kinds of mechanisms posited by contemporary neuroscientists are inadequate to explain animal cognition. Gallistel’s argument proceeds in roughly two steps. First, it is argued that currently recognized neural mechanisms of computation have the functional architecture of either look-up tables or finite-state machines, and hence the range of problems they can practically compute is radically constrained. Second, it is argued that there is a wealth of ethological evidence showing that animals routinely solve problems well outside the scope of problems that can be practically computed by look-up tables or FSMs. In particular, animals solve problems that necessitate storing a wide range of information, which can’t be anticipated in advance, and which is flexibly manipulated long after it is initially stored. In this section I’ve explicated the theoretical concepts underlying the first part of the argument. In the next section,
I will address the second part; I will review some of the behavioral evidence that Gallistel claims cannot be explained by contemporary neuroscience. Then, in Chapter 3, I will bring the two parts of the argument together and explain in detail why Gallistel holds that the mechanisms posited by contemporary neuroscientists are incapable of explaining the behavioral evidence.

### 2.3 The Computational Basis of Animal Learning

As we saw in section 2.2, Gallistel argues that there are certain functional requirements that any computing mechanism must satisfy in order to exhibit what we might call *computational flexibility*: the capacity to efficiently encode a wide range of novel information and store that information for an indefinite duration in a thermodynamically stable form, such that it can be fetched from memory when needed in a way that is not ‘pre-specified’ in the computational architecture. To exhibit such flexibility, Gallistel argues, a mechanism must be equipped with an addressable read-write memory that can store compactly coded symbols. Gallistel marshals a rich array of ethological evidence, from species across a range of different taxa, to argue that animals do in fact exhibit behavioral capacities that can only be explained by computationally flexible mechanisms. The purpose of this chapter is to present a small sample of this evidence, to illustrate the kinds of behavioral capacities at issue. I’ll begin by discussing the remarkable ability of the Saharan desert ant to return to its nest after long foraging expeditions, apparently by keeping a running estimate of the distance and direction in which it has travelled. I’ll then discuss the caching abilities of scrub jays. These birds seem to be able to remember the specific locations of hundreds of food caches over considerable periods of time; but not only that, they can apparently remember what they stored at each cache, when, and for how long, and can use this information to flexibly inform their subsequent foraging strategies. All of this evidence cries out,
claims Gallistel, for computational mechanisms equipped with an addressable read-write memory.

2.3.1 Path Integration in Insects

Sessile animals have the luxury of not needing to know where they are. But animals that are capable of voluntary movement must, in general, be able to determine their own location with respect to locations of behavioral importance; they must be able to navigate. Navigation is relatively easy when you can perceive familiar landmarks, but it becomes much more difficult when landmarks are scarce or imperceptible. Sailors who lose sight of land are in serious trouble unless they have some means of navigating other than piloting, or simply steering in relation to perceived landmarks. Far from shore, there are no landmarks. One method that human seafarers use to find their way through the vast, featureless expanse of the ocean is called ‘dead reckoning’, or ‘path integration’. As the latter name suggests, this involves keeping track of where one is in relation to a point of origin by successively integrating the directions and distances that one has travelled into an overall estimate of one’s location. As depicted in Figure 2.3, we can think of this mathematically as a matter of successively adding together the ‘local’ vectors that correspond to the directions & distances of each of the sub-paths that one has traversed into a ‘global’ vector that corresponds to the direction & distance from the point of origin to the current location. This global vector can then be inverted to obtain the ‘homing path’ that one must take in order to directly return home.

Path integration clearly requires some method of measuring how far, and in what direction, one has travelled. It also requires a way of keeping track of this information and performing certain calculations with it, such as vector addition. Human seafarers obtain the relevant measurements and perform the relevant calculations using external representational artifacts, such as compasses, sextants, and log-books. But as Darwin
(1873) pointed out, even people who lack such props seems to be capable of a kind of *implicit* dead reckoning; moreover, he suggested that non-human animals might be capable of dead reckoning too. Indeed, decades of subsequent research have amply borne Darwin’s suggestion out, and have provided a wealth of evidence of path integration in a wide range of species. Perhaps most notably, given their tiny brains, many insects are capable of path integration; and perhaps the most skilled navigator among the insects is the desert ant *Cataglyphus*, which is capable of successfully navigating one of the most inhospitable and featureless terrestrial environments on earth, the Saharan desert.

**Figure 2.3:** An example of path integration. Suppose a navigator traverses the path ABC, and arrives at the point labeled ‘Current position’. If she is able to determine the direction and distance of each segment of this path, i.e. the vectors, A, B and C, and is able to keep a running sum of these vectors, she can obtain the ‘global vector’ G. By inverting the global vector, she can then obtain the direction and distance she must travel in order to return to the point of origin. Adapted from Kubie & Fenton (2009).
The desert ant is capable of traveling hundreds of meters through harsh desert environments on foraging expeditions, and, upon finding food, of returning unerringly back to its nest. However, it does not do so by following its convoluted, circuitous outbound foraging path, as it would if it were following a scent trail; rather, it returns home in a straight line (see Figure 2.4 below). Moreover, it apparently does not rely on visible landmarks, as it is capable of returning to its nest even in sparse desert environments with very few landmarks. This suggests that desert ants employ path integration to return home. More conclusive evidence for this hypothesis is provided by displacement experiments, pioneered by Rudiger Wehner (e.g. 1986), in which an ant is allowed to leave its nest and forage until it finds a morsel of food, at which point it is captured by the experimenter and moved to a novel location. As mentioned, when an ant finds food in the wild, it immediately follows a straight ‘homing path’ back towards its nest. When displaced by an experimenter immediately after finding food, it does the same thing — but once it has travelled along its homing path to where its nest ‘should’ be, and finds no nest there, it begins the kind of haphazard search strategy it performs when lost in the wild. Much like ‘false belief’ experiments with human infants, displacement experiments seem to reveal that the desert ant can misrepresent, and hence represent, the location of its nest. Desert ant navigation is not bound by local environmental cues like landmarks or scent-trails, but involves computing a fallible representation of location using a method that looks very much like path integration.

Any mechanism capable of performing path integration, whether biological or engineered, must, as mentioned, be capable of measuring directionality and distance. This is a requirement imposed by the nature of the path integration computation itself, at, as it were, the ‘computational level’. Painstaking experiments, again largely pioneered by Wehner, have shed considerable light on the sources of directionality and distance information available to the desert ant. First consider directionality. Generally speak-
Figure 2.4: Displacement experiments evidence the capacity for path integration in the desert ant, *Cataglyphus*. (A) Outbound path. The ant leaves its nest at location N and traverses a path indicated by the solid line until it reaches a morsel of food at location F. The dotted line represents the *homing vector*, the direction and distance back to ant’s nest. (B) Upon reaching F, the ant is captured, then released at location R. It then traverses a path corresponding to the homing vector to where its nest ‘should’ be, i.e. location N’. Upon failing to find the nest there, it begins a systematic search procedure. Adapted from Wehner & Wehner (1986).

...
solar ephemeris — itself changes as a function of other parameters that depend on the specific environment in which the ant finds itself, in particular the latitude and the time of year. Thus in order for the ant to use polarized skylight as a reliable source of directionality information, it must in some sense ‘know’ how patterns of skylight polarization correlate with time of day in the local environment; that is, it must know the local ephemeris function. Experience-restriction experiments with ants and bees suggest that such insects are innately equipped with a ‘template’ of the solar ephemeris, which reflects the universal principle that the sun’s location at dawn is 180° from its location at sunset, and which gradually comes to reflect the local ephemeris function as the insect experiences local correlations between time of day and patterns of skylight polarization (Dyer & Dickinson, 1994; Wehner & Müller, 1993). To register such correlations, the insect brain clearly must have some way of matching information about skylight polarization patterns with information about time of day. Researchers speculate that this capacity is mediated by neurons in the central complex of the insect’s brain, which appear to integrate light polarization signals from photoreceptors with temporal signals from the insect’s internal circadian clock (Homberg, 2004).

As Gallistel points out, the process by which insects appear to learn the local ephemeris function is evocative of the Chomskyan view that language learners are innately equipped with certain universal principles that determine the space of possible languages, the parameters of which are fixed by the properties of the specific languages that the learner is exposed to (Chomsky, 1995). Indeed, claims Gallistel, this reflects the very nature of learning, understood as a process by which information from the world is extracted from experience and communicated to the brain: as information theory tells us, in order for a signal to communicate information, the receiver must somehow implicitly embody the range of possible messages that it might receive, what’s known as the support of the probability distribution over the messages. In Gallistel’s view, a domain-specific learning mechanism, qua biological information
receiver, embodies a support for a specific probability distribution that is fixed by the innate principles governing its development.

Having discussed the sources of directionality information available to the path integration mechanisms of the desert ant, let us now turn to the sources of distance information. While flying insects like bees seem to rely primarily on visually-derived signals to measure distance — specifically, the rate of optic flow — walking insects seem to rely more on ‘idiothetic’ signals originating from internally produced motor commands (Wehner & Srinivasan, 2003). This idea receives considerable support from recent experiments with *Cataglyphus*, which found that if an ant is trained to forage at a particular location, then has its legs artificially shortened or lengthened with pig bristles, when it is subsequently allowed to return to the foraging location, it will tend to either undershoot or overshoot, respectively (Wittlinger et al., 2006). The ant thus seems to determine the distance it has traveled on the basis of the number of steps it has taken: It is as if, on the training trials, the ant ‘counts’ the number of steps it takes to get to the foraging location. Then, on the experimental trial, the ant travels until the number of steps it has taken equals the number of steps it previously recorded. Since it now covers less (or more) ground, it undershoots (overshoots) the foraging location.

What does all of this show about the nature of the mechanisms mediating path integration in ant brains? According to Gallistel, it shows that such mechanisms must be capable of dynamically recording and updating values of behaviorally important variables, and of carrying those values “forward in time computationally accessible form”; that is, it must be capable of storing values for an arbitrary duration, in such a way that they can enter into computational operations when needed. For example, the ant’s path integration mechanism must apparently keep a running tally of the number of steps the ant has taken on an outward path; it must record this value when the ant reaches a foraging location; it must store this value in case the ant ever travels along
the same path; and if it *does* ever travel along the same path, the mechanism must retrieve the stored value from memory and compare it with the current tally of steps taken, so as to determine whether the ant has travelled the appropriate distance.

But calculating distance is just one aspect of path integration. The ant’s path integration mechanism must also be capable of measuring and storing directionality values. For such values to be accurate, the ant’s brain must have an accurate estimation of the local ephemeris function, derived from local correlations between polarization patterns and time of day. However, as Gallistel puts it, evidence for such correlations only arrives in ‘drips and drabs’, over the course of several days. It seems that the ant’s brain must record specific registrations between polarization patterns and times, and carry these forward in time, so that they may be integrated into a running estimate of the ephemeris: “different bits of information, gathered from experience at widely different times in the past, are computationally combined to obtain the explicit information... that informs current behavior” (Gallistel, 2008, p.233).

Moreover, in order for directionality and distance information to be behaviorally efficacious, it must be combined so as to calculate the vectors corresponding to each sub-section of the ant’s foraging path, and these vectors must continually integrated into a running estimate of the *global vector* from the ant’s nest to its current location. Finally, this global vector must be inverted so as to calculate the direction and distance that the ant must travel to in order to return home. “All of this,” writes Gallistel, “presupposes a symbolic memory, a memory capable of encoding information about the world and the animal’s relation to it and carrying that information forward in time in a computationally accessible form” (2008, p.234). Gallistel holds that none of the ‘connectionist’ mechanisms currently countenanced by neuroscientists could play the functional role of such a read-write memory, and hence that neuroscientists must look to as-yet undiscovered ‘classical’ mechanisms for explanations of cognitive capacities like dead-reckoning. I’ll discuss Gallistel’s arguments for this claim in subsequent
chapters, but next I want to consider another suite of behavioral capacities that, in Gallistel’s view, evinces the kind of computational flexibility than can only be explained by positing a read-write memory.

2.3.2 Food Caching in Birds

People are able to consciously recall specific episodes from their past: what they had for breakfast, what they were doing when the World Trade Center fell, and so forth. This capacity was first identified as a distinctive kind of memory on the basis of psychological evidence by Endel Tulving (1972), who called it ‘episodic memory’. Given considerations about the adaptive utility of such a capacity, and the evolutionary continuity between human and non-human brains, one might think that it would be remarkable indeed if episodic memory were unique to humans. However, one of the essential characteristics of episodic memory as Tulving defined it is the occurrent, conscious ‘noetic’ experience of recalling a specific episode (Tulving, 1985), and the kinds of data that have traditionally been accepted as legitimate sources of evidence for something’s being conscious in psychology have been restricted to verbal reports. Since animals can’t verbalize their experience, conclusive evidence of genuine episodic memory in non-human animals has been regarded as very hard, if not operationally impossible, to come by. However, in a practical spirit, we might put aside the question of whether non-human animals possess episodic memory as defined by Tulving, and ask instead whether they exhibit the functional capacities associated with episodic memory, regardless of any phenomenological corollaries — that is, whether they can record information about what, where and when an episode occurred, and to use that information to inform subsequent behavior. After Clayton & Dickinson (1998), let’s call this capacity ‘episodic-like’ memory.

Many animals hoard food when food is abundant, both to hide it from other creatures and to prepare for times of scarcity. But there’s little point hoarding food unless
you can remember *where* you’ve left it. Moreover, different food perishes at different rates, so it would presumably be adaptive to remember *what* you stored and *when* you stored it. This suggests that food-hoarding behavior might be a promising domain in which to look for a capacity for episodic-like memory. Among the most prolific animal hoarders are scrub jays, *Aphelocoma*, which ‘scatter hoard’ food in thousands of distinct caches over several square miles throughout the abundant autumn months, and are able to relocate those caches with a remarkable degree of accuracy weeks or even months later. Experiments in both the field and the laboratory have revealed that the jay’s ability to retrieve food is guided by specific memories of the spatial location of caches, which are formed in the brief moment when the bird initially caches the food (Shettleworth, 1995).

Nicola Clayton,24 Tony Dickinson and colleagues have investigated the content and the capacity of the scrub jay’s spatial memory in a series of remarkable laboratory experiments, which have established beyond reasonable doubt that scrub jays indeed possess episodic-like memory: they are able to remember where they stored a morsel of food, what kind of food it was, and when they stored it. These experiments used hand-reared scrub jays as subjects, which were allowed to cache food in ice-block trays filled with sand. In one of their most interesting experiments, Clayton et al. (2001) allowed jays to cache three different types of food: peanuts, crickets, or mealworms. The researchers had established in previous experiments that jays prefer crickets and mealworms over peanuts. However, crickets and mealworms decay far more quickly than peanuts, and the researchers had also established that jays prefer peanuts over rotten crickets and mealworms. Because the birds were hand-reared, the experimenters were able to control the information that the birds received about the decay rates of

---

24Nicola, assuredly a woman, is mistakenly referred to as ‘Nicholas Clayton’ by Gallistel & King (2010, p.213)
the various foods. The experimenters isolated this variable in an initial set of training trials, in which half the birds (the experimental group) were given evidence that the various foods decay at different rates, while the other half (the controls) were given no such evidence. In these training trials, both experimental and control birds were first allowed to cache peanuts (P) in one side of a caching tray, and were then, depending on the trial, were allowed to cache either crickets (C) or mealworms (M) on the other side. After a delay of either 4, 28, or 100 hours, the birds were allowed to freely retrieve food from either side of the tray in a recovery phase. Thus in this first training stage, each bird experienced six trial types: P/C and P/M trials, at each of the three delay periods.

The crucial experimental intervention occurred during the delay period of these training trials, when the experimenters surreptitiously replaced the crickets or mealworms that the experimental birds had cached with substitutes that had been artificially ‘rotted’. The specific schedule by which this occurred depended on the trial: in the 28-hour-delay trials, only the mealworms were replaced with ‘rotten’ counterparts, whereas in the 100-hour-delay trials, both mealworms and crickets were replaced. Thus after the delay period, when all the birds were allowed to retrieve the food they had cached, the experimental birds, but not the controls, received evidence that mealworms become inedible after 28 hours, and that crickets become inedible after 100 hours.

After this set of training trials, the birds underwent a series of test trials, which were identical to the training trials except that the researchers removed all the cached food during the delay period, so that the birds had no visual or olfactory cues about what they had stored when they subsequently attempted to recover their caches after the delay. The experimenters recorded where the birds directed their first searches, and where they directed the most searches. They predicted that, in line with the birds’ preferences and what the birds had learned about the differential decay rates of the various foods, experimental birds would preferentially search for food in mealworm
and cricket caches after a delay of 4 hours, but would tend to avoid mealworm caches after a delay of 28 hours, and cricket caches after a delay of 100 hours. On the other hand, the experimenters predicted that the control birds would preferentially search for mealworms and crickets after all time delays. The researchers’ predictions were borne out; the birds seemed to be able modify their foraging behavior on the basis of rich episodic-like memories that integrated information about the location of a given cache, the kind of food stored there, and the rotting rate of that particular kind of food.

Gallistel takes this to provide further evidence that animal brains are capable of storing a huge amount of novel information about behaviorally-relevant aspects of the world, which could not have been pre-specified in advance, for considerable periods of time, in such a way that it can be flexibly manipulated by computational operations. Consider, for example, just how much information jays brains can store: Gallistel (2008, p.236) holds that in the wild, jays create up to 30,000 distinct caches, and can retrieve food from each of them. Consider how long they can store it: Clayton’s experiments suggest at least 100 hours, and field observations suggest far longer, on the order of weeks or even months. Moreover, consider the computations they are able to perform with this information:

[T]he birds compute from the current date-time cache the time elapsed since they made that cache. They compare the time elapsed since that cache was made with the remembered time that it takes the contents of that cache to rot. If the computed time elapsed is greater than the remembered rotting time, they visit first the caches where they put the peanuts, even though they prefer (unrotted) meal worms to peanuts, and even though they made the meal worm caches more recently.

(Gallistel, 2008, p.236)

But perhaps most importantly, consider how flexible and unconstrained the jay’s representation of environmental variables is: for example, the jay is able to learn about specific locations it has never seen before, about the decay rates of various
foods it has never tasted before, and is able to flexibly combine this information in ways that it neither it, nor evolution, could have ever possibly anticipated. This dramatically underscores the importance of positing computational mechanisms that are not susceptible to the problems of pre-specification of combinatorial explosion when it comes to explaining the cognitive feats that many animals are capable of: there are so many different possible locations, types of food, rates of decay, and delays between caching and recovery that a jay might encounter — and so many different possible combinations of these factors — that “no architecture with finite resources [could] cope with this infinitude by allocating resources in advance to every possible combination” (Gallistel & King, 2010, p.217).

Indeed, Gallistel claims that the behavioral evidence from scrub jays, desert ants, and a host of other species, taken together with the ‘ineluctable’ constraints on practical computing mechanisms discussed in Chapter 2.2, conclusively demonstrate that the biological mechanisms that explain cognition are ‘classical’ in nature. Most fundamentally, if the organism’s brain is to receive information from the environment, it must encode that information in the form of readable symbols, i.e. symbols that can be decoded into specific messages. To efficiently encode the vast amount of information it evidently receives, the brain must employ a ‘compact’ coding scheme, analogous to the binary encoding of number. To store this information for hours, days, or months, symbols must be stored in a thermodynamically stable form. Consider, for example, what would happen if the ant’s memory for distance were unstable, and degraded over time: the ant would ‘think’ that it is getting closer to home, simply by standing still (Gallistel, 2008, p.231). Moreover, the symbols must be addressable so they can be fetched from memory and transported to the machinery that implements primitive computational operations when needed. But more than that, the addressing mechanism must support indirection so as to allow for a distinction between variables and their instances, for it is this distinction that obviates the problems of pre-specification
and combinatorial explosion: if the brain is equipped with ‘higher-order’ variables corresponding to general categories of experience like Food or Location, it needn’t have particular values — specific foods, locations, or combinations thereof — pre-specified within its computational architecture.

In short, the behavioral evidence “implies that the nervous system possesses a read-write memory mechanism that performs the same essential function performed by the memory registers in a computer” (Gallistel, 2008, p.229). Gallistel argues at length that none of the ‘connectionist’ mechanisms currently recognized by neuroscientists could perform the function of a read-write memory. I’ll present these arguments sympathetically in Chapter 3, then I’ll go on to critically evaluate them in Chapter 4. But before all that I should first identify their target. So in the next section I’ll discuss some of the history, experimental evidence, and theoretical concepts the provide the foundations for the connectionist orthodoxy in contemporary neuroscience.

2.4 Computing with Neurons

Animals are clearly capable of impressive feats of learning and memory, the richness of which has been obscured by behaviorist methodologies that treat animals as general-purpose stimulus-response learners. When we instead attempt to understand animal learning from an ethological perspective, in terms of the species-specific behaviors that enable an animal to successfully get by in the specific environment to which is adapted, it seems we must postulate mechanisms that represent features of that environment, and computational processes that enable those representations to effect ecologically appropriate behaviors.

But surely none of this is news to neuroscientists. Neuroethology, one of the main branches of contemporary neuroscience, is devoted to investigating the neurobiological mechanisms mediating the ecologically appropriate behaviors of animals. Moreover,
the discipline of neuroscience as a whole is replete with talk of information, represen-
tation and computation. As Christof Koch (1999) writes in the introduction to a
textbook subtitled Information Processing in Single Neurons, “The brain computes!
This is accepted as a truism by the majority of neuroscientists engaged in discovering
the principles employed in the design and operation of nervous systems” (p.1). Indeed,
applications of the notions of information and computation to the explanation of an-
imal behavior first appeared in the context of neuroscience, well before the discipline
of cognitive science had been established (Garson, 2003; Piccinini, 2004). Gallistel
surely knows all of this, and indeed cites the quotation from Koch (1999) I just men-
tioned. So why do he and his co-author Adam King (2010) write that “neurobiologists
are not focused on understanding the nervous system from a computational point of
view” (p.171)? One might be tempted to dismiss this claim as hyperbole, but Gallistel
has repeatedly, over the course of several publications, made claims to the effect that
effective information processing in nervous systems requires a functional architecture
that is not “consistent with our current understanding of neurobiological mechanisms”
(2008, p.227). What is his justification for such claims? I’ll address that question in
Chapter 3. Before doing so, I should at least provide a sketch of what our current un-
derstanding of neurobiological mechanisms amounts to. That is the task of the present
section. I’ll begin with a potted history of our understanding of neural structure and
function.

2.4.1 A Potted History of Experimental Neuroscience

Pioneering neuroanatomical work by Ramón y Cajal at the cusp of the 20th century
established the neuron doctrine, the view that neurons are discrete cells, connected to
each other at sites that Sherrington later called ‘synapses’, rather than fused together
in a continuous syncitium. In the middle of the 20th century, electrophysiological work
showed that when neurons reach a certain threshold level of internal electrical poten-
tial, they send ‘spikes’ of electrical activity along their axons, which — via mechanisms that were quite mysterious at the time — are transmitted across synapses to excite, or increase the electrical potential of, post-synaptic neurons. These spikes were said to obey the ‘all-or-none’ principle: they have a fixed amplitude that is not proportional to the degree of excitation that elicits them. What varies is not the amplitude of a spike but their frequency. The observation that the spiking frequency of a neuron is often proportional to the degree of sensory stimulation led the electrophysiologist Edgar Adrian (1928) to conclude that spikes carry information about stimuli, arguably the first application of the notion of information to explain neural function (Garson, 2003).

Controversy about the basic principles by which presynaptic spikes cause post-synaptic excitation was largely resolved by the 1950s, when most researchers had arrived at a consensus that synaptic transmission is primarily chemical rather than electrical. As subsequent research revealed, presynaptic spikes cause the release of chemical neurotransmitters, which, like a key opening a lock, open channels in the post-synaptic neuron that lead to changes in electrical potential.

Throughout this fertile period of discovery in the early part of the 20th century, neuroscientists were learning that for every hard-won generalization there seemed to be exceptions. For example, it became clear that some neurons are connected together in something like a syncitium after all, fused by electrical synapses or ‘gap junctions’; moreover some neurons communicate using graded electrical signals rather than ‘all-or-none’ spikes; and some neurons inhibit rather than excite post-synaptic neurons. The general lesson here is one that molecular biologists were also learning during roughly the same period: contrary to prevailing positivist models of explanation that were largely inspired by patterns of explanation found in physics, appealing to elegant, exceptionless laws to explain biological systems didn’t seem to shed much light on how

---

25 For an overview of, and references to, some of the relevant studies, see Bullock et al. (2005).
those systems work. To understand the capacities of contingent, evolved biological systems, it gradually became apparent that we must take those systems apart, examine their messy, variable inner components, and elucidate how those components interact so as to generate the capacity of interest.

I’ll make more of this point later, in Chapter 4, but let’s put it aside for now, and turn to consider how some of the aforementioned findings about neural structure and function might shed light on psychological phenomena. William James (1890 [1981]) developed a range of detailed and remarkably prescient speculations about how hypothetical changes in the nervous system might mediate learning — or as he put it, the formation of ‘habits’ — and seems to have been the first to employ the now widely-used term ‘plasticity’ in this connection: “the phenomena of habit in living beings are due to the plasticity of the organic materials of which their bodies are composed” (p.105). Perhaps most notably, he anticipated contemporary views about the principles by which neural plasticity occurs: “When two elementary brain-processes have been active together or in immediate succession, one of them, on reoccurring, tends to propagate its excitement into the other” (p.566). However, James had no notion of the mechanisms by which neural plasticity might occur.

The Italian neuropsychiatrist Eugenio Tanzi (1893) was perhaps the first to suggest that the connections between neurons discovered by Cajal might be modified by experience, and hence might be the locus of the neural plasticity mechanisms hypothesized to mediate learning (Berlucchi & Buchtel, 2009). Specifically, he speculated that learning and practice might elicit a kind of ‘ameboid' growth at synapses, which would increase the efficacy of transmission between neurons. Several of Tanzi’s contemporaries, including Lugaro, Cajal, and Sherrington, developed similar hypotheses, though Cajal held that in addition to changes in the efficacy of existing synapses, we must also posit the growth of entirely new synapses, for only the latter, in his view, could explain the acquisition of certain kinds of complex, long-lasting memories. In
any case, the idea that learning and memory involve some form of long-term synaptic plasticity soon became the predominant view in the psychology and neuroscience of the early 20th century.

However, this didn’t last long. Contrary to the widespread view that “the leading hypothesis among both theoreticians and experimentalists [since the early 20th century] has been that synaptic plasticity underlies most long-term behavioral plasticity” (Koch, 1999, p.308), the synaptic plasticity hypothesis has in fact seen several reversals of fortune over the years, facing many criticisms and competing hypotheses. One of the most eminent and vociferous critics was Karl Lashley, who in 1930 wrote that “there is no direct evidence for any function of the anatomical synapse: there is no evidence that synapses vary in resistance, or that, if they do, the resistance is altered by the passage of the nerve impulse”. It’s thus somewhat ironic that the work that probably did most to re-consolidate interest in the synaptic plasticity hypothesis was *The Organization of Behavior*, by Lashley’s student Donald Hebb (1949). Hebb proposed that short-term memories are instantiated by transient assemblies of neurons, which are permanently consolidated into long-term memory as a result of activity-dependent changes in the synaptic efficacy between neurons. Echoing James, he famously characterized the principle governing these activity-dependent changes as follows: “When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased” (p. 62).

Despite the widespread influence of Hebb’s conjectures, they were just that — conjectures — and didn’t provide any concrete experimental evidence of the existence, let alone the behavioral significance, of synaptic plasticity. However, such evidence did emerge; first slowly, then at a rapidly increasing rate. Some of the earliest evidence came from work by Eric Kandel and colleagues, who demonstrated that simple forms of conditioning in the sea slug *Aplysia* involve synaptic plasticity (Castellucci
et al., 1970), and went on to elucidate the cellular and molecular mechanisms of this process in exquisite detail, eventually earning a Nobel prize for his efforts. Research into the role of synaptic plasticity in more complex forms of learning in mammals was galvanized by the discovery of long-term potentiation (LTP) by Bliss & Lømo (1973). These authors showed that a brief, high-frequency ‘tetanic' stimulus applied to certain neural pathways in the rabbit hippocampus would increase the transmission efficacy of downstream synapses for hours, days, or even weeks. Subsequent neuropharmacological work began to untangle the molecular mechanisms underlying this phenomenon, showing, for example, that NMDA receptors are crucially involved in the induction of LTP. Simultaneous work in behavioral neuroscience began to elucidate the behavioral significance of LTP. One of the earliest and most compelling findings was obtained by Richard Morris et al. (1986), who showed that pharmacologically blocking hippocampal NMDA receptors prevented rats from learning how to successfully navigate a water maze. Since then, evidence for the ‘synaptic plasticity hypothesis', the view that activity-dependent synaptic plasticity is involved in learning and memory, has accumulated to the point where, as Martin & Morris (2002) recently put it, “few doubt that the general principle is correct” (p.609).

One of the most important set of findings concerning how synaptic plasticity is involved in learning and memory formation pertains to the timing-dependency of LTP. Research into this phenomenon was initiated by the discovery of long-term depression (LTD): prolonged, low-frequency stimulation was found to sometimes reduce the transmission efficacy of synapses (Dunwiddie & Lynch, 1978). Subsequent research showed that whether LTP or LTD occurs depends not only on the frequency of the eliciting stimulus, but also on the relative timing of spiking activity in pre- and post-synaptic neurons: if presynaptic spiking precedes post-synaptic spiking within a window of a few milliseconds, then LTP is induced, but if post-synaptic spiking occurs first, LTD occurs (Bi & Poo, 1998). This phenomenon has come to be called spike-timing depen-
dent plasticity (STDP), and is widely regarded as a direct experimental vindication of Hebb’s postulate that ‘cells that fire together wire together’ (Caporale & Dan, 2008).

2.4.2 A Potted History of Theoretical Neuroscience

To better understand the theoretical importance of Hebb’s postulate and its recent incarnation in STDP, we should retrace our steps to the beginning of our history of experimental neuroscience, and pursue a parallel path through the history of theoretical neuroscience. Theoretical neuroscience was pioneered by Nicolas Rashevsky, who sought to develop a ‘mathematical biology’ that “would stand in the same relation to experimental biology as mathematical physics stands to experimental physics” (1938, p.9). Rashevsky developed abstract models of networks of interconnected neurons using systems of differential equations. However, the first formal characterization of neural networks as performing computations, using the tools of discrete mathematics, appeared in the seminal (1943) paper by Warren McCulloch and Walter Pitts.26 McCulloch & Pitts showed how networks of idealized neurons could implement the primitive truth-functions of Boolean logic, and proved that such networks could be wired together to compute the truth-value of any finite Boolean expression.27 While this work is often said to have inaugurated the theoretical tradition that eventuated in the neo-connectionism of the 1980s, Boden (1991) points out that it can equally well be regarded as a progenitor of the ‘classical’ tradition in cognitive science via...

---

26For a defense of the claim that McCulloch & Pitts (1943) developed the first truly ‘modern’ computational theory of the mind-brain, in the sense that they were the first to employ formalisms from mathematical logic and computability theory to explain the operation of the mind-brain, see Piccinini (2004).

27However, contrary to a popular myth, McCulloch & Pitts (1943) did not prove that their networks have the computational power of Turing machines. They did assert that their networks, if equipped with a memory tape and tape-head, can compute all and only the numbers that a Turing machine can, but they did not attempt to prove this assertion. The myth that they did prove this assertion seems to have originated with McCulloch’s (1949) later mischaracterization of his own results (Piccinini, 2004).
its influence on theoretical computer science and computer engineering; most notably, Kleene (1956) generalized the McCulloch-Pitts model to develop the formalism for finite-state automata, and von Neumann (1945) used McCulloch-Pitts nets as a template for the digital logic circuits in his design for a stored-program digital computer. McCulloch-Pitts neurons essentially just are the logic gates in conventional computers.

While McCulloch & Pitts were ultimately interested in the operation of the mind and brain, they did not develop their model with an eye toward biological realism — indeed, they deliberately abstracted away from a great deal of neurophysiological detail known at the time — and it’s perhaps for this reason that their model was largely ignored by neurophysiologists (Piccinini, 2004). One of their most salient departures from biological realism was their assumption that their networks do not change over time, and hence are incapable of learning — for, in accordance with widespread views about the neural mechanisms of learning, they assumed that learning “require[s] the possibility of permanent alterations in the structure of nets” (McCulloch & Pitts, 1943, p.124). They did not regard the static nature of their networks as a problem for their project, since they were primarily interested in the synchronic question of whether their networks could in principle compute specific Boolean functions, not in the diachronic question of how their networks could come to compute such functions (ibid., p.117).

However, ignoring the problem of explaining how a network could learn to compute a given function doesn’t make it go away. Enter Frank Rosenblatt (1958) and his perceptron model. Although Rosenblatt was impressed by the rigor and mathematical sophistication of earlier models of neural computation such as McCulloch-Pitts nets, he felt that they drew too much inspiration from the infallible, deterministic systems of symbolic logic and digital computation to accurately reflect how the presumably fallible, probabilistic networks in the warm, wet environment of the brain actually store and process information. He sought to combine the mathematical precision of these earlier models with the neurophysiologically informed speculations of theorists like Hebb to
develop more biologically realistic models of neural computation — specifically, models that would reflect the kind of experience-dependent plasticity mechanisms thought to mediate learning. The result was what Rosenblatt called the ‘perceptron’. This model, which can be characterized not too inaccurately as the offspring of McCulloch & Pitts (1943) and Hebb (1949), is itself the intellectual forefather of much of the subsequent neural network theorizing in artificial intelligence in the late 20th century, and its inception perhaps marks the first significant fissure between the ‘connectionist’ and ‘classical’ traditions in cognitive science. Rosenblatt (1958) locates himself, along with others such as Hebb and Hull, within the broad tradition of British empiricism, and characterizes their conception of neural function as ‘connectionist’ in the sense that on their view, “information is contained in connections or associations” between neurons, as a “preference for a particular response” (p.387, emphasis in the original), instead of being stored as a “topographic representation”, which stands in “some sort of one-to-one mapping [with] the sensory stimulus” (p.386-7).

Rosenblatt made various exaggerated claims about the explanatory power of his perceptrons; for example, he claimed in a (1959) paper that “the perceptron has established, beyond doubt, the feasibility and principle of non-human systems which may embody human cognitive functions at a level far beyond that which can be achieved through present day automatons” (p.449). Claims like these generated a backlash amongst many researchers in the emerging field of artificial intelligence, who felt, contra Rosenblatt, that computational modeling of cognitive phenomena could proceed

---

28It should be noted that although Rosenblatt describes Hebb’s theories as an ‘inspiration’ for the perceptron, the main point of contact between the two researchers’ conceptions of learning in neural networks is simply that learning involves experience-driven changes in connection weights. The actual learning rule employed by perceptrons is error-driven, i.e. supervised, and hence quite different in character from the unsupervised learning rule proposed by Hebb. I’ll return to this distinction later in the chapter. It’s also worth noting here that although Rosenblatt intended his perceptrons to be more biologically realistic than McCulloch-Pitts networks, he nevertheless thought of them as “extreme simplifications” of biological neural networks (Rosenblatt, 1962, p.28).
largely independently of research into the messy details of how the brain works. At stake here was more than just the fate of certain theoretical and methodological views within cognitive science; access to limited governmental funding was on the line. The backlash was led by Minsky & Papert (1969), who established that perceptrons of a certain kind are incapable in principle of learning to solve an interesting class of computational problems. Their results punctured the initial bubble of enthusiasm for perceptrons, and played a role in diverting governmental funding away from neural network research. As a result, such research stagnated for much of the following decade.

However, theoretical work that started to emerge in the late 1970s — most notably, the development of the backpropagation learning rule, which I’ll describe in a moment — obviated many of the limitations described by Minsky & Papert, and generated a resurgence of interest in artificial neural networks. This resulted in a profusion of different models with varying degrees of computational power, biological realism, and practical utility. In the context of cognitive science, this resurgence of interest — which was driven in large part by the work of Rumelhart, McClelland, Hinton, and others, anthologized in what has come to be known as the ‘bible of connectionism’ (Rumelhart et al., 1986b) — manifested in the field of ‘neo-connectionism’, which attempted to model psychological capacities using artificial neural networks. These neural network

---

29 Of course, there was no lack of hype from this side of the computational tracks, either. For example, Simon & Newell (1959) (in)famously wrote that “It is not my aim to surprise or shock you... But the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create.” (p.6). This claim was highly contentious then, and isn’t much less contentious today, even in an era in which machines can win millions on Jeopardy.

30 Namely, one-layer perceptrons with a step activation function — I’ll explain what this means in the next section.

31 This field is sometimes simply called ‘connectionism’, especially in the literature on the so-called ‘classicism versus connectionism debate’. For reasons I will discuss in the following sub-section, I use ‘connectionism’ more broadly to encompass, roughly, any theory or model that attempts to explain aspects of the mind or brain by appealing to experience-dependent changes in plastic connections between neuron-like elements. I reserve ‘neo-connectionism’ to refer to the specific subspecies of connectionism that was prominent primarily during the 1980s, which used idealized neural network
models were usually highly idealized, and at a far remove from the details of actual neural circuitry. However, work on artificial neural networks also had an influence in theoretical neuroscience, where it fused with the largely independent Rashevskyan tradition of mathematical biophysics to form the field of computational neuroscience in the 1990s.

However, whether they found application in cognitive science, neuroscience, or even commerce and industry, these later neural network models generally operated according to much the same principles as Rosenblatt’s perceptron. To understand these principles, let’s look more closely at how these models work. Neural network models are directed graph structures consisting of a set of nodes, \( \{N_0, N_1, N_2, \ldots, N_n\} \), typically arranged into input and output layers, possibly with one or more intervening or ‘hidden’ layers. The nodes are joined by connections with modifiable weights, where the weight from node \( N_i \) to \( N_j \) is a numerical value \( w_{ij} \). Moreover, each node has a variable level of activation, \( a_i \). The nodes, weights, and activation levels are typically thought to correspond to neurons, synaptic conductances, and neural firing rates, respectively, though the assignment of a biological interpretation to a neural network model is not always straightforward, and depends upon the kind of model employed and the specific use to which it is put. However, at present, we’re just interested in a generic neural network model. An example of such a model, consisting of a single input layer and an output node, is depicted in Figure 2.5.

When a network receives a pattern of activation across its input nodes, activity flows through the network from one layer to the next, being transformed or ‘filtered’ by the connection weights along the way, until it reaches the output nodes. The models — most notably, multilayer perceptrons trained by backpropagation — to model psychological capacities. Thus, according to the usage that I will adopt throughout this dissertation, computational neuroscience is distinct from neo-connectionism, but is a subspecies of connectionism. On the other hand, McCulloch-Pitts nets do not fall under the rubric of ‘connectionism’ as I use that term.
connection weights thus determine which output activation pattern will be elicited by a given input activation pattern, and hence, when activation patterns are appropriately interpreted, which input-output function the network computes. But by grounding a disposition to elicit a particular output pattern, the weights also, in a sense, embody the information that is stored within the network. The weights in a connectionist network thus determine both how the network processes information, and what information the network stores. This is the basis of the widespread claim that, in contrast to conventional digital computers, connectionist networks lack an architectural distinction between processor and memory.

To get a sense of how the information processing capacities of networks are realized by the activities of individual nodes, consider again the network depicted in Figure 2.5, which has an input layer consisting of nodes $N_1, N_2, ..., N_n$, and an output layer consisting of a single node, $N_0$. Providing input to such a network involves setting the activity levels of the input nodes to specific values. This activity then propagates to the nodes in the next layer of the network, in this case to our single output node $N_0$. 

![Figure 2.5: A schematic depiction of a single-layer, feed-forward connectionist network consisting of $n + 1$ nodes. The activity level of node number $i$ (i.e., $N_i$) is denoted '$a_i$', and the weight of the connection from node $N_i$ to node $N_j$ is denoted '$w_{ij}$'.](image)
The activity level of this node, $a_0$, is determined by multiplying the activity level $a_i$ of each of node $N_i$ that is connected to $N_0$ by the corresponding weight of that connection, $w_{i0}$, adding these weighted activity levels together, and passing them through a particular function $F$, called an ‘activation function’. That is,

$$a_0 = F\left(\sum_{i=1}^{n} a_i w_{i0}\right)$$

Activation functions determine the activity levels of all ‘internal’ nodes in a network that aren’t directly driven by external input. The activation function employed by McCulloch & Pitts (1943) was a binary ‘step’ function — i.e. a function that returns 1 when incoming weighted activation reaches a certain threshold, and 0 otherwise — which they chose due to their belief that neural activity is ‘all-or-none’. In justifying their choice of activation function in this way, they seemed to be assuming that the function models individual spiking events in neurons. By contrast, later models developed by the ‘neo-connectionists’ of the 1980s employed a variety of continuous, nonlinear activation functions, typically sigmoidal. Insofar as these activation functions were given a direct, biological interpretation at all, they were assumed to model the spiking frequency of a neuron. As I’ll go on to discuss later, more recent ‘spiking neural network’ models have returned to the idea of representing individual spiking events in neurons, though they do so with far greater biological fidelity than McCulloch-Pitts nets.\(^{32}\)

I mentioned earlier that there were certain theoretical innovations in the 1970s and 80s that helped obviate some of the problems with simple perceptrons identified by Minsky & Papert (1969), and cleared a path for the groundswell of interest in neo-
connectionism. One of these innovations was the use of nonlinear activation functions. Another was the use of multilayer perceptrons — structurally complex networks, with one or more ‘hidden’ layers in addition to the input and output layers. These hidden layers enable networks to recognize higher-order, abstract features of their input. But perhaps the most important innovation was the introduction of a novel method of training neural networks, called back-propagation of error, or just backprop.\textsuperscript{33} I suggested earlier that Rosenblatt’s central contribution was to incorporate the capacity to learn into models of neural computation. He did so using an error-driven learning rule by which the weights of a neural network are adjusted so that the degree of mismatch between the network’s actual performance and its desired performance is gradually reduced.\textsuperscript{34} Backpropagation essentially generalizes the learning rule employed by Rosenblatt so that it can apply to multilayer perceptrons — as the name suggests, it allows errors to be propagated back from output layers through hidden layers, gradually adjusting connection weights along the way. While perceptrons equipped with nonlinear activation functions and multiple layers can in principle solve many of the computational problems discussed by Minsky & Papert, backpropagation was pivotal because it allows such perceptrons to learn to solve those problems.

The popularization of backpropagation rekindled interest in using neural networks as psychological models, and led to a range of practical applications in commerce and industry. But despite its theoretical and practical importance, backpropagation is often criticized for being biologically implausible. As Rumelhart et al. (1986a) themselves

\textsuperscript{33}While the history of backpropagation is somewhat contentious, it is widely agreed that the first description of backpropagation as a method for training is neural networks appears in Paul Werbos’s (1974) Ph.D. thesis. However, Werbos’ work languished in obscurity or many years, and backpropagation was subsequently (re)discovered by several independent researchers (Medler, 1998). Of these, Rumelhart et al. (e.g. 1986a) probably did the most to demonstrate the theoretical power of backpropagation, and its potential as a tool for psychological modeling.

\textsuperscript{34}An error-driven learning rule like this is said to be ‘supervised’, since it involves reducing the discrepancy between the actual performance of a system, and the performance desired by a ‘supervisor’ or teacher.
concede, backpropagation “is not a plausible model of learning in brains” (p.536). One reason is that backprop is a ‘supervised’ learning rule, which involves reducing the discrepancy between the actual and desired performance of a network. In the context of training artificial neural networks for practical purposes, it’s clear who defines the desired performance of the network: the humans who are training the net. However, in the context of biological neural networks, it’s not at all clear who or what would define the desired performance of the network; presumably not a little homunculus inside the brain! A second reason backpropagation is widely regarded as biologically implausible is that it requires a mechanism by which error signals propagate ‘backward’ up presynaptic axons, yet there are no widely recognized mechanisms by which this could be achieved.

In the 1990s, neural networks increasingly came to be employed not just as highly idealized, ‘phenomenological’ models of psychological capacities, but also as mechanistic models of specific neural systems and circuits. This motivated a search for more biologically plausible learning rules. Much of the search focussed on unsupervised learning rules, to avoid the problematic question of who or what would supervise the training of a biological neural network. We encountered the broad outline of one such rule in the previous section: Hebb’s postulate. In its most general form, this rule states that the increase in the weight of the connection from one node \(N_i\) to another \(N_j\) is proportional to the simultaneous activities of those two nodes \(a_i\) and \(a_j\), respectively. That is:

\[
\Delta w_{ij} \propto a_i a_j
\]

In such a simplistic form, Hebb’s rule is clearly inadequate. Since the activities of nodes cannot be negative, the connection weights in a network equipped with only this rule would monotonically increase until they reach saturation. For this reason, researchers have explored various ‘anti-Hebbian’ learning rules. Research into the
computational properties of Hebbian / anti-Hebbian learning was galvanized by the
discovery of spike-timing-dependent plasticity (STDP) in the 1990s, which, with its
balance of LTP and LTD, seemed to provide a plausible mechanism by which Hebbian
learning might be biologically realized. This dialog between computer scientists ex-
ploring the properties of biologically-inspired learning algorithms, and neuroscientists
investigating the physiological mechanisms that might instantiate those algorithms, has
continued over the past two decades under the rubric of computational neuroscience,
and has generated a profusion of biologically inspired learning rules. Summarizing
these is beyond the scope of the present project, but suffice it to say that many neu-
roscientists would agree that Rosenblatt’s quest for biologically realistic mathematical
models of the neural mechanisms that mediate learning and other cognitive processes
is well on its way to being fulfilled.

2.4.3 What is Connectionism?

Maass (1997) usefully distinguishes three ‘generations’ of neural network models, on
the basis of the kinds of activation functions that the basic computational units em-
ploy. The first generation comprises McCulloch-Pitts nets, which employ a binary step
function. The second generation comprises perceptrons, which employ continuous,
nonlinear (usually sigmoidal) functions. The third generation comprises spiking neural
networks, also known as ‘integrate-and-fire’ networks, which don’t employ a single ac-
tivation function, but rather several functions that model, in a relatively fine-grained
way, the membrane and threshold potentials of neurons. At the expense of losing some
of the precision of Maass’ original distinction, we can make some additional general-
izations about these three generations of neural networks, to help locate some of the
points made in the previous section within a broader historical and theoretical context.

35For a review, see, for example, Wörgötter & Porr (2005).
These generalizations are summarized in Table 2.4 below.\textsuperscript{36}

The contrast between the second and third generations is particularly salient in the present context. The neural networks of the second-generation were employed by the neo-connectionists of the 1980s as models of perceptual and cognitive capacities. Structurally, these networks typically had one or more hidden layers, and were often ‘feed-forward’ in the sense that the nodes in one layer were connected to nodes in the next successive layer, without any return connections back to themselves or to previous layers. However, some second-generation networks were ‘recurrent’, in the sense that they incorporated nodes with return connections, thereby forming cyclic networks. Recurrency itself wasn’t a novel innovation of second-generation networks; for example, McCulloch-Pitts nets incorporated recurrent connections. However, because of the continuous activation levels and nonlinear activation functions employed by second-generation networks, adding recurrency introduces complex temporal dynamics such as attractor states into their behavior, which can be analyzed using the tools of dynamical systems theory.

The introduction of recurrent connections into second generation networks increases their biological plausibility to some extent, since connectivity in cortical neural networks is known to be highly recurrent (Kandel et al., 2000). Even so, second generation networks are still at a considerable remove from actual, biological neural networks. One reason is that they typically employ backpropagation as a learning rule, and as we’ve

\textsuperscript{36}Note that this way of carving things up is highly schematic and somewhat procrustean. It is useful for present expository purposes, but shouldn’t be regarded as an attempt to provide a general overview of theoretical work on neural networks. Any responsible survey of that vast and technically sophisticated body of literature would address a range of issues and distinctions that I’ve had to ignore here. For an example of the procrustean nature of the schema I’ve provided here, it’s important to note that despite what Table 2.4 might suggest, spiking neural networks are not a recent innovation; on the contrary, precursors can be traced back to pioneering work by Lapicque (1907). Having said that, spiking models do have wide currency in contemporary work in computational neuroscience, in part because of their interesting theoretical connections with the relatively recently discovered phenomenon of spike-timing dependent plasticity (STDP), which I will discuss in more detail later.
<table>
<thead>
<tr>
<th>Generation</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network Type</strong></td>
<td>McCulloch-Pitts</td>
<td>Perceptron</td>
<td>Spiking neural network</td>
</tr>
<tr>
<td><strong>Activation Function</strong></td>
<td>Binary step function</td>
<td>Nonlinear, usually sigmoidal</td>
<td>Complex</td>
</tr>
<tr>
<td><strong>Architecture</strong></td>
<td>Single-layer recurrent</td>
<td>Multilayer, feed-forward or recurrent</td>
<td>Multilayer, usually recurrent</td>
</tr>
<tr>
<td><strong>Learning Rule</strong></td>
<td>None</td>
<td>Backpropagation</td>
<td>Biologically inspired (e.g. STDP)</td>
</tr>
<tr>
<td><strong>Field (Era)</strong></td>
<td>Cybernetics (40s-50s)</td>
<td>Connectionist psychology (80s-90s)</td>
<td>Comp. neuroscience (90s-2010s)</td>
</tr>
<tr>
<td><strong>Application</strong></td>
<td>Theoretical</td>
<td>Psychological simulation</td>
<td>Mechanistic modeling</td>
</tr>
</tbody>
</table>

Table 2.4: A summary of three generations of artificial neural network research.
seen, interpreting backpropagation as a literal model of neuroplasticity mechanisms is problematic. Another reason is that characterizing neural activity in terms of firing rates by using continuous activation functions elides over a great deal of electrophysiological detail concerning how membrane dynamics give rise to individual spiking events. For these and other reasons, it is difficult to interpret second-generation network models as mechanistic models of actual neural circuits or systems. Consequently, some second-generation researchers suggested that the ‘nodes’ and ‘connections’ within their models were not to be thought of as isomorphic to neurons or synapses, but rather to certain higher-order, functionally-individuated entities at the systems level (Smolensky, 1988). Others thought of their models more as idealized, phenomenological models, primarily intended to demonstrate the possibility that neuron-like systems exhibit certain cognitive capacities, rather than to elucidate exactly how they do so. For example, reflecting on the famous NETtalk model, which learns to transform written English text into speech, Churchland & Sejnowski (1990) write that “it would be surprising if the details of the model bore any significant resemblance to the way reading skills are represented in the human nervous system. NETtalk is more of a demonstration of certain network capacities and properties than a faithful model of some subsystem of the brain” (p.362).

Nevertheless, while I lack the space to argue the point here, I think that most neo-connectionist researchers took themselves to be contributing to a project whose long-term goal was the development of mechanistic explanations of cognitive capacities, couched in terms of specific circuits, systems, and processes in the brain.\footnote{I’ll talk more about mechanistic conceptions of (specifically cognitive) explanation in Chapter 4.} I think that the second-generation models they developed are best understood as what Piccinini & Craver (2011) call ‘mechanism sketches’. These models could not, given the state of knowledge about the neurobiology of cognition at the time, be anything
more than highly schematic models of notional mechanisms; however, they weren’t offered as mechanistic explanations *in themselves*, but rather as *sketches* of mechanistic explanations, to be modified and filled in by subsequent theoretical and experimental developments. Indeed, contemporary ‘naturalistic’ approaches in the philosophy of science that pay close attention to the historical vicissitudes of scientific research suggest that this is a general norm of knowledge production in the life sciences: schematic, idealized models enter a cyclical process whereby they are evaluated against constraints from neighboring fields, are appropriately revised, and if all goes well, eventually become sufficiently constrained as to provide literal, mechanistic explanations of the target phenomenon (Darden, 2002).

The third generation of neural network modeling, which emerged over the past two decades in cognitive and computational neuroscience, takes us closer to the goal of literal, mechanistic explanations of cognitive capacities (Eliasmith, 2003). As I mentioned above, second-generation models make a number of simplifying idealizations, such as modeling neural activity as firing rates. Of course, idealization *per se* isn’t problematic; any model must abstract away from *some* irrelevant details, or else end up as an analog of Borges’ perfectly accurate, but perfectly useless, map that’s as large as the kingdom it depicts. However, the past two decades have seen increasing experimental and theoretical evidence that the timing of individual spikes in many cases *is* in fact directly relevant, both to the dynamical behavior of a network, to the information that neural activity encodes, and to the computations that a network can perform.\(^{38}\) This has led to a profusion of theoretical work using spiking neural network models, which represent some of the biophysical parameters that determine a neuron’s membrane potential, and can thereby capture the temporal dynamics of a neuron’s spiking behavior. Complementing this work has been a wealth of research on temporal

---

\(^{38}\)See Abeles (1991), Rieke et al. (1996), and Maass (2001), respectively.
coding, i.e. encoding information in the timing of individual spikes, and on learning rules that are sensitive to spike timing, such as STDP. The spiking network models that have emerged from this groundswell of research offer a powerful and biologically realistic suite of modeling tools that are increasingly being used to elucidate the specific neural mechanisms that mediate various cognitive capacities (see, e.g., Eliasmith & Anderson, 2003; Rolls, 2007).  

Despite these innovations, third-generation neural network models are recognizably connectionist in roughly Rosenblatt’s (1958) sense: they comprise networks of nodes joined by weighted connections, which map input patterns of activity to output patterns, and which can be ‘trained’ to perform different input-output mappings via learning rules that change the weights of the connections on the basis of the input that the network receives. This is the sense of ‘connectionism’ that I will employ in the remainder of the dissertation, since I think it is the sense that Gallistel has in mind when he claims that the connectionist mechanisms posited by contemporary neuroscientists are incapable of explaining ‘interesting’ — that is, cognitive or intentional — aspects of animal behavior, and moreover because it is continuous with an earlier usage of ‘connectionism’ that precedes the neo-connectionism of the 1980s. To put us in a position to evaluate this claim, I’ll spend the remainder of this section attempting to clarify the character and commitments of connectionism in the sense at issue.

First, let’s contrast the notion of connectionism at issue with a much more lib-

---

39To reiterate note 36, the historical sketch I’ve provided in this section is highly schematic, intended only to illustrate some broad trends. For example, I mentioned there that although spiking neural network models are widely used in contemporary computational neuroscience, precursors can be traced back more than a century. Moreover, spiking neural networks are by no means the most biologically realistic models currently on offer in computational neuroscience; indeed, they are highly unrealistic compared to ‘compartment’ models that reflect the detailed geometry and biophysical properties of (parts of) individual neurons. However, compartment models are generally too complex to afford computationally tractable simulations of the behavior of populations of neurons, let alone entire systems-level networks, and hence generally employed as mechanistic models of psychological capacities. The hope is that spiking networks are realistic enough to capture the explanatorily relevant mechanisms that mediate such capacities.
eral notion identified by Piccinini (2008b). According to this liberal notion, virtually any computational system consisting of a set of computational elements joined together by signal-carrying connections counts as a connectionist network. This sense of ‘connectionism’ encompasses all three generations of neural networks I’ve surveyed, including McCulloch-Pitts nets. But as I mentioned earlier, McCulloch-Pitts nets are functionally equivalent to the logic gates that are components of conventional digital computing devices. “Because of this”, Piccinini points out, “there is a straightforward sense in which digital computers are connectionist systems” (ibid., p.313). The liberal sense of ‘connectionism’ that Piccinini is employing here is thus clearly not the narrower sense that Gallistel has in mind when he argues that the explanatory power of connectionist neuroscience is severely constrained. The central difference is that connectionist networks in the narrower sense that Gallistel takes umbrage with attempt to model learning in terms of the experience-dependent plasticity of connection weights, which Gallistel regards as a contemporary manifestation of a long-standing but fundamentally mistaken associative theory of learning.\footnote{See, in particular, Gallistel & King (2010, pages 155, 187 and 278).} Since connectionism in this narrower sense is the critical focus of the present dissertation, I’ll simply refer to it using the term ‘connectionism’, except in contexts in which the contrast between it and Piccinini’s more liberal sense is salient, in which case I’ll use the term ‘associative connectionism’.\footnote{I should emphasize that this is just a convenient label. I will go on to argue in Chapter 4 that the connectionist mechanisms posited by contemporary neuroscientists are not ‘associative’ in the objectionable sense that Gallistel has in mind.}

There’s an additional notion of connectionism whose extension falls under that of Piccinini’s liberal notion, but which should be distinguished from associative connectionism. In fact, we’ve already marked this distinction; we noted that the logic circuits in digital computers are connectionist in Piccinini’s liberal sense, but not in
the sense of associative connectionism. More generally, we can identify a sense of connectionism according to which the mechanisms that explain cognition are networks of interconnected nodes that, perhaps together with other components, operate according to broadly ‘classical’ principles: they comprise components and operations that correspond closely to the symbols, rules, variables, and memory registers of conventional computing machines. It’s tempting to refer to connectionism in this sense using the term ‘implementational connectionism’ after Pinker & Prince (1988) and Marcus (1998), since those authors seem to endorse the kind of view I have in mind; however, I think this terminology is potentially misleading, since the famous paper by Fodor & Pylyshyn (1988) has saddled the term ‘implementation’ with theoretical baggage that I doubt Marcus, at least, is prepared to carry. Fodor & Pylyshyn use ‘implementation’ to express a relation between *explanatorily autonomous levels of explanation*. They’re thus quite happy to allow that the neural mechanisms that mediate cognitive processes are connectionist in the associative sense of ‘connectionism’; they just think that mechanisms identified at the ‘neural level’ are strictly *irrelevant* to explaining cognitive capacities.\(^{42}\) In contrast, Marcus seems to use ‘implementation’ to mean something like *instantiate*. He argues that neural mechanisms might explain cognitive processes, but only insofar as they *instantiate* the components and operations of conventional computing mechanisms. Contrary to Fodor & Pylyshyn, he thus *denies*

\(^{42}\)Why? Because they hold that cognitive explanations are constitutively explanations that posit representations, and that mechanisms described at the neural level are not representations; as Fodor (1998a) later put it, “notions like... *representation* aren’t accessible” in “the language of neurology” (p.96). It’s not clear why one should think that cognitive explanations are *constitutively* representational; one might have thought that what makes explanations cognitive is that they explain cognitive capacities, and that whether or not such explanations necessarily invoke representations is an open empirical question. More importantly, it’s not clear why one should think that the ‘level’ (whatever exactly that means) at which one *describes* something (whatever exactly that means) has any implications concerning *what that thing is*. Unfortunately, I’ll have to shelve these questions for now. The point I wish to make here is simply that the notion of implementation employed by Fodor & Pylyshyn is bound up with idiosyncratic and controversial views about the nature of explanation that Marcus, Pinker & Prince, and other so-called classicists almost certainly do not intend to take on board when they employ the term ‘implementation’.
that the neural mechanisms that mediate cognition are connectionist in the associative sense of ‘connectionism’. So to avoid the unwelcome connotations of the term ‘implementation’, let’s call the kind of view that Marcus endorses — the kind of view described earlier in this paragraph — ‘classicist connectionism’.

This seems to be the kind of view that Gallistel endorses, too. Like Marcus, but unlike Fodor & Pylyshyn, Gallistel holds that cognitive capacities are to be explained in terms of mechanisms ‘at the neural level’, but that those mechanisms operate according to broadly ‘classical’ principles. Thus the distinction between classicist and associative connectionism will prove to be the axis around which debates in this dissertation turn. However, the contrast between classicist and associationist connectionism perhaps isn’t yet as sharp as we might like it to be, since we have no precise criteria by which to judge whether or not a biological mechanism operates according to ‘classical’ principles. Moreover, it’s not clear how many classical-style components it would take for a biological mechanism to count as classicist. Would Gallistel be vindicated if it turned out that in order to explain animal cognition we must appeal to connectionist networks that enter into states that qualify as ‘symbols’ on some reasonable interpretation of that term, but that do not contain any other ‘classical’ components? Presumably not; but to answer questions like these we’ll have to make judgement calls in specific cases, which, in the absence of precise demarcating criteria, will have to be informed by a general sense of the character of information processing in associative versus classical connectionist systems. Thus, to hone our judgment, I’ll close this chapter with a brief discussion of the character of information processing in (associative) connectionist networks.
2.4.4 The Character of Information Processing in Connectionist Networks

Information processing in connectionist networks is typically understood not in terms of the application of computational rules to symbols stored in memory, but rather in terms of the transformation of input activation patterns into output activation patterns. These patterns are often given a geometrical interpretation as vectors in a multi-dimensional vector space, whereby the $n$ nodes in a given layer correspond to the dimensions of an $n$-dimensional space, and the activity levels of those nodes determine the terminal point of a vector within that space. If we think of the connection weights in a network as a matrix of numbers, we can then understand computational operations within a network as the multiplication of an input vector with a weight matrix to obtain an output vector, i.e. as *vector-vector transformation*.\(^{43}\) This is regarded as *computation* in virtue of the fact that input and output vectors are taken to encode the arguments and values of a given function; the operative notion of computation here is the relatively broad one, endorsed by Gallistel and briefly discussed in Chapter 2.2.3, of ‘putting a physically realized function into action’. Churchland et al. (1990) articulate this conception of computation in the context of connectionism when they write that “a physical system computes a function $f(x)$ when there is (1) a mapping between the system’s physical inputs and $x$, (2) a mapping between the system’s physical outputs and $y$, such that (3) $f(x) = y$” (p.48).

As we’ve seen, the promise of connectionist networks is that they can *learn* to compute functions. Indeed, such networks have proven to be effective at both pattern classification and regression, i.e. learning to categorize inputs into distinct classes of outputs, and learning to assign inputs to real-valued outputs, respectively. In the con-

\(^{43}\)This idea appears to have originated with Pellionisz & Llinas (1979), and was subsequently introduced into the philosophical literature by Paul Churchland (1986).
text of neo-connectionist psychological modeling, the pattern classification capacities of networks are often treated as an analog of perceptual discrimination or semantic categorization capacities, and when networks are employed in this way their states are typically given a semantic interpretation, as concepts or other mental representations. Let’s now look a little more closely at how mental representations have been understood in the context of neural network theorizing.

When a multilayer perceptron has been trained via backpropagation to classify a subset of its inputs as falling under a given category, those inputs will tend to elicit similar activation patterns across the hidden layer of the network, in the sense that the activation patterns will correspond to points in the vector space of the hidden layer that ‘cluster’ together in a sub-region of that space. Training is thus said to ‘partition’ hidden-layer activation space into distinct regions, which constitutes how the network categorizes its input. Consider, for example, the network described by Gorman & Sejnowski (1988), which was trained to discriminate between sonar signals received from rocks and sonar signals received from naval mines. After training, ‘rock’ signals tended to elicit activation vectors that clustered together in one partition of hidden-layer activation space, whereas ‘mine’ signals tended to elicit vectors that clustered in another partition, as depicted in Figure 2.6. When a network like this is interpreted as a cognitive model, the partitions are often characterized as concept types, whereas particular activation patterns within those partitions are tokenings of those types.

The partitions within a network cannot simply be ‘read off’ from the structure of the network by a human observer, in the way one might imagine ‘reading off’ the symbolic representations stored within a classical mechanism. However, partitions can be revealed by various methods of statistical analysis. One of the earliest and

 Though, in Chapter 4 I’ll argue that this image of ‘reading off’ the representations stored within a computational system is deeply beguiling, even in the case of classical systems.
Figure 2.6: Partitions in the hidden-layer activation space of a multilayer perceptron trained to distinguish between sonar signals received from rocks and sonar signals received from naval mines. After training, ‘rock’ signals tend to elicit activation vectors that cluster together in one partition of activation space, whereas ‘mine’ signals elicit activation vectors that cluster in another partition. Signals of either type that are statistically more ‘prototypical’ of signals within the training data elicit patterns that occupy a more central location within a given cluster. From Churchland (1990), based on research by Gorman & Sejnowski (1988).

Most well-known uses of such a method was the use of hierarchical cluster analysis to reveal the categorical structure of NETtalk, the aforementioned network developed by Sejnowski & Rosenberg (1987) to transform written English text into speech. NETtalk achieves this feat by learning to map certain letters, which are presented to the network in the context of other letters, to a phonemic representation that can enter into a speech synthesizer. The categorical structure within the network that mediates this capacity was revealed by eliciting all letter-phoneme activation vectors in the hidden layer of the network, averaging the pairs of vectors that are closest together in activation space, and repeating this process with the resulting averaged vectors until only two averaged vectors were left. The outcome of this process is depicted by the tree
A crucial point lying just below the surface of the earlier discussion of activation-space representation is that the geometric interpretation of connectionist networks provides a measure of distance between activation vectors, and hence, when a network is successfully trained so that its partitions reflect semantically meaningful categories, provides a measure of semantic similarity. This helps to elucidate a number of psychologically interesting properties of connectionist networks. First, networks are capable of generalization: they can correctly classify novel input patterns that were not present in their training data. For example, the Gorman & Sejnowski (1988) network was able to distinguish between novel rock- and mine-derived sonar signals with > 90% accuracy. How does such generalization work? If a network receives a novel input activation pattern that is sufficiently similar to input patterns it has learned to classify as $P$, the novel input will elicit a hidden-layer pattern within a partition corresponding to the $P$ classification, which will then elicit an appropriate output pattern.

A second interesting property concerns prototypicality effects. We often regard some objects as being more representative or prototypical of a given category than others. Most of us would probably agree that a Labrador is more prototypical of the category DOG than a Chihuahua. Indeed, this is a pervasive feature of human cognition, as revealed by a range of behavioral measures (Rosch, 1978). To illustrate, most of us would probably be able to classify a Labrador as a dog faster than we would a Chihuahua. Clustering in connectionist networks provides a sketch of how this phenomenon might be cognitively realized. When a network has learned a certain
Figure 2.7: ‘Dendrogram’ produced by hierarchical cluster analysis of the hidden-layer of NETtalk, a connectionist model trained to transform written text into speech. The dendrogram reveals the hierarchical structure of partitions within NETtalk’s hidden-layer activation space. Many of these partitions correspond closely to phonological categories. From Sejnowski & Rosenberg (1987).

classification, such that a particular subset of its training data reliably elicit hidden-layer activation vectors that cluster within a specific region of activation space, input patterns that are statistically highly representative of patterns within that subset of training data will elicit hidden-layer vectors that occupy a central location within the cluster. Input patterns that are less representative will elicit hidden-layer vectors that
lie on the periphery of the cluster. They are the Chihuahuas of activation space.

Introducing recurrent connections into a connectionist network requires a slight modification to the notion of activation pattern representation, but it also helps account for a number of other psychologically interesting phenomena. As mentioned earlier, recurrent excitatory connections enable the activity within a network to feed-back upon itself, and to thereby sustain dynamic, reverberating patterns of activity across the entire network, as opposed to the static activation patterns within a single layer of a feed-forward network. These reverberating patterns can be analyzed using the tools of dynamical systems theory, as attractors in a multi-dimensional phase space, where — as with the vector spaces of feed-forward networks — each dimension corresponds to a node in a given network. Recurrent networks that are capable of sustaining reverberating patterns are thus often called attractor networks, and can be classified on the basis of the ‘shape’, i.e. manifold, that their attractors trace through phase space; thus, for example, we have point, line, ring, and even chaotic attractors.\footnote{For an overview of these different types of attractor network, and of the various explanatory uses to which they might be put, see Eliasmith (2005).}

The analog of the partitions in a trained feed-forward network are, in the context of attractor networks, basins of attraction: regions of phase space within which a trajectory will converge toward a given attractor. As in feedforward networks, when attractor networks are employed as psychological models, attractors and basins of attraction are often characterized as analogs of memories or mental representations.\footnote{John Hopfield (1982) pioneered the development of attractor network models, demonstrated how they could be analyzed using the tools of dynamical systems theory, and identified many of the psychological interesting properties of attractor networks that I’m about to discuss.}

One psychologically interesting property of attractor networks is that they exhibit pattern completion: if a network has ‘stored’ a given item of information, in the sense that it has formed a lasting disposition to enter into a particular dynamic pattern of activity upon receiving that item of information as input, then the subsequent pre-
sentation of a even a small fragment of that information to the network will elicit the entire activity pattern. This phenomenon is analogous to ‘content-addressable memory’, a type of memory employed in certain engineered computers that are designed for high-speed search applications, in which the presentation of an input datum will immediately return all memory addresses containing that datum within a single operation. Pattern completion and content-addressability seem to be pervasive features of human cognition, too; we can infer the identity of something that is visually occluded simply by seeing a small part of it, and a fragmentary hint of something we are trying to recall will often bring to mind the entire memory. Attractor networks thus promise to elucidate the neural mechanisms that mediate these cognitive capacities.\(^{47}\)

Another interesting property of attractor networks, as well as of feed-forward networks, is that they exhibit graceful degradation. There are at least two different ways of understanding this phenomenon, which aren’t always distinguished in the literature. According to one, graceful degradation is a close relative of the kind of pattern completion we just considered: a network is able to respond appropriately to incomplete or distorted input, to the extent that the input is similar to the data the network was trained on. Anyone who has suffered through a series of ‘syntax error’ messages when programming a computer will be all too familiar with the converse of this phenomenon. According to a second way of understanding graceful degradation, it pertains to a network’s resilience to damage: artificially lesioning a network by removing one or more connections or nodes will typically not eliminate the functionality of the network in one fell swoop, but will instead gradually degrade the global performance of the network in proportion to the number of connections or nodes that are removed. This phenomenon is analogous to Lashley’s infamous ablation experiments, which suggested that a rat’s

\(^{47}\)Indeed, as I’ll discuss in later chapters, researchers have had considerable success using biologically realistic third-generation attractor networks to model a range of specific memory systems in the brains of humans and other animals (see, e.g., Rolls, 2010).
memory of how to navigate a maze was not eliminated by lesioning any particular part of the rat’s brain, but was gradually diminished in proportion to how much of the brain was lesioned (Lashley, 1950). While different, both kinds of graceful degradation exhibit a kind of robustness and adaptability that is distinctive of many biological systems, but which contrasts with the ‘brittleness’ of many human-engineered mechanisms, which simply stop working when the proverbial spanner is thrown in the works.

Perhaps the most important property of connectionist representation, which to some extent underlies all the other properties we’ve considered so far, is their distributed nature: both activation vectors and attractor states are distributed across the activities of numerous different nodes in a network. Note that that isn’t necessarily to say that all the nodes within a network significantly contribute to a given activation pattern. For example, Sejnowski & Rosenberg (1987) report that on average only about 20% of the hidden units in NETtalk contributed to the activation patterns elicited by a given input; the activation patterns were, as it’s sometimes put, ‘sparsely distributed’. It’s also worth noting that not all connectionist networks employ distributed representations. In ‘localist’ networks, the activity of a given node is taken to have semantic significance — as indicating, for example, the presence or absence of a given semantic feature.

Another important qualification to be made here is that distributed representa-

48Lashley’s views about the distributed nature of the ‘engram’ are clearly evocative of the idea that memories are distributed across the nodes or weights of a connectionist network. However, the resemblance here is only superficial. Lashley held that memory traces are globally distributed across the entire cortex, rather than sparsely distributed across the synapses of specific populations of neurons, as most contemporary connectionists believe. Indeed, Lashley (1950) took the results of his ablation experiments to be “incompatible with theories of learning by changes in synaptic structure”, noting that “integration cannot be expressed in terms of connections between specific neurons” (p.176).

49Note that in the remainder of this section I’ll use ‘activation pattern’ to refer indiscriminately to both static patterns of activation in feedforward networks (i.e. vectors in hidden-layer activation space), and dynamic patterns of activity in recurrent networks (i.e. attractors in phase space).

50However, localist networks are arguably best understood as connectionist in Piccinini’s liberal sense, not in the stronger associative sense that is of primary interest here, since such networks lack many of the distinctive features of associative connectionist networks that we just surveyed.
tions or encodings per se are by no means exclusive to, or even especially distinctive of, connectionist networks. In fact, distributed coding is pervasive in communications technology, largely because of the fault tolerance it affords. However, there is an important kind of distribution that is distinctive of connectionist representation, which involves a kind of intermingling or ‘superposition’ of the resources used to instantiate different representational vehicles. This occurs into two importantly different ways: diachronically and synchronically. Diachronically, different activation patterns occurring at different times within the same network are constructed out of the same basic set of components — namely the activation levels of nodes — hence they overlap to the extent that they share the same components. This is the structural basis of the distance metric defined over vectors in activation space. Moreover, in a network that has been trained to partition its activation space in a way that reflects meaningful categories, it becomes the structural basis of a similarity metric defined over representations, such that, as van Gelder (1991) puts it, “similarities and differences among the items to be represented will be directly reflected in similarities and differences among the representations themselves” (p. 41). By virtue of this property, the relation between a connectionist representation and its content is often said to be non-arbitrary, in contrast with the alleged arbitrariness of the symbolic representations in a conventional computer: in order for an activation pattern representation to have a given content, it must (qua vector) occupy a specific location in the activation space of the network, and hence must (qua pattern) have a specific structure.

Synchronically, the resources used to store a given activation pattern within a network are intermingled with the resources used to store every other activation pattern within the same network. To understand this, we first need to understand how an activation pattern might be ‘stored’. Activation patterns are occurrent states of a network that are elicited by a particular input. Which pattern will be elicited by which input is determined by the matrix of connection weights in the network. The
weights thus embody a disposition for a given activation pattern to be elicited, and in this sense encode a ‘memory trace’ of that pattern. Now, some connection weights might play a more important role than others in determining which pattern is elicited by a given input, but typically no single weight will play a decisive role; which pattern is elicited is determined collectively, by many different weights. Correspondingly, a given set of weights might play a role in eliciting many different activation patterns. That is why destroying the weighted connections in a network results in the ‘graceful degradation’ in the global performance of the network, in proportion to the number of connections that are destroyed. So, in sum, both occurrent activation patterns and their memory traces are distributed; but, more importantly, they are superpositionally distributed, and that’s what explains many of their most interesting and distinctive properties.

Thus concludes our survey of some of the history and theory underpinning the connectionist orthodoxy in contemporary neuroscience. Of necessity, much was left out. But hopefully we now have a clearer view of the target of Gallistel’s criticisms. It is to those criticisms that I will now turn.


Chapter 3

Exorcising the Connectoplasm

Nature abhors a vacuum tube.
— Myron Glass, engineer at Bell Labs c.1940

3.1 Introduction

Gallistel discusses several interconnected objections to the tenability of explaining animal cognition within the connectionist framework I’ve just outlined. The purpose of this chapter is to present these objections as clearly and charitably as possible. For stylistic reasons, I’ll sometimes speak with Gallistel’s voice, omitting qualifiers like ‘in Gallistel’s view...’, but I shouldn’t be read as endorsing these objections; I’ll go on to reply to them in later chapters. Gallistel’s objections to the connectionist orthodoxy in contemporary neuroscience stem from what he takes to be a fundamental architectural limitation of connectionist networks: they lack a read-write memory. Because of this, he thinks, connectionist networks are effectively look-up tables or finite-state machines, and hence suffer the crippling computational inadequacies of those systems — most notably the problems of preSpecification and combinatorial explosion. Thus although connectionist networks might play some limited role in animal brains, just as,
say, look-up tables play an important but limited role in conventional electronic computers, they cannot play a prominent role in explaining the most interesting aspects of animal cognition — the aspects that are distinctive of intentional agency.

Why? Because, for Gallistel, the aforementioned architectural problems reflect a much deeper *conceptual* problem with contemporary connectionist neuroscience; specifically, with the theory of learning that he takes to lie at its heart. Gallistel sees a ‘conceptual chasm’ running the length of the history of psychology, separating psychologists into two irreconcilable camps, each with a characteristic way of understanding the nature of learning, and a corresponding way of understanding the functional architecture of the brain.¹ On one side of the chasm is ‘mainline’ or classical cognitive science, exemplified by the work of Chomsky, Fodor, Marr, Newell, and Pylyshyn, amongst others. According to this ‘computational/representational’ tradition, “learning is the extraction from experience of information about the world that is carried forward in memory to inform subsequent behavior” (Gallistel & King, 2010, p.187). More specifically, learning in this tradition is understood as a matter of extracting the values of environmental variables from experience and storing them in the form of physical symbols, which carry information about those values ‘forward in time’, such that the symbols can later enter into the computational operations that issue in motor commands, and thereby inform behavior. Since information can only be communicated when a receiver of that information possesses a prior probability distribution defined over the possible values of a given variable, the present view holds that the brain contains a range of domain-specific learning modules which, *qua* receivers, implicitly embody within their processing architecture a specification of the range of possible values of the variables that are relevant to their specific domain (Gallistel, 2000). In

¹These claims appear most explicitly on page 101 of Gallistel & King (2010), but also lie just below the surface at various places in Gallistel (2000), Gallistel (2006), and Gallistel (2008).
perhaps more familiar terms, the brain embodies innate principles that are filled in by the values of specific parameters derived from experience (Chomsky, 1995). However, the present view also allows that variables themselves can be learned; they needn’t be merely implicit in the architecture of a learning module, but can be stored explicitly as manipulable symbols, through indirect addressing. This story about learning has clear ramifications for understanding the computational architecture of the brain: it presupposes that the brain contains computational mechanisms which have access to an indirectly addressable read-write memory that can store symbols in a computationally accessible form. This, for Gallistel, amounts to saying that the brain has the computational architecture of a Turing machine.

On the other side of the chasm — clearly the dark side in Gallistel’s Manichean view of things — is a tradition with a long history, extending back to the Early Modern empiricists, if not to Aristotle centuries before, running through the associationist psychology of the 19th century, to the behaviorism of the early 20th century, eventually culminating in the connectionist orthodoxy of contemporary neuroscience. The common thread running through all these views is the idea that learning consists in the formation of associations between certain entities in the mind or brain on the basis of experienced environmental regularities. While the entities at issue have changed over the years — the Early Moderns wrote of associations between ideas, the behaviorists of associations between stimuli and response — the fundamental conception of learning has not. When neuroscientists report the most cutting-edge research on STDP, they’re essentially still appealing to “much the same rewiring rules that Locke suggested” (Gallistel & King, 2010, p.253). More broadly, the very notion that learning involves neural plasticity is not a pellucid deliverance of experimental observation, but is the product of a particular way of thinking about the nature of learning, the associative theory of learning: “changes in synaptic conductance are the physiologists’ conception of how the brain realizes the changes in the strengths of associative bonds” (ibid., p.xiii).
Gallistel defines the associative theory of learning as the view that learning consists in the “rewiring by experience of a plastic brain so as to make the operation of that brain better suited to the environment” (ibid., p.187). Fodor gives fuller expression to the view when he writes that,

The human mind is a blank slate at birth. Experience writes on the slate, and association extracts and extrapolates whatever trends there are in the record that experience leaves. The structure of the mind is thus an image, made a posteriori, of the statistical regularities in the world in which it finds itself.

(Fodor, 1998b, p.203)

As with the ‘computational/representational’ approach to learning, the associative theory is said to be wedded to a particular conception of the computational architecture of the brain. On the associative view, input from experience rewire a plastic brain so as to make it better adapted to the environment, but input isn’t recorded as a discrete symbol that might be manipulated by computational operations. To the extent that the notion of computation finds any purchase here at all, computation in an associative system is simply a matter of transitioning from one state to the next in a way that is determined by the current pattern of associations. These associations aren’t explicitly ‘read’; they just implicitly determine the sequence of states the system is disposed to transition through. We saw an instance of this general idea in our earlier discussion of connectionism: recall that computation in connectionist networks is typically understood as the transformation of an input vector into an output vector via a matrix of connection weights. These weights, or associations, can be rewired by experience so the state-transitions of the associative system better reflect properties of environment, but the fundamental computational architecture here is nevertheless that of a finite-state machine. Thus, insofar as contemporary neuroscientists attempt to explain learning in terms of associative plasticity mechanisms, they are “committed to the thesis that the brain has the functional architecture of a finite-state automaton, rather than that of a Turing machine: it lacks a read/write memory” (Gallistel & King,
2010, p.176). Further, because connectionist networks cannot store and manipulate symbolic representations, contemporary neuroscience “represents a resurgence of the behaviorist view that the brain does not really represent the external world” (Gallistel, 1998, p.11).

Gallistel is of course not alone in suggesting that connectionism, broadly construed, is a descendant of either associationism or empiricism or behaviorism (take your pick). Sometimes this suggestion is made by proponents of connectionism. For example, Rosenblatt (1958) writes that his ‘connectionist’ position, “which stems from the tradition of British empiricism, hazards the guess 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” (p.386). However, the comparison is usually made by opponents of connectionism, with a tangible underlying presumption that connectionism inherits certain flaws of its purported predecessors. For example, Papert (1988) sees connectionism as “behaviorism in computer’s clothing” (p.9). Pinker (1997) claims that “a perceptron... is a high-tech implementation of an ancient doctrine: the association of ideas” (p.113). Fodor & Pylyshyn (1988) do these authors one better by suggesting not only that connectionism is plagued by the same problems as the behaviorism of figures like Hull (p.49), but also that it gives rise to a psychology “not readily distinguishable from the worst of Hume and Berkeley” (p.64).² It’s clear that this is all supposed to be a very bad thing. But it’s often not clear what exactly is being claimed of connectionism when it is characterized as associationist, or empiricist, or behaviorist, and why

²It should be noted that although the references to Hull and Hume suggest that Fodor & Pylyshyn (1988) regard connectionism as a species of both behaviorism and empiricism, those authors do assert that “Classicists and Connectionists are all Representational Realists” (p.11), and that connectionism is officially “neutral on the Empiricism/Nativism issue” (p.60). So perhaps when they compare connectionism to the views of Hull and Hume, the relevant dimension of comparison is the associationism that all of these views purportedly share; perhaps we shouldn’t take Fodor & Pylyshyn to endorse the view that connectionism is a manifestation of behaviorism or empiricism after all. However, they do end their paper by castigating connectionism for reviving “the Empiricist idea that all learning consists of a kind of statistical inference” (p.63). It’s all a bit confusing.
exactly that is supposed to be problematic.

In a series of articles published over the last two decades or so,\(^3\) coming together in a recent book co-authored with the computer scientist Adam King (Gallistel & King, 2010), Randy Gallistel has presented a tightly interconnected set of reasons why a commitment to the associative theory of learning, and to the corresponding view about computational architecture, make trouble for contemporary connectionist neuroscience. These cluster into three broad themes — Representation, Computation, and Learning — which I’ll discuss below, in sections 3.3, 3.4, and 3.5 respectively.

At the center of Gallistel’s cluster of arguments is the claim that contemporary neuroscience is anti-representational. This might be news to connectionist psychologists and cognitive neuroscientists, who typically think of themselves as part of a cognitivist tradition that eschews behaviorism and embraces representations as the inner mediators of behavior. For example, the editors of the so-called ‘bible’ of the neo-connectionism of the 1980s were “explicitly concerned with the problem of internal representation” (Rumelhart et al., 1986b, p.121). Appeal to neural representations of “non-neural parameters, such as head velocity or muscle tension or visual motion” is ubiquitous in contemporary computational neuroscience (Churchland & Grush, 1999, p.157). Even some of the staunchest critics of broadly connectionist approaches agree that “Classicsists and Connectionists are all Representational Realists” (Fodor & Pylyshyn, 1988, p.11).\(^4\) So despite what he sometimes seems to suggest, Gallistel isn’t just reporting the views of connectionist neuroscientists; he’s making the substantive claim that, despite what they might think, the mechanisms they posit can’t effectively function as representations. His basis for this claim is that neural


\(^4\)Indeed, the central argument of Fodor & Pylyshyn (1988), that connectionist representations don’t have combinatorial structure and hence can’t explain cognitive systematicity, simply doesn’t make sense on the assumption that connectionists are anti-representationalists.
networks, as conceptualized by contemporary neuroscientists, lack a read-write memory. He marshals a range of engineering considerations to argue that none of the memory mechanisms currently recognized by neuroscientists, such as synaptic plasticity or reverberating neural circuits, could write information to a thermodynamically stable substrate in such a way that it could subsequently be read by computational operations, and, moreover, that none of these mechanisms employ efficient, compact information coding schemes. Thus these mechanisms face severe constraints on the amount of information they can encode, and the durations for which they can encode it. Because of this, they cannot sustain the kind of rich functioning homomorphisms that are required to explain animal cognition.

The lack of a read-write memory is also said to severely constrain the computational power of neural networks. As I mentioned earlier, Gallistel holds that by lacking a read-write memory, neural networks have the functional architecture of finite-state machines, and that such mechanisms suffer from the problems of pre-specification and combinatorial explosion. These problems are obviated in conventional engineered computers because they are equipped with an indirectly addressable read-write memory, which allows addresses in memory to be explicitly symbolized and therefore accessible to computational operations. This enables symbols to ‘point’ to locations within memory, which might contain a range of different symbols, thereby introducing a functional distinction between variables and their instances. This introduces a capacity for variable-binding, which allows for the flexible generation of complex symbolic structures that weren’t pre-specified in the functional architecture of the computational mechanism. However, Gallistel claims, neural networks cannot avail themselves of these benefits; such networks lack a read-write memory, so a fortiori, they lack an indirectly addressable read-write memory.

Much of the research in the connectionist tradition has of course focussed on the question of how neural networks can learn, and one might have though that this re-
search would speak to the problems of pre-specification and combinatorial explosion. However, Gallistel argues that this research presupposes an associative theory of learning that “fundamentally misrepresents the nature of learning” (Gallistel & King, 2010, p.278). The core of his complaint is that the associative theory characterizes learning as a mechanistic process by which associative bonds are gradually and automatically formed on the basis of experienced statistical regularities in the environment, which ignores a vast body of evidence from contemporary learning theory that, by contrast, learning is a rational process by which an animal actively forms and tests hypotheses about causal contingencies within its environment. For example, Gallistel points out that while conditioning has traditionally been understood as a process whereby associations are automatically formed on the basis of temporal pairing between stimuli and response, the consensus within contemporary animal learning theory is that temporal pairing is neither necessary nor sufficient for conditioning; what’s crucial is that the animal comes to predict that certain stimuli are contingent upon others. This and other evidence from contemporary learning theory suggests that animals form hypotheses about the causal structure of their environment, and update these hypotheses in accordance with broadly Bayesian principles. Gallistel argues that not only is this perspective antithetical to the associative theory of learning, it simply can’t be explained by the associative mechanisms recognized by neuroscientists, for Bayesian inference involves complex computations defined over probability distributions, yet connectionist networks can neither encode nor compute with probability distributions since, again, they lack an addressable read-write memory.

Gallistel & King (2010) make these objections concrete by showing how they apply to a specific neural network model that they take be an exemplar of contemporary connectionist theorizing in neuroscience: a model of path integration developed by Samsonovich & McNaughton (1997). I’ll discuss this model, and Gallistel’s criticism of it, in section 3.6. The upshot of these criticisms of contemporary connectionist
The computation of a representation seems to require a functional architecture that is not transparently consistent with our current understanding of neurobiological mechanisms... [T]he behavioral evidence for representation and computation, together with basic insights in computer science about the centrality of a read-write memory mechanism in physically realized computation, implies the existence of a neurobiological read-write memory mechanism... [T]he discovery of such a mechanism may someday have an impact on neuroscience comparable to the impact of the discovery of radioactivity on physics.

(Gallistel, 2008, p.227-8)

Just as Darwin’s estimates of the age of the earth initially seemed incompatible with what was then thought to be known about the availability of energy in the solar system, but were subsequently vindicated by the discovery of radioactivity, so too, in Gallistel’s view, will the behavioral evidence for the existence of a read-write memory mechanism be vindicated by future neuroscientific discoveries. Gallistel believes that it is futile to speculate, given our present state of ignorance, what those future discoveries will reveal. However, he suggests that the stability and energy efficiency of the mechanisms of genetic memory — DNA, RNA, and their supporting molecular machinery — make them promising candidates for the mechanisms of ontogenetic, read-write memory. I’ll summarize Gallistel’s speculations about the molecular basis of memory in section 3.7. However before getting on to Gallistel’s criticisms and their purported consequences, I should first discuss Gallistel’s conception of the relationship between contemporary neuroscience and associationism, and the scope of the problem that he sees with contemporary neuroscience, so we have as clear a view as possible about what exactly Gallistel is claiming.
3.2 The Scope of the Problem

Someone acquainted with the richness and breadth of contemporary neuroscience might be tempted to regard Gallistel’s arguments as a *reductio* of their own premises. For, one might think, the clear success of explaining learning and memory in terms of ‘connectionist’ mechanisms of neural plasticity constitute an existence proof that the conclusion of Gallistel’s arguments is false, or is overwhelmingly likely to be so. Since the discovery of LTP in the early 1970s, research into the mechanisms and functional significance of synaptic plasticity has become one of the most active and productive areas of neuroscience. A search of PubMed reveals on the order of 10,000 articles with ‘synaptic plasticity’ and ‘learning’ or ‘memory’ as keywords. Many of these articles report clear experimental evidence that synaptic plasticity plays an important role in learning and memory. While neuroscientists are often careful to emphasize that a strong form of the synaptic plasticity hypothesis — that synaptic plasticity is both necessary and sufficient for storing memory traces in the brain — has yet to be conclusively established\(^5\), very few neuroscientists doubt that synaptic plasticity plays an important role in various kinds of learning and memory in a range of different brain structures (Martin & Morris, 2002). In any case, there simply is no well-understood and experimentally vindicated neurophysiological mechanism *other* than some form of neural plasticity that could plausibly encode memory traces in the brain. Researchers have investigated various alternative mechanisms and all have proven to be empirically inadequate (Leiman & Christian, 1973). Unlike Fodor’s claim about the explanatory status of the Language of Thought Hypothesis circa 1975, it really is no exaggeration to

---

\(^5\)Some think that this is primarily for technical reasons. While there is a great deal of evidence that synaptic plasticity is *necessary* for memory storage, we arguably don’t yet have the technology to evaluate whether it is *sufficient*. That is, we can’t yet implant false memories by manipulating, *en masse*, the synapses within specific neural networks of a behaving animal’s brain (Martin & Morris, 2002).
say that connectionism is currently *the only game in town* when it comes to explaining the neural basis of learning and memory.

Gallistel agrees that, in a very thin sense, this is true. Indeed, it’s a truism: “No one believes that how the brain is wired is irrelevant to how it functions” (Gallistel & King, 2010, p.259). Thus Gallistel seems to endorse connectionism in the weak sense outlined by Piccinini (2008b) that I mentioned in Chapter 2.4, according to which *some* psychological capacities are explained, in part, by appealing to networks of interconnected neurons. More specifically, he seems to endorse what I earlier called *classicist connectionism*, the view that the various mechanisms recognized by neuroscientists, which are usually characterized in terms of experience-dependent plasticity, in fact operate in a distinctively classical manner. For example, Gallistel & King (2010) suggest that nothing *in principle* prevents neuroscientists from speculating that the brain is a “conventional computer built out of neurons” (p.246), in which symbolic variables are encoded in synapses (p.191), using a compact binary code that is decoded by neurons configured to behave as AND-gates (pp.260-1).

From Gallistel’s perspective, this view has numerous explanatory advantages. In particular, it avoids the various problems with ‘associative’ connectionism sketched above. So why, asks Gallistel, don’t neuroscientists take advantage of the power of classical computing mechanisms and embrace classicist connectionism? Because, in Gallistel’s view, they are convinced that the nervous system operates according to connectionist principles due to a prior conviction that learning is fundamentally associative. It is the strength of this conviction, rather than direct experimental evidence, that has given rise to the contemporary consensus within neuroscience that learning and memory involve synaptic plasticity; Gallistel holds that the actual evidence for the synaptic plasticity hypothesis is “murky and controversial”, and far weaker than many appreciate (Gallistel & King, 2010, p.178-9). Kuhn warned that deeply-held theoretical convictions can distort how experimenters interpret their observations, and
Gallistel seems to think that this is precisely what has happened in neuroscience: the associative theory of learning has provided a distorting lens through which neuroscientists have ‘seen’ evidence for a role of synaptic plasticity in learning and memory where there really is none. Moreover, peering through the lens of the associative theory of learning has, in Gallistel’s view, prevented neuroscientists from seeing mechanisms that lie outside of the purview of that theory, mechanisms of read-write memory.

Yet, Gallistel claims, there must be such mechanisms. He takes his claim that an indirectly addressable read-write memory is a necessary component of any mechanism capable of performing complex computations under real-world constraints to have the character of physical or even logical necessity, and to be more or less entailed by mathematics and computer science: “the most fundamental aspects of the functional structure of a computer are dictated by the logic of computation itself and that, therefore, they will be observed in any powerful computational device, no matter what stuff it is made of” (Gallistel & King, 2010, p.167). Moreover, he takes his arguments that none of the neurobiological mechanisms currently recognized by neuroscientists could play the functional role of a read-write memory to be decisive. His claim isn’t merely that we don’t yet have detailed, well-confirmed models of the neural mechanisms that mediate the kind of behavioral capacities discussed in Chapter 2.3. That’s uncontroversial. His claim, rather, is that to explain such capacities, neuroscience must undergo something like a Kuhnian paradigm shift: there must be “mechanisms in the nervous system not yet dreamed of in the philosophy of neuroscientists” (Gallistel, 2006, p.70).

It’s important to reflect for a moment on just how revolutionary this conclusion is. If Gallistel is right, then — despite decades of careful experimental and theoretical research, despite the publication of thousands of research reports apparently to

---

6 “[N]euroscientists think that [some form of synaptic plasticity] must be the mechanism of learning and memory [because] this hypothesis is consistent with what they take to be established psychological fact, namely that learning and memory are associative processes” (ibid., p.178).
the contrary, despite the laureating of Nobels — neuroscientists have, quite literally, learned *nothing* about the mechanisms of learning and memory. Nevertheless, Gallistel thinks that this state of affairs has precedent in the history of science. He takes the current epistemic situation in neuroscience to be analogous to that in molecular biology prior to 1953. Evidence from classical genetics suggested the existence of a molecular mechanism of heredity with two strange properties: the ability to replicate itself, and the ability to encode the structure of proteins. None of the mechanisms recognized by the molecular biology of the time exhibited these properties, and it was only by searching for a mechanism that *did* exhibit these properties that Watson & Crick (1953) were able to elucidate the structure and function of DNA. Neuroscience has yet to produce its Watson & Crick; the only thing preventing it from doing so, in Gallistel’s view, is a commitment to associationism.

One last point of clarification before moving on. The astute reader will have noticed a tension in the foregoing exposition of Gallistel’s views. On the one hand, I said that Gallistel endorses classicist connectionism, the view that currently recognized neural mechanisms operate according to classical principles. Yet on the other, I said that Gallistel believes that no currently recognized neural mechanisms could implement a read-write memory. This is in fact a tension in Gallistel’s own writings that he doesn’t explicitly address, but I think it can be easily resolved. I think Gallistel is best interpreted as claiming that although currently recognized neural mechanisms might implement the *processing elements* of a classical computer — that is, neural networks might be wired up like McCulloch-Pitts networks to implement logic circuits — we must look beyond currently recognized neural mechanisms to find the mechanisms of read-write memory. This interpretation isn’t transparently consistent with *everything* that Gallistel says; as we saw above, Gallistel suggests that synapses might ‘in principle’ implement the elements of a read-write memory. However I think it provides the most charitable interpretation of Gallistel’s claims taken as a whole, so I’ll assume
this interpretation in what follows. We’re now in a position to unpack Gallistel’s arguments against associative connectionism. I’ll begin with his objections to connectionist representation.

### 3.3 Problems with Connectionist Representation

Recall that in Gallistel’s view, for an animal to successfully get by in the world, its behavior must be appropriately informed by distal stimuli, and for that to be possible, the animal’s brain must be capable of encoding the various proximal stimuli impinging on the animal’s sensorium into a signal that can subsequently be decoded into a symbol that embodies explicit information about the distal stimulus. This decoding process requires inverting the initial encoding so that information conflated in the sensory signal is disambiguated and made explicit. It’s only when information is made explicit in this way — i.e. when specific messages about distal stimuli are physically embodied in thermodynamically stable but computationally manipulable symbols — that it can be written to memory and subsequently read from memory so as to enter into the complex computations that eventuate in appropriate behavior. Moreover, it’s only via causal chains like this, whereby messages are communicated from distal stimuli to internal symbol systems, that the functioning homomorphisms constitutive of representation can be sustained.

Thus it would seem that in Gallistel’s view, having a read-write memory is physically or even conceptually necessary for a mechanism to instantiate representations. This suggests the following first pass at what Gallistel takes to be the fundamental problem with the connectionist mechanisms recognized by contemporary neuroscientists, the problem from which all the other problems ultimately derive: connectionist networks lack a physical mechanism that could store explicit vehicles of information in a stable but computationally accessible form, i.e. they lack a symbolic read-write
memory. Hence they cannot comprise a system of symbols and computational operations defined over those symbols, hence they cannot be functionally homomorphic to systems of distal stimuli, hence they cannot in principle represent anything. Gallistel doesn’t state this objection explicitly, but he does come very close. For example, in a (2008) paper, he writes that “the lack of a symbolic memory is what distinguishes a neural net from a conventional computing machine” (p.235), and that a symbolic memory is “essential to computation and representation” (p.230, my emphasis) because “[a] representation, that is, a functioning homomorphism, exists only when a machine can [perform computations] on its symbols” (p.239). This strongly suggests, if not entails, that neural networks are strictly incapable of representing anything.

However, the aforementioned objection doesn’t seem to be a plausible reconstruction of Gallistel’s more considered views about connectionist mechanisms, for Gallistel’s account of representation doesn’t seem to require that representing systems incorporate a read-write memory. It’s true that Gallistel sometimes suggests in passing that representing systems must incorporate a read-write memory, as we saw in the quote from Gallistel’s (2008) paper in the last paragraph. However, when it comes to spelling out the details of his account of representation, Gallistel typically doesn’t identify a read-write memory as an essential component of a representing system. As we saw in Chapter 2.2, the central idea of his account is that a representation “is a system of symbols [that is] isomorphic to another system... so that conclusions drawn through the processing of the symbols in the representing system constitute valid inferences about the represented system” (Gallistel, 2001, p.9691). The essence of this view of representation is that “the brain creates models of the behavior-relevant aspects of the environment” (Gallistel, 1990a, p.27). There is no mention of a read-write memory here; neither the letter nor the spirit of Gallistel’s account seems to require that representing systems incorporate a read-write memory.

Moreover, several of Gallistel’s central examples of representing systems don’t in-
corporate a read-write memory in any obvious sense. For example, Gallistel & King (2010) illustrate the notion of a functioning homomorphism by discussing an electronic weight scale, which is held to mediate a functioning homomorphism between possible states of the scale and possible weights by virtue of the fact that “[t]he transducer in the scale that translates a weight into a bit pattern is engineered so that progressively heavier loads produce bit patterns that encode progressively higher numbers” (p.57). There’s no clear sense in which the scale needs a read-write memory in order to perform this function. Perhaps, appearances to the contrary notwithstanding, one might hold that this example is merely intended to illustrate Gallistel’s account of representation, and isn’t intended as a literal exemplar of that account. But consider another of Gallistel’s examples, which clearly is intended to be a literal exemplar of his view: “The circadian clock... is perhaps the simplest of all the well-documented functioning [homomorphisms] between brain processes and the world” (Gallistel, 1998, p.28). Circadian clocks are biochemical mechanisms that produce an endogenous oscillation with a period of roughly 24 hours, and hence are homomorphic with the Earth’s day-night cycle. But again, there’s no clear sense in which they incorporate a read-write memory.⁷

What seems to be crucial for whether something counts as a representing system in Gallistel’s view is whether it can perform computations on symbols: “What makes a representation are the symbols, together with the operations on them” (2008, p.238). But as we saw in Chapter 2.2, Gallistel allows that there are mechanisms that can perform computations on symbols without having a read-write memory; namely, look-up tables and finite-state machines. Recall that Gallistel’s objection to these mechanisms was that their computational power and efficiency is severely limited, not that they’re

---

⁷I’ll discuss circadian clocks, and why they count as literal exemplars of Gallistel’s ‘structural’ account of representation, in much more detail in Chapter 5.
incapable of computing *simpliciter*. So Gallistel would presumably allow that these mechanisms are at least *candidates* for standing in functioning homomorphisms with external systems; they’re not conceptually precluded from doing so. Indeed, since Gallistel seems to hold that representing systems are necessarily computational systems, and that computational systems that lack a read-write memory are *ipso facto* look-up tables or finite-state machines, it would seem to follow that, by his lights, electronic scales and circadian clocks are look-up tables or finite-state machines; but Gallistel holds that they nevertheless exemplify his conception of representation.

These considerations suggest a tempering of the objection that I tentatively attributed to Gallistel a few paragraphs ago. Despite his repeated claims that connectionist theorizing in neuroscience is part of an ‘anti-representational’ tradition, Gallistel is probably best interpreted as arguing not that neural networks are strictly *incapable* of representing anything, but that their representational power is severely limited. Here’s how the argument might go: computing mechanisms that lack a read-write memory are incapable of flexibly generating novel symbols and combining them into complex symbolic structures. Thus their expressive power is severely limited. In particular, they cannot instantiate representing systems that are sufficiently rich and complex to be homomorphic with the kinds of external world systems that animals are clearly behaviorally responsive to. Neural networks and other mechanisms recognized by contemporary neuroscientists lack a read-write memory, so even though they might in principle be capable of representation, and might in practice even play *some* etiolated representational role within the brain, they cannot play the role required to explain interesting forms of animal behavior. Again, Gallistel doesn’t ever spell out this objection in so many words, but I think that it’s the most charitable reconstruction of the argument implicit in his various complaints about neural network representation.

Still, why believe that it’s true? Neural networks certainly *seem* capable of ‘writing’ information to memory — that is, of encoding information within their connectivity
matrices via experience-driven plasticity. Moreover, they seem capable of ‘reading’ this information — that is, of using encoded information to elicit particular activation patterns that can steer subsequent computation. After all, artificial neural networks are routinely and successfully used for all sorts of practical applications, ranging from speech recognition to medical diagnosis to controlling autonomous robots, and their success at these tasks is presumably explained in part by their ability to read and write information. Why, contrary to appearances, should we think that they lack a read-write memory? What exactly is a read-write memory, anyway?

Gallistel’s claim that neural networks lack a read-write memory seems to be primarily motivated by a widely acknowledged structural feature of neural networks. Unlike conventional engineered computers, neural networks lack an architectural distinction between memory and processor: the connections of a network are said to both store information, and to determine the course of information processing within the network. As the authors of the ‘bible’ of neo-connectionism put it, “[k]nowledge is not directly accessible to interpretation by some separate processor, but it is built into the processor itself and directly determines the course of processing” (Rumelhart et al., 1986b, pp.75-76). We saw in Chapter 2.4 that this intermingling of memory and processor manifests in a couple of other distinctive characteristics of connectionist information storage. First, information in a connectionist network is typically not stored as an explicit, physically distinct vehicle of information; rather, it is stored dispositionally, as a tendency to elicit a particular output upon receiving a particular input. Correspondingly, information is not stored at a specific physical location, but is distributed superpositionally over many or even all of the connections in a network. Rumelhart (1981) summarizes these points when he writes that “information [in a connectionist network] is not stored anywhere in particular. Rather it is stored everywhere. Information is better thought of as ‘evoked’ than ‘found’” (p.3).

As we’ve seen, this manner of information storage is often said to confer a number
of advantageous computational properties, such as pattern completion, automatic generalization, and fault tolerance. But in Gallistel’s view, it’s disastrous. For although the input and output nodes of a network might be said to encode information in an explicit symbolic form, there’s nothing in the network itself that reads, writes or manipulates symbols. The most obvious candidate for a symbolic memory in a network is the matrix of connection weights, but, according to Gallistel, weights don’t instantiate symbolic values, they merely instantiate a “procedure that will generate different values given different input vectors” (Gallistel & King, 2010, p.285). Gallistel sees this as a contemporary manifestation of a traditional anti-representational, associative view of learning according to which learning consists in acquiring new procedural knowledge for getting by in the world, rather than in acquiring explicit propositional knowledge of facts about the world (ibid.). We’ve seen that it might be too strong to claim that neural networks are incapable of representing simpliciter; but by lacking an architecturally distinct read-write memory, they’re apparently incapable of representing the kinds of systems in the external world that animals can clearly represent. And surely that’s ‘anti-representational’ enough.

Before packing up and going home, connectionists might insist that although they don’t recognize a structurally distinct read-write memory, they do recognize mechanisms that could function as a read-write memory. On their behalf, Gallistel considers two such mechanisms. The first is the candidate I just mentioned, namely the weights of the connections in a neural network — in physiological terms, the synaptic efficacies of axonal projections. The second candidate is that symbols are encoded in reverberating activity in recurrent neural circuits. I’ll discuss the first second and the second first, just to make things interesting.
3.3.1 Symbols as Reverberating Circuits?

Given the number of neurons in an animal’s brain, and the number of synaptic connections made with other neurons, circular chains of neural connectivity — ‘recurrent circuits’ — are bound to be pervasive. Indeed, some have claimed that recurrent connectivity is a basic theme of cortical organization (Shu et al., 2003). Kubie (1930) was perhaps the first to suggest that recurrent circuits might support self-sustaining, ‘reverberating’ patterns of activity, and to speculate about their functional role. Arguing that the brain is a kind of digital computer, McCulloch (1949) suggested in passing that reverberating activity might play the role of a symbolic memory. Motivated more by considerations about the psychological character of memory, Hebb (1949) developed a model according to which short-term memories are instantiated by reverberating patterns of activity in transient ‘cell assemblies’, which are consolidated into long-term memory traces via synaptic plasticity mechanisms. Amit (1989) was one of the first to analyze reverberating neural activity using the tools of dynamical systems theory, characterizing such persistent activity patterns as attractors in the state-space of recurrent ‘attractor’ networks. Subsequent theoretical work has elaborated upon these themes, giving rise to a range of biologically realistic spiking network models that exhibit persistent activity patterns resembling experimental recordings from the neurons of animals performing working memory tasks (e.g. Zipser et al., 1993). Today, reverberating circuits are said to be “an integral part of the most plausible computational models of working memory” (Fuster, 2009, p.908).

That reverberating activity in recurrent circuits might plausibly serve as a mechanism of short-term, working memory — but not of long-term memory — was recognized early on. For example, Kleene (1956) pointed out that neurons in recurrent chains would become fatigued if they sustained reverberating activity over long periods of time. The problem is that reverberating activity is volatile, in the sense that the integrity of the information stored in the activity pattern requires a continuous
source of energy to maintain. Moreover, the amount of metabolic energy required to sustain neural activity is very high; on the basis of the known energy consumption of the cortex, Lennie (2003) estimates that fewer than 1% of cortical neurons could be concurrently active. But as we saw in Chapter 2.3, many animals seem to store information in a computationally accessible form for considerable periods of time; for example, scrub jays seem to store information about what they stored where, and to use that information hours, weeks, or even months later to inform their foraging strategies. Moreover, they store many, many such items of information. So it seems radically implausible, given the aforementioned considerations about cortical energy consumption, to suppose that animal brains precisely orchestrate the ongoing activity of a huge number of reverberating circuits over the time-scales necessary to account for the behavioral evidence. As we saw in Chapter 2.2, the symbols stored within read-write memory must be thermodynamically stable. The proposal that memory is instantiated by volatile reverberating activity fails to meet this criterion.

Gallistel & King (2010) echo Kleene’s point, and add a couple of points of their own. The signal in a reverberating circuit is presumably a volley of action potentials, which passes from one neuron to another over several synapses several times a second. Now, synaptic transmission is inherently noisy due to the stochastic nature of both neurotransmitter release and neurotransmitter reception. So in the absence of any mechanisms that reconstitute the signal after each synapse that it passes over — and, the authors claim, no such mechanisms are currently recognized — the fidelity of the signal will presumably degrade over time the longer it continues to circulate within the network.\footnote{See Gallistel & King (2010, p.185).} Perhaps it might be argued that the signal somehow degrades sufficiently slowly that the loss in fidelity is not problematic on the timescales required for very short-term memory. Indeed, as Gallistel & King (2010, p.185) point out, there is
considerable evidence that recurrent circuits in the oculomotor system are exhibit stable patterns of reverberating activity that encode eye position for up to 1-2 minutes. This is long enough for oculomotor control, but surely not for the long-term symbolic memories required to explain animal cognition.

There’s arguably a deeper problem in the vicinity, though. How much the signal degrades over time, and how problematic this is, will depend on how information in the reverberating signal is encoded. Contemporary neuroscience offers two main proposals about how information is encoded in neural signals: rate coding, in which information in encoded in the frequency of action potentials, and temporal coding, in which information in encoded in the fine temporal structure of action potentials. If the reverberating signal employs rate coding, then the problem of degradation over time is perhaps less damning, since, to use the terminology introduced in Chapter 2, rate codes are analog: the magnitude of the encoded value is proportional to the magnitude of the relevant aspect of the physical signal — in this case, to the frequency of action potentials. So small changes in spiking frequency will result in only small errors in the represented magnitude. But this property is precisely what makes rate coding unsuitable as a general proposal about how symbols are encoded in memory, since, recall, analog codes are not compact, and non-compact codes are highly inefficient. As Gallistel & King (2010) put it, analog codes “are not practical as the basis for a representational system whose symbols must each be capable of representing a large number of different possible states of the represented system” (pp.90-1). I’ll discuss this issue in more detail in section 3.3.3 below.

On the other hand, if reverberating signals employ temporal codes, which potentially are compact, the problem of the degradation of the signal looms. For small

---

9Gallistel & King (2010) don’t quite formulate the dilemma I’m about to pose. However, they do present elements of the dilemma at various different points, so I trust that they would endorse my formulation.
errors in the reverberating signal could potentially lead to dramatic errors in the rep-
resented value. To get a sense of the problem, compare unary (non-compact) and
binary (compact) encodings of the number 4: 1111 and 100, respectively. Now con-
sider the consequences of adding a stray 1 to the beginning of each of these strings:
the first string would then encode 5, while the second would encode 12. Small er-
rors introduced into compactly encoded signals can produce dramatic changes in the
magnitude of the encoded value; in this sense, compact codes are typically ‘brittle’. 
So given the noisiness of synapses, it’s implausible that reverberating circuits could
sustain compactly encoded information with high fidelity for even modest periods of
time.

Interestingly, many of these points are recapitulated in the history of computer
technology. Computer engineers distinguish between at least two different ‘levels’ of
memory: primary and secondary. The central difference is that only primary memory
is directly accessible by the CPU. In today’s computers, the role of primary memory
is played by RAM, whereas that of secondary memory is played by the hard disk.
But in the early years of computing technology, primary memory was implemented by
‘acoustic delay lines’, a device that stored a digital signal in the form of a mechanical
wave that circulated through a resonant medium such as a cylinder of mercury. This
mechanism is functionally analogous to a reverberating neural circuit, and, as in such
a circuit, information is stored in a volatile form. This in itself isn’t a problem for
a primary memory mechanism — indeed, the RAM in contemporary computers is
volatile — but it makes delay lines unsuitable for ‘long-term’, secondary memory. But
even as primary memory, delay lines were problematic, for they only allowed sequential
access to stored information; the processor would have to wait until the desired signal
circulated through the medium and reached the receiver (Copeland, 2008). Delay
line memory was soon eclipsed by faster and more efficient mechanisms of primary
memory, such as magnetic core memory and eventually the RAM of today’s integrated
circuits. Gallistel & King (2010) see in these technological developments a lesson for speculations about the neural basis of read-write memory: mechanisms that carry information forward in time by continually circulating a volatile signal through space are fundamentally unsuited to play the role of read-write memory.

### 3.3.2 Symbols as Synaptic Weights?

So much for reverberating circuits as a mechanism of long-term symbolic memory. But recall that they were never intended as such; in Hebb’s model, as in virtually all connectionist models since, long-term memory traces are held to be instantiated as patterns of synaptic weights. Let’s now consider whether that mechanism could plausibly serve as a read-write memory.

In Gallistel’s view, there’s a straightforward and conclusive reason why it could not: the information purportedly stored in patterns of synaptic activity simply cannot be read, hence cannot enter into computational operations. To see why, recall that on Gallistel’s information-theoretic view of world-brain communication, in order for a message to be communicated from a distal stimulus to the brain, it must first be encoded into a sensory signal, then decoded into an explicit symbol. This decoding process involves reconstituting the initial message from the various other messages that it might have been conflated with in the sensory signal; mathematically, it amounts to inverting the initial encoding function. Thus if we are to suppose that synapses are symbolic vehicles of information, “the mapping from experience to the state of the symbol must be invertible, that is, it must be in principle possible to deduce from the state of the synapse an aspect of the experience that produced it” (Gallistel & King, 2010, p.191).

The problem here is that on standard ways of understanding synaptic plasticity, the weight of a particular synapse is not fixed by one particular aspect of experience; it is continually buffeted by the ongoing stream of experience. This is a corollary of the
distributed, superpositional nature of information storage in connectionist networks. While this form of information storage is often said to confer various advantageous properties, like pattern completion and fault tolerance, Gallistel argues that, on the contrary, it precludes connectionist networks from storing information explicitly, in the form of readable symbols. For if the weight of a given synapse is determined by many experiential inputs, the value of any one of those inputs cannot be ‘read off’ from the value of the weight. In mathematical terms, the function that maps input values to the value of a given weight is many-one, hence that function cannot be inverted. The problem here isn’t so much that the information carried by a given synaptic weight cannot be read; it’s that there simply is no unambiguous, explicit information to be read. So, Gallistel claims, on the standard picture of information storage in connectionist networks, weights simply cannot function as symbolic vehicles of explicit information.\(^\text{10}\)

What if we were to abandon this picture and suppose instead that particular weights in a network are somehow fixed by determinate aspects of experience? This proposal would arguably jettison much that’s interesting and distinctive about connectionist theorizing, but insofar as it appeals to experience-dependent plasticity mechanisms, it would presumably still count as the kind of ‘associative’ connectionism that Gallistel objects to. Nevertheless, the proposal would not help. To see why, recall from Chapter 2.4 that the weight of a synapse essentially functions as a multiplicative constant that scales pre-synaptic signals. These scaled signals are added together in the integration zone of the post-synaptic neuron, and transformed by the neuron’s activation function to determine the post-synaptic signal. Using the formalism introduced earlier, the signal produced by a post-synaptic neuron \(N_0\) is \(a_0 = F(\sum_{i=1}^{n} a_i w_{i0})\), where the \(a_i w_{i0}\)

\(^{10}\)This argument appears most clearly in Gallistel (2008, p.237), and Gallistel & King (2010, pp.191-194).
are the scaled pre-synaptic signals. Now, all that the post-synaptic neuron ‘sees’ are these $a_i w_{i0}$. It has no independent access to the values of the pre-synaptic weights $w_{i0}$, and has no way of obtaining them from the scaled signals $a_i w_{i0}$, again for the simple reason that multiplication is a many-one function, hence isn’t invertible. Once the weights are multiplied by pre-synaptic activities, their values cannot subsequently be recovered. So even if synaptic weights were fixed by particular aspects of experience, the architecture of a neural network precludes them from being readable.\footnote{This argument appears most clearly in Gallistel (2008, p.238), and Gallistel & King (2010, pp.195-196).}

At this point, proponents of the connectionist orthodoxy in neuroscience might complain that Gallistel is tilting at straw. They might point out that information in neural networks is typically held to be embodied by distributed patterns of synaptic weights, not individual weights. That, remember, is what’s supposed to be so distinctive and interesting about connectionist schemes of information storage. Now, as we saw a couple of paragraphs back, Gallistel does discuss distribution, but he does so in the context of arguing that individual synaptic weights cannot encode explicit information. Why not hold instead that distributed patterns of weights encode explicit information? Gallistel discusses this possibility, and argues that it’s no more plausible than encoding information in individual weights. He begins by pointing out that distributive encoding of information — which he characterizes as the use of the same set of memory elements “to represent all of the values stored in a memory, with the state of every element contributing to the representation of every value” (Gallistel & King, 2010, p.284) — is far from unique to connectionist networks, and is in fact ubiquitous in contemporary communications and information technology. Encryption, for example, is widely used to send information securely across the internet, and employs a kind of distributive encoding. A standard method of encrypting a file is to take the
bit-string corresponding to the file, which can be understood as a number encoded in binary, and multiply that number with a large prime number. The resulting bit-string is distributively encoded in the sense that one must know each bit within the string in order to recover the original file; there’s a clear sense in which each bit contributes to the representation of the original file (ibid.).

This example suggests an objection of a kind that should by now be familiar: the whole point of encrypting information using a distributive coding scheme is to render it unreadable. Encrypted strings can only be decrypted, and hence read, by inverting the initial encryption operation — for example, by dividing the encrypted bit-string with the prime used to encrypt it. But without an encryption ‘key’, like the prime multiplicand, this is a computationally intractable problem. If the distributive codes used by connectionist networks are anything like those used in cryptography, the readability problem again looms large. Now, clearly not all distributed information is unreadable; if it were, then distributive coding wouldn’t be ubiquitous. But Gallistel holds that there are two properties of distributive coding schemes which make the view that explicit, readable information is distributively encoded in patterns of synaptic weights especially implausible. He illustrates these properties using an adaptation of the Paul Revere example discussed in Chapter 2.2. Recall Revere’s code for communicating the direction of the British attack: one lantern if by land, two if by sea. Now, suppose Revere had wished to communicate a second message: whether or not the British force is greater than 1,000 soldiers. He could have communicated both messages in a distributed fashion by hanging either one or two lanterns, on each of three successive occasions, in accordance with the code shown in Table 3.1 below. For example, if

\[12\]

At least, that’s the hope that much of today’s online commerce is riding on. The hope, in the vocabulary of computer scientists, is that P≠NP.

\[13\]

Of course, there are indefinitely many other, more efficient, codes he might have used to communicate these messages. The coding scheme here is not intended to be practical, but merely to illustrate some general properties distributive codes. The example is drawn from Gallistel & King (2010, p.193).
Revere were to hang two lanterns, followed by two, followed by two, this would communicate that the British are greater than 1,000 strong and are coming by land. This message is distributed over the three successive signal presentations.

<table>
<thead>
<tr>
<th>Signal number</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size/Route</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 1,000</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>&gt; 1,000</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.1: An example of a distributive code, used to communicate two messages: the size of the British force (whether or not greater than 1,000 strong), and the route via which they are approaching (whether by land or sea). These messages are distributed over three successive signals, each of which consists in hanging either one lantern or two. Hanging two lanterns on three successive occasions will, according to this coding scheme, communicate that British are greater than 1,000 strong and are coming by land. Adapted from Table 11.1 of Gallistel & King (2010, p.193).

The first point to notice about this coding scheme is that each signal within a sequence of lantern-hangings is determined by a distinct rule within the code. That is, there’s a distinct rule for mapping possible messages (combinations of army size and route) to signal types (number of lanterns) for each signal presentation in the sequence; that’s why Table 3.1 contains a distinct column for each signal presentation. Gallistel holds that this is a general property of distributive coding schemes. The problem that this poses for the view that information is distributively encoded in patterns of synaptic weights is that, for each synapse in a network, there must be a distinct rule that

---

14To read this from the table, notice that the first signal of two lanterns rules out two possibilities: that the British are less than or equal to 1,000 and are coming by land; and that the British are greater than 1,000 and are coming by sea. The second signal of two lanterns rules out a third possibility: that the British are less/equal to 1,000 and coming by sea. The only remaining possibility is that the British are greater than 1,000 and are coming by land. Note that in this case the third signal is redundant; a common property of distributed coding schemes, in virtue of which such schemes are often robust to noise. The third signal isn’t always redundant, however. To see why, try decoding what the signal sequence 1,1,1 communicates. If the third signal were always redundant, the code wouldn’t be distributive.
maps the network’s possible inputs to the weight of the synapse, and each rule must presumably be instantiated by some specific type of neurophysiological process. Thus “[t]here have to be as many different processes that change the strengths of synapses as there are synapses whose conductance must be changed” Gallistel & King (2010, p.195). However, Gallistel argues that the suggestion that there is a distinct type of neurophysiological process governing the fluctuations of each of the $1.5 \times 10^{14}$ synapses in the brain is radically implausible. One reason, he suggests, is that it would require a massive investment of physical resources. However, it’s not clear why a case in which every synapse involves a distinct type of plasticity mechanism would necessarily require more physical resources than a case in which every synapse involves the same type of mechanism; in both cases, the same number of token mechanisms would be involved. Perhaps a more compelling reason why the present suggestion is implausible is that it is, according to Gallistel, inconsistent with a core assumption of the associative theory of learning, namely that there are very general laws relating aspects of experience to changes in associative strength. As I’ll discuss in more detail in section 3.5, Gallistel holds that synaptic weights are widely held to be the neurophysiological realization of the associative bonds posited by the associative theory of learning. Thus “the assumption that the synaptic conductances themselves are symbols in a distributed encoding... is not consistent with the assumption that there is one or even a few rules... that relate experience to synaptic strength” (ibid., p.195).

The second point to notice about distributive coding schemes is that they take a lot of work to decode. We saw this in stark form when discussing encryption: sometimes decoding a distributively encoded signal is computationally intractable. But even in the best case, it involves a significant amount of computational work. One can get an intuitive sense of this by trying to decode a signal generated by the very simple code described in Table 3.1. The general point here is that complex encodings require complex computations to decode. Consider a relatively complex, compact code like
the binary encoding of number. Gallistel points out that binary coding is a kind of
distributive coding, in the sense that the value of a binary-encoded number is “spread
across the bits” of the binary string (Gallistel & King, 2010, p.193). To read the value
of a number encoded in binary, one must perform a relatively complex computation.\textsuperscript{15}
Contrast this with a number encoded via a simple code like unary: to read the value,
one simply counts. Now, the fact that complex encodings require complex computations to decode is not inherently problematic; indeed, as I’ll discuss in the next section,
Gallistel holds that the symbolic vehicles of explicit information in animal brains must,
in general, be encoded via efficient, compact codes. However, Gallistel claims, this fact
poses a problem for the view that symbols are instantiated by distributed patterns of
synaptic weights, for as we’ve seen, synaptic weights aren’t accessible to computation,
since weights and activities are conflated in the process of post-synaptic integration.
Claiming that symbols are instantiated by distributed populations of synapses rather
than individual synapses doesn’t obviate this problem; if anything, it exacerbates it.

\subsection*{3.3.3 The Inefficiency of Neural Codes}

So far we’ve focussed on problems with the \textit{mechanisms} of information storage cur-
rently recognized by neuroscientists. But as Shannon keenly recognized, the kinds
of mechanisms used to store and manipulate information are crucially dependent on
the schemes used to encode that information. This became apparent in the last two
sections, where issues about information coding were addressed in passing. We’ll now
turn our full attention to the issue of how information is encoded in neural systems.
That this issue might be a source of problems for the connectionist orthodoxy in neu-
roscience is suggested by Gallistel & King (2010) when they claim that neuroscientists
have simply \textit{ignored} the question of how explicit information might be encoded in neu-

\textsuperscript{15}Specified in note 4 in Chapter 2.2.1.
ral symbols: “The question ‘How could one encode a number using changes in synaptic conductances?’ has, so far as we know, never even been posed” (p.279; see also pp. viii, 78-9, and 170-1).

It’s hard to know what to make of this claim. The notion that the nervous system carries coded messages about distal stimuli had wide currency in neuroscience well before Shannon (1948) developed his quantitative account of information (e.g. Adrian, 1928). Moreover, Shannon’s insights were quickly incorporated into theoretical speculations about the nature of neural information coding (e.g. Mackay & Mcculloch, 1952). Today, questions about how the brain encodes information, and how it combines and manipulates that information as it carries out computations, have been said to lie at the heart of systems neuroscience (Averbeck et al., 2006).

Gallistel & King do qualify their claim by writing that while “the question of the code used by neural signals is a recognized question [in neuroscience]... the question of the code used by neural symbols is not” (p.78, my emphasis); remember that signals carry information through space, whereas symbols carry information forward through time, and it’s primarily the latter we’re interested in, since only symbols are directly involved in representation and computation. Now, the authors’ qualification is hard to square with the fact that they go on to discuss and evaluate various coding schemes that have wide currency in neuroscience as though they’re schemes for encoding information in neural symbols (see, e.g., ibid. p.91). So perhaps it’s best to take their claim that neuroscientists have ignored questions about how information might be encoded in neural symbols with a grain of salt, and focus on whether the schemes for symbolic coding that neuroscientists have proposed are at all plausible.

The earliest idea about how information might be encoded by neurons is the rate-coding hypothesis, the proposal that information is encoded in the firing rate of a neuron. We encountered this idea in Chapter 2.4, where we saw that it is assumed by most second-generation neural network models, which treat the functionally relevant
properties of nodes as a monotonic ‘activity level’, which is typically taken to model the firing rate of a neuron. As Gallistel & King (2010, p.76) acknowledge, there is abundant electrophysiological evidence that rate coding plays an important role in many systems within the nervous system, particularly those that are close to the sensory periphery. In Gallistel's terms, rate coding is clearly employed by neural signals. But we began to see why it is not plausibly employed by neural symbols in section 3.3.1, when we discussed the possibility that information might be encoded in the frequency of reverberating activity within a recurrent neural circuit. One of the problems we encountered there, recall, is that the range of values that a recurrent circuit could encode is proportional to, and constrained by, the dynamic range of frequencies that such a circuit could generate. Rate coding is thus a form of analog coding, and as such is highly inefficient. Qua analog, rate codes are analogous to unary codes, and as Gallistel & King (2010) put it, “no symbolization of quantity that has any appreciable power uses unary symbols” (p.100). Indeed, the authors suggest that we can reject out of hand any proposed scheme for encoding information in neural symbols that, like rate coding, obeys the analog principle, since, they claim, brains must be capable of encoding numbers as large as $10^{30}$, yet they only contain on the order of $10^{25}$ atoms. So even if, per impossible, all the atoms in the brain could be enlisted to encode numbers via an analogy code like unary, the brain still wouldn’t have the resources to encode the full range of values it can in fact encode. The example, of course, is fanciful, but it underscores just how inefficient analog schemes like rate coding are.

16Gallistel & King (2010) explicitly criticize the view that symbols might be encoded in the frequency of reverberating activity within a recurrent circuit, so they presumably think that a significant number of neuroscientists endorse that view. So, again, it really is quite puzzling why they hold that neuroscientists have ignored the question of how explicit information is encoded by neurobiological mechanisms.

17These claims appear in Gallistel & King (2010, p.90).
The example here suggests an additional problem with analog coding schemes, namely that analog symbols become increasingly hard to distinguish as their encoded magnitudes increase. Consider our analog scheme for encoding numbers in atoms. This affords a straightforward means of comparing the values of two symbols: just weigh the aggregates of atoms. To reiterate a now-familiar refrain, simple codes afford simple computational procedures. But, Gallistel & King (2010) claim, this comparison procedure decreases in accuracy as the encoded numbers increase: “no weighing procedure could distinguish the weight of $10^{29}$ atoms from the weight of $10^{29} + 1$ atoms” (p.90). The worry, as it manifests in the specific case of the rate coding hypothesis, is that as spiking frequencies increase, they become increasingly difficult to discriminate.

Now one might worry that these arguments are only effective against the rate coding hypothesis if we endorse the questionable assumption that brains must be capable of representing numbers as large as $10^{30}$, which the authors assert but give no specific reason to believe. But even if we reject this assumption, the arguments come together in a particularly problematic way when we consider the well-known physiological limitations on the rate at which a neuron can fire. Specific values vary considerably between different types of neurons, but the fastest that most neurons can fire is on the order of 100 Hz. So if neurons encode numerical information in firing rates, and send one message per second, they can only encode numbers up to approximately 100. Surely this is enough to rule out the rate coding hypothesis, even if the previous arguments weren’t.

So much for the rate coding hypothesis. Another commonly proposed coding scheme in neuroscience is what’s sometimes called ‘place coding’. The idea here is

---

18Gallistel & King don’t present this specific argument, but it is closely related to, and is a stronger version of, arguments that they do present, so I’ll assume that they’d endorse it. They do make a very similar point on page 77, but the issue there is about limitations on the number of messages that a neuron could communicate within a given period of time, whereas the present issue is about limitations on the value that could be communicated by a single message.
that a particular neuron within a population is maximally responsive to the specific value of a sensory parameter, or ‘feature’. Such neurons, often called ‘feature detectors’, are typically organized topographically, according to their preferred parameter value, thus the information they encode is determined by their ‘place’ within the population. As with the rate coding hypothesis, there is abundant electrophysiological evidence that place coding plays an important role in the brain. Some of the earliest evidence was provided by Hubel & Wiesel (1962), who identified ‘bar’ and ‘edge’ detectors in the cat visual cortex. Subsequent research has revealed topographic place coding to be a ubiquitous strategy for encoding information in sensory cortex. That it might play a role in later stages of sensory processing was suggested by Hubel & Wiesel, who proposed a highly influential scheme by which simple feature detectors combine hierarchically to give rise to increasingly complex feature detectors, perhaps eventually culminating in neurons that detect perceptual features, such as distal objects, rather than merely proximal sensory features. This idea, though not uncontroversial, has received support from evidence of neurons high in the visual processing hierarchy that are selective for specific categories of objects, such as faces (Desimone, 1991, e.g.). Note that place coding should be distinguished from place cells, which are specific neurons in the hippocampus that fire selectively when an animal is in a specific location within a given environment (O’Keefe & Dostrovsky, 1971). Nevertheless, place cells encode information via place coding, and hence help to illustrate the wide range of roles that place coding is thought to play in the nervous system.

Despite the abundant evidence for place coding, Gallistel & King claim that it cannot provide a general scheme for encoding information in neural symbols. This is because, in their view, place coding is an instance of what we earlier called a ‘nominal’ coding scheme. Recall that in such a scheme, the mapping from messages to symbols is entirely arbitrary, in the sense that there’s no compact, generative procedure by which symbols are generated from the messages they encode. Symbols aren’t constructed by
generative procedures on this scheme, but must be allocated values by a pre-specified labeling scheme. Moreover, recall also that the lack of compact procedures for generating nominal symbols entails the lack of compact procedures for operating on such symbols. Consider how we might construct a mechanism for performing addition on numbers that are encoded by arbitrary bit strings. First, we’d need to know in advance the bit strings that encode each of the numbers that we want the machine to operate on; and second, we’d need to code that information implicitly into the structure of the mechanism for each of the possible pairs of addends we want the mechanism to operate on. For a mechanism that computes any interesting restriction of addition, that’s going to be a lot of numbers. To put the general point in now-familiar terms, a mechanism that computes using a nominal code must have the functional architecture of a look-up table, and as such suffers from the problems of pre-specification and combinatorial explosion. Given that place coding is a form of nominal coding, the place coding hypothesis thus seems inadequate as a general account of how information is encoded in neural symbols.

So it seems that the arguments discussed in this section collectively underscore the points made in Chapter 2.2: the principles by which the nervous system encodes explicit information into neural symbols must be such that those symbols are both compact and constructible, much like the principles by which conventional digital computers encode numbers into binary strings. Indeed, echoing points made in their discussion of the problems with reverberating circuits, Gallistel & King suggest that the history of communications and information-processing technology holds lessons for speculations about neural coding. That history reveals an inexorable move towards mechanisms that encode, store and manipulate information digitally, suggesting that that is the optimal solution to “the ubiquitous problems of noise, efficiency of transmission, and precision control” (ibid., p.24). Since the brain is finely tuned by evolution to perform optimally given the physical constraints it faces, the authors hold that there’s every
reason to think that it too employs compact, efficient digital coding schemes.

### 3.4 Problems with Connectionist Computation

As we saw in the previous section, Gallistel is best understood as arguing that although the kinds of connectionist mechanisms recognized by contemporary neuroscientists might in principle embody representations, they can’t represent the complex kinds of external systems that animals are behaviorally responsive to. The reason for this, ultimately, is that they lack anything that could function as a read-write memory. But the immediate reason the lack of a read-write memory is problematic vis-a-vis the representational power of connectionist mechanisms is that it precludes such mechanisms from performing powerful and practical computations, and hence prevents them from participating in rich homomorphisms. As we saw in Chapter 2.2, Gallistel holds that computational mechanisms lacking a read-write memory have the functional architecture of finite-state machines or look-up tables, which suffer from the problems of pre-specification and combinatorial explosion. In this section, I’ll explore in detail how Gallistel thinks these problems plague the kind of connectionist mechanisms of computation recognized by neuroscientists.

However, as with Gallistel’s claims about connectionist representation and information coding, Gallistel’s claims about connectionist computation are slathered with a layer of rhetoric, so it will require some interpretative effort to get at the substantive claims beneath. For example, just as Gallistel sometimes seems to assert that connectionist mechanisms are strictly incapable of representation, he sometimes seems to assert that such mechanisms are strictly incapable of computation. In his (2008) paper, he writes that “mechanisms that perform [computational] operations... such as the arithmetic operations... and logical operations (e.g., AND and OR)” are “missing from our current conception of the nervous system” (p.239). However, just a
page later he writes that “[i]n many neural net models... combinatorial operations are implemented by a table-look-up architecture”, which might, for example, “effect the multiplication of all possible values of two variables” (p.240). Similarly, in his (2010) book, he writes that “neuroscience has no answer” to “the question of how the brain symbolizes simple quantities and its procedures/mechanisms for performing arithmetic operations on those quantities” (p.100), yet goes on to write that “the mechanism of signal transmission across synaptic junctions makes possible the implementation of the AND, OR, and NOT functions” (p.172), and claims that work by Koch (1999) shows “how experimentally demonstrated mechanisms within individual neurons are sufficient to implement a variety of computational operations”, including multiplication (p.174). Now, as we’ll soon see, Gallistel holds that there are various problems with these mechanisms for implementing computational operations; but it’s one thing to say that a mechanism implements computational operations poorly, and another to say that it doesn’t implement them at all.19

Moreover, when it comes to evaluating what the substantive problems with connectionist computation are supposed to be, it’s not entirely clear what Gallistel wants to say. On the one hand, Gallistel wants to say that, by lacking a read-write memory, neural networks have the computational architecture of a finite-state machine: “contemporary thinking about the neurobiological mechanism of memory tries to dispense with the tape and place all of the memory in the transition table” (Gallistel & King, 2010, p.xi). But on the other, Gallistel wants to say that neural networks have the computational architecture of a look-up table: “the neural network architecture im-

---

19The quotes in this paragraph aren’t presented in the spirit of gotcha journalism, intended to demonstrate that Gallistel has somehow contradicted himself. Perhaps these statements could be given an interpretation that would resolve the apparent tension between them. The point is that Gallistel hasn’t provided such an interpretation, or acknowledged the apparent tension, so it’s up to the reader to charitably resolve the tension. It seems to this reader that it’s most charitable to interpret claims to the effect that neuroscientists have ignored questions about neural computation, or that neural mechanisms are incapable of computation, as hyperbole for rhetorical effect.
plements arithmetic and other basic functions by table look-up of nominal symbols” \((ibid.)\). I don’t think that this is an irremediable equivocation, but it does raise an interpretative question about what exactly Gallistel is claiming. I think we can charitably resolve that question by focusing on how one’s conception of neural computation reflects one’s conception of the symbolic vehicles of neural computation. For reasons I’ll unpack in a moment, if one holds that information is stored in the synapses of a neural network according to a distributive code, then it’s natural to think that neural networks have the computational architecture of a FSM. On the other hand, if one holds that information is stored as reverberating activity via a rate or a place code, then it’s natural to think that neural networks have the architecture of a look-up table. Gallistel doesn’t explicitly line things up in this way, but it’s consistent with a lot of what he says, and it provides a convenient way of framing the issues.

First consider the picture of neural computation that emerges if we adopt the view that symbols are encoded by synaptic weights. As we’ve seen, Gallistel holds that synaptic weights are incapable of being read by downstream mechanisms, since their value is conflated with pre-synaptic activity levels in the process of post-synaptic integration. Thus experience-dependent changes in synaptic weight don’t reflect the encoding of explicit, readable information; rather, such changes merely ‘rewire’ the network, implicitly determining the states that the network is disposed to transition through. As a leading computational neuroscientist puts it, experience changes “the transition function governing how the machine switches from one state to the next” (Koch, 1999, p.470). Gallistel & King (2010) summarize the point as follows: “plastic synapses within neural networks [do not] encode the values of variables. What they encode are not the values themselves, but rather a procedure that will generate different values \textit{given different input vectors}. In such schemes, the memory elements do not themselves represent values... That’s why these schemes have a finite-state architecture” (p.285).
On the other hand, consider the picture of computation that emerges if we adopt the view that symbols are instantiated by reverberating activity in a recurrent network. One point that remained just below the surface during our earlier discussion of this conception of neural symbols is that although Gallistel denies that synaptic weights can function as symbols, he does allow that stable patterns of reverberating activity can. He allows that such patterns, \textit{qua} physically distinct, enduring entities, can enter into different states on the basis of causal relations with distal stimuli, and hence can encode information; that they can persist in those states over time, so can write information to memory; and that they can be causally efficacious within computational processes, hence can be read. As Gallistel (2011) puts it in a recent paper, patterns of reverberating activity “are not very good symbols, but symbols they certainly are” (p.254). One reason he thinks they’re not ‘very good’, recall, is that they’re volatile, hence implausible as a mechanism of long-term read-write memory. Another reason that’s more directly pertinent to the present issues is that the method by which they encode information is typically place coding; the specific value that a pattern of activity encodes is determined by its location within a population of neurons. But recall that in Gallistel’s view, place coding is a nominal coding scheme, which can only be effected by a look-up table. As I put it in Chapter 2.2, the non-compactness of symbols infects the procedures that operate on those symbols. To get at the point from another angle, recall from Chapter 2.4 that a purported virtue of the kind of recurrent attractor networks that support stable patterns of reverberating activity is that they exhibit pattern-completion, which can be understood as a form of content-addressable memory. But, Gallistel & King (2010) write, “[c]ontent-addressable memories are look-up tables in which the different possible results are accessed through parallel search... content-addressable memory requires a physically distinct memory location for every possible input string” (p.93-4).

So it seems that insofar as we cleave to the mechanisms of information processing
recognized by contemporary neuroscientists, we’re faced with a dilemma: either we sup-
pose that neural information is encoded by synapses, in which case that information is
merely implicit in the transition table of a finite-state machine, or we suppose that in-
formation is encoded by reverberating activity patterns, in which case that information
is explicit, but cannot employ compact coding schemes and can only be manipulated
by look-up tables. In either case, it seems, the computational architecture we’re stuck
with suffers from the problems of pre-specification and combinatorial explosion. But
how exactly do those problems manifest in the case of the computational mechanisms
recognized by neuroscientists, and how do they preclude such mechanisms from per-
forming powerful and practical computation? Gallistel presents a complex skein of
problems associated with pre-specification and combinatorial explosion. I think we
can disentangle those problems by weaving them around a conceptual framework that
consists of three grades of computational flexibility, each of which requires an increas-
ingly sophisticated form of read-write memory. I’ll discuss this tripartite framework
over the next three sub-sections.20

3.4.1 Constructability

Recall from Chapter 2 that the central problem with computational mechanisms that
lack an effective read-write memory, in Gallistel’s view, is that they allocate distinct
elements of computational machinery in advance of computing a given function, in
proportion to the total number input-output pairs that constitute that function. But
most functions of any interest are functions of multiple variables, which might take
indefinitely many possible values, so the total number of input-output pairs that con-

---

20 Again, although Gallistel doesn’t explicitly discuss the problems associated with pre-specification
and combinatorial explosion within the tripartite framework I’m about to present, the framework is
implicit in much of what he says, and provides a convenient structure around which to develop the
discussion.
stitute such functions will be astronomical. More to the point, the behavioral evidence we surveyed in Chapter 2 suggests that animal brains in fact do compute ‘interesting’ functions like these. For example, the brains of scrub jays can apparently store information about thousands of cache locations the jay has never encountered before, integrate that with information about what was stored at those cache locations and when, and perform computations on those integrated arrays of information to determine where the jay ought to search for food. It’s radically implausible that information about each of the possible combinations of specific locations, foods and caching times was actually instantiated by pre-specified circuitry within the jay’s brain, before the jay experienced a particular caching event. First, because natural selection couldn’t have anticipated contingent events like these, which might happen only once in the history of the universe, and, second, because the number of possible combinations of locations, foods and times making up these events would outstrip the number of elementary particles in the universe, let alone the number of neurons in the jay’s brain.

The upshot of these considerations is, as Gallistel & King (2010) put it, that “[a] computing machine can only have a finite number of actual symbols in it, but it must be so constructed that the set of possible symbols from which those actual symbols come is essentially infinite... This means that the machine cannot come with all of the symbols it will ever need already formed. It must be able to construct them as it needs them” (p.74, my emphasis).

Seen from this perspective, the problem of pre-specification as it pertains to neural networks is ultimately about the capacity to learn; or rather the lack it. This becomes clear when, echoing the previous quote, Gallistel & King write that “it is physically impossible that our brains have innate elements of their structure specific to every episode we might experience or specific to the variables of which the memory of an episode would be constituted. However, neural network models often end up assuming just that” (p.157). Issues about learning will take center stage in section 3.5, but the
point here underscores a computational aspect of learning that’s especially relevant given our present concerns. Recall that, for Gallistel, learning in the first instance consists in extracting behaviorally useful information from experience, which amounts to decoding the values of distal variables into computationally manipulable symbols. The considerations raised in the previous paragraph suggest that in order for a computational mechanism to be capable of manipulating many possible values, and flexibly combining them in unforeseeable ways, it must have access to a mechanism that is capable of constructing vehicles for those values on the fly — namely, a read-write memory mechanism. Thus by enabling a capacity for constructing symbols, which in a computational system of any appreciable power is necessary for learning, a read-write memory provides the most fundamental way of avoiding the pre-specification problem.

One final point of clarification before moving on to consider the next grade of computational flexibility. Gallistel is not generally averse to innately-specified computational mechanisms. On the contrary, as we’ll see in section 3.5, he emphasizes the importance of explaining learning in terms of innately-specified mechanisms that are adapted to solve specific computational problems. The objectionable kind of nativism that he thinks is exemplified by neural network theorizing is a nativism that pertains to specific, contingent values of environmental variables that might change over an animal’s lifetime, such as an ant’s current location with respect to its nest, or a jay’s preference for a specific type of food. For reasons we’ve discussed, these cannot plausibly be pre-specified within the computational machinery. On the other hand, Gallistel holds that some behaviorally important values are not contingent in this sense but universal, and hence much more likely to be built implicitly into the structure of the computational machinery. An example is the period of the earth’s rotation, which doesn’t change appreciably over biologically significant time-scales, and hence, perhaps unsurprisingly, is reflected in the genetically-specified endogenous period of the circadian clocks found in virtually all terrestrial organisms. It arguably makes little
sense for Mother Nature to build mechanisms for learning values like these if she can rely on them to stay constant. The general point here is quite familiar, and has been expressed by, among others, Shepard (1989): “in systems that have evolved through natural selection, the features of the world that are both biologically significant and absolutely invariant in the world have tended to become genetically internalized” (p. 104). In fact, as Shepard emphasizes, it is generally not the specific values that will become genetically internalized, but the abstract structure of the space of possible values. We saw an example of this in the experiments on how invertebrates learn the solar ephemeris function discussed in Chapter 2.3.1, which reveal that ants and bees seem to innately expect that the sun’s location at dawn is 180° from its location at sunset.

From Gallistel’s perspective, the upshot of these considerations is that explicit information about the contingent values of environmental variables cannot be innately specified, whereas implicit information about the range of possible values must be innately specified, within the structure of the computational mechanism that constructs symbols for the explicit values. As we’ll see in more detail later, read-write memory in Gallistel’s view is a malleable and domain-general mechanism of information-storage, whereas the computational mechanisms that extract and manipulate information stored in memory are innately specified, domain-specific learning modules: “The tape [of a Turing machine] has no pre-specified structure other than its uniform topology... By contrast, each state of the machine is distinct from each other state, and its structure is specific to a specific state of affairs (pre-specified)” (Gallistel & King, 2010, p.108-9).

### 3.4.2 Transportability

The second grade of computational flexibility — the second way a computational mechanism can mitigate the problems of pre-specification and combinatorial explosion —
is achieved by a mechanism that has access not merely to a simple, sequential read-write memory, but to an *addressable* read-write memory. To see why addressability is important, consider how complex functions of multiple arguments are computed in conventional computers like von Neumann machines. Such machines contain within their processing unit a stock of primitive, hard-wired computational operations that implement one- and two-argument functions. The two-argument functions can be recursively composed together to produce increasingly complex functions of indefinitely many arguments. This requires the processing unit to sequentially orchestrate the primitive operations over time and to store intermediate results in read-write memory. But perhaps more importantly, it requires the machine to *fetch* the arguments for a given function from potentially very different locations in memory, and *transport* them to the mechanism in the processing unit that implements the function. This is made possible by the fact that conventional computers have access to a read-write memory with *addresses*, which enable a symbol at an arbitrary location in memory to be found, and transported to the processing unit when needed.

The computational challenge posed by complex functions of multiple arguments is that of taming the combinatorial possibilities inherent in such functions. By orchestrating primitive computational operations over time, and transporting symbols from memory to those operations when needed, conventional computers can decompose complex functions into simpler, two-place functions, and thereby combat “combinatoric explosions with combinatorics” (Gallistel & King, 2010, p.136). However, the challenge here is not merely that of bringing the symbols that instantiate the arguments of two-place functions together with the mechanisms that implement those functions; what makes the challenge especially demanding from an engineering perspective is that symbols might be stored at arbitrary locations throughout memory and combined in ways that couldn’t have been anticipated when those symbols were acquired. Transportability isn’t just a matter of fetching symbols from memory to the computational
machinery, but of finding the relevant symbols and combining them together in flexible ways. That’s what makes the need for memory addressing mechanisms so acute.

This aspect of the challenge of computing complex functions is underscored by behavioral evidence from animal learning experiments. As Gallistel often puts it, the information an animal needs in order to behave adaptively in a given context typically arrives in ‘dribs and drabs’ throughout the animal’s experience, often long before it is needed. For a desert ant to navigate home, it must successively integrate the direction and distance it has travelled into an estimate of its current location with respect to home. This seems to require some form read-write memory, for without such a mechanism “there is no way to make the information gained from current experience accessible to future computation and behavioral control” (Gallistel, 2006, p.68). But sometimes the information required for adaptive behavior can be flexibly combined in ways that couldn’t have been anticipated when the information was acquired. For example, recall that a jay’s foraging strategy can be informed by comparing information about when certain caches were made, with information about how long the foods stored at those caches take to decay; supposing a jay recalls that its favorite food takes $x$ hours to decay, then if fewer than $x$ hours have elapsed since it cached that food, it will preferentially search those caches, but if more than $x$ hours have elapsed, it will search caches containing less desirable food. So it seems that “information about times of occurrence and the durations of intervals is in fact stored in a manner that makes it accessible to computation”, and that “interval information can be acquired before the time-of-occurrence information or after”, thus the two kinds of information “can be combined in ways that could not have been foreseen when the information was acquired” (Gallistel & King, 2010, p.217). This seems to demand not merely a read-write memory, but an addressable read-write memory, which enables information to be fetched from arbitrary locations in memory, and transported to the computational machinery when needed.
Given these considerations, Gallistel argues, it’s not clear how a mechanism like
a neural network — which doesn’t have an effective read-write memory, let alone an
*addressable* read-write memory — could tame the combinatorial possibilities inherent
in complex functions of multiple arguments. As we’ve seen, proponents and detractors
of connectionism alike agree that neural networks lack an architectural distinction be-
tween processor and memory; but that’s just to say that they don’t contain a unit that
performs primitive computational operations that interfaces with a structurally distinct
unit that stores symbols at specific addresses, which can be orchestrated together across
time so as to recursively compose simple functions into indefinitely complex ones. How
then do neural networks compute complex functions of multiple arguments? Perhaps
one might suggest that it’s parochial to assume that neural networks compute complex
functions in the same way as conventional computers, via functional decomposition,
and that instead they compute such functions *all at once*, by simply mapping input
values to output values. But, Gallistel claims, that’s tantamount to saying that neural
networks function as massive look-up tables. The suggestion would seem to require, for
example, that the jay’s brain contains pre-specified computational machinery specific
to every possible combination of location, caching time, food type, and decay rate that
it might ever experience. But that’s an astronomical number of possibilities, and a jay
will only ever *actually* experience a vanishingly small proportion of them throughout
its lifetime. The suggestion at issue thus raises the problems of pre-specification and
combinatorial explosion in spades.

Perhaps one might instead suggest that neural networks *do* compute complex func-
tions via functional decomposition, by incorporating several distinct, modular look-up
tables that compute simpler functions. But given that such mechanisms don’t contain
an addressable memory that can fetch symbols from arbitrary locations in memory
and transport them to these proposed look-up tables, it would seem that the look-up
tables would in effect have to *brought to the symbols*, by replicating them at each of the
various locations at which the symbols they take as inputs are stored. There wouldn’t, according to the present suggestion, be a distinct computational mechanism that performs a primitive operation such as addition on whichever symbols are transported to it; rather there would be many such mechanisms distributed throughout memory, that operate only on the symbols that reside at those memory locations. But that suggestion seems hard to square with the aforementioned point that symbols might be stored at arbitrary locations throughout memory, and might be operated on in ways that couldn’t have been anticipated when the symbols were acquired; moreover it would seem to entail a profligate waste of physical resources. Nevertheless, Gallistel holds, this is in fact how complex computational problems are tackled in many neural network models:

In many neural net models of even simple computations, such as those involved in dead reckoning, the combinatorial operations are implemented by a table-look-up architecture... There is a separate look-up table for each instance of a given kind of operation — a different table for each different case in which two variables must be added. A given table can effect the multiplication of all possible values of two variables, but it can operate only on those two variables... In short, there are as many different look-up tables as there are pairs of variables whose values may have to be combined.

(Gallistel, 2008, p.240)

Again, the problems of pre-specification and combinatorial explosion loom large.

3.4.3 Compositionality

Time for a quick recap. Even the most primitive form of read-write memory, in Gallistel’s view, provides a capacity that neural networks lack, namely the capacity to record novel values of environmental variables as they’re experienced, and to store those values for indefinite durations until they’re needed by the computational machinery. This first grade of computational flexibility constitutes a minimal capacity to learn, and thereby allows mechanisms with a read-write memory to overcome the most fundamental form
of the pre-specification and combinatorial explosion problems. However, primitive, sequential read-write memory is highly inefficient. A more efficient form of memory is provided by mechanisms capable of memory addressing. To make explicit a point that remained implicit in the previous section, addressability effectively allows the values of environmental variables stored in memory to be treated as values of those variables by the computational machinery. Mechanisms that allow the computational machinery to directly fetch and manipulate the symbol stored at a given address allow the value instantiated by that symbol to bind the variable instantiated by that address. This capacity enables a second grade of computational flexibility, namely the capacity to compute functions of multiple variables, the values of which might be acquired across many different experiences, and stored at many different locations throughout memory.

Even though these two grades of computational flexibility mitigate two fundamental aspects of the problems of pre-specification and combinatorial explosion, there’s a third aspect of these problems that even an addressable read-write memory cannot overcome. This, in effect, combines the first two aspects: it requires the capacity to learn & manipulate variables themselves. To achieve this capacity, Gallistel argues, an indirectly addressable memory is required, a memory that allows the symbols stored at a given address to be treated not only as the values of distal variables, but also as memory addresses and instructions for the processor. As I explained in Chapter 2.2.3, this allows the symbols stored at a given address to function as pointers to other addresses, and allows those pointers, or ‘variables of variables’, to be computationally manipulated. To see how this capacity might be pertinent to the question of how animal brains overcome the problems of pre-specification and combinatorial explosion, note that animals don’t merely need to learn the values of environmental variables, they also apparently need to learn the variables themselves. For example, wild scrub jays sometimes make up to around 30,000 caches during the fall, but during a lean season a particular jay might make far fewer than that. So just as specific cache locations are
contingent, the number of cache locations a jay might make is also contingent; in neither case does it seem plausible that the relevant information is innately pre-specified. As Gallistel & King (2010, p.157) put it, it seems radically implausible to suppose that, for example, a variable for the location of the 29,567th cache is pre-allocated in the jay’s memory, when a jay might only ever make 10,000 caches.

Instead of supposing that contingent variables like these are innate, as they effectively are in a mechanism with a non-indirectly addressable read-write memory, Gallistel & King argue that it’s much more plausible to suppose that higher-order, ‘universal’ variables, which can take contingent variables as their values, are innate. These universal variables would be analogous to Kantian categories of experience; in the case of the jay they would include categories like Location, Food Type, Time, and so forth. The authors hold that this functional distinction between (contingent) variables of (universal) variables requires an indirectly addressable read-write memory; universal variables are instantiated by a register at a specific address, containing a symbol that could be incremented to point to a series of other addresses, much like the register 80 in the example given in Chapter 2.2.3. This mechanism allows contingent variables, such as that for the location of the 29,567th cache, to be generated as needed, without being implicitly pre-assigned in the structure of memory; all that is implicitly pre-assigned is a universal variable corresponding to a general category like Location. Thus Gallistel & King hold that in order to avoid the full brunt of the pre-specification and combinatorial explosion problems, “we must postulate a mechanism by which a variable can be accessed by another variable (a variable of variables: indirect addressing)” (p.159).

A pointer to a series of other addresses in memory implements a simple data structure called an array. Once we have a mechanism in place with the capacity to generate arrays, claim Gallistel & King, we have the capacity to generate data structures of indefinite complexity: “once we have... indirection, we are unlimited in terms of being able to form hierarchies of variables” (ibid., p.159). Thus an indirectly addressable
memory seems to be a crucial component of any mechanism that is capable of flexibly generating representing systems that are homomorphic with complex, changeable represented systems in the external world.

However, it’s not at all clear how a mechanism with the computational architecture of a neural network could instantiate an addressable memory, let alone an indirectly addressable memory. An addressable memory requires a strictly ordered series of memory locations, each of which has a unique identifier, so that although the memory locations are arranged in sequence, they can be accessed ‘randomly’, or non-sequentially, by the computational machinery. But we’ve seen that the only state or mechanism within the purview of contemporary neuroscience that could function as a symbol within a read-write memory, in Gallistel’s view, is a pattern of reverberating activity in a recurrent network. Putting aside worries about the thermodynamic instability of reverberating patterns, it’s not clear how networks, understood as memory locations, could be arranged in sequence or uniquely identified in the manner required for an addressable memory. We saw in Chapter 2.4 that recurrent networks do function in ways that are analogous to a content-addressable memory, but we’ve also seen that Gallistel & King hold that content-addressable memories are look-up tables, and hence suffer from the problems of pre-specification and combinatorial explosion. Furthermore, even if neural networks could somehow be arranged to implement an addressable read-write memory, it would remain mysterious how they could implement indirection, since it’s not clear how a reverberating pattern of activity could encode an instruction to the computational machinery or a pointer to another location in memory. Thus it’s not clear how a mechanism with the architecture of a neural network could overcome the third grade of computational flexibility and flexibly generate complex data structures. In section 3.6, we’ll see a case study of how such a mechanism might be pre-wired to implement a complex data-structure, but we’ll also see in stark quantitative detail how this gives rise to the problems of pre-specification and combinatorial explosion.
3.5 Problems with Connectionist Learning

At this point, one might wonder why proponents of the connectionist orthodoxy in neuroscience couldn’t appeal to the capacity for neural networks to learn to obviate the problems of pre-specification and combinatorial explosion. After all, the phenomenon of learning has been at the heart of artificial neural network theorizing at least since Rosenblatt (1958) sought to improve upon the static networks developed by McCulloch & Pitts (1943), moreover experimental work on the relation between synaptic plasticity and learning is one of the most important and fertile areas of contemporary neuroscience. Indeed, Gallistel’s arguments might seem ironic to many, given the widespread perception that mechanisms with the architecture of a neural network have an advantage over ‘classical’ mechanisms when it comes to explaining learning. It’s precisely this perceived proficiency for learning that explains much of the interest and excitement surrounding the re-emergence of connectionism in the 1980s. To many researchers, ‘neo-connectionism’ seemed to provide an alternative to the overly nativist assumptions of classical information-processing models.

Bechtel (1985) expresses this prevalent attitude when he writes that although connectionism is situated squarely within the post-behaviorist, cognitivist tradition alongside classical information-processing psychology, it nevertheless shares with behaviorism an affinity for explaining learning, whereas learning in the classical information-processing tradition “has been a problematic phenomenon” (p.58). The reason for this, in Bechtel’s view, is that the mechanisms posited by classicists embody implausibly strong nativist assumptions; even if classicists try to explain learning in terms of extracting information from experience and storing it in memory, they allegedly have no story about how the computational mechanisms that operate on the information in memory could be learned, or even how those mechanisms could develop in ways that are responsive to an organism’s experience. By contrast, Bechtel argues, because connectionist networks lack an architectural distinction between memory and proces-
sor, connectionists have no trouble explaining how computational mechanisms could be shaped by experience.

Of course, to Gallistel, this attitude is anathema. It’s precisely the lack of a distinction between memory and processor that, in his view, precludes connectionist networks from providing a satisfactory explanation of learning. Gallistel holds that the distinction between memory and processor reflects an appropriate division of labor between those components of a cognitive mechanism that might plausibly be modified by individual experience, and those that are fixed by the genetic endowment of the species. But because connectionists collapse the distinction between memory and processor, the same mechanisms are forced to do double duty as the environmentally malleable repositories of information and the fixed, special-purpose mechanisms that manipulate that information. Thus although researchers in the connectionist tradition ostensibly endorse a strong form of empiricism according to which “the brain is relatively formless” at birth and is gradually rewired by experience via domain-general learning mechanisms to respond more adaptively to the environment, they are in fact unwittingly committed to a radically implausible form of nativism according to which the brain’s responsiveness to the environment is innately pre-specified in its computational architecture (Gallistel & King, 2010, p.252-3). This is because of the widely unappreciated architectural commitments of the associative theory of learning that, Gallistel thinks, is presupposed by the connectionist tradition. Learning as the associationist understands it involves the formation of associative bonds, but associative bonds cannot function as symbols in a read-write memory. Thus associationists, in Gallistel’s view, are committed to learning mechanisms that have the functional architecture of finite-state machines, and which therefore suffer from the problems of pre-specification and combinatorial explosion. So although the associative theory of learning has historically been motivated by empiricism, “[t]here is little appreciation of the radical nativism implicit in it: the necessity of providing structurally in advance for every
possible behaviorally relevant state of the world” (*ibid.*, p.177).

For Gallistel, then, the various problems with the computational mechanisms recognized by contemporary neuroscientists that were bruited in previous sections ultimately issue from the associative theory of learning that neuroscientists allegedly assume when they attempt to explain learning in terms of experience-dependent plasticity mechanisms. However, Gallistel holds that a commitment to associationism makes trouble for neuroscientists in various other, more direct ways, which I’ll survey in the remainder of this section. These specific problems with neural associationism are diverse, but they’re unified by a common theme, pithily expressed by Fodor (1983) when he writes that “[t]he traditional, fundamental, and decisive objection to association is that it is too stupid a relation to form the basis of a mental life” (p.82). Learning as the associationist sees it is a passive, mechanistic process whereby the brain is gradually rewired by the statistical structure of the environment; however, this view allegedly ignores a large body of evidence that learning is in fact an active, rational process whereby the agent generates and tests hypotheses about the causal contingencies in her environment. There are two aspects of this broad criticism of associationism, which will provide a convenient framework around which to structure my presentation of Gallistel’s specific objections to the neuroscientific incarnation of associationism: Anti-Representationalism and Empiricism.

### 3.5.1 Anti-Representationalism

Gallistel holds that neuroscientists are committed to an associative theory of learning because he thinks that they identify changes in synaptic conductance with changes in associative bonds. For example, he and King (2010) write that “[t]he... modification of one or more synaptic conductances is almost universally assumed to be the neurobio-
logical realization of the psychologist’s concept of an association” (p.212). However, to reiterate points made earlier, Gallistel denies that synaptic conductances can encode explicit information: the weight of a given synapse isn’t fixed by a determinate aspect of experience but is continually buffeted by the ongoing stream of experience, and even if it were, that weight couldn’t be recovered after the process of post-synaptic integration. The synaptic weights in a neural network don’t encode specific values, but rather procedures for generating values given certain patterns of input to the network. According to Gallistel, this is a manifestation of the anti-representationalism inherent in the associative theory of learning; according to associationism, learning rewires the brain so the learner’s behavior is better adapted to its environment, without recording the values of specific environmental variables. Gallistel here seems to identify the ‘procedures’ that are purportedly embodied by synaptic matrices in neural networks with the procedural knowledge discussed by psychologists, i.e. knowledge of how to perform certain motor skills. He distinguishes between the “indirect, opaque ‘knowing’ that is characteristic of finite-state machines”, and the “transparent symbolic knowledge” that is characteristic of mechanisms with a read-write memory, and holds that the former corresponds to procedural knowledge, whereas the latter corresponds to the ‘ordinary sense of knowing’ that we employ when we assert, for example, that a jay knows where and when it cached a morsel of food (Gallistel & King, 2010, p.101). Thus, in Gallistel’s view, associationism is inherently anti-representational in the sense that it characterizes learning as the acquisition of procedural knowledge for getting by in the world, rather than the acquisition of propositional knowledge of facts about the world.

\[21\] The claim that neuroscientists hold that synaptic conductances ‘are’, ‘realize’, or ‘embody’, associative bonds appears repeatedly throughout Gallistel’s writings. For a representative sample, see Gallistel & Gibbon (2001, pp.146-7), Gallistel (2008, pp.229-30), Gallistel & King (2010, pp.xiii, 177-8, 212, 282), and Gallistel & Matzel (2013, p.174).
It’s not always clear whether Gallistel holds that this is a view that associationists— and in particular, contemporary neuroscientists—themselves endorse, or whether it is a view that they are unwittingly committed to. On the face of it, many neuroscientists who study the role of synaptic plasticity in learning and memory do in fact seem to think that plasticity is involved in explicit, episodic memory; for example, some of the earliest and most compelling evidence for a role of synaptic plasticity in learning came from studies of episodic place learning in rodents (Morris et al., 1986). However, in a recent paper, Gallistel and his co-author Louis Matzel (2013) are explicit that regardless of what neuroscientists themselves might think, the associative framework they assume “may explain reflex modification phenomena, but it does not explain the learning of behaviorally important facts and the formation of data structures” (p.172). Moreover, extending arguments they’ve both developed elsewhere, they argue that conditioning, which is traditionally regarded as the epitome of a reflex modification phenomenon mediated by associative mechanisms, in fact doesn’t involve association at all. Thus the scope of plasticity-based, associative explanations of learning is very limited indeed.

In respondent (or ‘classical’) conditioning protocols, an unconditional stimulus (US) that has an innate tendency to elicit a certain behavioral response is repeatedly paired with a conditional stimulus (CS) that has no innate motivational valence, such that the CS eventually acquires the tendency to elicit the response. The traditional explanation of this process, the broad outline of which harks back at least to British empiricism, is that it involves the automatic, gradual formation of an association between the US and the CS on the basis of the temporal pairing between the stimuli. Gallistel holds that neuroscientific explanations of respondent conditioning presuppose this traditional explanation, and simply try to flesh it out by appealing to currently

---

recognized mechanisms of Hebbian plasticity such as long-term potentiation (LTP). The idea is that if neurons that are responsive to the US and the CS fire together, they gradually come to wire together. Thus, in Gallistel’s view, neuroscientific explanations of conditioning appeal to “much the same rewiring rules that Locke suggested” (Gallistel & King, 2010, p.253).

One source of problems for the view that the induction of LTP is the mechanism of respondent conditioning stem from the fact that that the temporal properties of LTP simply don’t reflect the temporal properties of conditioning at the behavioral level. For example, long-term potentiation is only ‘long-term’ relative to other forms of synaptic plasticity; the duration of LTP is measured in hours or days, whereas a conditioned response can last for months or years — sometimes for the duration of an animal’s lifetime. Gallistel & Matzel (2013) underscore this point by describing an experiment by Power et al. (1997), which, they claim, found evidence of LTP in the CA1 region of the hippocampus in rabbits that had been trained on an eyeblink conditioning task. They point out that the increased synaptic efficacy found in this experiment did not last more than 24 hours, whereas eyeblink conditioning is known to last for weeks or months. Similarly, Gallistel & Matzel point out that the time-scale of LTP induction does not reflect the intervals between CS and US presentations during which conditioned responses can be acquired. LTP occurs only within a very narrow time window between presynaptic and postsynaptic activity; in certain hippocampal cells, LTP only occurs if presynaptic activity occurs within a window of about 20 msec before postsynaptic activity, and in fact if presynaptic activity occurs within 20 msec after postsynaptic activity, long-term depression (LTD) occurs (Bi & Poo, 1998). By

---

23 Although I will focus on respondent conditioning in what follows, I assume along with Gallistel that the central points I’m about to discuss carry over to operant (or ‘instrumental’) conditioning. As Gallistel & Gibbon (2000) argue, these conditioning phenomena are fundamentally similar, at least with respect to the considerations that are salient here.
contrast, conditioning at the behavioral level can occur if the CS precedes the US by seconds, minutes, or in some cases — such as taste-aversion conditioning — even hours.

A deeper problem concerns not just the specific temporal properties of LTP, but the very idea that respondent conditioning involves the gradual acquisition of a conditioned response due to temporal pairing between stimuli. Gallistel & Matzel cite several studies from the contemporary associative learning literature suggesting that the interval between presentations of the CS and the US in respondent conditioning protocols per se doesn’t explain the number of presentations that are required for the acquisition of a conditioned response. Rather, what explains the rate of acquisition of a conditioned response seems to be the ratio of the interval between CS and US presentations to the background rate at which the US is presented, i.e. the CS-US / US-US ratio. Increasing the CS-US interval does not significantly alter the rate of conditioning so long as the US-US interval is correspondingly increased. In this sense, respondent conditioning is a time-scale invariant phenomenon; the crucial determinant of the ‘associability’ between two stimuli isn’t a specific temporal duration, but rather a unitless ratio. Thus the central problem with the view that LTP is the mechanism of conditioning isn’t that the temporal window of LTP induction fails to match the ‘window of associability’ for conditioning at the behavioral level, it’s that there simply is no window of associability.

How, then, is conditioning to be explained? Progress toward an answer to that question was made by a series of experiments beginning in the late 1960s that rev-

---

24 See, in particular, Gallistel & Gibbon (2000), and Gottlieb (2008).
25 Note that I am following Gallistel here in using the term ‘associability’ in a purely operational sense, to refer to the number of presentations of a CS and a US in a given conditioning protocol required to achieve a certain threshold level of conditioned responding, since this terminology is entrenched in the associative learning literature. Similarly, I am using ‘associative learning theory’ as a name, rather than a description, of a specific field of research that employs a distinctive experimental methodology, viz. the kind of conditioning protocols I’ve been describing. It should be clear by now that Gallistel denies that the (so-called) associability studied in (so-called) associative learning theory in fact depends on associative mechanisms.
olutionized associative learning theory. These experiments showed that, contrary to the traditional conception of conditioning, temporal pairing is neither necessary nor sufficient for the acquisition of a conditioned response. Rather, conditioning seems to depend on the subject’s perception of *contingencies* between the CS and US, such that the occurrence of a CS allows the subject to *predict* whether or when the US will occur. For example, experiments on inhibitory conditioning by Rescorla (1969) showed that when an US is only ever presented in the *absence* of a CS, such that there is a negative contingency between the CS and the US, the subject is conditioned *not* to exhibit the response elicited by the US. Conditioning in this protocol thus occurs even though there is no temporal pairing between the CS and US; temporal pairing is not necessary for conditioning. The most straightforward explanation of this phenomenon, it seems, is that the subject learns the contingency between the CS and the US, i.e. that the former predicts the absence of the latter. Conversely, experiments on ‘blocking’ by Kamin (1969) showed that when a subject initially acquires a conditioned response to a given CS, and subsequently undergoes conditioning trials in which that stimulus is presented together with a second CS, she will not acquire a conditioned response to the second CS. Thus conditioning to the second CS does not occur even though it is temporally paired with the US; temporal pairing is not sufficient for conditioning. Again, this phenomenon is straightforwardly explained if we suppose that conditioning is not an automatic, mechanistic process based on temporal pairing, but rather involves the subject’s perception of contingencies between the CS and the US. In blocking protocols, the occurrence of the second CS doesn’t provide any predictive information about the occurrence of the US over and above that which is already provided by the first CS.

One of the most important theoretical developments to emerge from the reconceptualization of the nature of respondent conditioning prompted by the experimental work of the late 60s was the model of conditioning developed by Rescorla & Wagner.
(1972). The Rescorla-Wagner model is a mathematical model of the circumstances under which conditioning occurs that attempts to lend quantitative precision to the intuitive idea that conditioning involves a reduction in the subject’s uncertainty about the extent to which the CS predicts the occurrence of the US. While the model has been highly influential, and successfully predicts phenomena like inhibitory conditioning and blocking, there are many other conditioning phenomena that it fails to predict. Most importantly, from Gallistel’s perspective, it fails to predict the time-scale invariance of conditioning. The reason for this is that although the model attempts to do justice to the view that conditioning is a cognitive phenomenon that involves the subject learning predictive relations between stimuli, it is still ultimately a version of an associative theory that presupposes a critical window of associability.

In a series of publications, Gallistel has argued that if we’re to truly do justice to the view that conditioning involves a reduction in uncertainty about contingencies between stimuli, and to the specific finding that conditioning is time-scale invariant, we must explicitly analyze conditioning in information-theoretic terms. Gallistel & King (2010) do just that, and derive an expression for the subject’s uncertainty about the timing of the US after the occurrence of the CS. I’ll prescind from the mathematical details of their derivation here and simply note the upshot: as suggested earlier, the most important factor in determining the amount of information that the CS communicates about the onset of the US is the ratio of the expected CS-US and US-US intervals in a given conditioning protocol. It’s this ratio that primarily determines the ‘associability’ of two stimuli. From their formal analysis, Gallistel & King conclude that in order for an animal to be capable of learning predictive relationships between stimuli, and hence to acquire conditioned responses, it must possess a specific type of learning mechanism, namely one that is capable of explicitly recording the values of

\[\text{26See, in particular, Gallistel & Gibbon (2001) and Gallistel & King (2010, pp.227-240).}\]
the intervals between those stimuli, and manipulating them to compute the relevant ratios. Conditioning, the epitome of an ‘associative’ learning process, seems to require *computation*, not mere association.

Gallistel holds that this computational perspective is inimical to contemporary neuroscience, which, he thinks, still cleaves to the traditional conception of conditioning as a mechanistic, reflex modification phenomenon mediated by temporal pairing — as evidenced by its emphasis on the role of Hebbian plasticity in learning. Not even more sophisticated incarnations of associationism inspired by the revolution in associative learning theory in the 1970s such as the Rescorla-Wagner model have had any impact in neuroscience, according to Gallistel: “contemporary neuroscientific research on the neurobiology of association formation proceeds as if contemporary associative theory, in which [phenomena like inhibitory conditioning and blocking play] a central role, did not exist” (Gallistel & King, 2010, p.231). Although the Rescorla-Wagner model and cognates are associative in the sense that they appeal to the formation of associative bonds within a certain temporal window, they require the values of those bonds to enter into certain computational processes such as summation, which is problematic on the assumption that associative bonds are realized by Hebbian synapses, since synaptic conductances cannot encode computationally accessible values. Yet, claims Gallistel, Hebbian synapses are the only mechanisms of learning that are currently recognized as neurobiologically plausible. While it’s true that the Rescorla-Wagner learning rule is in fact closely related to the backpropagation algorithm that was widely employed by the neo-connectionists of the 1980s,27 I pointed out in Chapter 2.4 that backpropagation is widely regarded as biologically implausible, even by its neo-connectionist proponents — in part because it is supervised by an ‘omniscient god outside the machine’, as

---

27To be more precise, the Rescorla-Wagner learning rule is formally equivalent to the delta rule, which is a special case of the backpropagation algorithm.
Gallistel puts it, and in part because there are no known mechanisms by which error signals could be propagated backwards through a network in the manner required by the algorithm. Neuroscientists don’t take backpropagation seriously as a mechanism of learning in actual neural networks, yet they don’t, according to Gallistel, recognize any other mechanisms that could explain conditioning as it is understood within contemporary associative learning theory. In sum, since the only kind of learning mechanisms that neuroscientists regard as plausible involve the automatic modification of synaptic conductance through temporal pairing, the only kind of learning that neuroscientists can explain involves non-cognitive changes in an organism’s reflexes and habits.\textsuperscript{28}

3.5.2 Empiricism

I mentioned earlier that in Gallistel’s view, conditioning is mediated by a learning mechanism that performs a specific set of computations on a specific body of information — specifically, a mechanism that computes the ratios of temporal intervals. This reflects a general point about Gallistel’s conception of conditioning, and of learning in general, that helps get us closer to heart of what he finds objectionable about the associative theory of learning that, in his view, is assumed by neuroscientists. Gallistel holds that conditioning is mediated by a special-purpose learning mechanism because he thinks that it is a special kind of computational problem, namely a problem in multivariate, non-stationary time series analysis.\textsuperscript{29} Such problems involve extracting predictive information about the temporal relations between possible events of interest from a body of data that might contain many potentially predictive events, which might with differing probabilities through different periods of time. Although traditional conditioning protocols obscure the ecological significance of conditioning understood as a

\textsuperscript{28}Gallistel doesn’t state this conclusion in so many words, but I think this is a reasonable extrapolation from many of his other explicitly-stated conclusions.

\textsuperscript{29}See Gallistel & King (2010, p.227).
problem in time series analysis, the capacity to solve this problem is of critical behavioral importance for any animal when, for example, foraging in the wild. For an animal to forage efficiently, it must be capable of learning which locations and sources of food are predictive of reward or punishment, under precisely which conditions.

Just as path integration is a behavioral capacity that helps an animal to solve a specific computational problem within its environment, and is mediated by a computational mechanism that reflects the form of that problem in its structure and function, the same is true, in Gallistel’s view, of conditioning. Indeed, Gallistel argues that this is an instance of a general point about learning that becomes clear when we take seriously the biological and computational nature of learning. Learning involves solving specific computational problems within an animal’s environment, problems that have a specific, mathematical form. Bracketing divine intervention or other miraculous saltations, any biological mechanism capable of solving a given problem must have been adapted by natural selection to have a specific structure and function, which reflects the form of that problem.30 This kind of adaptive specialization is a ubiquitous feature of evolved, biological mechanisms: “The foliated structure of the lung reflects its role as the organ of gas exchange... The structure of the hemoglobin molecule reflects its

30Gallistel cites Chomsky (1975) as an early and influential proponent of the view that learning mechanisms are specialized to solve specific computational problems. This is somewhat ironic, since — at least in the case of the specialized learning mechanism that Chomsky is most famous for advocating, the so-called ‘language organ’ — Chomsky does not argue for specialized learning mechanisms for the same reasons Gallistel does. Indeed, Chomsky (e.g. 1988) argues that the putative language organ is not an adaptive specialization, and seems to think that it did in fact appear as a result of a (miraculous, from a biological perspective) saltation. Pinker & Bloom (1990) provide a convincing and influential rebuttal to Chomsky’s arguments, ably defending the view that language evolved gradually via conventional Darwinian processes. In fact, contrary to the simplistic view that Darwinian adaptation necessarily involves modularity (in a strong sense of ‘modularity’, such as that articulated by Fodor, 1983), when one traces the changes that the evolution of language has wrought upon the brain by looking closely at the neuroanatomical substrate of linguistic capacities from a comparative perspective, the view that there is a strongly modular ‘language organ’ appears radically implausible. Language evolution seems to have involved widespread changes in a systems-level network comprising regions in temporal, parietal and frontal cortices (Aboitiz & Garcia, 1997). I’ll discuss Gallistel’s views about the relation between adaptation and modularity in more detail in Chapter 4.
function as an oxygen carrier” (Gallistel & King, 2010, p.219). Prima facie, it would be remarkable if adaptive specialization were not a feature of evolved, biological mechanisms of computation, too. From a biological perspective, then, we shouldn’t expect to find a general-purpose mechanism of learning that solves ‘the’ problem of learning, any more than we should expect to find “a general-purpose sensory organ, which solves the problem of sensing” (ibid., p.218). This is because, from a computational perspective, learning is a matter of solving specific computational problems, with a specific mathematical form, which must be reflected in the structure of the learning mechanism. So in Gallistel’s view, there’s no such thing as the mechanism of learning; rather, there are distinct, special-purpose learning modules for each of the computational problems that an animal faces within its environment.31

However, Gallistel holds that for centuries, learning has been conceptualized as a phenomenon whereby experienced regularities sculpt the mind or brain through domain-general processes of associative plasticity. As a testament to how deeply entrenched this conception of learning is, consider the widespread view, mentioned at the beginning of this section, that connectionist mechanisms have an advantage over classical mechanisms when it comes to explaining learning. From Gallistel’s perspective, this view evinces an unquestioned assumption about the essential nature of learning, namely that it involves the kind of associative plasticity exemplified by connectionist mechanisms. Gallistel holds that regardless of the theoretical framework within which this associative conception of learning has appeared — whether in Early Modern empiricism, behaviorism, or contemporary connectionism — it inherently characterizes learning as an unconstrained, experience-driven, domain-general process. Specifically, he holds that this is true of the associative conception of learning as it manifests in

31Gallistel’s discussion of these matters concerning the modularity of learning mechanisms appears in Chapter 13 of Gallistel & King (2010), which is drawn in large part from Gallistel’s earlier (2000) paper.
connectionist theorizing; he writes, for example, that theorists who develop neural network models “often are at pains to point out that the network has solved a problem in the absence of an initial structure tailored to the solution of that problem” (2000, p.1179), citing a model by Becker & Hinton (1992) that, those authors claim, discovers depth in random-dot stereograms without “prior knowledge of the third dimension” (ibid., p.161).

Proponents of the view that connectionist mechanisms are better poised to explain learning than classical mechanisms tend to argue that although it’s clear how, from a connectionist perspective, computational mechanisms and stored information could be shaped by experience, it’s entirely unclear how, from a classical perspective, computational mechanisms could ever be learned. But as I mentioned earlier, Gallistel denies that classicists should be moved by this purported problem, since he holds that the distinction between computational mechanisms and mechanisms of information storage reflects an appropriate division of labor between what’s innate and what’s shaped by experience. Indeed, he holds that by conflating this distinction, connectionists and other proponents of associative conceptions of learning have fundamentally mischaracterized the phenomena that presently stand in need of explanation, viz. learning and memory. Gallistel holds that these phenomena are quite distinct: “learning is the extraction from experience of behaviorally useful information, while memory is the mechanism by which information is carried forward in time in a computationally accessible form” (Gallistel & King, 2010, p.279). Correspondingly, he holds that we should expect these phenomena to be mediated by distinct mechanisms: learning is mediated by innate, special-purpose computational modules, whereas memory is mediated by a malleable, general-purpose mechanism — a mechanism of ontogenetic information storage that is biologically ‘universal’ in the same way that action potentials are a universal mechanism of communicating signals, and DNA is a universal mechanism of storing phylogenetic information. As Gallistel & King put it, “we see no more reason
to suppose that different mechanisms are required for different kinds of messages... than to suppose that different kinds of action potentials are required or different kinds of DNA” (ibid., p.287).

A crucial feature of the ‘computational/representational’ approach to learning, in Gallistel’s view, is that it provides a framework within which to understand how prior information is combined with incoming information to provide a statistically optimal posterior estimate of a specific aspect of the world. This prior information might be implicit in the innately-specified structure of the mechanism, or it might be explicit in the currently stored memory states that the mechanism has access to, but in either case the computational/representational approach provides a way of understanding how, in principle, this information might be integrated with incoming information, via computational operations, in a statistically optimal fashion. This is important because of an increasing appreciation in cognitive science over recent years that neural and cognitive processes are inherently probabilistic, noisy, and uncertain. One reason for this is that any physical mechanism is susceptible to a certain amount of thermodynamic noise, and the warm, wet environment of the brain is especially noisy; another is that the signals received at sensory receptors sometimes inherently underdetermine the information that is extracted from those signals — for example, computing a 3-dimensional visual representation from the signals provided by the 2-dimensional surface of the retina is widely recognized to be an ‘inverse problem’ that cannot be solved on the basis of the retinal signals alone.

How might the brain overcome these problems? Hermann von Helmholtz (1866 [1925]) famously suggested that visual perception might involve a kind of unconscious probabilistic inference, whereby the brain supplements the noisy and ambiguous visual signals it receives with background assumptions and hypotheses. Over the part two decades, developments in statistics and machine learning have provided a suite of precise, mathematically rigorous tools for understanding probabilistic inference, and
researchers in psychology and neuroscience have adapted these theoretical tools so as to extend von Helmholtz’s basic insight and account for the properties of a wide range of sensory, motor, and cognitive systems. At the foundation of these theoretical developments lies Bayes’ theorem — a trivial entailment of the axioms of probability theory which, given a certain interpretation of probability, achieves deeper resonance as a rule for optimal probabilistic inference. The picture emerging from the broadly Bayesian approach in contemporary cognitive science is that the brain is in the business of making probabilistic inferences about its environment by updating estimates about environmental properties on the basis of incoming evidence, in ways that are ‘rational’, or approximately Bayes-optimal.

From Gallistel’s perspective, the Bayesian approach fits naturally within the information-theoretic picture of world-brain communication and the corresponding computational/representational conception of learning, since information is fundamentally a probabilistic quantity, and the communication of information — the reduction in a receiver’s uncertainty — can only occur when the receiver has a prior probability distribution over the possible messages it might receive. However, Gallistel suggests that the Bayesian approach is antithetical to the associative conception of learning, since, to echo points made in the previous sub-section, associative learning according to Gallistel is an arational, mechanistic process that simply recapitulates the relations that were imprinted by experience; learned behavior, on the associative view, “is always recapitulative of the input-output conditions during learning” (Gallistel & Matzel, 2013, p.170). This is because associative architectures do not incorporate innately-specified learning mechanisms that implicitly embody within their structure prior probability distributions defined over a specific domain, which are required for Bayesian inference to get off the

32 For a review of the Bayesian approach in the context of psychology, see Chater et al. (2006). For a review of the approach in the context of neuroscience, see Knill & Pouget (2004).
ground. Only when a cognitive system incorporates domain-specific learning mechanisms can “the inferences [that the system] draws from limited experience go beyond what is justified by that experience” (Gallistel & King, 2010, p.226), and do more than merely recapitulate experienced regularities.

It’s important to note that the problem here is not merely that most associative architectures do not in fact incorporate domain-specific structures that specify the prior probability distributions required for Bayesian inference, such that associationists could avoid the problem simply by building domain-specific structure into their associative networks. Rather, Gallistel seems to think that the problem is inherent to associative architectures because, in his view, such architectures simply cannot encode probability distributions in the form of explicit symbols that could enter into the computations that mediate Bayesian inference. As he writes in his (2008) paper, “[a] net’s inability to specify a probability distribution over possible states of the world... makes it difficult to combine the evidence from different computational procedures” (p.235). Again, from Gallistel’s perspective, we run up against the fundamental limitations of a computational architecture that lacks a memory that can store explicit symbols for long periods of time in a computationally accessible form.

This is problematic because of the abundant evidence that learning involves updating an estimate of specific states of the environment according to broadly Bayesian principles. Indeed, to reiterate a theme from the previous sub-section, we find evidence of rational, probabilistic inference even in the traditional heartland of the associative theory of learning, namely research on conditioning. This fact has been obscured because traditional methods of analyzing the results of conditioning experiments typically involve averaging results across subjects to produce smooth, gradual learning curves, a practice that has lent credence to the associative view that conditioning involves gradual changes in associative bonds. However, Gallistel et al. (2004) argue that gradual learning curves are an artifact of averaging across subjects, and reanalyze the results
of several conditioning studies to show that the learning curves of individual subjects often display abrupt jumps, consistent with the subject making decisions on the basis of accumulated evidence about the presence or absence of certain contingencies within its environment. Again, even the epitome of a supposedly domain-general, automatic, associative learning process — conditioning — in fact seems to be a cognitive process by which a subject’s explicit representation of environmental contingencies is rationally updated on the basis of incoming evidence. In Gallistel’s view, this requires us to posit specialized learning modules that have access to a ‘universal’ read-write memory mechanism.

Summarizing the lessons that he takes the foregoing considerations to have for neuroscientific explanations of learning and memory, Gallistel (2000) writes that “[d]espite long-standing and deeply entrenched views to the contrary, the brain no longer can be viewed as an amorphous plastic tissue that acquires its distinctive competencies from the environment acting on general purpose cellular-level learning mechanisms” (p.1190).

### 3.6 A Case-Study of Connectionist Inefficiency

So far we’ve seen several serious accusations leveled against the connectionist mechanisms of computation posited by contemporary neuroscientists, but we’ve yet to see much direct evidence that those mechanisms are guilty. That is, we’ve been told in very general terms that connectionist mechanisms suffer from various crippling defects, but we we’ve yet to see how any specific model actually suffers from those defects. Thus, to convince a skeptical jury, Gallistel & King (2010) bring their arguments to bear on what they take to be an exemplar of the kind of connectionist theorizing they object to: the model of path integration developed by Samsonovich & McNaughton (1997). I’ll first sketch the most significant features of this model, then I’ll explain why Gallistel
& King find those features problematic.

### 3.6.1 The Samsonovich-McNaughton Model

The model developed by Samsonovich & McNaughton (1997) is a computational simulation of a path integration mechanism that was hypothesized to exist in the rodent hippocampus by McNaughton et al. (1996). This earlier hypothesis was motivated by a large body of electrophysiological evidence that the hippocampus in rodents, and presumably other mammals, functions as a ‘cognitive map’ of the animal’s environment (O’Keefe & Nadel, 1978). The most important of such evidence is the discovery that certain ‘place cells’ in the rodent hippocampus exhibit a high rate of firing whenever the animal is in a specific location within its environment (O’Keefe & Dostrovsky, 1971). Collectively, this evidence suggested to many that the hippocampus maintains a representation of the animal’s current environmental location that persists even in the absence of incoming perceptual information, such as in total darkness, but which can be updated on the basis of perceptual information, as well as self-motion or ‘idiothetic’ signals. McNaughton et al. (1996) built upon this idea by suggesting that the internal representation of current location might be instantiated by a stable pattern of activity within a population of place cells in the hippocampus, which, from a dynamical systems perspective, could be understood as an attractor in a two-dimensional state-space. The current location of the attractor in the state-space, on this view, reflects the current location of the animal within a planar environment — just as the water levels in the buckets in the hydraulic path integration mechanism described in Chapter 2.2.2 reflect the current location of the boat in a planar environment. Samsonovich & McNaughton (1997) implemented this idea computationally using a relatively biologically realistic network of spiking neurons. The crucial question they faced was how to update the current location of the attractor on the basis of incoming perceptual and idiothetic information in such a way that it continued to reflect the animal’s current location.
The basic architecture of the mechanism they proposed is depicted in Figure 3.1. The mechanism looks complicated, but there are only a few components we need to focus on for present purposes. The first component is the cognitive map itself, labeled $P$ in the diagram. This can be understood as a ‘chart’, a population of place cells arranged conceptually on a two-dimensional plane such that the location of given place cell in the plane corresponds to the ‘place field’ for that cell, i.e. the location within a specific environment for which that cell is most responsive.\textsuperscript{33} Due to the pattern of connectivity between the neurons in the population represented by the chart, the population supports a stable gaussian ‘bump’ of self-sustaining neural activity — the aforementioned planar attractor.

The second major component is the tripartite structure labeled $H', R, H$. Collectively, this structure serves to compute the current compass heading of the animal. The sub-component $H$, like $P$, is an attractor network that supports a stable bump of neural activity, but in this case the bump is an attractor that occupies points on a \textit{ring}, rather than points on a \textit{plane}. The current location of the activity bump in the ring attractor reflects the current compass heading of the animal, i.e. the angle $\theta$ with respect to north. $H$ is thus identified with a population of ‘head direction cells’, cells that are maximally responsive when an animal’s head is oriented in a particular direction, which have been found in various mammalian brain structures that are closely connected to the hippocampus (Taube et al., 1990). By means of a mechanism that I won’t describe here, the entire tripartite structure $H', R, H$ updates the current

\textsuperscript{33}Note that this two-dimensional arrangement of place cells in a chart is simply a way of conceptualizing the relations between place cells, i.e. in such a way that they reflect relations between the environmental locations for which the cells are responsive. It doesn’t reflect any assumptions about the \textit{anatomical} organization of the cells. Moreover, a particular arrangement of place cells in a chart is relative to a particular environment; in a different environment the same population of cells might be arranged differently. This reflects the phenomenon of ‘remapping’: the electrophysiological finding that there are typically no correlations between the place fields for different place cells across different environments (O’Keefe & Nadel, 1978).
Figure 3.1: The attractor network model of path integration in the rodent hippocampus developed by Samsonovich & McNaughton (1997). A ‘bump’ of self-sustaining neural activity within a population of hippocampal place cells \((P)\), understood as an attractor in a two-dimensional state space, serves as an estimate of the animal’s current location within a planar environment. The movement of the bump of activity is driven by signals about the animal’s current heading, provided by head direction cells \((H)\), and by signals about the animal’s current speed, provided by efference copies from the motor system \((M)\). For details about how these signals are combined, see the text. From Fig. 2 of Samsonovich & McNaughton (1997, p.5902).

location of the activity bump in \(H\) by integrating signals about the rate at which the animal’s head is turning provided by \(H’\), so as to compute the distance that the bump travels around \(H\). That takes care of the directionality computations required by a

\[34\]

I won’t describe this mechanism in detail because it works according to essentially the same principles as the mechanism by which the activity bump in \(P\) is updated so as to reflect the animal’s current position, which I’ll describe in the following paragraph. My purpose here is simply to outline the general principles by which these mechanisms work; to that end, I’ll just focus on the central mechanism, viz. the one that performs path integration.
path integration mechanism, but what about distance information? Such information is provided by the structure labeled $M$ in the diagram, which provides self-motion signals, presumably derived from efference copies of motor commands.

The three components I’ve just described come together in the component that actually performs the path integration computation, namely $I$. This component consists of a stack of two-dimensional layers of neurons, where each $I$ layer corresponds to a particular compass direction $\theta$, and projects to the $P$ layer via connections that are slightly displaced in that direction (the connections labeled $W^{IP}$ in the diagram). The $P$ layer projects directly back to each $I$ layer via reciprocal connections that are not displaced (labeled $W^{PI}$). The elements of the $H$ array project to each layer of $I$ in such a way that the current bump of activity in $H$, which corresponds to the animal’s current head direction, selects one layer of $I$ as active — the one corresponding to the current direction, $\theta$ — and inhibits all the others.\textsuperscript{35} The current bump of activity in $P$ excites a region of the currently active $I$ layer, and this excitation returns to the $P$ layer via the $W^{IP}$ connections, driving the bump in $P$ in the direction associated with the active $I$ layer. Moreover, the $M$ array projects excitatory connections to all the layers of $I$ and thereby modulates the gain of the excitation returning from the currently active $I$ layer to the $P$ layer in a way that is proportional to the current speed of the animal. Thus directionality signals from $H$, and speed signals from $M$, drive the activity bump through the two-dimensional state space encoded by $P$ in a way that reflects the direction and speed of the movement of the animal through its environment.

At least, that was the hope. And, indeed, the numerical simulations of the Sam-

\textsuperscript{35}It’s worth noting that head direction cells don’t constitute a true compass, since they don’t signal geocentric compass directions, but rather the direction of the animal’s head relative to its current environment (Taube et al., 1990). This subtlety is largely tangential to the focus of the text, which concerns the mechanisms by which directionality signals update the representation of location.
sonovich & McNaughton (1997) model suggested that model lived up to expectations. For example, the simulation showed that when the network is initialized with a burst of random activity, a stable bump of activity spontaneously forms in the $P$ component, and that when the network is stimulated by activity in the $H'$ and $M$ components in a way that mimics the head-turning and self-motion signals produced by an animal moving in a circle at a given speed, the activity bump traces a circle through the planar state-space encoded by $P$, thereby performing path integration. However, Gallistel & King (2010) hold that the model exemplifies the various problematic features of currently recognized mechanisms of neural computation that were surveyed in the last few sections, and illustrates just how problematic those features are when it comes to explaining relatively simple cognitive capacities like path integration.

3.6.2 Problems

Let’s first consider how the the Samsonovich-McNaughton model exhibits the allegedly problematic features canvassed in previous sections. To begin, let’s consider how in-

---

36See Fig. 10D in Samsonovich & McNaughton (1997). An interesting aspect of the Samsonovich-McNaughton model that is reflected in this figure, which unfortunately I’m unable to discuss in the text because it is largely orthogonal to the main points that I wish to cover, pertains to the fact that the model — and indeed any path integration mechanism — accumulates error over time. Path integration provides an estimate of current location, and that estimate is only as accurate as the signals it is derived from. Directionality and distance signals will inevitably be degraded by various sources of noise, and any errors introduced by noise will accumulate at each stage of the path integration process. Human navigators, when they can, mitigate the problem of path integration error by incorporating information about observed landmarks into their estimate of current location, and it is widely believed that animal navigators do the same. Samsonovich & McNaughton explore how landmark information might be used to update the estimate in their model, and thereby correct PI-based errors, via the component depicted in Figure 3.1 that I haven’t discussed yet, namely $V$. This is understood to be a population of neurons that encodes the total sensory information that the animal receives from a particular point of view, i.e. a specific location and orientation within a given environment. The authors don’t simulate this component in detail, but they do show that input from this component can shift the activity bump within $P$, thereby showing how, in principle, sensory information might help to correct PI-based errors. For example, they show that if input from $V$ is sufficiently strong, it can force the activity bump to abruptly jump from one location in $P$ to another, as if the animal suddenly realizes, on the basis of incoming sensory information, that it’s not where it thought it was.
formation is coded within the network. The network clearly exemplifies the strategy of encoding information as a stable pattern of reverberating activity within a recurrent neural network that was discussed in section 3.3.1. What is the scheme by which information is encoded? Note that the frequency of the reverberating activity isn’t directly relevant to the information encoded by the activity bump; rather, it is the location of the activity bump within a population of neurons that determines the specific value that the bump encodes. Thus the network employs a place code, which, recall, Gallistel takes to be an instance of a nominal information coding scheme. To put the point slightly differently, the Samsonovich & McNaughton network uses an entire population of neurons to encode a single variable, where the specific value of that variable is determined by the locus of activity within that population.

At this juncture, it’s worth reiterating a point I made in section 3.4, namely that Gallistel does allow that reverberating patterns of activity such as those exhibited by the Samsonovich-McNaughton network count as symbols, and that the attractor networks that sustain such networks count as a read-write symbolic memory. Somewhat confusingly, Gallistel & King (2010) introduce their discussion of the Samsonovich-McNaughton model by writing that it exemplifies a computational architecture “that does not have a symbolic memory” (p.242), and that “lacks a read/write memory” (p.243), yet a few pages later they assert that “activity in a reverberatory neural circuit can serve as a memory” (p.245), and in particular that the recurrent circuitry in the Samsonovich-McNaughton network “constitutes a read/write memory” (p.259). This latter assertion seems to be their considered view, and indeed the correct view,

---

37 It’s worth noting, though, that the frequency of the activity bump in fact does play an important functional and explanatory role within the model. For example, one of the central results of the model was that the shape and movement of the activity bump was modulated by the background theta rhythm, qualitatively reproducing an experimentally observed phenomenon known as ‘theta phase precession’.

38 In a later paper, Gallistel (2011) writes that all neural network models of path integration, including the Samsonovich-McNaughton model, “are symbolic models” (p.254).
for stable patterns of activity in attractor networks clearly seem to satisfy the criteria that Gallistel sets out for something’s being a symbol; perhaps most importantly, they are, as Gallistel & King allow, “accessible to computation” (p.249).

But to return to the question at hand: how are those symbols computationally manipulated? Consonant with the fact that the symbols within the network are reverberating patterns of activity, Gallistel & King hold that they are manipulated by a *look-up table*, instantiated by component $I$. The neurons in this component are said to function as ‘three-legged’ AND-gates, which receive one input from the $P$ component, one from $H$, and one from $M$, thus selecting the row, column, and layer of the look-up table instantiated by $I$. Now, this isn’t an entirely accurate way of describing the functionality of the $I$ component. Although localized regions of the $P$ component do project to localized regions of the $I$ component, the $M$ component, for example, diffusely modulates the gain of the entire $I$ component. Moreover, there’s no literal, theoretically substantive sense in which the neurons in $I$ function as AND-gates. The most important reason is that there is no global oscillatory signal in the network that precisely synchronizes the inputs to a neuron in $I$, ensuring that they all line up at the same clock cycle so that the conjuncts they purportedly instantiate are all part of the same conjunction; that’s just not how spiking neurons work. Nevertheless, it seems clear that the manner in which $I$ manipulates the activity pattern in $P$ is loosely analogous to the operation of a look-up table.

So far we’ve seem *that* the Samsonovich-McNaughton model exemplifies some of the allegedly problematic features of neural network theorizing identified by Gallistel & King, but we haven’t yet addressed the question that we opened this section with, which concerned *how* those features are problematic. On the face of it, the model seems to do what it was designed to do, so given its apparent explanatory success, one might take it to show precisely the *opposite* of what Gallistel & King take it to show, namely that the allegedly problematic features of neural networks aren’t really
problematic at all. However, Gallistel & King discuss several specific ways in which these features make trouble for the Samsonovich-McNaughton model, which promise to prevent this *tollensing* of their attempted *ponens*.

Let’s first consider why the manner in which symbolic vehicles of information are instantiated in the Samsonovich-McNaughton network might be problematic. We’ve seen that although Gallistel concedes that patterns of reverberating activity in an attractor network count as symbols in a read-write memory, he holds that they’re ‘not very good symbols’, his central reason being that they’re not thermodynamically stable. This general point manifests in a few specific ways in the context of the Samsonovich-McNaughton model. First, Gallistel & King point out that reverberating patterns of activity are transient, ephemeral entities that only emerge and persist under a highly specific range of physiological conditions; whether or not a stable pattern of activity will form in an actual neural network, and for how long, depends on the connectivity pattern in the network, the balance of excitation and inhibition, how much noise there in the network, and a range of other specific, quantitative details. Gallistel & King argue that the fact that a stable pattern formed in the simulated network developed by Samsonovich & McNaughton is “at best weak evidence that such a mechanism exists and actually works in any real brain” (*ibid.*, p.258). One reason is that the Samsonovich-McNaughton model is a set of complex nonlinear differential equations that doesn’t have an analytic solution, so the only way to determine what the model will do is to simulate it numerically. But the process of programming and running a highly complex simulation of this kind might introduce many errors, so it’s difficult to be sure that the behavior of the simulation accurately reflects the behavior of the model it simulates.\(^{39}\)

Moreover, partly in order to mitigate some of the computational complexity in-

\(^{39}\)For these objections, see Gallistel & King (2010, pp.247-9).
volved in simulating the model, Samsonovich & McNaughton introduce several idealizing assumptions into the model itself. Gallistel & King focus on three in particular. First, they point out that the rectangular arrangement of neurons in the $P$ and $I$ arrays raises the question of what happens when the activity packet reaches the boundaries of the rectangle. Samsonovich & McNaughton address this question by holding that the $P$ and $I$ arrays in fact have the topology of a torus; in other words, the rectangle is curled into the shape of a tube, and then again into donut, so that the boundaries are periodic — when the activity bump exits one side of the rectangle, it re-enters on the other side, like Pac-Man trying to escape his maze. Gallistel & King hold that this is a quick and dirty fix that precludes the representing structure, i.e. the toroidal $P$ array, from being homomorphic with the represented structure, i.e. the Euclidean plane: “points in the represented world that are maximally distant from each other are in this symbolic world maximally close” (p.255).

The second idealizing assumption is that in order to keep the computation tractable, inhibition was not modeled explicitly in the way the activity of the excitatory spiking neurons in the model was; rather, it was treated as a global constant that provided an upper limit on the total number of neurons firing at a given time. But, to echo points made in the preceding paragraph, the specific balance of inhibitory and excitatory activity in a recurrent network is widely regarded as a crucial determinant of whether, and for how long, an stable pattern of activity will form, so, Gallistel & King argue, it’s not clear how the Samsonovich-McNaughton model would behave if inhibition were modeled in more detail.

The third idealization also concerns the parameter sensitivity of the model, specifically the sensitivity of the model to various sources of biophysical noise. Although some of the simulations of the model incorporated a degree of noise, Gallistel & King complain that Samsonovich & McNaughton do not discuss the effects of manipulating noise levels in detail. But again, the level of noise in a recurrent network is a deter-
minant of whether, and how long, stable patterns of activity will form. The authors hold that this issue is especially pressing given theoretical results by Maass & Sontag (1999) showing that “introducing even modest amounts of noise into a net can destroy its functionality” (p.256).

All of these issues raise questions that are significant in their own right, but they are especially salient given the background question of whether the activity bump in the Samsonovich-McNaughton network could persist long enough to account for the behavioral evidence. Samsonovich & McNaughton reported simulations of their model of up to six seconds, over which time the activity bump remained stable, but Gallistel & King point out that ant foraging expeditions can last on the order of half an hour, and rat foraging expeditions last roughly ten times that (p.262).

Another problem related to the manner in which information is encoded in the Samsonovich-McNaughton network concerns the efficiency of the coding scheme employed. In many of the simulations of the model, the $P$ component is initialized to contain on the order of 50,000 neurons. Assuming that we can individuate the values instantiated by the network neuron-by-neuron, on the basis of the maximally active neuron at the center of an activity packet — not a very accurate assumption since activity packets are diffusely spread out over large populations of neurons — the $P$ component could represent at most 50,000 different locations. Gallistel & King point out that a binary string of only 16 digits could encode more information than that ($2^{16} > 50,000$), and indeed they suggest that such a string could in principle be instantiated by the activity of 16 neurons if each neuron were recurrently connected to itself in such a way that it could function as a binary switch, whereby excitation would put the neuron into a self-sustaining ‘on’ state, whereas inhibition would put it into

---

40This number includes the excitatory pyramidal cells modeling in the network, but not the inhibitory interneurons that, as I mentioned earlier, were not explicitly modeled.
a quiescent ‘off’ state. In such a scheme, they suggest, locations could be compactly encoded as $x$-$y$ coordinates, with the first 8 neurons encoding the $x$-coordinate, and the second 8 encoding the $y$-coordinate (ibid., p.260). It’s not clear how seriously the authors intend this suggestion given the concerns they raise about the timescales over which reverberating activity can persist, and given that the scheme they propose would appear to be dramatically more susceptible to the concerns they raise about parameter sensitivity and noise than the Samsonovich-McNaughton model, but in any case their suggestion seems to underscore how profligate the latter model is. In further support of this point, Gallistel & King note that the total number of neurons in the Samsonovich-McNaughton model is on the order of $10^6$, which is approximately the total number of neurons in the ant brain, yet ant brains are of course capable of doing far more with their neurons than just path integration.

When we turn to consider problems associated with the manner in which computations are implemented in the Samsonovich-McNaughton model, the theme of profligacy persists. First note that the look-up table mechanism that implements the path-integration computation in the model, component $I$, requires a very large number of precisely wired neurons to perform a relatively simple computation. But the central problem that Gallistel & King see with this mechanism is that it performs only that computation, and only using the specific variables that it is pre-specified to interface with, such as $M$ and $P$. The model, as it stands, cannot perform many of the computations related to path integration that are evidenced in the behavioral literature, such as computing the direction and distance back to the animal’s point of origin, or of recording specific locations of behavioral interest and using that information to compute an optimal foraging strategy. But even ants are capable of recording the location of a source of food and subsequently navigating back to that location (Collett et al., 1999). Now, the Samsonovich-McNaughton model was really only intended to test the plausibility of implementing an attractor-based cognitive map in spiking neurons, so
it’s perhaps unfair to complain that it doesn’t in fact perform certain computational tasks that it wasn’t designed to perform. But the concern that Gallistel & King are raising concerns how the model could perform those further tasks. They point out that although the $P$ component might be imagined to interface with another computational mechanism that, for example, computes the homing vector, it would have to be pre-wired to do so. The values of the variable instantiated by $P$ are not generally accessible to computation.

We can see how this point relates to the theme of profligacy by focusing on the question of how the Samsonovich-McNaughton network might be capable of recording and manipulating information about the location of specific landmarks. There is in fact a component of the network that is relevant to this capacity, which I haven’t yet discussed in detail.\footnote{I discuss this component, and the explanatory role that it plays within the S-M model, in note 36.} The component labeled $V$ in Figure 3.1 is understood to be a population of neurons that encodes the total sensory information that the animal receives from a particular point of view, i.e. from a specific location and orientation within a given environment. We might suppose that the location of a specific landmark is recorded in long-term memory by somehow combining the locus of activity in $V$ with the current locus of activity in $P$. But here we allegedly encounter the ramifications of a mechanism that fails to satisfy the third grade of computational flexibility discussed in section 3.4.3, the capacity to flexibly generate variables and compose them into data structures. Recall that in Gallistel’s view, the only remotely plausible candidate for a symbol in a neural network is a reverberating pattern of activity. But generally speaking, only one stable pattern can exist within a given network at a time. So, Gallistel & King argue, the only way in which a mechanism with the Samsonovich & McNaughton architecture could record the location of a specific landmark would be for it to replicate the components $P$ and $V$, along with the current loci of activity
within those components, to obtain a new network, \( \{p_1, v_1\} \). To record the location of a second landmark, the mechanism would have to generate another network, \( \{p_2, v_2\} \). And so-on for each landmark that the mechanism records.\(^{42}\)

It’s not at all clear how this process of generating new variables as needed — i.e. generating entirely new populations of neurons \textit{de novo} — could be implemented using mechanisms recognized by neuroscientists. To echo points made earlier, all of the possible variables for locations would presumably have to be pre-allocated. But each of these variables \( \{p_n, v_n\} \) is instantiated by roughly 500,000 neurons.\(^{43}\) Jays can apparently remember up to around 10,000 cache locations, but even if we conservatively assume that they can remember only 1,000 locations with high fidelity, we’re already dealing with approximately \( 5 \times 10^8 \) neurons. Problems of profligacy become far more acute when we consider how to perform computations on the values of these variables. Many animals are capable of traveling from one arbitrary location that they’re experienced to another. This capacity is presumably mediated by computations defined over the values of the variables for locations recorded in memory. In the kind of architecture exemplified by the Samsonovich-McNaughton model, this would presumably involve distinct look-up table mechanisms, analogous to component \( I \), for each of the possible pairs of location variables — in fact, each of the \textit{ordered} pairs, since it’s once thing to compute the range and bearing from Location 1 to Location 2, and another to compute the reverse trip. So that’s \( 2 \binom{1,000}{2} = 999,000 \) different look-up tables, each consisting of approximately 270,000 neurons.\(^{44}\) Together with the neurons required to store the

\(^{42}\)Concerning notation: I’m using letters rather loosely here to denote both networks and the variables they instantiate. Uppercase letters denote specific networks, which correspond to a given variable, and lowercase letters denote networks that are currently in a specific attractor state, where such states instantiate the current value of a given variable. So \( P_n \) denotes the network that instantiates the variable for the \( n \)th location, and \( p_n \) denotes the current value of that variable.

\(^{43}\)It’s hard to say for sure, since Samsonovich & McNaughton (1997) don’t explicitly model the \( V \) component.

\(^{44}\)In many of the simulations of the Samsonovich-McNaughton model, the \( I \) array comprises six layers, each containing 45,000 neurons.
variables, that’s around 270 billion neurons, almost three times as many as are in the human brain, just to perform a relatively simple navigation computation.\textsuperscript{45} However, that’s based on a relatively conservative estimate of the number of locations an animal might remember; the real problem here is that the number of neurons required scales exponentially with the number of locations. On the basis of these considerations, Gallistel & King conclude that the computational architecture assumed by Samsonovich & McNaughton “does not appear to be physically realistic” (p.262).

3.7 Memory Amongst the Molecules?

The upshot of the theoretical and experimental considerations raised in Chapter 2 was that “the behavioral evidence for representation and computation, together with basic insights in computer science about the centrality of a read-write memory mechanism in physically realized computation, implies the existence of a neurobiological read-write memory mechanism” (Gallistel, 2008, p.228). The upshot of the arguments presented earlier in this chapter is that none of the mechanisms currently recognized by neuroscientists could function as an effective read-write memory. But the central problem with contemporary neuroscience, as Gallistel sees it, is not so much that it has yet to find a mechanism of read-write memory, as that it is looking in entirely the wrong place. Because they are convinced that the nature of learning and memory is fundamentally associative, neuroscientists simply aren’t looking for mechanisms with the kind of properties needed to explain the behavioral evidence; they’re not looking for

\textsuperscript{45}The estimate of the number of neurons required by the Samsonovich-McNaughton model that I’ve provided here is based on a calculation described by (Gallistel & King, 2010, p.264), however it employs estimates of the relevant numerical values that I think are fairer to Samsonovich & McNaughton, and that are generally more accurate. Although the absolute value of the number I’ve obtained is vastly smaller than the number obtained by Gallistel & King, it licenses the same conclusion regarding the plausibility of the model, thus I think the estimate here is argumentatively stronger than that provided by Gallistel & King.
mechanisms that could store symbolic vehicles of explicit information in a thermody-
ynamically stable form, at discrete addressed locations in memory, such that they could
be fetched from memory and transported to the mechanisms that implement primit-
tive computational operations when needed. The associative plasticity hypothesis that
guides their research “fundamentally misrepresents the nature of what it is we should
be looking for” (Gallistel & King, 2010, p.278).

Nevertheless, Gallistel & King are optimistic that neuroscientists will eventually
realize the error of their associationist ways and come to appreciate that “brains must
possess a mechanism for carrying information forward in time in a computationally
accessible form”, writing that this “elementary insight will someday transform neu-
roscience” (ibid., p.287). They believe that it is futile, given our present state of
ignorance, to speculate about the details of the neurobiological mechanisms of read-
write memory — in the same way that the biochemical nature of the gene was, until
Watson & Crick (1953) came along, “beyond the power of pre-1953 biochemists to
imagine”, they hold that the neurobiological nature of read-write memory is beyond
the power of present-day researchers to imagine (ibid., p.281). However, they do offer
some suggestive speculations about the general direction in which researchers should
look for the mechanisms of memory. Specifically, drawing inspiration from Feynman’s
famous point that ‘there’s plenty of room at the bottom’, they suggest that researchers
ought to look down, at the molecular level, rather than the cellular level of neural cir-
cuitry. Feynman’s point, as Gallistel & King interpret it in the present context, is that
the smaller the scale at which read-write memory can be mechanistically implemented,
the more efficient it will be; implementing memory at the molecular level would be less
metabolically expensive, would allow a higher density of information storage, would
allow signals to be transmitted faster since they’d have to traverse shorter distances,
and so forth.

One obvious question about this proposal concerns how signals transmitted by
neural spike trains could be decoded into symbols instantiated at the molecular level. Although we might grant that it is unfair to expect Gallistel & King to provide specific details about the mechanisms by which this inter-level decoding process is supposed to be achieved, if they fail to provide us with an inkling of how this process could be achieved, it surely wouldn’t be unreasonable for us to reject their molecular memory hypothesis, for it wouldn’t be clear what the content of that hypothesis is. Gallistel & King do go some way toward addressing this concern by suggesting that molecular memory might be implemented by ‘settable molecular switches’ analogous to rhodopsin, a photopigment in photoreceptor cells that can occupy distinct thermodynamically stable states. They write that it is “not hard to imagine how a spike train releasing transmitter onto a metabotropic receptor could set in motion an intracellular molecular cascade that changed the settings of intracellular molecular switches” analogous to rhodopsin (ibid., p.280). There’s a sense in which this is true. Molecular biologists have painstakingly elucidated the details of various receptor-mediated intracellular processes that quite naturally fall under the description of ‘molecular cascades that set molecular switches’. Such research provides us with an understanding of what counts as a molecular switch-setting process, and allows us to imagine other processes of the same kind. One of the most well-studied examples of such a process is in fact that which mediates LTP; the induction of LTP involves an intracellular molecular cascade that ‘sets’ various molecules in certain thermodynamically stable states. However, the ‘settable switches’ involved in LTP, light transduction, and other such processes are presumably not the kind of settable switches that Gallistel & King have in mind. For Gallistel & King, molecular switches must not only be capable of occupying different thermodynamically stables states, those states must be capable of encoding the specific values of environmental variables. But our question was precisely about how this latter property could be physically realized, and that question isn’t resolved by appealing to our understanding of processes that don’t exhibit the relevant
Perhaps more interestingly than their talk of molecular switches, Gallistel & King suggest, alternatively, that molecular read-write memory might be implemented by the subunits of macromolecules, such as the nucleotide sequences in nucleic acids like DNA or RNA. Nucleotide sequences are highly thermodynamically stable and can store huge amounts of information in a highly compact form for tens of thousands of years, hence they satisfy the central criteria for an effective read-write memory. Gallistel & King are forthright that they “have no idea what the machinery might look like that would transcribe information from a spike train into a nucleotide sequence” (p.280). However, they argue that this shouldn’t give us license to dismiss the hypothesis that read-write memory is implemented by nucleotide sequences, since this hypothesis explains something that currently accepted hypotheses about the neurobiological mechanisms of memory do not, namely how a biological mechanism could store vast amounts of symbolized information in a thermodynamically stable form. Proof of this principle was recently provided by Goldman et al. (2013), who were able to encode all of Shakespeare’s sonnets, a color JPEG photograph, and a snippet of audio from Martin Luther King’s ‘I Have a Dream’ speech — estimated to be a total of $5.2 \times 10^6$ bits — in DNA. Thus the nucleotide hypothesis, “wild as it is, passes the test that the synaptic plasticity hypothesis fails” (Gallistel & King, 2010, p.280).

While it might pass the coding test, what about the computational flexibility test? Remember that Gallistel holds that we must posit not just a garden-variety serial mechanism of read-write memory, but an indirectly addressable read-write memory mechanism. How might we incorporate indirection and addressability into a nucleotide-based system of read-write memory? Gallistel & King don’t address this question

---

46It’s worth noting that the nucleotide sequences that Goldman et al. (2013) used to encode information were addressable, but the addresses they employed weren’t interpretable by any known intracellular mechanisms. Rather, they were designed to be interpreted by the scientists and instru-
directly, but they do make some comments in other contexts that are suggestive of how they *might* address the question. In the context of arguing that any effective memory mechanism must support a distinction between symbols in memory that function as addresses, and symbols that function as data, they point out that the molecular mechanisms of phylogenetic memory — DNA, RNA, and their auxiliary molecular machinery\(^\text{47}\) — support just such a distinction, in the form of the distinction between *promoters*, DNA sequences that serve to identify specific genes and initiate the process of transcribing and translating them into polypeptides, and *coding sequences*, DNA sequences that actually encode the structure of polypeptides. Gallistel & King hold that this distinction *just is* the distinction between memory addresses and data, writing, for example, that the phylogenetic memory mechanisms implement a kind of “indirect addressing that makes possible the hierarchical structure of the genome” (p.169). While they don’t explicitly endorse the view that the hypothetical nucleotide-based system of ontogenetic memory employs essentially the same molecular mechanisms of indirect addressing as the nucleotide-based system of *phylogenetic* memory, it seems reasonable to tentatively impute to them that view.

Nevertheless, such an imputation must be tentative, since, as mentioned, Gallistel refuses to make any specific conjectures about the neurobiological mechanisms of read-write memory. However, he is confident that there must *be* such mechanisms, that such mechanisms are completely alien to contemporary neuroscientists, and that the eventual discovery of such mechanisms “may someday have an impact on neuroscience comparable to the impact of the discovery of radioactivity on physics” (2008, p.228).

---

\(^{47}\)I should note that it’s actually highly controversial within some circles of theoretical biology, most notably developmental systems theory (Oyama et al., 2001), to ‘privilege’ the causal role of nucleic acid in developmental and evolutionary explanations in the way that I appear to be doing here. I wish to remain officially neutral about these debates, and write of ‘auxiliary’ molecular machinery purely for expository convenience.
In the next chapter, I'll turn from exposition to criticism, and argue that Gallistel's confidence about these matters is misplaced.
Chapter 4

Beyond Good and Evil

Psychologists and neurophysiologists chart the same bay — working perhaps from opposite shores, sometimes overlapping and duplicating one another, but using some of the same fixed points and continually with the opportunity of contributing to each other’s results.

— Donald Hebb, *The Organization of Behavior*

4.1 Introduction

I agree with Gallistel that the results of ethological experiments provide significant constraints on theories about the neural mechanisms underlying animal cognition. By providing precise characterizations of the cognitive capacities of animals, they not only identify what is to be explained, but also provide broad constraints on how those capacities are to be explained. As Gallistel rightly emphasizes, it is unclear how to account for the results of Clayton et al.’s (2001) experiments with scrub jays, for example, without supposing that jay brains are somehow capable of flexibly integrating large amounts of information from various modalities, recording that information for considerable periods of time, in such a way that it can be retrieved when needed so as to inform the jay’s foraging strategies.
But Gallistel holds that the ‘top-down’ behavioral evidence can do more. He holds that such evidence, together with the ‘ineluctable’ constraints on the space of powerful and practical computing mechanisms that he thinks are imposed by physics and mathematics, allow us to make abductive inferences about the specific character of the neural mechanisms of animal cognition. In particular, given his view that the purportedly associative mechanisms recognized by contemporary neuroscientists cannot satisfy the constraints on powerful and practical computing mechanisms, he holds that ‘top-down’ evidence entitles us to reject virtually everything that neuroscientists believe they have learned, ‘bottom-up’, about the neural mechanisms of learning and memory: “If behavior is the last court of appeal, then there are mechanisms in the nervous system not yet dreamed of in the philosophy of neuroscientists” (Gallistel, 2006, p.70).

Here I disagree. Viewed from the perspective of the history and philosophy of science, Gallistel’s purely top-down methodology is not promising. Progress in developing mechanistic explanations of complex systems typically proceeds through the piecemeal integration of constraints on the structure and function of a mechanism that are provided by fields working at different levels of analysis (Craver, 2005). Gallistel discounts constraints provided by the entire discipline of neuroscience because he thinks that the hegemony of the associative theory of learning has forced neuroscientists to interpret their results in terms of the neural plasticity hypothesis, while preventing them from exploring alternatives. However, this across-the-board Kuhnian constructivism is generally implausible, and in the present case it’s simply not true. Several alternative mechanisms to neural plasticity have been proposed over the years, but none have withstood close empirical scrutiny.1 In particular, the molecular memory hypothesis

---

1 As Craver (2003) concludes after surveying the state of research into neural memory mechanisms in the mid-20th century, “[o]ne cannot simply assume that researchers... viewed synaptic plasticity as the only, or even the most plausible, hypothetical explanation of memory” (p.172).
that Gallistel & King (2010) tentatively endorse was explored in a series of experiments by Hydén (1973), McConnell (1962) and others throughout the 1960s and ‘70s. This research provided some tantalizing and highly publicized early results suggesting that, for example, conditioned reflexes in flatworms might be transferred through cannibalism, but these results were not borne out by more careful replications. The current status of neural plasticity as the most plausible hypothesis about the mechanism of learning and memory simply reflects normal science working as it should.

Of course, Gallistel thinks he has principled theoretical reasons for adopting the top-down methodology that he does. He thinks that computer science has revealed fundamental constraints on the physical realization of computation that can only be satisfied by mechanisms with the ‘functional architecture of a Turing machine’\(^2\) or a read-write memory, and that by lacking anything that could function as a read-write memory, neural networks are mere finite-state machines, and suffer from various crippling computational limitations as a result. But these claims have an appearance of theoretical authority they’re not entitled to, since Gallistel is simply misusing the relevant theoretical concepts. Computability theory, the proper home of automata-theoretic notions like Turing machines (TMs) and finite-state machines (FSMs), deliberately abstracts away from physical resource limitations such as time, space and energy so as to limn the boundaries of which functions can be computed in principle; it imposes few constraints on how functions are in fact computed by physical mechanisms. Turing machines and FSMs are essentially just a formalism for expressing functions in intension, and their state diagrams needn’t map directly on to the physical structure of a mechanism that implements a given function. Moreover, to the extent that it makes any sense to identify physical computing mechanisms with automata-theoretic devices

\(^2\)“[A] mechanism functionally equivalent to the tape in Turing’s abstract conception of a general purpose computing machine... is essential to computation and representation” (Gallistel, 2008, p.230).
like FSMs, all physical computing mechanisms ‘are’ FSMs in the relevant sense — even those equipped with a read-write memory. So the fact, for example, that the number of states of a FSM increases exponentially with the number of input-output mappings that it computes doesn’t entail that a mechanism that implements the FSM is faced with a combinatorial explosion of physical resources; some FSMs, such as conventional computers, are capable of avoiding such combinatorial explosions. So Gallistel’s argument that neural networks face a combinatorial explosion problem in virtue of the fact that they are FSMs is a non sequitur.

When Gallistel claims that powerful and practical computing mechanisms must have the ‘functional architecture of a Turing machine’, this is just a picturesque way of saying that they must have access to a read-write memory mechanism, and doesn’t receive any theoretical support from computability theory. Nevertheless, it’s surely true that for neural networks to explain the relevant behavioral evidence they must have some way of recording, manipulating, and retrieving information; they must have a ‘read-write memory’ in a broad functional sense. But Gallistel has a particular structural conception of how read-write memory must be realized, because he takes certain traditional ways of thinking about human-engineered computers to provide constitutive constraints on what it is for a mechanism to store and process information. But again, this presupposes an explanatory methodology that doesn’t seem at all promising; just as progress in developing mechanistic explanations of complex systems involves integrating evidential constraints from various different fields, it also involves adapting theoretical tools and concepts to new explanatory domains. (Shannon, 1956) himself perceptively expressed this point when he wrote that researchers outside of communications engineering who were ‘jumping on the bandwagon’ of information theory “should realize that the basic results of [information theory] are aimed in a very specific direction, a direction that is not necessarily relevant to such fields as psychology”, and that adapting the tools of information theory to these other fields “is
not a trivial matter of translating words to a new domain, but rather the slow tedious process of hypothesis and experimental verification” (p.3). Indeed, this slow — but by no means tedious — process is precisely what has been going on in neuroscience for more than half a century. Contrary to Gallistel’s claim that in the “contemporary literature on the neurobiology of memory, there is no discussion of [questions about neural coding]” (Gallistel & King, 2010, p. viii), and that “neurobiologists are not focused on understanding the nervous system from a computational point of view” (ibid., p.171), research into neural coding and information processing, using information-theoretic tools, has been at the forefront of mainstream neuroscience for decades.³

Strangely, in a large body of critical work that purports to show that neuroscientists doesn’t even ask, let alone provide plausible answers to, questions about information coding and processing in the nervous system, Gallistel doesn’t address this research. There’s no mention, to take just one example, of the long research tradition, directly inspired by Shannon’s (1948) insights, of attempting to understand how the structure and function of neural circuitry might be adapted to the statistical structure of the environment so as to enable the efficient coding of sensory stimuli — a tradition running from Barlow’s (1961) pioneering ‘efficient coding hypothesis’, to work on natural scene statistics (Simoncelli & Olshausen, 2001), to contemporary work on predictive coding (Huang & Rao, 2011). Why does Gallistel turn a blind eye to this research? And why is so much of his discussion of contemporary neuroscience colored by partisan rhetoric, such as his claim that “[n]eurobiologists are math agnostics” (Gallistel, 1998, p.9) who deny that the brain is a computer because it allows them to “stay in their comfort zone [and] avoid mastering [the] confusion” of ideas like indirect addressing (Gallistel & King, 2010, p.158)? I suspect the reason is that Gallistel is bewitched by a background picture of a Manichean struggle between the forces of computationalism on

---

³For a recent review of such research, see Dimitrov et al. (2011).
the one hand, and the anti-representational associationists on the other. Someone who fought in the front lines against the behaviorists during the cognitive revolution might find this picture compelling. But, in fact, it’s a caricature of contemporary cognitive science, inaccurate both in detail and broad outline. Behaviorism as a methodology and as a metaphysics has been well and truly defeated. Neuroscientists seek to explain distinctively mental capacities such as perception, memory, imagination and even consciousness by appealing to internal representational mechanisms, using conceptual tools inherited from an earlier generation of computationalists; it’s just that their tools have been honed against the experimental realities they’ve encountered.

At least, that’s what I propose to argue in the present chapter. The purpose of this chapter is to respond to Gallistel’s criticisms of the mechanisms posited by contemporary neuroscientists, and to argue that there is every reason to be optimistic that something like those mechanisms might very well explain the cognitive capacities that are distinctive of intentional agency. There’s no reason — at least, none provided by Gallistel — to suppose that neuroscience must undergo a radical Kuhnian revolution. As mentioned, I think that any of Gallistel’s criticisms are superficially compelling because they are dramatized by a background picture of a Manichean struggle waged throughout the history of psychology between the forces of computationalism and associationism. I will thus begin, in sections 4.2 and 4.3, by attempting to dispel this bewitching image. Certainly there was a sociological division between two distinct research traditions in the field of artificial intelligence in the latter half of the 20th century, driven in large part by competition for government funding, and to some extent that division percolated into cognitive psychology via the kinds of computational models that psychologists employed. With a bit of retrospective historical reconstruction we can even project this division onto much older debates about the mind. But to depict this division as a grand chasm running the length of psychology, between theoretically homogeneous research traditions, obsurses significant differences between
researchers within those purported traditions, as well as significant similarities across the traditions. Identifying some of those differences and similarities is worthwhile in its own right, to reveal the richness and complexity of the history of cognitive science, but it will also serve to highlight some important themes that I’ll pick up again in sections 4.4 and 4.5, where I engage with Gallistel’s criticisms more directly. In section 4.4 I address the theoretical foundations underlying Gallistel’s criticisms, and argue that they have an appearance of rigor and legitimacy that they don’t deserve, since they misappropriate certain theoretical concepts from computability theory. Because of this, I argue, Gallistel’s criticisms don’t constitute objections to the possibility of explaining animal cognition in terms of currently recognized neural mechanisms, but simply raise empirical questions about how such mechanisms might flexibly store and manipulate information. Then, in section 4.5, I draw from a wealth of recent work in theoretical and experimental neuroscience to provide a sketch of some ways in which those empirical questions might plausibly be answered. I conclude that, contra Gallistel, there is reason to be optimistic that neuroscience has the resources to explain cognition after all.

4.2 Dissent in the Computationalist Ranks?

Let’s begin our investigation of the Manichean struggle between computationalism and associationism by evaluating the ideological purity of the computationalist rank and file. I’ll focus on the most prominent of the researchers identified by Gallistel as belonging to ‘mainline’ computationalist cognitive science, as well as the researchers Gallistel identifies as having provided the conceptual and theoretical foundations of the computationalist tradition.

---

4See (Gallistel & King, 2010, p.101).
Let’s start by considering the views of the figure who many take to be the progenitor of computationalism: Alan Turing. The view that Turing was himself a computationalist has become part of the folklore of cognitive science, taken for granted with little critical scrutiny. The view is often expressed by Jerry Fodor, who, for example, attributes to Turing the view that “cognitive mental processes are operations defined on syntactically structured mental representations that are much like sentences” (2000, p.4). However, there’s little to no evidence that Turing actually believed this. Those who seek to provide such evidence almost invariably cite Turing’s (1950) landmark paper ‘Computing Machinery and Intelligence’. In that paper, Turing famously argued that there are no good reasons why an appropriately programmed digital computer couldn’t mimic human cognitive capacities. However, he doesn’t explicitly address, let alone assert, the view that human cognition is explained by mechanisms with the functional architecture of a digital computer. These views are clearly quite different; the assertion that digital computers might be capable of thinking no more implies that biological brains have the architecture of digital computers than the assertion that jet planes might be capable of flying implies that birds have turbine engines. The only statements in Turing’s (1950) paper that bear on the question of whether he was a computationalist in Gallistel’s sense in fact suggest that he wasn’t. Turing writes that “[t]he nervous system is certainly not a discrete-state machine” (p.451), but that this needn’t in principle preclude a discrete-state, digital computer from simulating human cognition. This at least gives the impression that Turing thought that the neural mechanisms of human cognition operate according to quite different principles from the

5Indeed, as Piccinini (2008a) and Wells (2004) point out, the view that sententially structured mental representations are stored within a read-write memory that is functionally equivalent to the tape of a Turing machine is anachronistic and hard to square with the theoretical motivations for Turing’s (1936) machine. Recall from Chapter 2.2.3 that, for Turing, the tape of his machines is analogous to the paper upon which a human computor writes when carrying out an effective procedure; that is, it’s distinct from, and external to, the machine itself.
mechanisms of digital computation. This impression is reinforced by comments that Turing makes in a letter to the cyberneticist W. Ross Ashby, likely written around 1946, which again suggest that Turing thought that a digital computer might usefully model neural processes, but in a way that needn’t directly reflect how such processes actually work: “although the brain may in fact operate by changing its neuron circuits by the growth of axons and dendrites, we could nevertheless make a model... in which the actual construction of the [digital computer] did not alter, but only the remembered data” (Copeland, 2004, p.375).

What’s more revealing than these incidental comments is a research report written by Turing in 1948 that remained unpublished until 14 years after Turing’s death, and languished in obscurity for many decades thereafter. Only recently has it become widely appreciated that Turing’s (1948 [2004]) paper presaged many of the key ideas of subsequent artificial neural network theorizing (Copeland & Proudfoot, 1996; Teuscher, 2004). In that paper, Turing discussed various different kinds of ‘unorganized machines’: networks of randomly and recurrently connected Boolean nodes. Turing’s unorganized machines were similar in many respects to the Boolean networks of McCulloch & Pitts (1943), however they differed in at least one crucial respect: as I mentioned in Chapter 2.4, McCulloch & Pitts ignored the question of how their networks might change over time and thereby learn, but for Turing (1948 [2004]), this question was paramount. One important class of unorganized machine incorporated a means by which an external agent might modify the connections and thereby organize the machine to compute certain functions. These machines are connectionist in the sense that, as Copeland & Proudfoot (1996) puts it, they are “functionally equivalent to [a network] in which the stored information takes the form of new connections” (p.367). However, Turing was dissatisfied with his specific conception of how an unorganized machine might become organized, and envisaged the development of ‘teaching policies’ that might be incorporated into the machine, thereby enabling it to self-
organize; Turing thus presciently anticipated the subsequent development of learning algorithms like backpropagation.

It is important to note that unorganized machines weren’t of purely academic interest to Turing; he regarded them as highly simplified and idealized models of the nervous system (Turing, 1948 [2004], pp.418), and his interest in them was motivated by the conviction that “the cortex of the infant is an unorganised machine, which can be organised by suitable... training” (ibid., pp.423-4, my emphasis). As he later put it, “the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets” (Turing, 1950, p.456). So there’s a strong case to be made that the purported progenitor of computationalism was not in fact a computationalist, but was both an empiricist and an associationist. Moreover, although Turing (1936) certainly laid the conceptual and theoretical foundations for our contemporary ways of thinking about computation, Piccinini (2004) makes a convincing case that the first theorists to build a computational theory of mind and brain upon those foundations were in fact McCulloch & Pitts (1943), the figures who are widely regarded as the forefathers of contemporary connectionism. The dichotomy between computationalists and connectionists is beginning to look a little blurry.

Perhaps the battle lines might become clearer if we look at the work of Claude Shannon. But Shannon was an engineer’s engineer, and, unlike Turing, had little interest in exploring the implications of his specialized theoretical work in communications engineering for broader issues about the nature of mind or brain. In an interview he gave in 1984, he explained that “he was attracted more to problems than to applications” (James, 2009). Indeed, as I mentioned earlier in this chapter, Shannon was wary of directly translating his theory of information to address problems outside of the specific domain for which it was developed, writing, for example, that “[i]t is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field” (Shannon, 1953b, p.105).
However, in an invited review of the state of play in computer science research in the middle of the 20th century, Shannon (1953a) does speak to some of the broader implications of that research. Like Turing (1950), he expresses optimism about the use of digital computers for simulating game playing, language translation, and other “complicated functions associated with the human brain” (Shannon, 1953a, p.1235). But, again like Turing, he doesn’t endorse the computationalist view that the human brain operates in much the same way as a digital computer. On the contrary, he emphasizes various ways in which biological brains and engineered computers differ, foreshadowing many of the complaints that a later generation of connectionists would make against ‘classical’ computational systems, such as their alleged fault intolerance and lack of adaptability. He ends the article with a series of ‘challenges’ for future computer scientists, concluding with what he presumably took to be the most important: “How can a computer memory be organized to learn and remember by association, in a manner similar to the human brain?” (Shannon, 1953a, p.1241). This all suggests that if we’re going to peg Shannon into any holes, it ought to be the one Gallistel calls ‘associationism’. But I don’t mean to hammer too hard on this categorization, either in Shannon’s case or Turing’s. My point, really, is not to object to the way Gallistel implicitly categorizes these theorists, but to object to Gallistel’s categories. To fit Shannon or Turing into a dichotomy between ‘computationalism/representationalism’ on the one hand and ‘connectionism/associationism’ on the other is procrustean and anachronistic.

Perhaps the dichotomy might come into focus if we look at a later era of theorists, who work squarely within the field we now know as cognitive science. After all, although Gallistel holds that Turing and Shannon provided the conceptual foundations for the ‘computational/representational’ tradition in psychology, he doesn’t explicitly identify them as proponents of that tradition, so perhaps we shouldn’t expect them to toe the party line. Let’s turn instead, then, to two of the most prominent researchers
who Gallistel *does* identify as belonging to the ‘computational/representational’ tradition: David Marr and Noam Chomsky. Surely *they’re* representationalist-cum-computationalists if anyone is.

Let’s start with Marr. Early in his career, Marr did pioneering work in theoretical neuroscience. He published three seminal papers that brought a unified theoretical perspective to bear on a tangle of neuroanatomical and neurophysiological detail, to explain the function of three major brain areas: the cerebellum (1969), neocortex (1970), and hippocampus (1971). This ‘unified perspective’ was essentially that “the brain’s central function is statistical pattern recognition and association” (Edelman & Vaina, 2001, p.9256). For example, Marr (1969) argues that the cerebellum mediates the acquisition of motor skills via plastic synapses on Purkinje cells, which operate according to broadly Hebbian principles.6 The core idea of his (1970) theory of the neocortex is that the cortex forms ‘classificatory units’, or concepts, on the basis of statistical redundancies in experience. In his (1971), he argues that hippocampus is what we’d now call an attractor network that functions as a content-addressable memory, again in ways that are mediated by Hebbian synaptic plasticity mechanisms.

Now, it is true that Marr’s views changed significantly later in his career. In the publication for which he’s most famous, his landmark book *Vision* (1982), he expresses reservations about the direction of his earlier work, and similar work in neuroscience around the same time, suggesting that it was more descriptive than explanatory. He emphasizes that if we’re to fully understand cognitive capacities we must look at them from a *computational* point of view — his famous ‘computational level’. But it would be a profound misinterpretation of Marr (1982) to suppose that he was objecting to the kinds of *mechanisms* that he had posited in his earlier work and was now advocating

---

6 “If a parallel fibre is active at about the same time as the climbing fibre to a Purkinje (output) cell which the parallel fibres makes synaptic contact, then the efficacy of that synapse is increased towards some fixed maximum value” (Marr, 1969, pp. 455456).
the view that the neural mechanisms of cognition were closely analogous to those of a digital computer. His emphasis on computation was *methodological* not *mechanistic*; his point was that to fully understand a cognitive process, we must understand the nature of the information-processing task that it is solving — i.e. the computation it is performing, where ‘computation’ is here understood not as some particular type of physical mechanism, but rather as computability theorists understand the notion: namely, as a function specified in intension. The computation identifies *what* needs to be explained — the cognitive capacity, characterized as a computable function — whereas the neural mechanisms explicate *how* that capacity is to be explained. It is only by integrating these two perspectives, according to Marr, that we arrive at a complete explanation of the capacity that is our explanandum. Unimpeachable methodological advice to be sure, but nothing here or anywhere else in Marr’s (1982) book suggests that he thought that the neural mechanisms of cognition operate in a manner that is closely analogous to a digital computer; indeed, for all his methodological concerns about his earlier work, he still thought, for example, that his theory of Hebbian-based motor learning in the cerebellum might very well be correct (p.14).

Finally, let’s turn to Noam Chomsky, whose famous critique of Skinner’s theory of language provided impetus to the cognitive revolution (Chomsky, 1959), and who went on to write books with titles like ‘Rules and Representations’ (Chomsky, 1980). Surely *he* is opposed to the view, allegedly endorsed by behaviorists and connectionists alike, “that the brain does not really represent the external world in any interesting sense of the term *representation*” (Gallistel, 1998, p.11)? And surely *he* would agree that representations in this interesting sense are comprised of symbols, where “[t]he

---

7Marr was, in effect, advocating the kind of methodology I sketched at the beginning of the chapter, of integrating ‘top-down’ and ‘bottom-up’ constraints to provide a complete mechanistic explanation of a given phenomenon — the kind of methodology that Gallistel seems to eschew when, for example, he emphasizes that behavior is the ‘court of last appeal’ when it comes to explaining cognition (Gallistel, 2006).
**essence** of a symbol is that it denotes or refers to something other than itself” (*ibid.*, p.8, my emphasis)? Well, the truth isn’t so straightforward. Representations certainly play an important explanatory role in Chomsky’s work in theoretical linguistics, but in his recent writings Chomsky makes it clear that the explanatory role they play has nothing to do with any sort of semantic relation like reference that they bear to entities in the external world. According to Chomsky, the representations that he and other theoretical linguists posits shouldn’t be understood relationally, as representations of anything; they’re simply internal states that are manipulated in the course of linguistic comprehension and production. As he puts it, “The internalist study of language... speaks of ‘representations’ of various kinds, including phonetic and semantic representations at the ‘interface’ with other systems. But... we need not ponder what is represented... . Accessed by performance systems, the internal representations of language enter into interpretation, thought, and action, but there is no reason to seek any other relation to the world” (Chomsky, 2000, p.160). Indeed, Chomsky holds that this point about the explanatory role of representations applies generally to the representations posited by theories in cognitive science. For example, discussing theories of visual perception, he writes:

There is no meaningful question about the ‘content’ of the internal representations of a person seeing a cube under the conditions of the experiments... or about the content of a frog’s ‘representation of’ a fly or of a moving dot in the standard experimental studies of frog vision. No notion like ‘content,’ or ‘representation of’ figures within the theory, so there are no answers to be given as to their nature. The same is true when Marr writes that he is studying vision as “a mapping from one representation to another...” (Marr 1982: 31) — where ‘representation’ is not to be understood relationally, as ‘representation of’.

(Chomsky, 2000, p.159)

So Chomsky doesn’t seem to be a representationalist after all; he explicitly rejects the view that representations, in the ‘interesting sense of the term’ that Gallistel has
in mind, play an explanatory role in cognitive science. Moreover, although the notion of computation features prominently in Chomsky’s work in theoretical linguistics, the notion he employs is, like Marr’s, the notion employed by computability theorists: that of a function specified in intension. Chomsky has repeatedly emphasized over the course of his career that grammars, qua computational entities, specify what a speaker computes, not how they compute it. So Chomsky isn’t obviously a computationalist in Gallistel’s mechanistic sense either. Indeed, none of the researchers I’ve just surveyed — purportedly paragons of the ‘computational/representational’ tradition — seem to be proponents of the specific kind of representationalist-cum-computationalist view that Gallistel endorses.

Much more could be said about these issues, and where various theorists’ allegiances lie, but I hope to have shown at least that it’s problematic to assume that there is a theoretically homogeneous ‘computational/representational’ research tradition in psychology, unified by a specific and substantive set of theoretical convictions that is inconsistent with the way contemporary neuroscientists think about the neurocomputational mechanisms of learning and memory. Insofar as there is a distinction between a computational/representational tradition and an allegedly anti-representational, associative tradition in psychology, it seems to be not so much a theoretical distinction as a political one, useful for drumming up support when funding is tight, but not so much for capturing the substantive empirical commitments of psychological theories. I’ll attempt to substantiate this claim further in the following section.

---

8See Egan (2010) for a detailed discussion of these points.
9See Matthews (1991) for a detailed discussion of this point.
4.3 Manufactured Consent Amongst the ‘Associationists’

The computationalists appear to be in disarray. Let’s now look on the other side of the trenches. This time I won’t focus on specific authors so much as general explanatory trends. Is there really such a thing as an ‘anti-representational tradition’, unified by a specific set of theoretical convictions, running from the British empiricism of the 17th century, through the associationist psychology of the 19th century, to the behaviorism of the 20th century, finding its apotheosis in contemporary neuroscience? Is it true, to focus on a few of Gallistel’s specific claims, that contemporary neuroscience “is committed to the thesis that the brain has the functional architecture of a finite-state automaton” (Gallistel & King, 2010, p.176), that “[h]istorically, much of the inspiration behind this pervasive thesis about learning and memory has been empiricism, the belief that the brain’s structure is largely determined by its experience” (ibid., p.177), and that “extreme empiricists tend to be anti-representational” (Gallistel, 2001, p.9693)?

Putting aside for now the questions of whether neuroscience is committed to the thesis that the brain has the functional architecture of a finite-state automaton, or whether indeed there’s any sense to be made of that thesis — questions I’ll pick up again in section 4.4 — it seems pretty clear that insofar as there is any sense to be made of that thesis, McCulloch & Pitts (1943) were the first to endorse it: they were the first to explain the operation of the brain in computational terms, using a formalism that was subsequently generalized by Kleene (1956) into what we now know as the finite-state machine model (Piccinini, 2004). However, the thesis that McCulloch & Pitts developed wasn’t directly ‘about learning and memory’; although they posited

\[^{10}\text{For expressions of the view that there is in fact such a tradition, see Gallistel (1998, p.11), Gallistel (2001, p.9693), Gallistel (2008, p.227), Gallistel & King (2010, pp.xiii-xv, 56, 101, 194-6, 242, 252).}\]
networks ‘with circles’, they didn’t explicitly characterize recurrency as a mechanism of memory,\textsuperscript{11} and they pointedly denied that their nets were capable of changing over time, and hence of explaining learning. Thus there doesn’t seem to be any clear sense in which their thesis was ‘inspired by empiricism’ either. Indeed, it’s noteworthy — if one follows Gallistel in assuming that there’s some sort of tension between empiricism and conventional notions of computation — that McCulloch reputedly thought of the project that he and Pitts were engaged in as “treating the brain as a Turing machine”.\textsuperscript{12}

However, when Gallistel writes that neuroscientists are committed to the thesis that ‘the brain has the functional architecture of a finite-state automaton’, the thesis he has in mind isn’t, I think, the thesis that arguably most straightforwardly and literally fits that description, the thesis that originated with McCulloch & Pitts (1943). Rather, I think, it’s the thesis that I dubbed ‘associative connectionism’ in Chapter 2.4, according to which cognitive processes are explained in terms of the operation of networks of nodes joined by connections that are modified by experience-dependent plasticity.\textsuperscript{13} Is this thesis committed to, or at least inspired by, empiricism, understood here as the view that “the brain’s structure is largely determined by its experience”, which in extreme forms “tends to be anti-representational”?

It is true that one of the researchers who contributed most to the theoretical development of associative connectionism, Rosenblatt (1958), explicitly locates himself and other ‘connectionists’ such as Hebb (1949) within the tradition of British empiricism, and writes that on their view, the neural networks involved in cognitive processes are,

\textsuperscript{11} Though it is worth noting that McCulloch (1949) did later suggest something along these lines.

\textsuperscript{12} This statement is attributed to McCulloch by von Neumann (1951, pp.32-3). Recall from Chapter 2.4 that, contrary to popular mythology, McCulloch & Pitts (1943) did not prove that their networks are computationally equivalent to Turing machines. They assert, without proof, that their networks would be equivalent to Turing machines \textit{if equipped with a tape}. In any case, the present issue pertains not to the \textit{content} of the view developed by McCulloch & Pitts (1943), but to the motivations behind it.

\textsuperscript{13} As I mentioned above, I’ll evaluate the relationship between this thesis and the thesis that the brain has the functional architecture of a finite-state automaton later, in section 4.4.
at birth, “largely random, subject to a minimum number of genetic constraints”, and are shaped over time through exposure “to a large sample of stimuli” (p.388). Moreover, he writes that on the connectionist view, information is not stored as “coded representations” with “some sort of one-to-one mapping between the sensory stimulus and the stored pattern” (p.386). These certainly seem like clear expressions of the empiricist and anti-representationalist views that Gallistel opposes. But when evaluating the significance of Rosenblatt’s comments, we should keep in mind that it would be hasty to draw generalizations about the theoretical commitments of a given research tradition on the basis of isolated comments from a single representative of that tradition. Turing’s (1948 [2004]) assertion that “the cortex of the infant is an unorganised machine” (p.423) shouldn’t, on its own, license us to conclude that classical computationalism is committed to empiricism. To determine the theoretical commitments of a given research tradition, we should look closely at widespread assumptions that are crucial to the explanatory function of the theories and models that are representative of that tradition.

Keeping that in mind, it’s noteworthy that the model that Gallistel holds up as a paragon of the allegedly empiricist, anti-representational tradition of associative connectionism that he takes umbrage with — namely, the Samsonovich & McNaughton (1997) model of path integration in the rodent hippocampus — is, by Gallistel’s own lights, neither empiricist nor anti-representational. This comes out explicitly in Gallistel & King (2010), albeit in an almost accusatory manner, as if it is somehow a failing of the model that it doesn’t exhibit the flaws it presumptively ought to exemplify. For example, Gallistel & King write that in the Samsonovich-McNaughton model, “the trigonometry, the multiplication, and the addition are mirrored in the structure of the neural circuit that carries out the combinatorial operation” (p.251), and that this structure appears to be innately pre-specified — which one might take to be an objection, given the authors’ extensive discussion of the ‘pre-specification problem’, until
the authors point out that they “mention this not as any sort of criticism, but as an indication of what we believe to be the ineluctable nativism inherent in any representational machinery” (p.252). Similarly, although the authors initially assert that the Samsonovich-McNaughton model exemplifies an architecture that “does not have a symbolic memory” (p.242), shortly thereafter they write that the reverberating circuits in the model do function as a symbolic memory (p.245), but complain that when connectionist theorists use reverberating circuits in their models, “they are bringing [a symbolic memory] in unannounced through the back door” (p.246).

So although Gallistel & King write that they discuss the Samsonovich-McNaughton model because it is “representative of a large class of models” (p.265) — presumably models within the tradition of associative connectionism, which is the critical focus of their book — they apparently don’t hold that the model is representative of that tradition in respect of being empiricist and anti-representational. Nevertheless, it seems clear that they do hold that the tradition of associative connectionism is, aside from this apparently recherché exception, generally committed to or at least deeply influenced by empiricism and anti-representationalism. To evaluate this view, let’s focus on each of these purported commitments in turn, focussing first on empiricism, understood here as the view that the structure of the mechanisms involved in cognitive processes is largely determined by experience.

4.3.1 Empiricism

The characterization of associative connectionism that I gave earlier, as essentially involving explanatory appeal to experience-dependent changes in the weights of the

---

14If this isn’t yet clear, consider these quotes from Gallistel & King (2010): “the neural network tradition is by and large an empiricist one” (p.252), and “empiricists [feel] uncomfortable with the assumption that the brain symbolically represents aspects of the world” (p.253). See also the numerous citations listed in note 10.
connections between nodes within a network, might suggest empiricism in the sense at issue, but it patently doesn’t entail it. Even if the experience-dependent plasticity of connection weights is an explanatorily important feature of connectionist models, it doesn’t follow that all of the structural elements of a connectionist network are determined by experience; perhaps the large-scale architecture of the network, rather than specific connections, is highly genetically constrained, or perhaps experience plays a necessary but non-specific role in the innately-guided maturation of a network. The authors of the ‘bible’ of neo-connectionism, Rumelhart et al. (1986b), express this general point in a discussion of where connectionist models fall on the spectrum between nativism and empiricism when they write that although some connectionist networks might contain structural elements that are largely genetically determined, and hence fall on the ‘rabidly nativist’ end of the spectrum, others might be much more environmentally labile, and hence fall on the ‘rabidly empiricist’ end. Thus, they argue, connectionist theorizing in an of itself is “quite agnostic about issues of nativism versus empiricism” (p.139). Indeed, even some of the harshest critics of neo-connectionism agree with Rumelhart et al. that connectionism as such is “neutral on the Empiricism/Nativism issue” (Fodor & Pylyshyn, 1988, p.60, emphasis in original).

The kind of view that Rumelhart et al. express most sympathy for is one that occupies a middle ground between radical nativism and radical empiricism, according to which cognitive mechanisms are experientially malleable, but develop in ways that are constrained by genetic factors. This general kind of view, which attempts to marry the most plausible aspects of nativism and empiricism, is also endorsed by Gallistel, who, as I explained in Chapter 3.4, holds that the functional distinction between computational mechanisms and read-write memory embodies a plausible division of labor between cognitive mechanisms that are innately determined and those that are shaped by experience. Expressed at this level of generality, the kind of view at issue is of course platitudinous; Rumelhart et al. describe it as ‘painfully obvious’. It is an instance of
what has been called the interactionist consensus in the nature versus nurture debate. Everyone agrees that phenotypic traits develop through complex interactions between both genetic and environmental factors.

A seminal exploration of how this kind of uncontroversial interactionist view might be developed into substantive hypotheses about cognitive development within a connectionist framework was undertaken by Elman et al. (1996), who distinguished three ways in which innate factors might interact with experience to constrain the development of a connectionist network. The first involves innate representations. We’ve seen that explicit representations within a connectionist network are typically understood as occurrent patterns of activity across the network, and that the patterns that emerge in response to a given input are shaped by the connection weights in the network. While individual connection weights, understood as synaptic efficacies, are unlikely in general to be genetically specified — there are roughly $1.5 \times 10^{14}$ synapses in the human brain, yet only around $2 \times 10^4$ protein-coding genes in the human genome\textsuperscript{15} — they might nevertheless be genetically constrained so as to fall within a certain range, or to be malleable only at certain developmental stages. This relates to a second class of developmental constraints, concerning timing. Elman et al. provide several examples of how the timing of certain developmental events, perhaps due to genetically-controlled maturational processes, can significantly influence how a connectionist network learns.

A third class of constraints concerns the architecture of a network: the pattern of connectivity within the network, the number and organization of layers, the number of nodes within each layer, the activation function used, and so forth.

This third class of constraints is especially noteworthy in the present context, since a misunderstanding of the relationship between these kinds of architectural properties and learning rules in connectionist systems seems, at least in part, to motivate Gal-

\textsuperscript{15}For these estimates, see Pakkenberg & Gundersen (1997) and the IHGSC (2004), respectively.
listel’s insistence that the connectionist tradition is excessively empiricist. Gallistel repeatedly characterizes learning as it is understood within the connectionist tradition as ‘rewiring’; in particular, he and King frequently describe supervised and unsupervised learning rules in connectionist systems as ‘rewiring rules’, such as when they write that unsupervised learning involves “much the same rewiring rules that Locke suggested” (p.253). At the very least, this is a misleading choice of terminology, for the term ‘wiring’ brings to mind the connectivity pattern within a network, such as which nodes are connected to which, whether there are any recurrent connections or layers, and so forth, yet learning rules typically have no influence whatsoever on connectivity patterns; they change the weights of connections, but not the architecture of the connections themselves.\footnote{An important caveat is that in some abstract and highly idealized neural network models, structural connections are not explicitly modeled, but are implicit in the weight matrices of the nodes. An example of such a model is the well-known self-organizing map (Kohonen, 1982). In such models, the ‘weight of a connection’ from one node to another might be zero, which could be interpreted as the absence of a connection between those nodes. Nevertheless, learning within such a model might modify the weight of that ‘connection’, and in that way change the connectivity pattern in the network. However, this doesn’t undermine the point in the text, which is that insofar as neural network models are given a biological interpretation, there is an important distinction to be made between changes in the structural connectivity between neurons, and changes in the weights (i.e. synaptic conductances) of those connections, a distinction that is conflated by talk of ‘rewiring’.}

That the ‘rewiring’ terminology manifests a substantive confusion rather than a merely misleading choice of words is evidenced when Gallistel & King ask whether the learning rules commonly employed within the connectionist tradition could produce a network with ‘representational structure’ that reflects a specific computational problem, as exemplified by the Samsonovich-McNaughton model: “The question is whether, if we start with a randomly connected net, and assume one of the currently popular unsupervised learning algorithms or the very widely used back-propagation algorithm, and we stipulate some, one hopes, half-way plausible experiences, will we end up with a net structured like the one in Figure [3.1]?” (p.253). The answer is of course no,
but this doesn’t reveal the Samsonovich-McNaughton model to be an exception that somehow proves the rule that connectionist networks are in general excessively empiricist, since the same point holds of any connectionist network; learning rules such as Hebbian plasticity and backpropagation just aren’t in the business of changing the connectivity patterns of networks. They modify weights, not connections.

Indeed, in many connectionist models, the connectivity pattern of the network is specified from the outset, with little consideration of how the network came to be wired up that way. Perhaps this is in part what motivates Gallistel & King when they claim that connectionist networks are in fact ‘radically nativist’ (p.177); remember that on their view, one of the central problems with the connectionist mechanisms posited by contemporary neuroscientists is that they exemplify a finite-state architecture, which suffers irremediably from the problem of pre-specification. On the face of it, this claim is hard to square with the authors’ repeated suggestions that connectionist theorizing is radically empiricist. One might attempt to resolve this apparent tension charitably, in a way that is consistent with the authors’ intended conclusions, by interpreting the authors as arguing that although connectionists think of themselves as belonging to the empiricist tradition, they are in fact mistaken, and are unwittingly committed to nativism. This interpretation is consistent with some things that Gallistel & King say, but not with another of their most central objections to connectionism; recall that Gallistel & King argue that a central problem with the connectionist orthodoxy in

---

17For example, Samsonovich & McNaughton (1997) write that “the prewiring of the multichart attractor map... is assumed without consideration of how this occurs” (p. 5904), and Gallistel & King (2010) take this to provide evidence of “the ineluctable nativism inherent in any representational machinery” (p.252). However, aside from the issues that I’ll go on to discuss in the text, it would be a mistake to suppose that when connectionist researchers leave it unexplicated how the connectivity patterns in their networks are established, they’re making a positive commitment that those connectivity patterns are ‘innate’ in the strong sense of being pre-specified. Any model has to leave some questions unanswered, and although questions about the ontogeny of connectivity patterns are challenging, the mere fact that they are sometimes bracketed doesn’t imply that they are presumed only to be answerable in terms of genetic ‘pre-specification’ — whatever exactly that might mean.
contemporary neuroscience is that it characterizes learning as domain-general process by which statistical regularities in the environment rewire the brain. So Gallistel & King really do seem to want to claim that the connectionist tradition is in fact both radically nativist and radically empiricist. Now, there is a way of resolving this apparent tension, but it is not one that is consistent with the kind of conclusion the authors wish to endorse: one might hold that the sense in which connectionist networks are both ‘nativist’ and ‘empiricist’ is precisely the same sense in which the cognitive mechanisms envisaged by Gallistel — and indeed everyone else who endorses the interactionist consensus, viz. everyone — are both nativist and empiricist, namely they comprise some structural elements that are genetically constrained, and some structural elements that are environmentally malleable.

Indeed, the aspects of a connectionist network that are most plausibly genetically constrained are arguably the macro-scale architectural properties such as connectivity patterns, which plausibly embody implicit information about ‘universal’, phylogenetically constant features of an animal’s environment, whereas the aspects that are most plausibly environmentally malleable are the connection weights that determine specific representational states, which might explicitly encode ‘contingent’, ontogenetically variable features of an animal’s environment. The picture that emerges, which has been endorsed in broad outline by central proponents of connectionism including Bechtel & Abrahamsen (1991), Elman et al. (1996) and Rumelhart et al. (1986b), looks very similar indeed to the picture developed by Gallistel, according to which there is a developmental division of labour between innate, special-purpose computational processors and a malleable, general-purpose read-write memory. However, it differs in that it needn’t presuppose a rigid dichotomy between specific components of a computing mechanism that are ‘innately fixed’ and components that are ‘environmentally malleable’.

This point comes out most clearly when we turn our attention from the highly ide-
alized models of 1980s neo-connectionism to the messy details of contemporary neuroscience. There we find that the distinction between connections and connection weights fractionates into a heterogeneous variety of mechanistic components and processes that vary across cell types, that operate at different developmental stages, and that interact in complex ways depending to varying degrees on both genetic and environmental factors. For example, to return to the earlier point that talk of ‘rewiring’ obscures the distinction between connections and connection weights, that distinction itself is becoming increasingly difficult to map cleanly on to specific neurobiological mechanisms.

Connections are generally thought to model synapses, which typically occur at structurally complex junctions between axons and dendritic spines. Throughout most of the twentieth century, neuroscientists assumed that these synaptic structures, and the neural circuitry they mediate, stabilized during early development and remained largely unchanged throughout adulthood. Talk of neural plasticity typically pertained to changes in the efficacy of synapses, not to large-scale structural changes in synapses themselves. However, it is now widely recognized that new neurons can grow in the adult brain, and that the dendritic spines of neurons, old and new, are continually extending and retracting, forming new synapses and eliminating old ones. Neuroscientists now distinguish between structural as well as synaptic plasticity, though the two phenomena are intimately intertwined. Structural changes in dendritic spines can be driven by experience — via, for example, mechanisms of LTP — and there is considerable evidence that such changes are involved in long-term memory; however, there is also evidence that structural plasticity in spines is often driven not primarily by

---

18This, incidentally, is one reason among many why it is puzzling that Gallistel holds that neuroscientists have traditionally assumed that the brain is an organ of unconstrained plasticity, capable of being ‘rewired’ by experience. The truth is that neuroscientists have until quite recently generally assumed that neural circuitry in the adult brain is static. As Leslie & Nedivi (2011) recently put it, “The adult brain had long been considered hardwired, incapable of the structural remodeling seen during development... In the past few decades, these long-held views have gradually shifted to accommodate a tremendous flexibility of neuronal form and function in the adult brain” (p.224).
experience but by various endogenous factors instead (Kasai et al., 2010; Holtmaat & Svoboda, 2009). In short, contemporary connectionists have the explanatory resources to account for traits that are environmentally malleable or innately fixed.

However, the foregoing considerations suggest a deeper objection to Gallistel, concerning the very cogency of the distinction between traits that are environmentally malleable and those that are innately fixed — and, by extension, the theoretical importance of the distinction between empiricism and nativism. So far I have employed these distinctions without critical scrutiny, to engage with Gallistel on his own terms. However, there is a long tradition of skepticism about the cogency and explanatory utility of those distinctions, which poses a challenge to articulate the concepts presupposed by those distinctions that I think we ought to take very seriously. We can see the outline of the challenge when we look again at Gallistel’s definition of empiricism: “the belief that the brain’s structure is largely determined by its experience” (Gallistel & King, 2010, p.177, my emphasis). If we interpret ‘determined’ very strongly so as to exclude inherited influences, then I take it that few researchers, if any, have ever taken this view seriously. It’s unclear how experience alone could determine anything, without operating on some physical substrate that shapes and constrains the effects of experience. Even the arch-empiricist of the 20th century, B.F. Skinner, held that for an organism to acquire a conditioned response, it must have an inherited “susceptibility to reinforcement by certain kinds of consequences” (1981, p.501). Conversely, I take it that few researchers, if any, have ever taken seriously the thesis that the brain’s structure is determined exclusively by inherited influences. Even the most ardent nativists allow that experience plays a necessary role in, for example, triggering certain developmental processes. The challenge, then, is to articulate a notion of experiential or inherited ‘influence’ that is weaker than a very strong sense of ‘determination’, but which preserves a substantive and theoretically fruitful distinction between acquired and innate traits. In the absence of such a distinction, the claim that contemporary
connectionist neuroscience is ‘empiricist’ or ‘nativist’ would seem to be devoid of empirical content.

However, the closer we look at the details of contemporary biology, such as the mechanisms of neural plasticity mentioned earlier, the more difficult it is to distinguish between innate versus acquired traits. Reflecting on the difficulties of articulating a notion of innateness grounded in biology, Griffiths (2002) writes that “In molecular developmental biology innateness seems as antiquated a theoretical construct as instinct and equally peripheral to any actual account of gene regulation or morphogenesis. In behavioral ecology, some authors regard the innateness concept as irretrievably confused” (p.70). Biological notions that might seem to provide fruitful ways of explicating the notion of innateness are patently unsuccessful in this regard. For example, it has long been recognized as a fallacy to identify the quantitative measure of heritability with the intuitive notion of innateness (Lewontin, 1974); heritability is a statistical measure of how much of the phenotypic variability in a population can be attributed to genetic differences, and says nothing about the extent to which genetic factors play a causal role in the development of a trait in an individual. If all the individuals in a population were branded with a lightning-bolt-shaped scar on their foreheads, that trait would have high heritability, but we should presumably ought not count it as innate. Attempts to explicate the notion of innateness in terms of the causal role that genes play in the development of a trait have also proven to be problematic. Genes play a crucial role in the development of all traits — even the experience-dependent induction of LTP involves protein synthesis and hence the expression of genes — and proposals to identify a special role that genes might play in the development of innate traits by appealing to the idea that information about such traits is encoded ‘in’ the genes have arguably failed to pick out a role that is unique to genes as opposed to environmental factors (Griffiths & Gray, 1997).

The problem here is not that there is no notion in the vicinity of the traditional
notion of innateness that is rigorous enough to play a robust role in scientific explanations; it’s that there are too many. Mameli & Bateson (2006) identify no less than twenty-six distinct definitions of innateness, and argue that at least eight of them pick out genuine properties that might play a robust explanatory role. However, they argue that in the absence of strong empirical reasons to think that these properties generally cluster together — reasons that, in their view, have yet to be provided — we should not speak of innateness simpliciter, since this risks running together various properties that might well be distinct, and thereby introduces the possibility of fallacies of equivocation. Simply because a trait is ‘innate’ in the sense of being a Darwinian adaptation, it might not follow that it is ‘innate’ in the sense of being developmentally canalized or insensitive to environmental perturbation. While it is beyond the scope of the present work to develop and defend this line of argument in detail, I do think that Mameli & Bateson (2006), along with many other researchers in the tradition of Lehrman (1953) who have expressed skepticism about the intuitive notion of innateness based on careful examination of the details of biology, raise a serious challenge for anyone who would place theoretical weight on the distinction between nativism and empiricism, to show how the distinction can bear that weight.

In sketching this skeptical challenge I have focussed on the notion of innateness, however I should not be read as endorsing anti-nativism, if that is to be understood as empiricism, or the contrary of nativism. Rather, the skeptical challenge threatens to undermine both nativism and empiricism as theoretically substantive and fruitful ways of characterizing cognitive and neural development. This is important to highlight, since theorists often pay lip service to the interactionist consensus about the nature / nurture debate, but then fail to take the lessons of interactionism seriously by endorsing one side or other of the dichotomy between nativism and empiricism. For example,

---

19 See, for example, Bateson (1991), Griffiths (2002), and Lewontin (1974).
Samet & Zaitchik (2012) seem to endorse interactionism when they write that “in the Nativism-Empiricism debate we are often dealing with ideology, not theory”, but then, strangely, spend the rest of their article arguing that “Nativism is experiencing a resurgence” in cognitive science, characterizing innateness using the unhelpful metaphor of ‘the factory settings for babies’, and dismissing any further attempts to clarify the notion as mere ‘philosophical bookkeeping’. One is thus left to infer that the content of their paper is ideology, not theory. The skeptical challenge is a call to do better: to imbue ‘nativism’ and ‘empiricism’ with theoretical content.

The need to meet this challenge is especially pressing in the present context. Recall that on Gallistel’s view of cognition, the computational architecture of the mechanisms that extract explicit information from experience are innately pre-specified, but the symbolic vehicles of explicit information are not. Gallistel takes this to be an expression of the ‘mainline’ computationalist tradition championed by self-described rationalists and nativists such as Noam Chomsky and Gary Marcus. However, it’s noteworthy that Gallistel’s view would in fact seem to fall within the purview of Chomsky’s (1965) conception of empiricism: “empiricist speculation has characteristically assumed that only the procedures and mechanisms for the acquisition of knowledge constitute an innate property of the mind” (p.51). Similarly, Gallistel’s view would seem to be an instance of a view that Marcus (2001) describes as “a sort of stripped-down nativism”, according to which “‘architectural’ aspects of brain organization are innate, but ‘representations’ are not” (p.369). Marcus attributes this view to Elman et al. (1996), who are widely perceived to be paragons of contemporary empiricism. So even amongst central proponents of mainline computationalism, which is widely thought to be distinctively pro-nativist and anti-empiricist, there appears to be no consensus about what ‘nativism’ and ‘empiricism’ mean.

---

What’s more problematic is the continued insistence on using these terms in the absence of any agreed-upon criteria for their usage. Marcus (2004) pays close attention to the kind of biological evidence about the role of genes in development that is often marshaled in support of skepticism about the notion of innateness, and argues that although this evidence undermines crude forms of genetic determinism, it leaves more sophisticated versions of nativism untouched. However, as Mameli & Papineau (2006) point out, it’s not clear why this more sophisticated sort of view ought to be classified as nativist; the empirical commitments of the view are similar to other views, based on similar evidence, that have been labeled empiricist, such as the view developed by Quartz & Sejnowski (1997). There are no clear empirical grounds for calling one view nativist and the other empiricist, and even if such grounds were to be provided by stipulation, it’s not clear what explanatory purpose this would serve; the categories of nativism and empiricism just don’t seem to be useful tools for elucidating the complexities of neuro-cognitive development. Indeed, in a strange twist, this is the conclusion that Marcus (2004) eventually reaches, after spending his book ostensibly defending a version of nativism. While the content of Marcus’ view is substantive and interesting, his characterization of it as nativist seems to be more ideologically motivated than theory-driven — an expression of political affiliation rather than empirical content. As he eventually recognizes, a sophisticated view of neuro-cognitive development isn’t a version of either nativism or empiricism, but one that repudiates that sterile dichotomy.

To relate these points back to the theme of this sub-section, the central problem with the claim that connectionist neuroscience is committed to empiricism isn’t so much that it’s false, as that it’s not even false; there’s no clear sense to be made of what it means. However, one might, in the spirit of Samet & Zaitchik (2012), reply that this is all just philosophical bookkeeping; we can recognize empiricism when we see it, and even though connectionist theorizing isn’t necessarily empiricist, when we
look at extant connectionist models, we see that, by and large, they embody empiricist assumptions. In accordance with this attitude, Gallistel (2000) writes, recall, that “people doing neural net modeling... are often at pains to point out that the network has solved a problem in the absence of an initial structure tailored to the solution of that problem”, citing a paper by Becker & Hinton (1992). Now, there’s a grain of truth to this; many connectionist models, especially those developed within the neo-connectionist tradition of the 1980s, do lack prior structure that is specific to the entities within their task domain. For example, a standard type of model in the neo-connectionist literature is a multilayer perceptron in which each of the nodes in a given layer is connected to each of the nodes in the next, and all of the connections are initialized to be random. However, to suppose that models of this type were developed to push an empiricist agenda is to ignore the specific theoretical and explanatory purposes for which such models were in fact developed.

Some models of this type were developed in order to address questions in theoretical computer science concerning which functions can, in principle, be learned on the basis of evidence. Proponents of these connectionist models were in fact pursuing the same project as researchers who are widely perceived to be critics of connectionism, such as Minsky & Papert (1969): that of charting what connectionist networks are theoretically capable of doing. Although the 1969 *Perceptrons* book is most famous for its negative results concerning the limitations of perceptrons, Papert (1988) later emphasized that he and Minsky did not think of their work as *attacking* perceptrons, but as contributing to our understanding of them, pointing out that more than half of their book reports positive results about the capabilities of perceptrons. As Papert rightly points out, “A real understanding of what a mechanism can do carries [implications] about what it can not do” (p.8).

Other models of the type at issue were developed in order to shed light on certain perceptual and cognitive capacities, specifically categorization. But categorization
quite often involves learning about what Gallistel calls ‘contingent’ features of the environment, features that might vary considerably from one generation to the next and hence are not evolutionarily stable. As Gallistel rightly points out, it’s unlikely that learning mechanisms have evolved specifically to track such contingent environmental features. For example, carburetors are presumably contingent, so it’s unlikely that animal brains are specifically adapted to categorize carburetors as such; the capacity to categorize carburetors is presumably a manifestation of a more general capacity for categorization. Thus it is unsurprising that many connectionist models of categorization start with relatively few constraints that are specific to the entities within their problem domain; by Gallistel’s own lights, this just reflects the nature of the capacity to be explained. This is not to deny what seems overwhelmingly likely, that some animal brains are specifically adapted to categorize particular environmental features, but just to assert that animals have a general capacity for categorizing contingent features of their environment, and that’s what relatively unconstrained connectionist models of categorization are attempting to elucidate. Keeping the point of the previous paragraph in view, we can see relatively unconstrained connectionist models of categorization not as expressions of an empiricist agenda, but as an attempt to reveal the extent and limitations of a general capacity for categorization.

A complementary point is that although many connectionist models embody relatively little prior structure that is specific to their task domain, many connectionist models embody quite a lot. For example, Elman et al. (1996) point out that although “connectionism is often seen as being quite opposed to modularist theories... there is no necessary reason why connectionist models should not be modular”, and that it is “therefore ironic that most current models are in fact highly task-specific and single-purpose” (p.392). One important type of modular connectionist network is the mixture of experts architecture developed by Jacobs and colleagues (e.g. Jacobs et al., 1991b). Networks of this type contain several different sub-networks that are architecturally
distinct, which compete during learning to process different subsets or aspects of the training data, such that different sub-networks specialize, becoming ‘experts’ at processing the data that is most suited to their particular architecture. As Jacobs et al. (1991a) point out, networks of this type are “well suited for incorporating domain knowledge” (p.219). In an interesting demonstration of how this might be so, Jacobs & Kosslyn (1994) trained a network consisting of two sub-networks on a visual categorization task, where one sub-network contained nodes with large overlapping receptive fields while the other contained nodes with small non-overlapping receptive fields, and found that the former sub-network became an expert at processing shapes, whereas the latter became an expert at processing metric spatial relations. As the authors point out, this model sheds light on how the functional specialization of the dorsal and ventral visual systems might become established; but, perhaps more interestingly given present concerns, it also illustrates a general point about the role of ‘innate’ constraints in the development of functional specialization. In mixture of experts networks, expertise isn’t ‘pre-specified’ within the architecture of sub-networks; nevertheless, there’s a relatively clear sense in which ‘innate’ architectural properties crucially contribute to the development of that expertise.

This point should lead us to reconsider facile judgments about whether connectionist models are ‘empiricist’ or ‘nativist’. For example, although it is easy to see the Becker & Hinton (1992) model of stereopsis from Gallistel’s perspective, as the epitome of a model that solves a problem ‘in the absence of an initial structure tailored to the solution of that problem’, the model in fact comprises several different modules that receive input from distinct but overlapping parts of the perceptual field, along with an internally-generated learning signal that registers coherence between the outputs of the modules. Together these seem to be just the thing for solving the problem of
stereopsis, which is essentially a problem in visual image registration. So from a certain perspective, the model seems to be about as nativist as it’s plausible for a model to be; although it has “no prior knowledge of the third dimension” (Becker & Hinton, 1992), and depends on experience to obtain that knowledge, it is innately equipped with just the right kind of structural and functional properties needed to extract that knowledge from experience. One might say that experience plays a merely supporting role in the maturation of the network’s knowledge; alternatively, one might insist that it’s only with a specific kind of experience that the knowledge can be learned. I think that the right response, however, is to point out that these ‘merely’s and ‘only’s are simply incidental to an explanation of how the knowledge is acquired; we gain nothing by classifying the model as empiricist or nativist, and nothing forces us to do so except ideology.

Finally, I think that these considerations show that the ‘we know nativism when we see it’ attitude that we’ve been engaging with is ultimately self-defeating. Insofar as researchers are guided in their search for innate traits by an intuitive picture of innateness according to which such traits are ‘pre-specified’ by a genome that functions as a sort of deterministic blueprint, then they will be guided by a crude caricature that doesn’t accurately depict the processes involved in the development of any trait, and will thus fail to appreciate the important ways in which genes in fact do contribute to development. To capture what’s true and important about nativism, we need a new set of conceptual tools, provided by molecular and developmental biology, that are too precise to render the bright line that traditionally demarcates nativism from empiricism. So not only is there no sense in which connectionism is theoretically

---

21It’s also worth noting here that the learning rule in the Becker & Hinton model is derived from information-theoretic principles; it effectively maximizes the mutual information between the outputs of the modules. This is just one counterexample amongst many to Gallistel’s repeated suggestions that researchers in the connectionist tradition are ignorant of, or unfriendly to, information theory.
committed to empiricism, the view that connectionist models tend in practice to be empiricist relies on a distorted picture of genetic contributions to development that nobody with nativist sympathies ought to accept.

4.3.2 Anti-Representationalism

Nevertheless, it is true that contemporary neuroscientific explanations of learning and memory are by and large manifestations of associative connectionism in the sense that I spelled out earlier, namely explanations that appeal to experience-dependent changes in the weights of connections between nodes within a network. So although, as I’ve argued, such explanations don’t necessarily or even characteristically depict experience as playing a deterministic role in cognitive development, they do accord experience an important explanatory role, thus we might grant that there is some sense in which they count as empiricist. But is it true that “extreme empiricists” in this sense “tend to be anti-representational” (Gallistel, 2001, p.9693)? Or that by virtue of exemplifying this kind of empiricism, “neural net theorizing represents a resurgence of the behaviorist view that the brain does not really represent the external world” (Gallistel, 1998, p.11)?

If one adopts an historical perspective, these claims might strike one as prima facie puzzling. For amongst the most plausible candidates for the title of ‘extreme empiricist’ are surely the British Empiricists of the Early Modern period, such as Locke and Hume, who were thoroughly committed to the existence of mental representations, or what they called ‘ideas’. Indeed, Fodor (2003), otherwise a resolute critic of anything smacking of empiricism or associationism, has argued that Hume’s theory of ideas is an early manifestation of the kind of representational theory of mind that is often said

---

22 I won’t articulate what exactly that sense might be, since I take that to be my interlocutor’s job. But for reasons discussed in the previous section, I take it that no cogent version of empiricism will be in tension with the view that genes and other endogenous factors play a crucial role in cognitive development.
to provide the conceptual foundations of contemporary cognitive science: “For Hume, as for... cognitive science, the mind is preeminently the locus of mental representation” (p.8). Moreover, the associationist psychologists of the 18th and 19th centuries such as Bain and Spencer also appealed to the association of ideas, and hence to mental representations.²³ Thus there seems to be no general tendency for empiricists to be anti-representationalists, and at the very least there’s no truth to the claim that “associative theories have always been anti-representational” (Gallistel & King, 2010, p.282, my emphasis). On the contrary, throughout much of Western history, empiricists have been representationalists. Indeed, according to what is arguably the most historically legitimate sense of ‘empiricism’ — which, note, is quite different from Gallistel’s conception of empiricism as the view that the brain’s structure is largely determined by experience — empiricism is the view that all knowledge derives from experience. Since knowledge on this view constitutively involves relations between ideas, and ideas are a species of mental representation, empiricists in the relevant sense are necessarily representationalists.

Of course, one might reply by pointing out that this traditional sense of ‘empiricism’ is quite different from the allegedly anti-representational sense that’s under consideration, according to which connectionist theorizing counts as empiricist by virtue of the fact that it invokes experience-dependent plasticity mechanisms. The fact that empiricists in the traditional sense are representationalists doesn’t entail that connectionists tend to be. Quite so, but that just serves to undermine the view that the various approaches in the history of psychology that might reasonably be described as ‘empiricist’ are manifestations of a theoretically homogeneous anti-representational and associationist research tradition. Contrary to Gallistel’s suggestion that any psychological

---

²³ “The connection between two states of consciousness occurring in succession, can very well represent the connection between two external phenomena occurring in succession” (Spencer, 1895, p.513).
theory that posits experience-dependent associative relations of one kind or another is *ipso facto* anti-representational, the question of whether such a theory counts as representationalist or not really does depend on the specifics of the theory — such as whether the associative relations that it invokes obtain between *representations*, as in traditional Empiricism.

Still, for all I’ve said so far it might be the case that connectionist theorizing does exemplify a kind of empiricism that tends towards anti-representationalism. Perhaps when Gallistel writes of ‘extreme empiricists’, he has in mind the behaviorists of the 20\textsuperscript{th} century, and perhaps his thought is that behaviorism and connectionism presuppose a common form of empiricism that, in extreme forms, manifests in skepticism about the explanatory import of representations. This line of thought might superficially seem plausible, since representations in the relevant sense are states or processes that are *internal* to the cognitive system, and behaviorists are widely regarded as having repudiated internal states or processes as explanatorily redundant. However, it would be a mistake to suppose that any empiricism that is common ground between connectionism and behaviorism also provides the grounds for the behaviorist’s skepticism about representations. The only sense in which connectionism and behaviorism are both empiricist is the sense in which they both emphasize the role of experience in explaining learning and cognitive development. But there’s no clear reason why such an emphasis on experience should lead to anti-representationalism. More to the point, it is not the behaviorist’s emphasis on *experience* that motivates her anti-representationalism, but her methodological strictures. To bring out this point, notice that the earlier talk of ‘internal’ states or processes was ambiguous; cognitive representations are ‘internal’ in the sense that they are parts or components of the cognitive system, whereas the internal states that behaviorists dismissed as explanatorily otiose were, in the first instance, ‘internal’ in the sense of being introspectively accessible. It is true that some ‘radical’ behaviorists such as Skinner (1953) also repudiated internal
states in the first sense, and held, for example, that neurophysiological states are irrelevant to psychological theorizing. But in either case, this aversion to internal states was due to a methodological assumption that a properly scientific psychology must only appeal to publicly observable behavior, an assumption that connectionists simply don’t share. It wasn’t extreme empiricism that motivated Skinner’s anti-representationalism, but an extreme conception of scientific explanation.24

These points will resonate throughout the remainder of the dissertation, so it’s worth pausing for a moment to unpack and examine them more closely. Consider the view that I’ve been calling ‘behaviorism’. I have in mind here the research program in psychology, rather than the view in philosophy about the meanings of mental-state terms that is sometimes called analytic behaviorism. Psychological behaviorism emerged at the beginning of the 20th century out of frustration with the introspective methodologies employed by an earlier generation of psychologists, which were thought to be too subjective and unreliable to provide the foundations of a rigorous experimental science. If psychology were to find a place alongside other serious sciences, it was thought, it would have to avoid speculating about internal states and restrict itself to replicable methodologies that delivered publicly observable evidence; specifically, to the experimental manipulation and measurement of behavior.

As suggested earlier, the most general and distinctive feature of behaviorism was its avoidance of states that are ‘internal’ in the sense of being introspectively accessible. However, from the beginning, there was a tendency amongst many behaviorists to avoid speculating about any kind of state that is internal to the cognitive system. Accompanying this was a certain attitude about the nature and subject matter of

24Someone might point out that Skinner’s philosophy of science was strongly influenced by logical positivism, so there is a sense in which his anti-representationalism was motivated by a kind of empiricism. This is true, but irrelevant to the point in the text, which is that his anti-representationalism wasn’t motivated by any form of empiricism that behaviorism has in common with connectionism.
psychological theorizing: psychology was thought to involve the elucidation of law-like generalizations about behavior, for purposes of prediction and control. For example, in his ‘Behaviorist Manifesto’, Watson (1913) writes that “psychology as the behaviorist views it is a purely objective experimental branch of natural science. Its theoretical goal is the prediction and control of behavior” (p.158). Skinner did more than anyone to articulate the conceptual underpinnings of this attitude, defining behaviorism in his (1969) book as “a philosophy of science concerned with the subject matter and methods of psychology” (p.22). Skinner drew his philosophy of science largely from Ernst Mach, who defended an instrumentalist view according to which scientific explanation consists in the description of law-like functional relations between observable variables. Speculation about unobservable entities was not interpreted realistically on this view, but heuristically, tolerated only to the extent that it proved useful in elucidating functional relations.

In accordance with this view, Skinner held that behavior is to be explained by describing functional relations between behavioral variables, rather than by positing unobservable ‘inner’ states — whether those states are the introspectable contents of consciousness or physiological states of the brain. This is not because Skinner denied what seems incontrovertible, that behavior is accompanied by conscious or neurophysiological states, but rather because he held that these publicly unobservable states are explanatorily otiose: “The objection to inner states is not that they do not exist, but that they are not relevant in a functional analysis” (Skinner, 1953, p.35). Skinner, like Mach before him, was ultimately opposed to the mechanical philosophy originating with Descartes and others, which attempted to explain natural phenomena by speculating about the physical states and causal processes that mediate those phenomena; for Mach and Skinner, there is nothing more to causality than functional relations.
between variables.\textsuperscript{25} As Skinner (1990) wrote later in his career, “What the brain does is part of what must be explained... We cannot find answers to questions of that sort in the body-cum-brain itself, observed either introspectively or with the instruments and methods of physiology” (p.1206).

The problem with the radical behaviorist’s conception of psychological explanation is encapsulated by an aphorism widely attributed to Chomsky: \textit{to call psychology a behavioral science is like calling physics a science of meter readings}. The most pressing aspect of the problem here is that, contrary to Mach and Skinner, mere descriptions of functional relations between variables simply don’t satisfy one of the central desiderata of scientific explanations: the deepening of our understanding of natural phenomena. Mach suggests that descriptions of functional relations, \textit{qua} law-like generalizations, provide us with understanding of phenomena by enabling us to predict them; when we can predict heretofore unexplained phenomena using law-like generalizations, “we feel at home with them, they no longer perplex us, they are explained” (p. 7, emphasis in the original). But this conflates prediction with understanding; even if we can reliably predict a phenomenon, we might still fail to understand \textit{why} or \textit{how} it occurs. For example, we might be able to reliably predict the tides without having any inkling of the moon’s gravitational influence on sea levels.\textsuperscript{26}

This is widely recognized to be one of the central failings of the kind of covering-law model of explanation that Mach’s and Skinner’s views exemplify. It is beyond the scope of the present work to enumerate the failings of the covering-law model, but it should suffice to say that contemporary philosophers of science now regard it as little more than an historical curiosity. Since the latter 20\textsuperscript{th} century, in response to problems with the covering-law model, philosophers of science have emphasized that descriptions of

\textsuperscript{25}For an articulation and defense of the view that Mach’s and Skinner’s conceptions of explanation were driven by an opposition to mechanism, see Chiesa (1992).

\textsuperscript{26}The example is from Cummins (2000).
causal mechanisms contribute to a genuine understanding of natural phenomena over and above the mere capacity to predict and control them. In an early and influential expression of this attitude, Railton (1978) wrote that “[k]nowing enough to subsume an event under the right kind of laws is not... tantamount to knowing the how or why of it... explanations making use of true, general, causal laws may legitimately be regarded as unsatisfactory unless we can back them up with an account of the mechanism(s) at work” (p.208). Over the past two decades, this attitude has coalesced into something of a consensus in the philosophy of science, according to which explanation, especially in the life sciences, proceeds by providing mechanistic explanations.27 A mechanism on this view is a system comprised of certain component parts, which are functionally organized in such a way that they interact to produce a regular outcome. A mechanistic explanation of such an outcome involves localizing the components of the underlying mechanism and elucidating their functional organization, so as to reveal how their interaction produces the outcome.

The foregoing points are vividly expressed in the context of psychological explanation by Cummins (2000), who argues that the law-like generalizations found in psychology do not, contra Skinner, provide explanations, but simply provide a systematic characterization of the behavioral phenomena that stand in need of explanation. He dramatizes the point with a rhetorical question: “Does the Law of Effect explain why giving a pigeon Pigeon Chow whenever it pecks a key increases the rate of key pecking? Or does it just restate the phenomenon in more general terms? Surely the correct moral to draw here is that the Law of Effect is an explanandum, not an explanans” (p.119). While Cummins doesn’t explicitly articulate his argument in the context of the ‘new mechanistic consensus’, and the body of his work suggests that

27For influential expressions of this ‘new mechanistic consensus’ in the philosophy of science, see Bechtel & Richardson (1993); Glennan (1996), and Machamer et al. (2000).
he in facts holds that psychological explanation is purely functional and autonomous from explanations in terms of underlying neural mechanisms, there are good reasons, again beyond the scope of the present work, to think that psychology is ultimately subject to the norms of mechanistic explanation. For example, Piccinini & Craver (2011) argue that functional analyses shouldn’t be regarded as complete explanations, but rather as mechanism sketches — incomplete or elliptical mechanistic explanations whose structural details are to be filled in by subsequent discoveries.²⁸

In any case, if we grant that explanation in psychology is at least often guided by the regulatory ideal of elucidating the mechanisms that mediate cognitive capacities, radical behaviorism seems deeply problematic. By eschewing internal states — specifically, the components of neural mechanisms²⁹ — it fails to elucidate the ‘how or why’ of cognitive capacities. Note that the problem here is ultimately a consequence of the explanatory strategy at the heart of radical behaviorism, and is only derivatively a consequence of the behaviorist’s anti-representationalism. Indeed, we see this dependency when we look at other incarnations of anti-representationalism in cognitive science. Contrary to Gallistel’s repeated suggestions that the main source of anti-representationalism in cognitive science is neo-connectionism and neuroscience,³⁰ one would be hard-pressed to find researchers in these areas who explicitly avow anti-representationalism.³¹ Instead, the dominant current of anti-representationalist sentiment comes from researchers in robotics (Brooks, 1991), computer science (Beer, 1995), and developmental psychology (Thelen & Smith, 1996), who align themselves with dynamical and embodied approaches in cognitive science, which, many have pointed out,

---

²⁸See also Wright & Bechtel (2006).
²⁹States are of course a different beast from components, but I don’t think that anything substantive hangs on this ontological subtlety, at least for present purposes.
³⁰See, for example, Gallistel (1998, p.11), Gallistel & King (2010, pp.56, 101, 196).
³¹Among the most plausible candidates include Freeman & Skarda (1990) and Globus (1992), though it’s not clear whether these researchers disavow all varieties of representation, or just those that are associated with the ‘classical’ computationalist or information-processing paradigm.
seem to employ the explanatory norms of the covering-law model.\textsuperscript{32} Although the use of tools from dynamical systems theory is certainly not in tension with the pursuit of \textit{mechanistic} rather than covering-law explanations — the Samsonovich & McNaughton (1997) model is a case in point — it is notable that proponents of \textit{mechanistic} dynamical explanations tend not to be shy about positing representations (Zednik, 2011).

Moreover, it is important to note that those behaviorists who did not follow Skinner in repudiating \textit{all} internal states — namely, the ‘mediational’ or ‘stimulus-organism-response’ behaviorists such as Hull (1930), Osgood (1956) and Tolman (1938) — posited internal states that pretty clearly play a representational role. For example, contrary to Gallistel’s claim that Hull (1930) is a member of an “anti-representational tradition”,\textsuperscript{33} Hull in fact posited a type of state that would seem to qualify as a representation in precisely Gallistel’s sense. Hull writes that when an organism has learned to successfully interact with the world, “the world in a very important sense has stamped the pattern of its action upon [the organism]... in such a way that a functional parallel of this action segment of the physical world has become a part of the organism... . The organism has thus acquired an intimate functional copy of the world sequence, which is a kind of knowledge” (pp.106-110). A \textit{functioning copy} would seem to be a functioning homomorph in everything but name.

Now, it is true that meditational behaviorists were less inclined to use the term ‘representation’ than the subsequent generation of cognitive scientists, and were generally shy of anything smacking of the mental. But this is more a reflection of sociological factors than theoretical commitments; before computing technology provided a way of conceptualizing how mental representation might be mechanized, there was a pervasive worry that talk of the mental would be interpreted as an acceptance of immaterial,

\textsuperscript{32}See, for example, Bechtel (1998) and Walmsley (2008).

\textsuperscript{33}Gallistel & King (2010, pp.xiv, 101).
non-naturalistic phenomena — Hull, for example, is careful to note that “[n]o spiritual or supernatural forces need be assumed to understand the acquisition of... knowledge” (p.106). In any case, as Ramsey (2007) rightly points out, when evaluating whether the posits of a theory count as representations, we shouldn’t look at the language in which the theory is clothed, but to the explanatory role that the posits are taken to play.\(^{34}\) Once again, even Jerry Fodor, scourge of associationists, holds that the internal states posited by meditational behaviorists play a representational role; although he holds that such states, qua associative, cannot explain systematic cognitive capacities, he allows that they count as representations all the same.\(^{35}\) Arguably, then, the meditational behaviorists were making speculative first attempts to sketch the mechanisms of cognitive capacities. In doing so, they inevitably found themselves positing internal representations. What enabled them to do so was freedom from the fetters of the radical behaviorist’s covering-law conception of explanation.

The crucial point here, though, is that connectionism, from its inception to its current incarnation in contemporary neuroscience, is distinguished by a self-conscious and systematic attempt to discover the neural mechanisms of cognition. For example, Hebb (1949) expressed “profound disagreement” with the Skinnerian view that explanation “is ultimately a statement of relationships between observed phenomena” (p.xiv), and held that psychology and neuroscience are ultimately pursuing the same explanatory project: “The problem of understanding behavior is the problem of understanding the

---

\(^{34}\)I will discuss Ramsey’s point, and the conclusion about the the representational bona fides of contemporary connectionism that he tries to draw from it, in Chapter 5.

\(^{35}\)I’m interpolating a bit. Fodor (1975) writes that meditational behaviorists like Hull “postulate mechanical... linkages between psychological states and assume that the links are forged by whatever laws determine the strength of habits, and that these “[i]nternal representations... are supposed to be associated to Ss and Rs in just the way that Ss and Rs are (supposed to be) associated to one another” (p.173). Fodor & Pylyshyn (1988) later argue that the internal representations posited by neo-connectionists cannot explain systematic cognitive capacities (or can, but only by implementing a classical cognitive architecture) precisely because they are associative in the same way the states posited by meditational behaviorists allegedly are.
total action of the nervous system, and *vice versa* *(ibid.)*. Ignoring the injunction on speculative theorizing imposed by his mentor Lashley, Hebb developed neurophysiologically informed hypotheses about the mechanisms mediating cognitive processes such as thinking that proceed independently of immediate sensory stimulation. Specifically, he held that such processes might be mediated by *phase sequences*, or cell assemblies that are associatively linked together, which play an explanatory role that is analogous to Hull’s *functioning copies*, and would seem to count as representations for the same reason.

Similarly, Rosenblatt (1958) developed his perceptron as a mechanistic model of learning. He emphasized that each of the parameters in his model is “is a clearly defined physical variable, which is measurable in its own right, independently of the behavioral and perceptual phenomena which we are trying to predict” (p.406), and went on to perceptively articulate one reason this affords an explanatory advantage over learning theories that focus on law-like relations between behavioral variables:

> Previous quantitative learning theories, apparently without exception, have had one important characteristic in common: they have all been based on measurements of *behavior*, in specified situations, using these measurements (after theoretical manipulation) to predict *behavior* in other situations. Such a procedure, in the last analysis, amounts to a process of curve fitting and extrapolation, in the hope that the constants which describe one set of curves will hold good for other curves in other situations. While such extrapolation is not necessarily circular, in the strict sense, it shares many of the logical difficulties of circularity, particularly when used as an ‘explanation’ of behavior.

*(Rosenblatt, 1958, p.406)*

---

36I think it’s transparently false that the parameters of perceptron models are ‘clearly defined physical variables’. There was, for example, a great deal of debate amongst neo-connectionists and their detractors about how to assign a physical interpretation to the parameters of the multilayer perceptrons they employed, as I discussed in Chapter 2.4.3. Nevertheless, I think that many neo-connectionist perceptron models were clearly *mechanism sketches*, developed with the intention of contributing to a broader project of elucidating the mechanisms of cognition. ‘Biological realism’ was clearly a regulatory ideal for Rosenblatt, and the neo-connectionists who followed in his footsteps.
So it seems clear that early connectionists, and their heirs in contemporary neuroscience, adopt norms of explanation that are quite different from those that motivate the radical behaviorist’s anti-representationalism. This undermines the suggestion, bruited earlier, that connectionism is anti-representational for the same reason that ‘extreme empiricists’ like Skinner were. What other reasons might there be for thinking that connectionism is continuous with an anti-representational tradition in psychology? One alleged reason that Gallistel often repeats concerns the influential view within neo-connectionism that processing in neural networks ought to be understood as ‘sub-symbolic’. Gallistel takes this view to be a disavowal of representations: “Ambivalence about representation appears in connectionist modeling circles, where the signal processing that occurs is sometimes said to be ‘sub-symbolic’” (Gallistel & King, 2010, p.56). But this simply conflates different notions of symbol; Gallistel’s error is to suppose that his somewhat idiosyncratic notion of a symbol, discussed in Chapter 2.2.2, is what proponents of sub-symbolic processing like Smolensky (1988) were intending to disavow. But on the contrary, the symbols that the sub-symbolicists had in mind were discrete tokens with representational contents that reflect the way a competent language-user might describe a given task domain; they were, as Clark (1989) put it, ‘semantically transparent’. Sub-symbols, on the other hand, “correspond semantically to fine-grained features below the level of the concepts consciously used to describe the task domain” (Smolensky, 1988, p.1). Sub-symbols thus have representational content, but a content that is subtler, finer-grained and more context-sensitive than the concepts expressed by words in a natural language. The activation of a node in a network that has been trained to recognize dogs, for example, carries semantic significance, but not of a kind that corresponds to a context-independent lexical concept like is hairy,

37See also, for example, Gallistel (1998, p.11), Gallistel & King (2010, pp.56, 194, 196).
or has a tail.\footnote{Again, even the staunchest critics of connectionism recognize this point; as Fodor & Pylyshyn (1988) put, “sub-symbolic states do have a semantics, though it’s not the semantics of representations at the ‘conceptual level’”. However, Fodor & Pylyshyn fail to recognize the broader significance of this point when they argue that the ‘sub-symbolic paradigm’ is simply a revival of the traditional Empiricist idea that concepts are defined out of simpler, atomic concepts. In fact, the sub-symbolic paradigm can be seen as a way of rejecting that idea; the promise of sub-symbols is that they reveal how a conceptual representation comprised of sub-symbols might be atomic, in the sense of not being definable in terms of components that themselves have conceptual content, without being simple, in the sense of lacking any semantic structure at all.} The node’s activation might signify a very subtle or holistic aspect of the network’s input at one time, and a different aspect at another. Nevertheless, this doesn’t rule out that a sub-symbol in the neo-connectionist’s sense counts as a symbol in Gallistel’s sense. At least ostensibly, a sub-symbol is computationally-manipulable vehicle of information that plays a role in representational structures; as Smolensky writes, “a sub-symbolic system can generate veridical representations of the environment [as] a result of extracting information from the environment and internally coding it in its weights through a learning procedure” \textit{(ibid., p.16).}

To some extent, we’re covering old ground. I argued in Chapter 3 that connectionists almost overwhelmingly take themselves to be representationalists, and that Gallistel is best understood not as arguing that the mechanisms posited by connectionists cannot (in fact) function as representations, but that they cannot function \textit{well enough} to explain cognitive capacities that are distinctive of intentional agency. Gallistel of course adduces several specific engineering considerations in support of this conclusion, which I will critically evaluate in Chapter 4.5. My aim here, though, has been to evaluate the general claim that connectionism belongs to an anti-representational, associative tradition, and inherits the presumptive flaws of earlier manifestations of that tradition. I have so far argued that connectionism is not objectionably empiricist or behaviorist; I will now address whether it is objectionably associationist.
4.3.3 Associationism

As I pointed out in the previous section, there is a continuity between mediational behaviorism, neo-connectionism and contemporary neuroscience, in the sense that the mediational behaviorists, like the later connectionists, speculated about the existence of internal states, which — eventually — might feature in mechanistic explanations of cognitive capacities. Indeed, many of the mediational behaviorists count as connectionists in that they posited mechanisms of neural plasticity to explain learning. Thus Hull (1943) writes that “the process of learning consists in the strengthening of certain [innate receptor-effector] connections as contrasted with others, or in the setting up of quite new connections” (pp.68-9). This approach stands in stark contrast with that of radical behaviorists. For example, Watson (1914) wrote disparagingly of attempts to explain learning in terms of neural plasticity: “psychologists talk quite volubly about the formation of new pathways in the brain, as though there were a group of tiny servants of Vulcan there, who run through the nervous system with hammer and chisel, digging new trenches and opening old ones” (p.166). I suggested that this perspective is problematic because it cannot elucidate the why or how of cognitive capacities, but that this problem is only derivatively about representations, and doesn’t impugn mediational behaviorism or the connectionism that came after.

Still, this doesn’t show that mediational behaviorism is without problems. Indeed, even though the mediational behaviorists provided sketches of internal states and processes that might explain cognition, their sketches were simplistic in the extreme. For example, note that Hull seems to have held that the neural plasticity that mediates learning occurs directly between receptors and effectors, rather than amongst intervening populations of neurons, and, relatedly, that it occurs ‘automatically’, without cognitive mediation.39 Given what we now know about the complexities of learning,

39See Hull (1943, p.69). Note, however, that it’s not entirely clear what Hull means when he holds
that seems radically implausible as an explanation of learning in general. Let’s provisionally call whichever features of the mediational behaviorist’s explanations of learning are objectionably simplistic ‘associationist’. The question before us, then, concerns what those features are, and whether contemporary manifestations of connectionism in cognitive & computational neuroscience exhibit them.

To address the first part of the question, let us turn to Gallistel’s characterization of associationism. Recall that, in Gallistel’s view, the ‘conceptual chasm’ separating the anti-representational and computational/representational traditions in psychology reflects two fundamentally different ways of thinking about the nature of learning: “In the first story, learning is the rewiring by experience of a plastic brain so as to make the operation of that brain better suited to the environment... In the second story, learning is the extraction from experience of information about the world that is carried forward in memory to inform subsequent behavior” (Gallistel & King, 2010, p.187). Upon reading this, one might wonder: why couldn’t the rewiring of the brain be the mechanism by which information is extracted from experience and carried forward in memory? Bracketing for the moment Gallistel’s concerns about the efficacy of neural plasticity as a mechanism of information processing and storage, why must we think of the two stories of learning as being conceptually incompatible, rather than simply as characterizations of the same phenomenon at different levels of analysis — one more functional, the other more mechanistic?

One point to make on Gallistel’s behalf is that the first, associative, story allegedly conflates two phenomena that are actually quite distinct: learning and memory. Recall that, according to Gallistel’s favored computational story, “learning is the extraction from experience of behaviorally useful information, while memory is the mechanism that learning occurs ‘automatically’. On one available reading, he simply means that it occurs without the intervention of an internal homunculus guiding the process of neural plasticity, which is surely unobjectionable.
by which information is carried forward in time” (Gallistel & King, 2010, p.279). But note that learning can’t just be the extraction of behaviorally useful information; if such information were not ‘carried forward in time’, nothing would have been learned. The process of extracting behaviorally useful information from experience would seem to describe perception rather than learning; learning is arguably better thought of as the process by which information extracted via perception is stored in memory such that it can guide subsequent behavior.40 This characterization of learning needn’t conflate learning with memory, so long as we distinguish the process of learning from the persisting state of memory that it results in. This brings us back to our question of why neural plasticity couldn’t play a role in the mechanisms that explain how this process of learning takes place. Again, where’s the alleged incompatibility between the two stories?

A clue might be found in the phrase ‘better suited to the environment’ that Gallistel employs in his characterization of the associative story. Gallistel holds that proponents of the anti-representational, associative tradition regard “all forms of learning as the learning of procedures” (Gallistel & King, 2010, p.101), and suggests that such procedural learning provides a kind of implicit knowledge that helps an organism get by in the world, without providing the kind of explicit knowledge of facts about the world that we attribute when we say, for example, that a scrub jay knows where and when it cached a morsel of food (ibid., p.100). Gallistel seems to have in mind here the distinction employed in psychology and neuroscience between procedural memory on the one hand and explicit or declarative memory on the other,41 and seems to be suggesting that the associative theory of learning only countenances procedural memory, whereas

40There is a widely-studied phenomenon of perceptual learning; see Fiorentini & Berardi (1980) for seminal work. But even this wouldn’t be learning properly so-called if it didn’t result in some lasting behavioral change.

41See (Anderson, 1976) and (Cohen & Squire, 1980) for seminal discussions of this distinction.
declarative memory is proprietary to the computational theory. Moreover, recall that the two theories about learning are supposed to correspond directly to views about cognitive architecture; correspondingly, Gallistel seems to identify the distinction between procedural and declarative memory on the one hand with, on the other, the distinction between information that is implicitly stored in the architecture of computational mechanisms and information that is explicitly stored in read-write memory. He writes, for example, that the “indirect, opaque ‘knowing’ that is characteristic of finite-state machines” should be distinguished from the “transparent symbolic knowledge” that is characteristic of machines with a Turing machine architecture (ibid.). Thus when Gallistel claims that the weights of a neural network do not instantiate explicit symbols but merely a “procedure that will generate different values given different input vectors” (ibid., p.285), it would seem that the ‘procedures’ he has in mind just are the procedures acquired during procedural learning.

This all strikes me as deeply wrongheaded. First, out of a concern for good ontological hygiene, we should distinguish conceptually between states or properties that are properly attributable to the entire person or cognitive system, and states or properties that are properly attributable to sub-personal mechanisms. When a person learns how to tie her shoelaces, the knowledge she acquires is ‘procedural’ in a sense that is quite different from any ‘procedures’ that are carried out by the mechanisms that explain her newly acquired capacity. Suggesting that these processes are procedural in the same sense is like suggesting that it is a person’s stomach that eats. Moreover, even if we revise the claim about the relation between the procedural /

---

42 When I write of the personal and sub-personal levels, I am not employing ‘person’ in a full-blooded sense according to which attributions of personhood have implications about personal responsibility and so forth. Rather, ‘person’ in the present sense just picks out an entire cognitive system as that system, with a distinctive set of capacities and dispositions, rather than as a mere aggregate of its parts. Perhaps ‘animal’ and ‘sub-animal’ would be a better choice of terminology, but the person-talk is ingrained.
declarative distinction and the distinction between cognitive architectures from one about *identity* to one about explanation or implementation, the claim would still be deeply problematic. The claim would seem to be that vehicles of explicit information implement states of declarative memory, and that it’s *by virtue of their being explicit*, in the sense of being computationally manipulable, that they’re explicit in the sense of being declarative. This would seem to entail that the information encoded by symbols is *constitutively* declarative, in the sense of being available at the doxastic level; indeed, given standard construals of the nature of declarative memory, it would seem to entail that such information is constitutively available to conscious awareness. But this is radically implausible when we remind ourselves of some of the systems that Gallistel counts as symbolic systems, which pretty clearly do not support a ‘doxastic level’, and which certainly are not *conscious*. Even if we allow that the symbols in the hydraulic path-integration mechanism depicted in Figure 2.2 count as memories in *some* sense, they’re surely not *declarative* memories. Moreover, the claim at issue would seem to entail, for example, that ‘mainline’ computational cognitive science is incapable of explaining procedural memory; that, given the ‘explicit’ read-write memory mechanisms it posits, it cannot explain the acquisition or exercise of ‘implicit’ knowledge, such as the knowledge one has of how to tie one’s shoes. This seems like far too heavy an entailment to hang on an unsupported assertion.

So, along with his apparent conflation of functional and mechanistic characterizations of learning, and of personal and sub-personal levels of explanation, Gallistel seems to conflate two radically different senses of *explicitness*: one that pertains to the

---

43 For example, according to Schacter’s (1987) influential account of the distinction between implicit and explicit memory, “explicit memory is revealed when performance on a task requires conscious recollection of previous experiences” (p.501).

44 Arguably, the claim at issue would also foreclose the possibility of a classical implementation of a procedural semantics. Extant views such as Johnson-Laird (1977) would seem to constitute a *reductio* of this claim.
computational accessibility of a sub-personal vehicle of information, and one that pertains to the doxastic or conscious accessibility of personal-level memories. Of course, Gallistel isn’t alone in this. Even those far friendlier to connectionism, such as Clark & Toribio (1994), have claimed that connectionist networks “embody a powerful kind of ‘knowing how’”, “without explicitly representing that knowledge”, where the relevant sense of explicit representation is said to involve symbol strings in a “declarative code” (p.403). Claims of this general tenor are in fact ubiquitous in discussions of connectionist schemes of information storage.45 But, again, procedural and declarative memory are states or capacities of entire cognitive systems, and should not be confused with the sub-personal mechanisms that instantiate or explain those states or capacities. I see no reason to think, for example, that the exercise of one’s ‘implicit’ knowledge of how to tie one’s shoelaces proceeds without the involvement of sub-personal vehicles of ‘explicit’ information; on the contrary, tying one’s shoelaces is a highly complex sensorimotor process that undoubtedly involves the orchestration of numerous perceptual and motor states that count as explicit in Gallistel’s sense. The general lesson in the background here, which will come into the foreground in Chapter 5, is that although it is tempting to think of explicit symbols as being distinctively ‘mentalistic’, or as being involved in sophisticated cognitive processes, they needn’t be; they might, for example, just be states in a hydraulic path-integration mechanism.

Nevertheless, even if we abandon the view that the distinction between procedural and declarative memory transparently maps onto a distinction between cognitive architectures, we might still hold that what’s distinctive of the ‘associative’ story about learning is that, in contrast with the ‘computational’ story, it tends to focus on relatively simple forms of procedural learning, and/or that it tends to posit sub-personal mechanisms that embody implicit rather than explicit information. Now, this is not

---

plausible as a *general* characterization of associationism that is intended to encompass British empiricism or even 19th century associative psychology, since those pre-20th-century forms of associationism posited associative relations between conscious, declarative *ideas*.\(^{46}\) However, this characterization might serve for present purposes, viz. to identify a sense of associationism that picks out certain problematic aspects of mediational behaviorism. For as the earlier quote from Hull suggests, when the mediational behaviorists speculated about the role of neural plasticity in learning, they arguably *did* tend to focus on relatively simple skills or habits, and posited simple monosynaptic pathways between receptors and effectors that presumably do not qualify as symbols. Our question, then, is whether contemporary neuroscience counts as associationist in the present sense.

I take it that the answer is clearly ‘no’. From its inception, research into the role of LTP in learning has focussed largely on declarative memory. As mentioned previously, some of the earliest and most compelling evidence that LTP is involved in learning came from studies of spatial learning in rodents; for example, (Morris et al., 1986) showed that blocking hippocampal NMDA receptors pharmacologically prevents the induction of LTP, which in turn prevents rats from learning the location of the platform in a water maze. As Gallistel (1990a) himself emphasizes, this kind of spatial memory is a crucial component of episodic memory, one form of declarative memory. Moreover, it involves structures in the hippocampus which have been known, since seminal work by Scoville & Milner (1957) with the severely amnesic patient H.M. to be crucially implicated in declarative memory. Indeed, LTP was discovered in the hippocampus Bliss & Lømo (1973), and the majority of subsequent research on the behavioral role

---

\(^{46}\)One of the central tenets of British empiricism, of course, is that memory is a revivification of perception. This is a way of articulating the view that memories are *episodic* — a species of declarative memory. For example, Locke (1689 [1975]) took memory to be a power of the mind “to revive Perceptions, which it has once had, with this additional perception annexed to them, that it has had them before” (p. 150).
of LTP has focussed on the role of hippocampal LTP in declarative memory.

Furthermore, contemporary neuroscientists frequently posit sub-personal vehicles of explicit information, in the form, for example, of reverberating patterns of activity in attractor networks. We’ve seen that although Gallistel sometimes suggests that these states are not symbols, his considered view is that they in fact are.\(^{47}\) Notably, reverberating patterns of activity have been hypothesized to play a role in another form of non-procedural, personal-level memory, viz. working memory. The basic idea here goes back at least to Hebb (1949), who posited reverberating cell assemblies “to account for the delay, between stimulation and response, that seems so characteristic of thought” (p.xvii). Hebb’s sketch of a mechanism has since been filled in by a substantial amount of physiological detail concerning, for example, the role of NMDA-receptor-mediated synaptic plasticity in regulating the stability of reverberating patterns (Wang, 2001). However, the outline has remained the same: recurrent networks are today “an integral part of the most plausible computational models of working memory” (Fuster, 2009, p.908).

This discussion foreshadows a couple of critical points that I’ll return to in Chapter 4.5 when I engage with Gallistel’s ‘engineering’ arguments against the explanatory power of the mechanisms posited by contemporary neuroscientists directly. First, note that in a book that is ostensibly about the role of memory mechanisms in explaining cognition, Gallistel & King (2010) do not explicitly differentiate the various different kinds of memory that psychologists and neuroscientists have disentangled over the past sixty years; the book does not contain the terms ‘procedural’, ‘declarative’ or ‘working’ memory. Indeed, the authors seem to think that there is one type of memory that is mediated by one type of ‘universal’ read-write mechanism: “we see no more reason to suppose that different mechanisms are required for different kinds of messages...\(^{47}\)

\(^{47}\)See my discussion of this point in Chapter 3, sections 3.4 and 3.6.
than to suppose that different kinds of action potentials are required or different kinds of DNA” (ibid., p.287). One might have thought that the reason is clear: as I just suggested, there is a vast body of experimental evidence that the brain contains several dissociable memory systems, which play various different functional roles, and which are mediated by distinct neural structures and processes. One salient drop in this ocean is a seminal neuropsychological study by Knowlton et al. (1996), which reported a double dissociation between the aforementioned declarative system mediated by the hippocampus, and a procedural system involved in learning habits that is mediated by the striatum. Other studies, using a variety of complementary methods ranging from experiments with animals to neural imaging studies with humans, have corroborated this distinction and elucidated structural and functional distinctions between several other memory systems — such as those involved in emotional memory, perceptual learning, and priming (Squire, 2004). On the basis of these kinds of dissociations, it is not unreasonable to argue, as Griffiths (1997) has done in the case of emotion, that there is in fact nothing that answers to our pre-theoretical notion of a unitary faculty of memory.  

The second, closely related, point is that Gallistel and colleagues seem to think that memory will be reductively identified with a specific and relatively simple kind of mechanism. We saw an aspect of this idea earlier, when I pointed out that Gallistel & King (2010) seem to identify declarative memory states with symbols in a read-write

---

48 See Craver (2004) for a percipient discussion of some of the subtleties involved in developing this kind of ‘kind-splitting’ argument.

49 As will become clear shortly, I am using ‘reduction’ in the technical sense articulated by Nagel (1961), which pertains to a specific relation between theories that obtains when the claims in one theory can be deduced from the claims in the other in virtue of bridge laws that establish identities between the entities posited by the two theories. There is a more liberal notion of ‘reduction’ abroad, which obtains when a given phenomenon is explained in terms of its underlying mechanisms (e.g. Sarkar, 1992). For reasons I’ll go on to discuss, I think that contemporary neuroscience provides — or is on its way to providing — reductive explanations of learning and memory in the latter sense, but not the former.
memory. Moreover, Gallistel & Matzel (2013) seem to foist their reductive aspirations onto their neuroscientific interlocutors when they attribute to neuroscientists “the hypothesis that LTP is the mechanism of memory” (p.192, my emphasis). Strangely, they do this in the face of repeated assertions by their interlocutors that this is not in fact an hypothesis they endorse. For example, recall from Chapter 3.5 that Gallistel & Matzel argue that LTP cannot be identified with the mechanism of conditioning, since (a) the time-scales of LTP induction do not reflect the CS-US intervals that are conducive to conditioning at the behavioral level, and since (b) the time-scales of LTP persistence do not reflect the duration of conditioned responses, which can last for months or years. In a (2002) review paper, Martin & Morris address these points in the context of defending the view that LTP plays an important role in memory mechanisms. They note that the first point is only problematic for that view “if a very simple circuit is assumed, in which CS-US associations occur online at the level of individual synapses. It is less problematic if the time-scale for information representation in a given brain region is different from that pertaining to events as they happen” (p.610). They go on to discuss a possible mechanism by which this might occur in the context of taste-aversion conditioning (p.625). Addressing the second point, Martin & Morris note that according to a widely-endorsed model, “the hippocampus plays a time-limited role in information storage, with the cortex being the ultimate repository of many kinds of memory”, and hence that “it would be premature to reject synaptic plasticity as a memory mechanism” merely because of its limited duration, since “[h]ippocampal LTP may need only last long enough... to permit completion of a slower neocortical consolidation process” (p.601). Gallistel & Matzel seize upon these quotes as an inadvertent indictment of ‘the LTP hypothesis’, writing that Martin & Morris have conceded that “the temporal properties of LTP do not explain the temporal properties of behaviorally measured association formation”, and that “the persistence of LTP does not explain the persistence of associative learning” (p.172). Their tone is triumphal, but they fail
to note that they are in fact expressing agreement with Martin & Morris, and that the LTP hypothesis they attribute to Martin & Morris, according to which ‘LTP is the mechanism of memory’, is not the LTP hypothesis that Martin, Morris, and other contemporary neuroscientists actually endorse. According to that hypothesis, LTP is not identical to memory, but plays an important role in a complex, multi-level memory mechanism: “Memory is a property of the entire organism whereas plasticity is a property of synapses; the circuit-level operation of a neural structure will not necessarily be reflected, in miniature, by the operational characteristics of its synaptic components” (Martin & Morris, 2002, p.610).

As Craver (2003) describes in a history of the hypothesis that LTP is involved in learning and memory, many neuroscientists throughout the 1950s and ‘60s did attempt to identify learning and memory processes with synaptic plasticity in relatively simple neural circuits, in ways that are reminiscent of the classical Nagelian model of reduction (Nagel, 1961). A representative example of this attempt is the model of classical conditioning developed by Gardner-Medwin (1969), depicted in Figure 4.1. Oppenheim & Putnam (1958) capture the ambient spirit of reductionism when they write that the neural network models of McCulloch & Pitts (1943) and others promise “to give hypothetical micro-reductions for memory” (p.20). However as Craver points out, these reductive aspirations were reconceptualized over subsequent decades; researchers...
first came to view synaptic plasticity as a simple example or experimental model of learning or memory, and then, spurred by the discovery of LTP in 1973, came to view synaptic plasticity as a component in a multilevel mechanistic explanation of memory.

This is evidenced in the report by Power et al. (1997) that Gallistel & Matzel (2013) cite in support of the point that there is a discrepancy between the duration of LTP and that of a conditioned response; Power et al. are quite explicit that the synaptic potentiation they observed in the hippocampus is probably too brief to be LTP, but that it plausibly reflects the first phase of an extended consolidation process whereby transient potentiation is subsequently “localized to a much more specific pattern or set of synapses” (pp.1186-7). This suggestion is consistent with contemporary views about declarative memory, mentioned in the earlier quotes from Martin & Morris (2002), according to which novel memories are rapidly encoded by the hippocampus and stored there in a labile state, until they are gradually consolidated throughout the cortex where they are stored in a more permanent form. LTP and other forms of synaptic plasticity are thought to play a crucial role in this process, but given what is known about the malleability of synaptic potentiation, nobody thinks that this role will be straightforward, and there is ongoing theoretical and experimental work investigating how synaptically-mediated memory traces might persist in the face of noise, interference from other memories, and reconsolidation after memory retrieval. Regardless of how this work pans out, it is clear that, just as contemporary biologists hold that genes play role in explaining the inheritance and development of phenotypic traits without identifying genes with traits, contemporary neuroscientists hold that synaptic plasticity plays a role in explaining learning memory without straightforwardly identifying

---

51 Contra Gallistel & Matzel (2013), who claim that Power et al. (1997) observed LTP.
52 For an early and influential computational model of this process, see Mcclelland et al. (1995). For a contemporary review of the neurophysiological evidence supporting the hypothesis, see Frankland & Bontempi (2005).
learning and memory processes with synaptic potentiation. The synaptic learning hypothesis that Gallistel and colleagues impute to contemporary neuroscientists is a view that few neuroscientists have endorsed since the 1950s.

Figure 4.1: A reductive model of classical conditioning that identifies the functional components of a conditioning protocol (US, CS, R) with the activities of single neurons joined by plastic synapses in a simple circuit. From Gardner-Medwin (1969, p.917).

Let us now turn to one final way of articulating the view that contemporary neuroscience is objectionably associationist. This is again implicit in the expressions that Gallistel uses to characterize the two stories about learning: ‘rewiring by experience’ in the case of the associative story, and ‘extracting from experience’ in the case of the computational story. This choice of prepositions suggests that a distinctive feature of associationism for Gallistel is that it characterizes learning as a passive, mechanistic process driven purely by environmental stimuli, whereas a distinctive feature of the computational story is that it characterizes learning as an active, cognitive process whereby an agent interprets incoming information and acts on it in ways that are

54Two other ways of articulating this view can be summarily dismissed. First, Fodor (1998b) characterizes associationism as the view that the structure of the mind is “an image, made a posteriori, of the statistical regularities in the world in which it finds itself” (p.203). I argued that contemporary neuroscience is not objectionably empiricist in section 4.3.1. Second, Gallistel often characterizes associationism as the view that learning doesn’t involve the acquisition of representations. I argued that contemporary neuroscience is not anti-representational in section 4.3.2.
constrained by rational norms. Indeed, as I discussed in Chapter 3.5, this seems to be the distinction that is driving many of Gallistel’s specific arguments against extant neuroscientific explanations of learning. Fodor (1975) invokes much the same distinction when he argues that the associationism of mediational behaviorists like Hull and Osgood is ‘wrong in every way it can be’: “Internal representations are typically paired with what they represent by computational (rather than associative) processes. That is, representations are not elicited but, as it were, assigned; and which representation is assigned is determined by calculations which rationally subserve the utilities of the organism” (p.173). Similarly, Clayton et al. (2006) distinguish between associative explanations that appeal to mechanistic processes, and rational explanations that invoke intentional processes of practical inference.

So perhaps this distinction provides us with a conception of associationism that encompasses mediational behaviorism and which, for reasons discussed in Chapter 3.5, is problematic. Our question, then, is whether contemporary neuroscience counts as associationist in the present sense. To evaluate this question we must do some further ground-clearing, for, again, the conception of associationism we have before us is grown over with an entanglement of personal and sub-personal issues. To begin,

---

55 One might worry about how this criticism squares with Fodor’s (1990) theory of content, which would seem to explicate what it is for representations to be ‘paired with what they represent’ precisely in terms of their dispositions to be ‘elicited’ by what they represent. It is, perhaps, open for Fodor to reply that this worry trades on an equivocation between different senses of ‘pairing’. The sense invoked in Fodor’s theory of content pertains to the metaphysical relation that obtains between a representation and its content, whereas the sense invoked in his criticism of associationism arguably pertains to the functional role that representations play in cognition; the fact that my horse-concept is about horses might be constituted by the fact that it is elicited by horses (in the right way), but this doesn’t exhaust the role that it plays in my cognitive economy — my horse-concept might be elicited by horses in the course of horse perception, but it might also be activated ‘offline’ whilst thinking about the noble equine. Perhaps that’s what Fodor has in mind when he writes of representations being ‘assigned’. This response might defuse the present worry, but it is a bit puzzling. Talk of representations being ‘paired with what they represent’ would certainly seem to be about the relations they bear to their contents rather than the functional relations they bear to other representations; and in any case, the whole point of the representational states that, say, Hull posits is that they play a role in offline cognition. That’s their explanatory raison d’être.
first note that distinctions like \textit{active} versus \textit{passive}, or \textit{rational} versus \textit{arational}, are distinctions that apply, at least in the first instance, to cognitive systems or their capacities — they apply ‘at the personal level’. Yet when Gallistel, Fodor, Clayton and others discuss associationism, they seem to have in mind some specific type of sub-personal mechanism. However, if we are attempting to grapple with relatively simple sub-personal mechanisms like synaptic plasticity or the functionally-individuated read-write memory mechanisms envisaged by Gallistel, it’s not at all clear that notions like \textit{active} or \textit{rational} find purchase. Indeed, ascribing these notions to sub-personal mechanisms strikes me as a kind of explanatorily-impotent animism. Instead of asking whether certain sub-personal mechanisms are active or rational, it seems to me that we should instead ask whether they play a role in \textit{explaining} active, rational, or otherwise distinctively cognitive learning capacities.

This point is a prophylaxis against a distressingly widespread confusion. Neuroscientific explanations of learning are clearly mechanistic in the sense that they are guided by the norms of mechanistic explanation (Craver, 2007). But this is an unobjectionable sense of ‘mechanistic’ that also subsumes the kind of computationalist explanations of learning that Gallistel envisages.\textsuperscript{56} It would be a mistake to suppose that because neuroscientists seek ‘mechanistic’ explanations of learning, they thereby characterize learning as ‘mechanistic’ in the sense of being an arational, reflex-like process. The mistake here, again, is to conflate different senses of ‘mechanism’: one that

\textsuperscript{56} Extant ‘classical’ computationalist models in cognitive science do not posit specific, structurally localized mechanisms, but computationalist modeling efforts are arguably guided by the regulative ideal of providing mechanistic explanations; as Piccinini & Craver (2011) put it, computationalist models are ‘mechanism sketches’. This is reflected more in the practice of computationalist modeling than in the methodological edicts of philosophical commentators such as Fodor & Pylyshyn (1988), who claim that explanations of cognitive capacities are independent of implementing mechanisms. In any case, it is clear that for Gallistel, cognitive explanations are ultimately mechanistic. It is thus slightly disingenuous for Gallistel (2011) to write that all the models of path integration he knows of, “including neural net models” (p.254), are symbolic; this suggests that there are \textit{non-connectionist} models of path integration that posit specific, localizable mechanisms, which is of course false.
pertains to sub-personal mechanisms and another that pertains to the personal-level capacities that the mechanisms are invoked to explain. This kind of mistake often manifests in the public media, as hand-wringering about the possibility that biology will someday reveal us to be unconscious zombies, in thrall to our brains or genes; or, more generally, as the worry that ‘naturalistic’ explanations will somehow drain the world of the magical properties presented in its manifest image. But a mechanistic explanation of learning, for example, needn’t entail that there are no rational processes of learning, any more than a mechanistic explanation of digestion need entail that there are no processes of digestion. We digest because of peristalsis and pepsin, not in spite of such things. Unfortunately the error here infects even otherwise serious scientific discussions of animal cognition. Thus Clayton et al. (2006) write that “the issue of whether an animal is psychologically rational turns on... whether [its] behaviour is caused by psychological mechanisms or by intentional processes (p.199, my emphasis). Of course, Clayton et al. don’t mean to suggest here that intentional process somehow transcend underlying mechanisms; my complaint here is that the dichotomy that Clayton et al. presuppose is mischaracterized, misleading, and doesn’t bear the theoretical load that they and others place on it — in particular, it doesn’t license inferences about the specific character of the mechanisms that mediate learning.57

This complaint might come into sharper focus when we consider how this kind of dichotomy manifests in debates within the ethological literature. Though they’re often not recognized as such, because of the misleading dichotomies in terms of which they’re framed, debates about whether certain animal behaviors are genuinely ‘rational’ or ‘cognitive’, or merely ‘mechanistic’ or ‘associative’, are generally not about whether those behaviors are ultimately to be explained by some specific type of neural

57This complaint is echoed by Papineau & Heyes (2006) when they write that “a bald assertion that some piece of cognition is ‘rational rather than ‘associative is no substitute for a specific hypothesis about the mechanisms responsible” (p.192).
mechanism, but are rather about what exactly the behavioral capacity to be explained is — and, in particular, about whether it is relatively ‘simple’ or more ‘complex’. One sees this, for example, in debates about whether apes are genuinely capable of representing and reasoning about the psychological states of others (Tomasello et al., 2003), or whether the results that allegedly demonstrate this capacity can be explained in ‘behavioristic’ terms (Povinelli & Vonk, 2004). One also sees this in Clayton’s painstaking food caching experiments with scrub jays, which are carefully designed to rule out ‘associative’ explanations (Clayton et al., 2001). That these debates are about nature of behavioral capacities rather than neural mechanisms is evidenced by the fact that they are addressed and resolved by conducting behavioral experiments. Although the terms in which these debates are expressed might seem to be picking out a specific type of mechanism, they are too under-specified to do this; their function, instead, is simply to characterize the relevant behaviors as being more or less sophisticated. It is certainly true that Clayton’s results provide strong constraints on theorizing about the mechanisms that explain scrub jay caching, by showing that such mechanisms must be capable of integrating information about what was cached, where, and when. But at most these constraints provide a highly schematic mechanism sketch, which might in principle be filled in by a range of different specific mechanism types. The fact that jays have the ability to ‘rationally’ store and retrieve food in ways that are mediated by memories and motivational states surely doesn’t, on its own, rule out the possibility

---

58 Hence the frequent use of terms like ‘simple’ or ‘complex’ to qualify the opposing positions in these debates: Clayton et al. (2006) write of a “simple associative account” of jay caching (p.213), Penn et al. (2008) write of “more complex, information-processing” accounts of social cognition in primates (p.116). The theoretical backdrop here is of course Morgan’s (1903) injunction to interpret animal behavior as the outcome of the simplest possible ‘psychical faculty’ (p.53). That psychical faculties might be ‘simpler’ or more ‘complex’ is intuitively compelling, but it’s unclear how to make the idea precise. Arguably, the idea rests on exactly the kind of anthropomorphism — grounded in our sense of which cognitive processes are difficult or valuable — that Morgan was attempting to repudiate. Why think that Clever Hans’ ability to tap his foot on cue is simpler than a capacity to count, in a sense of ‘simpler’ that isn’t merely a parochial vestige of a pre-Darwinian scala naturae?
that mechanisms of ‘associative’ neural plasticity might play a role in explaining this ability; just as evidence that neural plasticity does play a role in explaining this ability wouldn’t somehow reveal that jays lack the ability.

To express the point in a way that resonates with many of the previous points, we shouldn’t simply assume that the distinction between relatively simple, reflex-like behavioral capacities and more complex, cognitively mediated capacities — however that distinction is ultimately cashed out — transparently maps on to the distinction between the connectionist mechanisms posited by neuroscientists and the ‘classical’ computational mechanisms imagined by Gallistel and others. This point underscores the real question in the vicinity, which is why connectionist mechanisms cannot explain complex, cognitively mediated capacities. One might have thought that this question was addressed in Chapter 3.5, but if one reviews the arguments presented there, one will see that they largely presuppose that neuroscientific explanations are not directed at explaining complex, cognitively mediated capacities. This is a variant of the view I considered earlier, that neuroscience is directed at explaining only procedural learning, and can be dismissed just as easily; a glance at a textbook on cognitive neuroscience reveals that neuroscientists at least aspire to explain how perception, memory, imagery and other intentional states interface with motivational and affective states to guide rational action.

Again, the real question here is why we should think that those aspirations are forlorn. To bring that question into sharper focus, we should ask what it is that distinguishes genuinely computational mechanisms from the putatively ‘associative’ mechanisms posited by neuroscientists, in virtue of which the former, but not the latter, can explain complex, cognitively mediated capacities. Gallistel, Fodor and others seem to think that there is a clear demarcation between mechanisms that perform association-like processes and genuinely computational processes. But there certainly seem to be many processes performed by conventional computational mechanisms that
might reasonably be called ‘associative’. *Linked lists* are one of the most fundamental data structures implemented by conventional computers, yet it doesn’t seem like a misuse of the language to describe each element in a linked list as being ‘associated’ with the next; indeed, a common type of linked list is called an ‘associative array’. *Programmable logic devices* are ubiquitous in contemporary electronics, but contain reconfigurable digital circuits that, again, it doesn’t seem perverse to describe as ‘associative’. Here, ‘association’ is simply used to describe the formation of a linkage between two elements of a mechanism, and is entirely neutral about whether the mechanism mediates a simple, reflex-like capacity at the personal level. Why not think that the connectionist mechanisms posited by neuroscientists are merely associative in this sense? To reply that such mechanisms cannot explain complex cognitive capacities would patently beg the question.

To make these considerations more concrete by relating them to neural and cognitive phenomena we’re now familiar with, consider the kind of functional role that Gallistel accords computational mechanisms in the explanation of episodic memory and path integration:

Gallistel (1990)... adduced evidence that spatio-temporal indexing is the mechanism by which the brain knits together the diverse aspects of experience computed by the many different problem-specific modules that are implied by the neuroanatomy and electrophysiology of the cortex. On this hypothesis, the brain binds the remembered color of an object to its remembered shape on the basis that the physically separated... memories of the object’s color and its shape have the same spatio-temporal index. They have the same spatio-temporal index because they were perceived in the same place at the same time. This hypothesis maintains that episodic memory, that is, the ability to reconstruct an experience in all of its diversity, depends fundamentally on the representation of the spatial and temporal location of the elements of the experience.

(Gallistel & Matzel, 2013, pp.188-9)

This *knitting* and *binding* is presumably a computational process in Gallistel’s eyes, but why not think that it might also be mediated by synaptic plasticity and other mech-
anisms recognized by neuroscientists? Indeed, there’s a considerable body of evidence that it in fact is. I mentioned earlier that on the received view of the neuroanatomy and neurophysiology of episodic memory, an initial labile memory trace mediated by the hippocampus is subsequently consolidated throughout neocortex. I also mentioned that the hippocampus is thought to play a crucial role in spatial representation, as the locus of a cognitive map that is updated by perceptual and idiothetic information. As Gallistel suggests in the above quote, we might see these different perspectives on hippocampal function as complementary if we suppose that the hippocampus functions as a spatiotemporal index to modality-specific information encoded in various different cortical regions. What Gallistel doesn’t mention is that this view was developed in the 1980s by Teyler and colleagues specifically in order to integrate emerging findings about the neurophysiology of LTP with what was then understood about the anatomy of the hippocampus and the neuropsychology of episodic memory. The view continues to guide neuroscientific research into episodic memory, and has been amply supported by a wealth of experimental evidence (Teyler & Rudy, 2007).

The hippocampal indexing theory developed by Teyler & Discenna (1986) proposes that as an animal actively explores its environment, incoming sensory information is processed by specialized circuits in primary sensory cortex, then bound together by multisensory association cortex, generating a distributed pattern of activity across several cortical sites that encodes what the animal perceives. Via projections from association cortex to hippocampal structures, this activity pattern drives activity in the hippocampal cognitive map that encodes the animal’s location, thereby imbuing the animal’s perception with spatiotemporal context; what the animal perceives is integrated with information about where and when. Synapses in the circuits that mediate these activity patterns are potentiated, and if the content of the animal’s perception is sufficiently salient, this labile pattern of potentiation undergoes a process of consolidation, forming a long-term memory trace distributed throughout various cortical
structures. When, for example, the animal subsequently experiences combinations of sensory states and spatial contexts that are similar to those that initially produced a memory trace, the corresponding activity in the hippocampus might project to cortical areas and elicit the memory trace. In this way, the hippocampus indexes information encoded in the cortex, and thereby mediates the retrieval of episodic memories.

This view is supported by many intertwined strands of evidence, most of which is beyond the scope of this dissertation. However, two strands are especially noteworthy, since they concern the computational roles of knitting and binding presupposed by the hippocampal indexing theory, and how those roles might be played by mechanisms that are familiar to neuroscientists. First, consider how occurrent activity in specialized cortical circuits might be bound together and reciprocally modulated by activity in the hippocampus. Since the 1980s, evidence has accumulated that oscillations in the gamma range (25–100 Hz) serve to bind together the activity of cortical circuits, and that gamma oscillations correlate with conscious perception.\(^{59}\) More recent work has explored possible function relations between cortical gamma and hippocampal theta oscillations; in particular, researchers have suggested that the hippocampal activity might temporally coordinate activity in the cortex through ‘cross-frequency coupling’ involving phase synchronization between theta and gamma oscillations.\(^{60}\) Sirota et al. (2008) provided direct evidence of this by recording single-cell activity and local field potentials in various hippocampal and neocortical areas of the behaving rat. They found that local, transient gamma oscillations in the cortex were phase-locked to hippocampal theta oscillations, thus demonstrating that “hippocampal theta oscillations can effectively link... sporadic and spatially distinct local gamma oscillations” (p.694).

How, then, might an occurrent pattern of activity distributed throughout the cortex

\(^{59}\)See Fries (2009) for a review.

\(^{60}\)See Lisman & Jensen (2013) for a review.
and entrained to hippocampal activity be laid down as a memory trace in the way envisaged by the hippocampal indexing theory? That LTP and subsequent consolidation processes play a crucial role in this is evidenced by a vast body of research, but one recent study is especially noteworthy in this context. Using a methodology similar to that developed by Clayton & Dickinson (1998), Fellini & Morellini (2013) found that mice were able to modify their behavior on the basis of memories about whether they had encountered a potential mate at a specific place and time, viz. memories about what they experienced, where, and when. Interestingly, the formation of what-when-where memories seemed to up-regulate the hippocampal expression of Arc/Arg3.1, a gene implicated in potentiation and memory consolidation. Moreover the injection of the amnesic drug anisomycin into a mouse’s hippocampus immediately after it had experienced a mate at a specific time and place seemed to prevent the formation of a unified what-when-where memory; the mouse’s memory became unknit.

The mechanisms just surveyed, which have been among the central focal points of research in mainstream neuroscience over the last three decades, would seem to play the kind of computational role that Gallistel thinks is necessary to explain episodic memory and other distinctively cognitive capacities. In the various articles and books of his that I’ve cited, Gallistel doesn’t explain why these mechanisms cannot play such a role; a fortiori, he doesn’t mention the relevant bodies of research. It would thus seem reasonable to suppose, at least provisionally, that the mechanisms posited by contemporary neuroscientists can explain distinctively cognitive capacities.

I think we are now in a position to reject the background picture that is motivating many of Gallistel’s arguments, of a Manichean struggle between associationist and computationalist research traditions in psychology. One can of course trace various historical currents weaving through the complex tributaries of psychology, and I don’t

---

61 See Wang & Morris (2010) for a review.
mean to deny that there is *some* respect in which, for example, British empiricism is continuous with the neo-connectionism of the 1980s. However, these historical currents are multifarious, and criss-cross each other in complicated ways. Moreover, they often reflect sociological and cultural influences more than substantive theoretical commitments. It is true that there was a division between ‘classicist’ and ‘connectionist’ research traditions in artificial intelligence and cognitive science throughout the 1970s and ‘80s. The division was driven in part by competition for limited funding, but also reflected substantive theoretical disagreements. However, it’s a dramatic oversimplification to depict this division as a yawning theoretical chasm running the length of the history of psychology. More importantly, this caricature simply doesn’t substantiate strong claims about the theoretical limitations of contemporary neuroscience. Perhaps, if one squints hard enough, one might see certain similarities between the views of early behaviorists and the views of contemporary neuroscientists. But similarities are cheap; there might also be similarities between the views of certain behaviorists and the views of computationalists. I’ve pointed out, for example, that Hull seems to have endorsed a notion of representation that is very similar to Gallistel’s. The question we must ask ourselves, though, is whether these similarities are in a respect, and to a degree, that allow us to infer that the presumed flaws of one research tradition carry over to the other. I have argued that in the case of contemporary neuroscience and its alleged fellow travelers in an anti-representational, associationist tradition, they do not. To evaluate the explanatory adequacy of contemporary neuroscience, we must evaluate it on its own terms. To impugn it on the basis of its superficial resemblance to certain problematic views is just, as it were, to think by association.
4.4 Confusions About Computability

Now that the background picture of a Manichean struggle between associationism and computationalism has fallen away, we can begin to identify and evaluate Gallistel’s specific arguments against the explanatory adequacy of contemporary neuroscience on their own terms. However, our expository work is not yet done, since although we have divested Gallistel’s arguments of their misleading rhetoric, they are still shrouded in certain theoretical considerations about the nature of computation that lend them an appearance of plausibility they’re not entitled to. Thus, in this section, I’ll peel away these theoretical considerations. Then, in the next, we’ll finally be in a position to examine the substantive arguments that lie beneath.

Considerations about the nature of computation provide the theoretical foundation of Gallistel’s arguments against the explanatory adequacy of the connectionist mechanisms posited by neuroscientists. Gallistel argues that such mechanisms cannot function as an effective read-write memory, and therefore have the functional architecture of a finite-state machine rather than a Turing machine, along with all the concomitant problems with such architectures such as the problems of pre-specification and combinatorial explosion. Gallistel thus applies theoretical constructs like Turing machines — which have their proper home in computability theory, the theory of which functions can in principle be computed — to physical computing mechanisms. Accordingly, he takes his assertion that the computing mechanisms instantiated in the brain have the architecture of a Turing machine to have a strength and legitimacy that is consonant with the rigor and certainty associated with methods of proof employed in computability theory; specifically he takes this assertion to have the character of a physical or mathematical law: “the most fundamental aspects of the functional structure of a computer are dictated by the logic of computation itself... therefore, they will be observed in any powerful computational device, no matter what stuff it is made of” (Gallistel & King, 2010, p.167). Since neuroscientists allegedly assume that the
brain has the architecture of a finite-state machine, Gallistel (2008) holds that they must implicitly believe that “the brain escapes the limitations and requirements that computer scientists believe are imposed by mathematics, logic, and physics” (p.227).

Finite-state machines are of course abstract mathematical entities, not physical mechanisms, which is why they are amenable to mathematical proof. Thus it’s not obvious precisely which relation Gallistel has in mind when he claims that the ‘connectionist’ mechanisms posited by neuroscientists have the functional architecture of finite-state machines. Sometimes he seems to be claiming that connectionist mechanisms have the computational power of FSMs, in the precise sense employed within computability theory, viz. that connectionist mechanisms and FSMs compute precisely the same set of functions. For example, Gallistel & King (2010) write that in a physical mechanism with the architecture of a FSM, “the absence of a readable memory radically limits computational power” (p.177), and that the mechanism of LTP “does not give us even the computing power of a finite-state machine” (p.180). This might be dismissed as a loose usage of ‘computational power’, but in other contexts it’s hard to deny that when Gallistel characterizes a computational mechanism as having the functional architecture of a given formal model of computation, he really does mean that the computability-theoretic results pertaining to that model apply to it. For example, Gallistel (2011) writes: “Problems that may be solved without writing to symbolic memory are called regular problems. There are many such problems; however, there are many others that are routinely solved by animals and that do demand all the components of the Turing machine. They cannot be solved by a physically realizable finite-state machine, which cannot write to symbolic memory” (p.113).

If Gallistel really does think that the connectionist mechanisms posited by neuroscientists have the computational power of FSMs, he doesn’t explain why. To even make sense of this idea, we must first assume a canonical characterization of connectionist mechanisms in terms of a formal model of computation that is amenable
to the methods of proof employed in computability theory. One such characterization is the neural network model developed by McCulloch & Pitts (1943), which was subsequently generalized by Kleene (1956) into what we now know as the finite-state machine. Perhaps that is what Gallistel is thinking of when he suggests that connectionist mechanisms have the computational power of FSMs. But there have been significant developments in the field of neurally-inspired models of computation in the last seventy years, and McCulloch-Pitts nets are just one model amongst many; there is no single, canonical model of neural network computation, but rather many different models corresponding to different ways of idealizing and formalizing the properties of connectionist mechanisms, and these different models have very different computational power and complexity profiles.

Šíma & Orponen (2003) provide a helpful taxonomy of various different neural network models of computation, based on different formal properties such networks might possess — such as their architecture (feedforward versus recurrent), time representation (discrete versus continuous), state type (binary versus analog), network size (finite versus infinite), and so forth — and they use this to structure a summary of results about the computational power of these models. As they point out, there are many different neural network models with the computational power of Turing machines, including many that fall on the biologically realistic end of the spectrum of idealizations, such as certain kinds of spiking network (Maass, 1996). Indeed, there are notional neural network models that are more powerful than Turing machines (Siegelmann & Sontag, 1994). Gallistel provides no reason at all to think that connectionist mecha-

---

62 It is important in this context to distinguish carefully between a model of computation and a computational or mathematical model. A model of computation is an abstract characterization of a computational process, expressed using a formalization that is amenable to the methods of proof employed in theoretical computer science, used to analyze the computational power or complexity of that process. A computational or mathematical model is a mathematical characterization of an empirical system, perhaps implemented as a computer simulation, used to assimilate, organize, and guide experimental research into that system.
nisms are best described by a model of computation with the power of FSMs rather than some more powerful model.

One might reply that these results about the power of neural network models are based on idealizations, and therefore don’t apply to real-world mechanisms. For example, it is widely pointed out that the Siegelmann & Sontag (1994) results concerning the super-Turing capacities of analog networks depend on the assumption that the weights of such networks can have continuously many distinguishable values, an assumption that isn’t true of any physical system, since all such systems are susceptible to some degree of noise. Much the same point holds of many of the proofs of Turing universality in neural networks, which assume that the relevant networks can have indefinitely many units; clearly an unrealistic assumption if we suppose that these neural network models are implemented by physical mechanisms. But, of course, one could raise precisely the same point about Turing machines — as many have. The reason Turing machines are computationally universal is that they have access to a tape with indefinitely many cells, which is clearly unrealistic from the perspective of physical implementation. But to make too much of the idealizations made by models of computation like Turing machines, FSMs, and artificial neural networks would be

---

63 Gallistel & King (2010) raise a related worry when they question whether the reverberating states in the attractor network model of path integration proposed by Samsonovich & McNaughton (1997) could persist for long enough to explain the behavioral data, given theoretical results by Maass & Sontag (1999) allegedly showing that “introducing even modest amounts of noise into a net can destroy its functionality” (Gallistel & King, 2010, p.256). The citation of this result is puzzling, for at least two reasons. The first is that Maass & Sontag were investigating whether a certain type of highly idealized neural network model of computation is capable of recognizing certain languages in the face of noise, not whether the kind of biologically realistic spiking attractor networks employed by Samsonovich & McNaughton are capable of sustaining attractor states in the face of noise. These are very different problems, employing very different kinds of neural network models (cf. the distinction between models of computation and computational models discussed in note 62), and there’s no reason to think that the Maass & Sontag results have any bearing on questions about the stability of attractor network states. The second reason the citation of the Maass & Sontag result is puzzling is that there is a large theoretical and experimental literature on the stability of attractor network states, which Gallistel & King don’t address. I’ll discuss this second point in more detail in section 4.5.

64 For example, Rumelhart et al. (1986b) write that “real biological systems cannot be Turing machines because they have finite hardware” (p. 118).
to misunderstand the content and goals of the field of computability theory to which those models belong. The whole point of computability theory is to abstract away from physical resource constraints such as time, space, and energy, so as to limn the boundaries of which functions can be computed in principle; it imposes few constraints on how functions are in fact computed by physical mechanisms.

The proposal that connectionist mechanisms have the computational power of FSMs presupposes that such mechanisms have a unique formal characterization in computability-theoretic terms. We’ve seen reason to doubt that, assuming there were such a characterization, it would have the power of FSMs as opposed to Turing machines or even more powerful formalisms. The present considerations about the role of idealization in relating formal models of computation to physical mechanisms call into question the assumption that there is a single best computational formalization of connectionist mechanisms. Without constraints on idealization, a moderately complex physical mechanism might support many different formal characterizations, with different computational power or complexity profiles. As Eliasmith (2002) puts it, “matter simply has too many input, output and state transition relations to know which are the right ones for a given analysis” (p.6).

However, the present considerations suggest not just that it’s difficult to map computational models on to physical mechanisms for the purposes of analyzing their computational power, but something more problematic: that it’s irrelevant to the project of developing mechanistic explanations of cognition. The question of whether a given type of computational mechanism, under some suitable idealization, could in principle compute a given function for all of its arguments, is just not a question that cognitive scientists need to address for the purposes of explaining specific cognitive capacities. The kinds of questions that arise when evaluating whether a mechanism might explain a given cognitive capacity concern whether mechanism can in fact compute ecologically relevant partial functions under ecologically realistic time constraints. But these
are questions about which computability theory is silent. That’s not to impugn computability theory, just to point out that that theory has a distinctive set of explanatory goals and methodologies that are largely orthogonal to the project of explaining cognition. To illustrate the point here using a toy example, note that even if it could somehow be shown that a connectionist mechanism has the computational power of FSMs, and hence cannot decide a mirror language like $A^nB^n$ for all its elements, this would have precisely no bearing on the plausibility of the mechanism if the capacity that it were invoked to explain simply involves the computation of some finite subset of the language. Indeed, no physical mechanism could decide $A^nB^n$ for all its elements, since that language is infinite, and computing it would eventually outstrip the resources of any finite mechanism. Turing machines can compute $A^nB^n$ only because, qua abstract entities, their tape gives them access to infinity. But the same is true of many abstract neural network models, which can access infinity via their weights or nodes.

In a telling passage, Gallistel superficially seems to address this point, but in doing so reveals that when he writes of connectionist mechanisms having the functional architecture of finite-state rather than Turing machines, he is in fact not making a claim about computational power, nor even employing computability-theoretic notions. Gallistel & King (2010) write: “Arguments that the brain can’t be a Turing machine (due to its infinite tape) but instead must be a weaker computational formalism are spurious — what requires the Turing machine architecture is not an issue of unbounded memory, it is an issue of being able to create compact procedures with compact symbols” (p.109). Note that the ‘spurious argument’ the authors allude to is not the argument that I just developed; in arguing that neural mechanisms cannot ‘be’ Turing machines since Turing machines have an infinite tape, my point was not that neural mechanisms must instead ‘be’ a weaker computational formalism such as a FSM; rather, my point was that assertions of identity between physical mechanisms and formal models of computation are, strictly speaking, incoherent. Used literally, in accordance with
the theory from which they derive their meaning, ‘Turing machine’ and ‘finite-state machine’ refer to abstract mathematical objects, not physical mechanisms.

The passage from Gallistel & King suggests that when they write of ‘Turing’ and ‘finite-state machines’, they are not using these terms literally. Note that they blithely jump from writing about the properties of Turing machines (infiniteness), to writing of certain functional properties of mechanisms with a Turing architecture (efficiency), without noticing that their em-dash marks a conceptual chasm between mathematical abstracta and physical mechanisms. Indeed, despite their frequent allusions to computing power and their discussions of computability-theoretic results, I think it’s clear that when Gallistel & King claim that connectionist mechanisms have the functional architecture of finite-state rather than Turing machines, they mean that such mechanisms lack a certain functionally-individuated mechanistic component — a read-write memory. This interpretation should of course come as no surprise; it involves a straightforward reading of ‘functional architecture’, it is evidenced throughout the authors’ (2010) book — such as when they write that “Contemporary neuroscience is committed to the thesis that the brain has the functional architecture of a finite-state automaton, rather than that of a Turing machine: it lacks a read/write memory” (p.176) — and it makes sense of many of their central arguments. But note that if this is the correct interpretation of the authors’ use of ‘Turing machine’ and ‘finite-state machine’, they are simply using those terms in an idiosyncratic way, and the claims they make using those terms are not legitimized by computability theory. Still, they wouldn’t be the first to illicitly draw legitimacy from computability theory by appropriating these terms. Gallistel’s fellow traveler Zenon Pylyshyn writes:

The difference between an extremely complex device characterized merely as proceeding through distinguishable states (but not processing symbols) and what I call a ‘computer’ is precisely the difference between a device viewed as a complex finite-state automaton and one viewed as a variant of a Turing machine. I shall use the term finite-state automaton as a general way of talking about any device whose operation is described without reference to the application of rules to sym-
bolic expressions. Thus, in this view, the ‘new connectionist’ machines described by Anderson and Hinton (1981) are finite-state automata characterizations. (Pylyshyn, 1984, pp.70-71; citations provided therein)

Of course, one can use terminology however one likes, but one cannot expect that one’s preferred usage will preserve the pattern of inferences licensed by established usage. To describe a certain type of physical mechanism as having the ‘functional architecture’ of a finite-state machine, then object to it on the basis of computability-theoretic results about finite-state machines properly-so-called has, as Russell once put it, all the advantages of theft over honest toil. A mechanism with the functional architecture of a finite-state machine is just a mechanism that lacks a read-write memory, and has nothing in particular to do with the precise notion of a finite-state machine in computability theory. The same goes for Turing machines. It is tempting to associate mechanisms that have a read-write memory with Turing machines, because of the beguiling visual metaphor of a memory tape, but that association has no theoretical content. When computability theorists discuss the symbols on a Turing machine’s tape, and Gallistel writes of the symbols stored by read-write memory, they both use the word ‘symbol’, but they mean quite different things by it. Symbols in the computability theorist’s sense are just discrete, meaningless tokens, whereas symbols for Gallistel are potentially analog vehicles of information. Moreover, the fact that the dominant model of computational universality came to be one that can readily be visualized in mechanistic terms is a theoretically uninteresting accident of history; there are many models of computation that are equivalent to Turing machines, but which cannot be given a mechanistic gloss. Models of computation are effectively just formal devices for expressing functions in intension.

Gallistel, Pylyshyn and many others seem to think that Turing machines are a kind of blueprint for computational mechanisms, which capture the essence of what it is to
be such a mechanism.\textsuperscript{65} They suggest that researchers who are invested in exploring computational mechanisms that don’t look much like Turing machines are computational naïfs, ignorant of elementary results in theoretical computer science, results that have the character of mathematical or physical law. In particular, researchers investigating neural computation pay “too much attention to neuroscientific speculations about the neural mechanisms that supposedly mediate computation and not enough to well-established results in theoretical and practical computer science concerning the architecture required in a practical computing machine” (Gallistel & King, 2010, p.ix). These claims would baffle the legions of theoretical computer scientists and computer engineers who are investigating not only neural computation, but range of other ‘natural’ or ‘unconventional’ computational systems such as quantum computers (Gershenfeld & Chuang, 1998), chemical computers (Paun, 1998), analog computers (Siegelmann & Fishman, 1998), and many others (Calude et al., 1998). Indeed, some of the most knowledgeable and prolific theoretical computer scientists working today are active contributors to the field of computational neuroscience. It is strange, to say the least, to accuse these researchers of computational naïveté. Moreover, it is true that, beyond the theorems of computability theory, there are results about the nature of computing mechanisms that have something approaching the character of physical law,\textsuperscript{66} but Gallistel doesn’t mention them, and they don’t in any way support his conclusions about connectionist mechanisms of computation.

So, to reiterate, theoretical notions and results from computability theory seem to be a red herring when trying to identify and evaluate Gallistel’s arguments against

\textsuperscript{65}See, for example, Pylyshyn (1989), who writes that “[i]n Turing’s original theoretical machine, and in every real digital computer, a distinction is made between the processor and the memory” (p.56). This claim, incidentally, is quite false. Digital computers with microchips like FPGAs that contain a mixture of logic gates and RAM blocks are ubiquitous.

\textsuperscript{66}See, for example, pioneering work by Bennett (1973), Landauer (1961) and others on the thermodynamics of reversible computation.
extant models of neurocomputational mechanisms. We can see this in stark form when we observe that look-up tables, which in Gallistel’s view are the means by which most connectionist mechanisms perform computations, can in principle implement any function, even uncomputable functions; a look-up table could ‘compute’ the halting function simply by listing all the values of the function. Of course, Gallistel would presumably reply that such a look-up table would suffer from the problems of pre-specification and combinatorial explosion in spades; but that just goes to show that the relevant issues here have nothing to do with computability or associated theoretical notions like Turing or finite-state machines; rather, they have to do with certain functional requirements on practical computing mechanisms, namely that they require some mechanism for flexibly storing and manipulating information.

Again, we already knew that was the focal point of Gallistel’s arguments. The foregoing might thus begin to seem like a quibble about how to express that point — a pedantic injunction on using the term ‘finite-state machine’. However, the issue here is not merely terminological. Even though Gallistel’s arguments against connectionist computational mechanisms are ultimately not about computational power, but about the functional properties of such mechanisms, they nevertheless still trade on a conflation between the properties of mechanisms and abstract computational models. Recall the outline of Gallistel’s argument: the number of states of a FSM increases exponentially with the number of input-output mappings that it computes; thus, since connectionist mechanisms are FSMs, they cannot compute functions of any moderate degree of complexity without suffering from the problems of pre-specification and combinatorial explosion.

As it stands, this inference rests on an identification between connectionist mechanisms. But even if it could be re-expressed in terms of a less metaphysically obtuse relation such as implementation, the substance of the inference would still rest on the assumption that there is a transparent mapping from the abstract states of a finite-state
machine to the functional components of a connectionist mechanism. That assumption might seem innocent, since the state diagram of a FSM superficially resembles the wiring diagram of a neural network; it is tempting to identify the states and transition arrows of the former with the units and connections of the latter. But that would involve an egregious misunderstanding. If we’re interpreting the state diagram of a FSM as a representation of the function computed by a given physical system, the states correspond to global states of the system, not isolable functional components. Again, a FSM is just a compact mathematical description of a function, and the states & transitions used to describe a function needn’t map transparently on to the components of a physical system that, under some appropriate idealization, computes the function.

We can see this by considering a computational mechanism with the ‘functional architecture of a Turing machine’ in Gallistel’s sense, viz. a mechanism with a read-write memory. By virtue of its memory, the mechanism might be able to occupy combinatorially many distinct states, and to implement relatively complex functions. Nevertheless, the function that the mechanism implements will be describable by some FSM; remember, issues about computational universality here are moot. Given the complexity of the function, the FSM will have vastly many states. But, note, the number of states of the FSM does not directly correspond to the number of isolable functional components of the mechanism that implements the FSM. Some mechanisms, such as conventional engineered computers, implement FSMs without suffering from the problems of pre-specification and combinatorial explosion. So Gallistel’s argument that connectionist mechanisms suffer from those problems simply by virtue of implementing FSMs is a non sequitur.

To explain animal cognition, it certainly seems that we must posit mechanisms that are capable of flexibly storing and manipulating information. But in appropriating concepts and results from computability theory, Gallistel provides no reason to think that the mechanisms posited by contemporary neuroscientists are incapable of
flexibly storing and manipulating information. To examine the explanatory adequacy of connectionist mechanisms, we must examine them on their own terms, without distorted preconceptions about how computing mechanisms ‘should’ work. That is what I propose to do in the next section.

### 4.5 Problematizing the ‘Problems’

As we’ve seen, Gallistel’s arguments against extant neurocomputational mechanisms present a somewhat hazy and shifting target. Sometimes it seems as though Gallistel is arguing that such mechanisms are strictly *incapable* of encoding explicit information or of representing anything; on a closer look it appears that he’s arguing just that they’re ‘not very good’ at encoding explicit information, and that the representations they can instantiate are relatively meager. In some places Gallistel seems to hold that the central problem with connectionist mechanisms is that they must embody an implausible amount of innately pre-specified structure; in other places, Gallistel seems to hold that the central problem with connectionist mechanisms is that they presuppose an associative theory of learning that allows no role for innate constraints.

Moreover, Gallistel’s arguments are often guided by certain sociological or theoretical preconceptions that turn out on close inspection to be inaccurate or confused. *Pace* Gallistel, neuroscientists are not guided by the explanatory norms that motivated the behaviorist’s anti-representationalism; on the contrary, they are expressly investigating the mechanisms that mediate distinctively cognitive capacities like perception, memory, decision-making, and so forth. In doing so, they posit mechanisms that are explicitly characterized in computational terms, and analyze these mechanisms using mathematical tools derived from information theory and related fields. Although

---

67See, for example, Gallistel (2011, p.254).
there is a sense in which Gallistel & King (2010) are right when they claim that “the language and conceptual framework for symbolic processing is alien to contemporary neuroscience” (p.101), this sense involves a specific conception of a symbol as a discrete, word-like vehicle of conceptual content; neuroscientists are quite familiar with symbols as Gallistel conceptualizes them, namely as vehicles of explicit information, though Gallistel doesn’t recognize this because of his tendency to focus on certain ‘classical’ ways of talking and thinking about computation.68

This tendency often leads Gallistel to force connectionist mechanisms into classical categories, without examining the details of those mechanisms to determine whether they’ll fit. Consider, for example, Gallistel’s suggestion that neural circuits might function as digital logic gates, in the way that McCulloch & Pitts (1943) envisaged.69 This was recognized to be physiologically implausible even in 1943; neurons are noisy beasts, and might fire even when they don’t receive simultaneous spikes from all of their pre-synaptic ‘conjuncts’, and in any case there is no global clock signal in the nervous system to ensure that spikes line up in the same clock cycle.

Of course, Gallistel thinks that theoretical & practical computer science show that classical categories limn the very nature of computation, and impose ‘ineluctable’ constraints on computing mechanisms. But I argued in section 4.4 that theoretical computer science is in no way inconsistent with connectionist conceptions of computation, and indeed, many of the conceptual innovations that gave rise to contemporary models of neural computation derived directly from theoretical computer science. Gallistel also argues for the hegemony of classical categories by appealing to the development of

68Cf. the discussion of ‘sub-symbols’ in section 4.3.2.

69This suggestion appears at least twice in Gallistel & King (2010): first when they claim that the neurons in the I layer of the Samsonovich-McNaughton model of path integration function as three-legged AND-gates (p.251), and second when they consider a hypothetical mechanism by which neural information might be encoded via a compact coding scheme, then decoded by neural circuits “configured to behave as AND gates” (p.261).
computing technology: the fact that the functional architecture of conventional digital computers has remained unchanged over the past 60 years “suggests that this architecture is strongly constrained by the nature of computation itself” (Gallistel & King, 2010, p.145). One might think that this suggestion is about as compelling as an analogous suggestion about the QWERTY keyboard and the nature of typing. But even if the reasons the von Neumann architecture has become ubiquitous aren’t completely arbitrary, as indeed I think is plausible, those reasons might be multifarious, and speculating about the nature of such reasons doesn’t seem like a very sound basis upon which to argue that the entire field of neuroscience is fundamentally misconceived.

In any case, I agree with Gallistel that there are good reasons to suppose that animal brains contain a read-write memory mechanism — if we understand read-write memory purely functionally, as some sort of mechanism for storing and flexibly manipulating information. My worry is that Gallistel assumes too specific a structural conception of read-write memory, informed by the image of memory mechanisms in conventional human-engineered computers. In seeking to explain the cognitive capacities of animals by positing computational mechanisms, neuroscientists are not beholden to a parochial picture of computation that we get from computer engineering; they just need to explain the relevant capacities. Moreover, in replying to Gallistel’s arguments that the mechanisms posited by neuroscientists are fundamentally incapable of explaining those capacities, we are not required to provide complete, well-confirmed neuroscientific explanations of those capacities in full dress; cognitive and computational neuroscience are still very much in their infancy, and as yet we have nothing approaching a complete mechanistic explanation of cognition; our question is whether something like the kinds of mechanisms that neuroscientists currently posit might plausibly explain cognition. We should keep these points in mind as we now turn, finally, to evaluate Gallistel’s specific arguments against extant models of neural computation.
4.5.1 Information & Representation

The most fundamental question of course is whether currently envisaged neurocomputational mechanisms can function as a read-write memory — whether they can encode, store, and manipulate vehicles of explicit information. Gallistel’s claim that they cannot is the foundation of all his other objections. If neurocomputational mechanisms cannot store and manipulate symbols, they cannot compute with them, hence they cannot generate complex representational structures, hence they acquire information about complex relations in the environment — they cannot learn. Thus if we can show that currently envisaged neurocomputational mechanisms can store and manipulate symbols, we will have done much to defang Gallistel’s objections.

First, let’s consider whether the neurocomputational mechanisms recognized by neuroscientists can store information in a form that is both explicit, so that it can play a causal role in computation, and thermodynamically stable, so that it can persist for behaviorally realistic durations. The most obvious candidate for such a mechanism is synaptic plasticity, which has been implicated as a memory mechanism for more than a century. However, Gallistel argues that although synaptic plasticity might embody information for long durations, that information is bound up with presynaptic activity and can’t be extricated by postsynaptic processes; synaptic memory is stable but inexplicit. Another candidate is reverberating activity in recurrent networks. Gallistel allows that reverberating activity patterns are physically distinct, causally efficacious entities, but argues that there are open questions about whether they can persist in a stable form for appreciable durations, and that even if they can, they are volatile and hence metabolically expensive; reverberating memory is explicit but unstable.

However, Gallistel doesn’t consider the possibility that these memory mechanisms might be importantly complementary, even though this has been the received view in neuroscience since at least the publication of Hebb’s (1949) *Organization of Behavior*. The idea is that volatile, transient reverberating activity patterns are ‘stored’ in a
stable form by patterns of synaptic potentiation within a recurrent neural network, in the sense that these synaptic patterns embody a persisting disposition for the network to elicit and sustain a given reverberating pattern upon receiving a certain input. This input might be a fragmentary version of the pattern that initially imprinted the pattern of synaptic potentiation; in this way, a recurrent network can recall reverberating patterns through pattern-completion, functioning as a kind of content-addressable memory, as described in Chapter 2.4.4.

According to this picture, patterns of synaptic potentiation store information implicitly, but in an entirely unobjectionable sense of ‘implicit’: they simply store information in a dormant or non-volatile manner. This is in fact closely analogous to the manner in which conventional computers store information in non-volatile secondary memory, such as a hard disk. The component bit-strings that make up a file stored on your hard disk might be arbitrarily distributed throughout various physical locations on the disk, and only ‘store’ the file by virtue of their disposition to interact with the CPU in the right way. Although it is common to think of secondary memory in a conventional computer as a kind of filing cabinet, with distinct folders for each file, a more accurate metaphor would be one according to which fragmentary paragraphs of each file are arbitrarily distributed throughout various different folders, with labels indicating the location of the next paragraph. Files aren’t, as it were, just sitting there, waiting to be found, they are generated though the interaction of bit strings with the CPU. An analogous point holds of the information stored by patterns of synaptic potentiation. What is important in both cases is that the information that is stored ‘implicitly’ can be made explicit, and play a causal role in computational processes. In the case of synaptic potentiation, this is not achieved as Gallistel assumes, through a process that extricates the values of individual synapses from postsynaptic activity, but rather through a process by which dispositions embodied by patterns of synaptic
potentiation are realized.\textsuperscript{70}

It is strange that, in their extensive discussion synaptic potentiation and reverberating activity as mechanisms of memory, Gallistel & King (2010) overlook the possibility that these mechanisms might be interrelated; not only because of the ubiquity and historical significance of this idea, but also because the component pieces of the idea float just below the surface of their discussion without ever quite breaking the surface. I’ve mentioned repeatedly now that although Gallistel & King deny that synaptic weights encode explicit values, they allow that weights encode a “procedure that will \textit{generate different values} given different input vectors” (p.285, my emphasis). Thus they apparently allow that synaptic weights might embody a disposition to evoke vehicles that encode explicit values. Moreover, given their view that reverberating activity patterns are \textit{bona fide} vehicles of explicit information, it seems reasonable to impute to them the view that among the vehicles of explicit information that are generated by synaptic weights are reverberating activity patterns. Thus Gallistel & King would seem to \textit{implicitly} endorse the view that synaptic potentiation and reverberating activity might be complementary memory mechanisms, though they explicitly discuss these mechanisms as if they were entirely independent.

Another reason it is strange that Gallistel overlooks the functional relationship between synaptic potentiation and reverberating activity is that, as I suggested above, this relationship is also reflected in the memory mechanisms of conventional engineered computers. Looking more closely at this relationship will reveal an important lacuna in Gallistel’s conception of read-write memory. Recall that the von Neumann architecture embodies a distinction between \textit{primary} and \textit{secondary} memory. Primary memory

\textsuperscript{70} As David Kirsh (1990) points out, there is a pervasive tendency in cognitive science to conceptualize the explicitness of information in terms of the “bewitching image of a word printed on a page” (p.350). Kirsh deftly dispels this image, and replaces it with the \textit{process}-based conception of explicitness that I employ in the text.
stores information in a thermodynamically volatile form such that it can be quickly manipulated by the CPU, whereas secondary memory contains information in a stable, non-volatile form that is suitable for long-term storage. If your computer has ever crashed or lost power before you were able to save the document that you had been working on, you will have experienced the frustrations of losing information stored in primary memory before it could be saved to secondary memory. Similarly, if you have ever been interrupted whilst performing mental arithmetic before you could obtain and memorize the result, you will have experienced an analogous frustration. Indeed, the distinction between primary and secondary memory in computer engineering was explicitly modeled on the distinction between the psychological phenomena of working and long-term memory. While it is too simplistic to simply identify working memory with reverberating activity in recurrent networks, and long-term memory with patterns of synaptic potentiation, we have seen that reverberating activity and processes of synaptic plasticity & consolidation play essential explanatory roles in contemporary mechanistic models of working and long-term declarative memory, respectively.\textsuperscript{71}

This contrast between primary volatile memory and secondary stable memory reflects an important functional distinction that is not captured by Gallistel’s emphasis on a unitary, universal read-write memory mechanism. For Gallistel, read-write memory does double duty as a mechanism that directly interfaces with computational mechanisms to mediate the flexible manipulation of information \textit{and} as a mechanism that stores information in a thermodynamically stable form for indefinite durations. This is reflected in, for example, Gallistel’s suggestion that thermodynamic stability of molecules like RNA and DNA make the genetic machinery a plausible substrate for read-write memory. But flexibility and stability are conflicting desiderata; for a vehicle of information to be quickly and flexibly manipulated, it must in general be

\textsuperscript{71}For reviews of relevant literature, see Fuster (2009) and Wang & Morris (2010), respectively.
thermodynamically unstable, so that it can quickly transition from one stationary state to another without requiring a large investment of energy to move the system out of equilibrium. That is why RAM in a conventional computer is volatile, and susceptible to power disruptions.

It is worth reflecting on the implications of this point for the proposal that the genetic machinery provides a plausible read-write memory mechanism. The phosphodiester bonds in DNA are extremely strong, and allow DNA sequences to persist unchanged for thousands of years, which is surely in part why evolution has harnessed DNA as a vehicle of readable phylogenetic information. But that is precisely why DNA would, prima facie, be unsuitable as a vehicle of writeable ontogenetic information; breaking and forming phosphodiester bonds on the timescale required to explain fast, flexible working memory processes would presumably require vast amounts of metabolic energy, which one might have expected molecular biologists to have noticed by now. Indeed, the inflexibility of DNA is biochemical basis of the central dogma of molecular biology — it’s why you can’t train a zebra to be a monkey. Perhaps there might be some mechanism by which the apparent inflexibility of DNA could be overcome; but this would require a revision to the conceptual repertoire of molecular biologists as radical as the revision that Gallistel thinks is needed in neuroscience. In any case, in the absence of a specific proposal about the mechanism by which DNA could function as a flexible read-write memory, and in the absence of decisive reasons to think that currently recognized neural mechanisms could not, the genetic memory hypothesis does not presently seem to be one that we need to take seriously.\textsuperscript{72}

\textsuperscript{72}RNA might be marginally more plausible as a mechanism of flexible read-write memory, since RNA nucleotides have an hydroxyl group that catalyzes the breaking of phosphodiester bonds, making this process less metabolically costly, and thereby making RNA more thermodynamically flexible. That is why RNA sequences degrade quickly, but DNA sequences can persist for thousands of years. Still, even if RNA is more flexible than DNA, the central point still holds that there is no known mechanism by which this flexibility could be harnessed so as to rapidly encode specific items of information. I mentioned in the introduction to this chapter that there was a burst of experimental work on
From the present perspective, then, to complain that reverberating patterns of activity are not thermodynamically stable, and that patterns of synaptic potentiation are not computationally efficacious, is just to conflate two different functional roles that memory mechanisms might play. Nevertheless, Gallistel might hold that although reverberating activity and synaptic plasticity might play complementary roles in principle, in practice neither plays their role very well. Indeed, as we saw in Chapter 3.6, Gallistel & King (2010) argue that although the simulations of the Samsonovich-McNaughton exhibited the formation and persistence of a reverberating pattern, this is “at best weak evidence that such a mechanism exists and actually works in any real brain” (ibid., p.258), since the simulations involved various simplifications and idealizations. They point out that in real brains, whether a reverberating pattern forms and persists depends on a very specific range of physiological parameters. But this is neither here nor there; Gallistel’s argument requires decisive evidence that reverberating patterns in attractor networks cannot function as memory states, not the absence of decisive evidence that they can. At the very least, it requires evidence that neuroscientists have simply ignored difficult empirical questions about formation and persistence of reverberating attractors. But, on the contrary, there is ongoing theoretical and experimental research into the conditions under which reverberating patterns form and persist, and the mechanisms by which they are manipulated. This work explores “how sustained activity can be stable in the presence of noise and distractors, how different synaptic and voltage-gated conductances contribute to persistent activity, how neuromodulation could influence its robustness, how completely novel items could be maintained, and how continuous attractor states might be achieved.”

molecular memory mechanisms in the 1960s and ’70s, which explored the possibility that memory might be encoded in DNA or RNA sequences (e.g. McConnell, 1962). This work was problematic not only because the initially promising results failed to be replicated, but also because no plausible mechanism was ever proposed by which memory could be encoded by nucleotides.
Researchers have in fact reported direct neurophysiological observations of reverberating activity patterns in mammalian cortex in vitro, and demonstrated how those patterns can be manipulated by synaptic inputs (Shu et al., 2003). More interestingly, given the earlier suggestion about the functional relationship between synaptic plasticity and reverberating patterns of activity, researchers recently provided strong support for the view that reverberating patterns are stored dispositionally by patterns of synaptic potentiation in a content-addressable manner by perfusing slices of rat hippocampus with a voltage-sensitive dye, stimulating the preparation at two different sites, thereby eliciting a pattern of electrical activity across the slice as revealed by the dye, then inducing LTP in the preparation using a burst of theta stimulation. They showed that subsequent stimulation at one of the original sites was sufficient to elicit the complete pattern, which apparently had been stored by the induction of LTP (Jackson, 2013). These observations were made in vitro, but together with the large body literature from which they were drawn, they would seem to provide reason to be optimistic that reverberating patterns of activity play an important functional role in the brains of awake, behaving animals. Gallistel doesn’t discuss any of this literature, so it is unclear what reason he has to be pessimistic.

Similar points can be made about the functional role of synaptic plasticity. As we saw in Chapter 3.6, Gallistel & King argue that ‘long-term’ potentiation is generally not long enough to encode long-term memories. But there is a vast body of theoretical and experimental research into the mechanisms by which the potentiated synapses

---

73 It should be noted that the experimenters did not specifically address the question as to whether the electrical patterns of activity they revealed were stable attractors. However, their results are certainly consistent with that possibility. Moreover, the specific region of the hippocampus they investigated, the CA3 region, is known to be densely recurrent, and has long been hypothesized to function as an autoassociative attractor network. This hypothesis originates with Marr (1971), and is one of the central motivations for the Samsonovich & McNaughton (1997) model.
mediating labile memory traces might be consolidated into long-term memory (Sossin, 2008), the role that reconsolidation processes might play after long-term memory traces have been reactivated during memory retrieval (Dudai, 2006), the mechanisms that allow memory traces to persist over long durations despite interference from other memory traces (Leibold & Kempter, 2008), and the role that various forms of non-synaptic neural plasticity might play throughout all of these processes (Mozzachiodi & Byrne, 2010). Again, Gallistel doesn’t mention any of this research, so it’s not clear what reasons he has for setting his face against it.

However, even if we grant that currently recognized neurocomputational mechanisms might be capable of storing and manipulating explicit information, we might still wonder how efficiently that information is encoded. Gallistel vacillates between the claim that neuroscientists have simply ignored the question of how information might be encoded in neural vehicles, and the claim that the answers neuroscientists have provided to that question presume radically inefficient coding schemes. Still, either claim might appear to be problematic — but only if one were to ignore the fact that research into the efficient coding of information in the nervous system using quantitative information-theoretic tools has been at the heart of theoretical and experimental neuroscience since the 1950s (e.g. Mackay & Mcculloch, 1952). One of the most seminal exemplars of such research is the ‘efficient coding hypothesis’ of Barlow (1961), who proposed that the structure of neural circuitry is adapted to the statistical structure of the environment, so as to maximize mutual information between neural vehicles of information and environment stimuli, thereby minimizing the neural resources needed to encode those stimuli. This idea has rippled throughout neuroscience and inspired a vast body of apparently fruitful research; as a recent reviewer put it, the “principle is extremely effective in predicting the empirical characteristics of classical receptive fields and provides a principled explanation for sparse coding and the segregation of processing streams in visual hierarchies. It has been extended to cover
dynamics and motion trajectories and even used to infer the metabolic constraints on neuronal processing” (Friston, 2010, p.131).

Against the backdrop of a body of research that uses rigorous, quantitative tools to explore how specific neurons or neural circuits actually encode information, research which has apparently elucidated many otherwise puzzling properties of the nervous system, Gallistel’s sweeping claims about how the nervous system must encode information, based on abstract efficiency considerations that are unconstrained by facts about the source of the information and what it is being used for, seem singularly uncompelling. Contrary to Gallistel’s claim that the coding schemes employed by the nervous system must be highly efficient, compact schemes like binary that are capable of encoding an effectively unlimited number of values (Gallistel & King, 2010, e.g. pp.74, 284), we shouldn’t expect, for example, that ants capable of foraging over distances of tens to meters are equipped with a path-integration mechanism that can encode distances of billions of kilometers.

In fact, contemporary developments in the research tradition inspired by the efficient coding hypothesis suggest a more resounding critique of Gallistel’s assumptions about neural coding, for they threaten to replace the simplistic, feedforward picture of world-brain communication that Gallistel assumes (depicted in Figure 2.1) with a more nuanced model of neural information processing that pays more than lip-service to the view that the brain is a probabilistic mechanism operating in an uncertain world. According to the predictive coding hypothesis and the broader ‘Bayesian brain’ framework from which it derives, the brain does not construct structured representations of the world simply by accumulating sensory information that percolates bottom-up through a hierarchy of increasingly sophisticated processing modules. Rather, the brain attempts to optimize generative statistical models by predicting, top-down, incoming sensory information. Sensory information on this view is not simply a signal that passes passively through a communication device like a radio or a television; rather, it
encodes prediction errors that the brain actively uses to update its generative models. According to the complementary *probabilistic coding hypothesis* within the Bayesian brain framework, neural vehicles of information do not encode specific values of environmental variables, but rather probability distributions defined over those variables. There is extensive theoretical and experimental work on the mechanisms of predictive and probabilistic coding, but again, Gallistel doesn’t discuss any of it. Gallistel & King (2010) do briefly mention a review of probabilistic coding by (Knill & Pouget, 2004), however they do not discuss how their claims about the radical implausibility of contemporary models of neural coding relate to the many specific proposals about the mechanisms of neural coding discussed within that paper. Those proposals would *appear* to describe plausible mechanisms by which information is efficiently encoded by neurons, and Gallistel provides no reason to think that those appearances are merely illusory.

### 4.5.2 Computation

Still, even if explicit information could be efficiently encoded in the ways suggested, how might the vehicles of that information be computationally manipulated? Information is impotent if it can’t be manipulated. Gallistel identifies three increasingly sophisticated ways in which computational mechanisms might flexibly manipulate information, and thereby mitigate the problems of pre-specification and combinatorial explosion. In his view, these three grades of computational flexibility can only be achieved by certain distinctively ‘classical’ computational mechanisms — most importantly, an indirectly addressable read-write memory. I don’t think that it behooves neuroscientists to find identify direct neural correlates of these mechanisms, but I do think that they must show how the psychological capacities allegedly mediated by these mechanisms can in

---

74See Huang & Rao (2011) and Pouget et al. (2013), respectively.
fact be explained in neuroscientific terms.

The first grade of computational flexibility concerns how novel vehicles of information might be quickly constructed and stored in such a way that they can be accessed as needed. Gallistel’s central argument for the claim that neural mechanisms cannot construct novel vehicles of information is based on his assumption that neural networks have the functional architecture of a finite-state machine, which have a fixed, ‘pre-specified’ set of states. But we’ve seen that that assumption is unfounded; there is no theoretically substantive sense in which neural networks literally are FSMs, and moreover they don’t have the functional architecture of FSMs either, since by Gallistel’s own admission there’s every reason to think that they can embody vehicles of explicit information.

Gallistel attempts to support the claim that neural networks have the architecture of FSMs by quoting a prominent computational neuroscientist as holding that input to a neural network changes “the transition function governing how the machine switches from one state to the next” (Koch 1999, p.470, quoted by Gallistel & King 2010, p.176). But if anything that undermines the claim that neural networks have the functional architecture of FSMs, since FSMs have a fixed transition function, and a fixed set of states. The field of neural computation has, from its inception, focussed on how neural networks might acquire novel states and transitions as a function of their input. In particular, a propos of previous points about the most plausible vehicles of explicit neural information, there has been a great deal of research into how biologically realistic recurrent attractor networks might acquire novel states, or attractors, and store them in such a way that they can be readily accessed (Rolls, 2010). I’ll postpone questions about the principles, or learning rules, by which attractor networks might acquire novel states until the next subsection on learning; the relevant point here is that attractor networks have the capacity to construct such states, store them, and quickly access those states as needed. Recall from Chapter 2 that attractor networks
function as a kind of *content-addressable memory* in which fragmentary input can elicit states that are stored in the synaptic matrix of the network. This recall of states from memory can occur extremely quickly (Battaglia & Treves, 1998), and the number of states that can be reliably recalled scales favorably with the number of states stored in the network (Risau-Gusman & Idiart, 2005). Thus attractor networks would seem to avoid the most fundamental manifestation of the problems of pre-specification and combinatorial explosion.

Gallistel & King (2010) objects to ‘learned content-addressable memory nets’ on the basis that they are look-up tables, and require “a physically distinct memory location for every possible input string” (p.94). This inference is another instance of Gallistel’s tendency to straightforwardly *identify* certain neuroscientific mechanisms with familiar constructs from theoretical computer science or computer engineering, without looking closely at how such mechanisms actually work. As with all such inferences, it carries the risk of equivocation; the fact that attractor networks function in ways that are analogous to a content-addressable memory doesn’t entail that they have all the properties of the engineered content-addressable memories that we’re familiar with. When we look more closely at how attractor networks work, it is clear that they can in fact map different inputs to the same attractor state; that is the basis of their apparent capacity for generalization. In an earlier paper, Gallistel (2006) recognizes this, but strangely seems to spin it as another objection to attractor networks: “The number of distinct stable attractor states is... much less than the number of distinct inputs. Thus, many different inputs cause the system to settle into any given stable attractor states” (p.65). Gallistel would seem to object to attractor networks *regardless of their properties*; but that just suggests that we needn’t take his objections too seriously. Indeed, given the large body of rigorous quantitative research into the storage and recall capacities of attractor networks issuing from seminal work in statistical physics by Hopfield (1982), and the many apparently promising applications of attrac-
tor network models to explain a diverse range of behavioral capacities (Eliasmith, 2005; Rolls, 2010), in order to develop a serious objection to the explanatory adequacy of attractor networks one would presumably need to critically engage with this research. But, again, Gallistel doesn’t mention any of it.

The second grade of computational flexibility concerns how information might be integrated across time and space. Even if we have an account of how novel vehicles of information are constructed and retrieved as needed, this doesn’t explain how ‘dribs and drabs’ of information received in different areas of the brain at different times might be integrated into a unitary representation with structured content. An important perspective on this problem is provided by recent work on probabilistic inference in the brain that I hinted at at the end of the previous sub-section. This work is one manifestation of the broader influence that Bayes’ rule has had in cognitive science in recent years. Bayes’ rule provides a general framework for understanding how probabilistic information from different sources is integrated together with prior information to form a statistically optimal estimate of the probability of a given state or event; when the sources of information are Gaussian distributions over continuous variables, and estimates are made across time by integrating previous estimates, this process is called Kalman filtering.

There is abundant psychophysical evidence that animals integrate information from different sensory modalities into a robust multisensory percept in a roughly Bayes-optimal fashion (Ernst & Bülthoff, 2004), and research into the neural mechanisms that mediate this kind of probabilistic inference is currently among the most active areas of computational neuroscience. There are several models of how probabilistic information might be encoded in neural activity, and corresponding proposals about how that information is combined to perform probabilistic inference (Pouget et al., 2013). One popular family of models employs linear probabilistic population coding, a scheme by which a potentially nonlinear probability function is encoded as a linear
combination of (radial) basis functions, each of which is implemented by the tuning function of a given neuron in a radial basis function network, such that the activity level of a given neuron reflects the degree to which its corresponding basis function contributes to the probability function (Zemel et al., 1998).\(^7\) This method of encoding and decoding probabilistic information in a radial basis function network is depicted in Figure 4.2. One advantage of this coding scheme over other schemes for encoding probabilistic information is that it makes probabilistic inference computations relatively straightforward; computing the product of two likelihood functions to obtain a posterior distribution simply involves *summing* the activities of the populations encoding the likelihoods. Other schemes for encoding probability distributions are more straightforward, but make probabilistic inference computations more complicated.

When radial basis function networks are equipped with recurrent connections, they can support attractor dynamics. Deneve et al. (2001) describe how basis function networks with attractor dynamics can implement the kind of probabilistic inference computations that seem to be involved in multisensory integration; in particular, they show how such networks can, at least in principle, perform two crucial aspects of multisensory integration: *coordinate transformation* and *reliability estimation*. Sensory information from a given modality is encoded in a frame of reference specific to that modality, and different modalities employ different frames of reference. Thus before information from different sensory modalities can be integrated, it must be remapped into a common frame of reference; this requires coordinate transformation. Moreover, different sensory modalities have different levels of reliability, and reliability can change dramatically across different contexts; for example, vision is generally more re-

\(^7\) Any continuous function can be approximated to an arbitrary degree of precision by the summation of a set of *basis functions*. Perhaps the most famous set of basis functions is the *Fourier basis*, which consists of a set of sinusoids of increasing frequency. Radial functions are, intuitively, functions with a single peak or trough. Radial basis functions are radial functions that provide a basis for decomposing some continuous function.
Figure 4.2: The method of encoding and decoding probability functions via linear probabilistic population coding in a radial basis function network. A probability function is decomposed into a sum of radial basis functions, which are implemented by the tuning curves of the neurons in the network. The activity level of a given neuron reflects the extent to which the corresponding basis function contributes to the composition of the probability function. From Fig. 1 of Zemel et al. (1998, p.406).

liable than audition for localizing objects in daylight, but not at night. Thus it would be adaptive for multisensory integration processes to flexibly assign levels of credence to sensory modalities depending on their reliability; this requires reliability estimation. Basis function networks excel at coordinate transformations, and the attractor dynamics provided by the addition of recurrent connections allows such networks to perform optimal reliability estimation. In more recent work, Denève et al. (2007) have extended this model to implement Kalman filtering, thus allowing the model to optimally integrate dribs and drabs of information over time.

This work on probabilistic inference elucidates the computational principles by

---

76Specifically, it allows them to closely approximate maximum-likelihood estimation (Deneve et al., 2001).
which information might be integrated across time and space, but it doesn’t address the specific physiological mechanisms by which distant populations of neurons distributed throughout the brain might dynamically coordinate their activities with one another. As I mentioned in section 4.3.3, it is widely thought that the transient synchronization of gamma oscillations in distant neural populations mediates communication between those populations. However, it has not been entirely clear how synchronization allows the specific information encoded by a population of neurons to be communicated, nor how synchronization is modulated so as to dynamically route information flow from one population to another. Again, these questions delineate one of the most active areas of contemporary computational neuroscience, and have inspired a number of promising proposals. For example, Akam & Kullmann (2010) describe a mechanism by which a simple spiking network with a recurrent layer of interneurons that receives projections from several distinct neural populations can function as a filter that gates the transmission of population-coded information from the various afferent populations depending on whether those populations exhibit oscillatory activity. The authors’ simulations of their mechanism show that switching a given population from an asynchronous to an oscillatory state “allows accurate selective transmission of population-coded information” (p.308).

Thus recent work on neural computation suggests that the mechanisms of synaptic plasticity, probabilistic population coding, attractor networks, and long-range neural synchronization might play importantly complementary roles, collectively allowing the brain to flexibly integrate information across time and space into a unified structured representation, thereby mitigating the problems of pre-specification and combinatorial explosion.

77See, for example, Fries (2009), Kumar et al. (2010), and Wildie & Shanahan (2011).
4.5.3 Learning

The previous points about the neural mechanisms of probabilistic inference are also pertinent when evaluating Gallistel’s arguments against contemporary neuroscientific conceptions of learning. Most of those arguments are based on the assumption that neuroscientists in fact presuppose a simplistic (non-cognitive and arational) associative theory of learning. That assumption is motivated by the view that when neuroscientists propose that synaptic plasticity phenomena like LTP play a role in learning and memory, they are attempting to reduce learning to synaptic potentiation. For example, Gallistel & King (2010) claim that changes in synaptic conductance are “almost universally assumed to be the neurobiological realization of the psychologists concept of an association” (p.212), and Gallistel & Gibbon (2001) claim that “[n]eurobiologists suppose that [new associations] are modified synapses between neurons” (p.146, my emphasis). However, these authors do not provide any textual evidence for their claims. In fact, as is clear from histories of the development of the hypothesis that LTP plays a role in learning and memory (Craver, 2003), and from contemporary scientific reviews of that hypothesis (Martin & Morris, 2002), neuroscientists since the 1970s have characterized LTP and other forms of neural plasticity as playing an important role in multilevel memory mechanisms, not as being identical to specific memories.

I’ll return to this point in a moment, but the point I wish to emphasize here is that if we drop the assumption that neuroscientists presuppose a simplistic associative conception of learning, Gallistel’s arguments that the mechanisms neuroscientists posit are incapable of explaining anything but simplistic associative forms of learning such as ‘reflex modification phenomena’ are remarkably thin on the ground. One such argument, suggested by Gallistel & Matzel (2013), is that extant neuroscientific mechanisms are incapable of explaining rational processes of optimal probabilistic inference guided by innate prior constraints, and are restricted to explaining arational, mechanistic processes that are merely “recapitulative of the input-output conditions” imprinted by experi-
ence (p.170). Why? Gallistel & Matzel don’t say, but a hint is provided when, in an earlier paper, Gallistel (2008) claims that neural network architectures are incapable of encoding probability distributions, and hence are incapable of performing the computations over such distributions that constitute probabilistic inference.\textsuperscript{78} This is because neural networks allegedly cannot encode explicit information. But we’ve seen that the general claim that neural networks cannot encode explicit information is false, by Gallistel’s own admission, and that there is a large body of research into precisely how neural networks can encode probabilistic information so as to implement probabilistic information, which Gallistel simply doesn’t discuss. In fact, the closest brush that Gallistel has with the relevant literature is a passing reference to a paper by Knill & Pouget (2004) in the context of claiming that “it is natural to model neural computation in a Bayesian framework” (Gallistel & King, 2010, p.37). This claim would seem to be an implicit retraction of his (2008) claim, and a concession that neural networks can perform probabilistic inference.

The only other direct argument that Gallistel makes in support of the claim that currently recognized neural mechanisms are incapable of explaining distinctively cognitive forms of learning is that the only biologically plausible learning rules that those mechanisms can implement are incapable of explaining even ‘classical’ respondent conditioning. Currently recognized mechanisms of synaptic plasticity can allegedly only implement \textit{unsupervised} learning rules like Hebbian learning, which simply imprint an associative relation on the basis of a fixed ‘window of associability’ between stimuli. By contrast, respondent conditioning is a \textit{time-scale invariant} phenomenon that depends on the animal learning a predictive relation between stimuli. Supervised learning rules like backpropagation might \textit{in principle} be capable of explaining respondent condition-

\textsuperscript{78} “A net’s inability to specify a probability distribution over possible states of the world... makes it difficult to combine the evidence from different computational procedures” (Gallistel, 2008, p.235).
ing, according to this argument, but they are radically biologically unrealistic.

Gallistel is correct that respondent conditioning is far more cognitively sophisticated than psychologists have traditionally characterized it as being, and has probably done more than anyone to demonstrate this. But, keeping in mind that processes of synaptic plasticity like LTP needn’t be identified with the formation of associative relations at the behavioral level, there are in fact open empirical questions about the learning rules that synaptic plasticity might mediate, and the broader role that synaptic plasticity might play in learning phenomena (McClelland, 2006). Several biologically realistic variants and extensions of fundamentally Hebbian learning algorithms have been developed, some of which incorporate error-correcting feedback processes, such as the Contrastive Hebbian Learning Algorithm developed by O’Reilly (1996). Even backpropagation is arguably not as biologically implausible as it is widely assumed to be, as there is neurophysiological evidence that LTD induced at certain synapses can propagate back up the axon of the presynaptic neuron to affect the synapses on its dendrites, in ways that are consistent with the backprop algorithm (Fitzsimonds et al., 1997).

Perhaps more importantly though, the last couple of decades have seen an explosion of research into the neurophysiological correlates of various formal models of reinforcement learning developed in the fields of machine and animal learning. Perhaps the most important of these is the temporal difference learning algorithm (TD) developed by Sutton (1988). This is effectively an extension of the Rescorla-Wagner model (RW), and has been used to model certain features of respondent conditioning that are not captured by RW (Sutton & Barto, 1990). Evidence that the TD algorithm might be implemented in the brain was provided by experimental evidence that the phasic firing of dopamine neurons in the primate midbrain encodes the difference between the actual reward that an animal received from a stimulus and the reward that it expected to receive, in the manner required by TD learning (Schultz et al., 1997). In the wake of this
evidence there have been ongoing efforts to elucidate the role of dopamine in reinforcement learning, which have amply supported the hypothesis that phasic dopaminergic firing encodes the reward prediction error signal of TD.\(^7^9\) Important support for this hypothesis was recently provided by Tsai et al. (2009), who showed that the elicitation of phasic firing in dopaminergic neurons in the midbrain of rats using optogenetic methods was sufficient to induce a conditioned preference for a specific location in the rat’s environment.

Given that research into the neurophysiological mechanisms of TD and other reinforcement learning algorithms inspired by the Rescorla-Wagner model has been at the center of mainstream neuroscience for decades, Gallistel’s claim that “contemporary neuroscientific research on the neurobiology of association formation proceeds as if contemporary [i.e. post-Rescorla-Wagner] associative theory... did not exist” (Gallistel & King, 2010, p.231) is puzzling.\(^8^0\) What’s more puzzling is that Gallistel & Matzel (2013) discuss some of the neuroscientific literature on temporal difference in an apparently approving manner, suggesting that they agree that this research might elucidate the mechanisms of conditioning. Certainly, they don’t provide any criticisms of neural implementations of TD as an explanation of conditioning. So given the aforementioned literature, and absent any criticisms, it would seem to be reasonable to believe that currently recognized neural mechanisms might implement the kind of reinforcement learning algorithms that plausibly mediate conditioning. Thus both of Gallistel’s direct arguments against the explanatory adequacy of the learning mechanisms recognized by contemporary neuroscientists would seem to be problematic — and indeed,

\(^7^9\)See Schultz (2013) for a recent review.

\(^8^0\)Another reinforcement learning algorithm that owes a debt to the Rescorla-Wagner model and that has found correlates in neurophysiological mechanisms is the Pearce-Hall (PH) model of respondent conditioning (Pearce & Hall, 1980). Although the evidence that both the TD and PH algorithms are implemented in the brain might appear to conflict, there is every reason to think that these learning processes are in fact complementary (Roesch et al., 2012).
Gallistel himself seems to recognize as much since he seems to implicitly agree that currently recognized neural mechanisms are capable both of performing probabilistic inference and of implementing powerful forms of reinforcement learning.

Finally, to tie together some of the points about the neural mechanisms of learning and memory discussed in this chapter, it might be helpful to consider some concrete examples of how, contrary to some of Gallistel’s suggestions, neuroscientists in fact are self-consciously invested in the project of explaining distinctively cognitive forms of learning and memory. An important recent trend in the literature on neurophysiological correlates of reinforcement learning algorithms is to characterize the neural mechanisms of reinforcement learning in terms of what in machine learning is called *model-based learning*, which involves learning on the basis of a representation of the structure of the learning problem. The distinction between model-based and model-free versions of reinforcement learning algorithms has been applied in the context of neuroscientific explanations of learning to explicate the distinction between distinctively cognitive, goal-directed learning and habitual or procedural learning (Daw et al., 2005). As we’ve seen, there are good reasons to suppose that at least some forms of conditioning are probably best characterized as distinctively cognitive, model-based forms of learning, and in particular that conditioning might often involve learning certain spatial and temporal relations between events in an animal’s environment, such as where and when certain food rewards appear, in ways that are mediated by episodic memory. Interestingly, given the role of the hippocampus in ‘knitting together’ the perceptual elements of episodic memories with their background spatiotemporal context as discussed in section 4.3.3, there is increasing evidence that the hippocampus plays a crucial role in model-based reinforcement learning (Johnson et al., 2007). In an interesting recent experiment that is consonant with this hypothesis, Tse et al. (2007) set out to examine how an animal’s set of prior expectations or ‘mental schemas’ about certain relations in the environment might influence its capacity to subsequently learn about
those relations, in the context of the background theory that hippocampally-mediated episodic memory traces undergo a usually slow process of consolidation throughout the modality-specific cortical networks that mediate the various components of the memory. They first trained rats to have certain expectations about which kinds of smells occur in which locations, and found that rats were able to quickly learn novel smell-location combinations that were consistent with their expectations. This was reflected in a very fast process of neocortical consolidation revealed by neurophysiological probes, suggesting that “systems consolidation can occur extremely quickly if an associative ‘schema’ into which new information is incorporated has previously been created” (p.76). Neuroscientists thus seem to be on their way toward explaining distinctively cognitive forms of learning.

I recognize that much of the foregoing discussion of the literature on neural mechanisms of representation, computation and learning has been superficial. This is partly out of necessity, due to space constraints, but largely reflects the nature of my project. I am waving my hands, but not because I am engaging in sleight of hand, trying to pass off a schematic account of some possible neurocomputational mechanisms as a concrete proposal about how such mechanisms actually work. Rather, I am gesturing at a large body of research that at least appears to provide the rudiments of what might one day be developed into complete explanations of cognitive capacities. To argue that such research cannot begin to explain cognition, one would presumably need to identify specific aspects of this research and explain why they’re problematic. Gallistel just doesn’t seem to be looking in the right direction.
Chapter 5

Deconstructing Structural Representations

A picture held us captive. And we could not get outside it, for it lay in our language and language seemed to repeat it to us inexorably.

— Ludwig Wittgenstein, * Philosophical Investigations*

5.1 Introduction

So far we’ve seen that although Gallistel sometimes *seems* to argue that the kinds of mechanisms recognized by contemporary neuroscientists cannot store and manipulate information, and hence cannot participate in the functioning homomorphisms that are constitutive of representation, when we inspect his arguments closely it becomes clear that the most he could be arguing is that such mechanisms are not very *efficient* at storing and manipulating information. For example, although he once asserted that the Samsonovich & McNaughton (1997) model of path integration exemplifies a computational architecture “that does not have a symbolic memory” (Gallistel & King, 2010, p.242), he points out in a recent paper that *all* neural network models of path
integration, including the Samsonovich-McNaughton model, “are symbolic models” (Gallistel, 2011, p.254).\(^1\) His complaint about the Samsonovich-McNaughton model is not that it lacks symbols, but that the symbols it employs are “not very good” (ibid., p.254). However, we just saw in the previous chapter that when it comes to evaluating whether a hypothesized biological mechanism plausibly explains a given cognitive capacity, bald assertions about whether the mechanism is ‘any good’ that are unconstrained by detailed considerations about the precise nature and limitations of the capacity to be explained, the phylogenetic and physical resource constraints that the mechanism might face, and so forth simply don’t cut any ice.

So Gallistel’s arguments don’t live up to their promise of showing that contemporary connectionist neuroscience is somehow anti-representational. An argument that might live up to that promise, in a way that’s highly congenial to Gallistel’s views, is provided by William Ramsey (2007). While Gallistel attempts to argue that the mechanisms recognized by neuroscientists cannot physically function as representations, Ramsey argues that they cannot function as representations in an explanatory sense. Ramsey develops his argument for purposes that are ultimately antithetical to Gallistel’s, for he thinks that contemporary neuroscience can plausibly explain cognitive capacities despite the fact that it doesn’t posit representations. His argument thus poses a challenge to representationalism: the view, at the hub of the cognitive revolution, that cognitive capacities can only be explained by positing mental representations. Nevertheless, both Ramsey and Gallistel agree that contemporary neuroscience posit mechanisms that, despite what their proponents might think, are not genuine representations.

---

\(^1\)One might think that it would be most charitable to assume that Gallistel simply changed his views about whether the S-M model contains a symbolic read-write memory between his (2010) book and the later (2011) paper. However, just a few pages after the quote mentioned in the text from the (2010) book, Gallistel & King assert that the S-M model contains circuitry that “constitutes a read/write memory” (p.259). So the most charitable interpretation seems to be that Gallistel is simply unclear in his own mind about whether or not the S-M model — and neural network models more generally — contain a read-write memory.
To properly understand Ramsey’s argument, we should look at it against the backdrop of previous anti-representationalist arguments. Since the mid-nineties, proponents of various non-classical theoretical perspectives in cognitive science, such as situated robotics and dynamical systems theory, have argued that large swaths of cognition can be explained without appealing to representations at all. However, these approaches have tended to be most successful when explaining relatively simple sensorimotor skills, and it’s unclear whether they’ll scale up to explain the kind of stimulus-independent cognitive capacities, such as reasoning and planning, for which representational explanations have seemed most compelling. Moreover, proponents of these approaches often target their anti-representationalist arguments at the kind of syntactically structured symbolic representations invoked by early ‘classicist’ cognitive theories, yet neglect to identify fully general conditions for something’s being a representation, so it is unclear whether their objections apply to representations as such, or merely representations of a certain kind.\(^2\) Together these considerations motivate the following kind of worry: even if (say) a dynamical explanation of a stimulus-independent cognitive capacity were forthcoming, why not think that it simply provides an interesting new way of thinking about representation?

In his 2007 book, *Representation Reconsidered*, William Ramsey develops an anti-representationalist argument that seems to sidestep this sort of concern. He argues that the central threat to representationalism comes not from some imagined future non-representational cognitive explanation, but from existing theories of ‘offline’, stimulus-independent cognitive capacities. He carefully examines the explanatory roles that representations have been taken to play within these extant theories and argues that some of these roles simply aren’t representational in nature. In particular, he argues

\(^2\)For a canonical statement of the first sort of worry, see Clark & Toribio (1994). For a canonical statement of the second sort of worry, see Bechtel (1998)
that theories within connectionism and cognitive neuroscience typically work by positing states that *detect* or reliably respond to some distal entity. These states, which Ramsey calls ‘receptor representations’, are characterized as representations by both connectionists and their most strident critics, but according to Ramsey there’s nothing distinctively representational about the explanatory role that they play; all sorts of mechanisms function in essentially the same way without us having any inclination to call them representations. The only reason we continue to call receptors ‘representations’, claims Ramsey, is because of conceptual inertia left over from the cognitive revolution; just as cosmologists still *talked* about Ptolemaic celestial spheres even after the Copernican revolution rendered such spheres redundant, connectionists still *talk* about representations despite the fact that representations play no explanatory role within their theories.

Ramsey allows that *some* cognitive theories posit states that play a genuinely representational role. Specifically, he holds that ‘classical’ theories typically work by positing internal *models* that instantiate the same abstract structure as some external system, which are manipulated within the cognitive architecture so as to enable the agent to successfully reason about or interact with the external system. According to Ramsey, these models or ‘structural representations’ — which, as we’ll see, can be identified with Gallistel’s *functioning homomorphisms* — function as genuine *stand-ins* for the external system. However, Ramsey holds that classical theories are on the decline in contemporary cognitive science, while associative connectionist approaches, in the sense I articulated in Chapter 2.4, are on the ascendance. Thus, he argues, the cognitive revolution is returning full-circle to a kind of neo-behaviorism: beneath the guise of representation-talk, contemporary cognitive science provides explanations in which “cognitive representations play no real explanatory role” (Ramsey, 2007, p.226). Further, while Ramsey makes no claims about whether connectionist explanations will ultimately prove to be *successful*, he argues that if they do, our commonsense concep-
tion of ourselves as intentional agents with internal mental states that are about the world, will have to be radically revised.

I don’t think the distinction between structural representations and receptors lines up with the distinction between classicism and connectionism in the way Ramsey thinks it does. The idea that a trained neural network embodies the same abstract structure as its task domain is implicitly or explicitly reflected in a great deal of connectionist research, and, I think, has some claim to being the ‘official’ meta-theoretical view about how connectionist systems work. However, I think there are deeper reasons why Ramsey’s anti-representationalist and eliminativist arguments fails, which reveal problems with the structural conception of representation that both he and Gallistel take for granted.

In this chapter, I largely prescind from debates about whether or not connectionist neuroscience posits structural representations, and focus on the cogency of the distinction between the ‘receptor’ representations that are supposedly distinctive of neuroscientific explanations, and the ‘structural’ representations that are purportedly the domain of classicism. I begin, in Section 5.2, by discussing what Ramsey thinks it takes for a given type of theoretical entity to qualify as a cognitive representation. Then, in Sections 5.2.1 and 5.2.2 turn to consider the two purportedly different kinds of theoretical entities — structural representations and receptors, respectively — to examine whether they meet Ramsey’s criteria for representationhood. In Section 5.3.1 I turn from exposition to criticism, and argue that when the two aforementioned concepts of representation are properly explicated, in ways that Ramsey is sympathetic to, there turns out to be no distinction between them; anything that satisfies the conditions of being a structural representation satisfies the conditions of being a receptor, and vice-versa. There only appears to be a distinction between structural and receptor representations because the former are tacitly conflated with introspectively accessible ‘mental models’ that plausibly mediate stimulus-independent cognitive capacities
such as memory and imagination. However, I argue in Section 5.3.2 that, contrary to widespread views in philosophy and psychology, there’s nothing distinctively mental about structural representations, for they’re to be found in all sorts of mindless systems such as plants. I conclude that the notion of structural representation might pick out a kind of theoretical entity that deserves to be called a ‘representation’, and that — despite what Ramsey and Gallistel think — this posit plays important explanatory roles in both classical cognitive science and connectionist neuroscience, but that on its own it cannot explain the kind of stimulus-independent cognitive capacities that we think of as distinctively mental, capacities that distinguish intentional agents from mindless systems. However, it doesn’t follow that neuroscience somehow has anti-representational or behaviorist tendencies; on the contrary, research into the mechanisms of ‘mental models’ is currently a thriving area of contemporary neuroscience.

5.2 The Job Description Challenge

Suppose a renal physiologist told you that kidneys use representations to filter blood. How, you might reasonably ask. Upon being told about nephrons, antidiuretic hormones and such, you might reply: That’s a perfectly good explanation of how the kidney functions to filter blood, but why should we consider that a representational function? That, surely, is the right kind of question to ask of representational claims about kidneys, and it’s exactly the kind of question that Ramsey asks of representational claims in cognitive science. While cognitivist theories employ representation-talk, we shouldn’t take that talk at face value if we’re to understand the ontological commitments of such theories. Instead, we should ask whether the posits of those theories play a genuinely representational explanatory role. Of course, to answer such questions we need an account of what it is to play such a role. We need what Ramsey (2007) calls a job description for representations.
To ask what it takes for something to satisfy a representational job description is to ask the ontological question of what it takes for something to be a representation; but by framing this question in terms of job descriptions, Ramsey is indicating that he’s adopting a particular methodological stance. He seeks not to develop a univocal analysis of representation that encompasses all and only the things we happen to call representations, for as he rightly points out, we apply ‘representation’ to such a congeries of things that such a project is surely forlorn. Instead, his project is located within the tradition of Quinean naturalistic ontology; he proposes that we look to specific cognitive theories and examine the explanatory roles that putative representations play within those theories. This approach leaves it open that those roles might fall under quite different types — that there might be quite different job descriptions for representations. It also leaves it open that a given explanatory role isn’t genuinely representational after all. How might we decide whether a given role is representational? Ramsey holds that we must be guided at least to some extent by our intuitive, pre-theoretical notions of representation. While he doesn’t endorse a fully descriptivist account of theoretical reference, he does hold that any proposed demarcation between those explanatory roles that are distinctively representational and those that are not must be continuous with our ordinary ways of thinking about representation, for otherwise the states that purportedly play a representational role would be representations in name only, like the alleged ‘representations’ in kidneys.

While Ramsey thinks that our commonsense notion of representation is a cluster concept, and hence isn’t amenable to analysis, he thinks that three aspects of the notion are central, and impose (perhaps defeasible) criteria for something’s satisfying a representational job description. First, and most importantly, representations represent something. We ordinarily think of both our representational mental states, such as beliefs and desires, and external representational artifacts, such as maps and words, as being about some object, property, or state of affairs. In the Brentanian jargon,
we think of representations as being directed at an intentional content. Second, while this phenomenon of aboutness or intentional directedness seems to involve a kind of relation, if so, it must be a very special kind of relation, in which the distal relatum — the intentional object of the representation — needn’t actually exist. To put it less tendentiously, it seems central to our conception of representations that they can misrepresent. Finally, we ordinarily think of the content of a representation as being somehow relevant to the causal role that the representation plays. If I represent someone as a friend, I might issue an affable ‘Hello’, but if I represent the person as my nemesis, I might engage in a very different, more sinister suite of behaviors.

Though Ramsey argues that we must begin with commonsense when identifying a job description for representations in cognitive science, he argues that we cannot end there, for although we have a venerable story about how external representational artifacts might misrepresent or play a role that’s relevant to their content — roughly, they’re interpreted by us — no such story will do for the sub-personal representations posited by cognitive scientists. The problem of course isn’t just that there are no little homunculi interpreting our sub-personal representations, it’s that positing such homunculi would lead to an explanatory regress. So showing that the sub-personal states posited by a cognitive theory satisfy a genuinely representational job description isn’t just a matter of showing that they exhibit the various puzzling intentional properties of representations ascribed by common sense; it’s also a matter of showing that the theory is able to explicate, in broadly mechanistic or ‘naturalistic’ terms, how those states exhibit those properties. This is what Ramsey calls the job description challenge.

Over the next two sections I’ll look at two broad families of cognitive theories, and evaluate how they fare against the job description challenge. But before moving on, I’ll say a few words to clarify the nature of that challenge. Through much of the late twentieth century, philosophers of mind sought to ‘naturalize semantics’ by providing
informative, non-circular, and broadly ‘naturalistic’ conditions for an intentional state to have a particular semantic content — where having a content $C$ is here understood as standing in a representation relation with $C$.\textsuperscript{3} This project looms large over the philosophical landscape, so is apt to be mistaken for the job description challenge, however the two projects are quite different: one is about the metaphysics of the representation relation, whereas the other is about the ontology of representational vehicles. These projects are related, for as we’ve seen, one intuitive constraint on something’s being a representation is that it has intentional content, but nevertheless they’re distinct, and failing to recognize them as such can only lead to confusion.

As Ramsey and others have pointed out, this confusion is apparent in one of the most influential attempts to naturalize semantics: Ruth Millikan’s (e.g. 1984) teleosemantics. Millikan’s view is complex and nuanced, but the central idea is that the content of a representational state is whatever has the biological function of eliciting that state. Ramsey points out that Millikan often seems to regard her view as an account of representation \textit{as well as} an account of content, by counting any state that has the function of being elicited by something as a representation of that thing. For example, she holds that because adrenaline flow has the function of readying the body for situations that require strenuous activity, it \textit{represents} such situations (Millikan, 1984, p.116). Yet it’s not clear why readying the body for action is a distinctively representational function, any more than filtering the blood is. By thinking that her theory of content can do double-duty as theory of representation, Millikan seems to cast the net of representationhood too wide, over areas on which it has no explanatory purchase.\textsuperscript{4} Indeed, Ramsey points out that this indifference to the question of what

\textsuperscript{3}The main players in this project include Dretske (1988), Fodor (1990), and Millikan (1984).

\textsuperscript{4}Others have made essentially the same point about the profligacy of Millikan’s view. For example, Allen & Hauser (1993) complain that Millikan’s view entails that “some interactions between trees can have content attributed to them” (p.88), and Sterelny (1995) expresses concern that on Millikan’s view, “it will turn out that saliva represents food” (p.256).
representations are endemic throughout the literature on naturalized semantics, where theorists typically enquire into the conditions for something’s having content, without asking what the nature of that something might be. Whether or not this gets the proper order of explanation backwards, or is simply a convenient way of carving up the problem space, the question of representationhood is clearly distinct from the question of content, and it’s the former that we’re interested in here.

5.2.1 Structural Representations

Now we know the ground rules, we can judge how various competitors fare against the job description challenge. Let’s begin by considering theories in the ‘classical’ tradition of cognitive science. Ramsey holds that when we examine classicist theories we find a recurring explanatory pattern: such theories typically attempt to explain cognitive capacities within a given task domain by positing neurally-encoded symbolic systems that instantiate the same abstract structure as some distal system within that domain. A given capacity is explained by showing how the internal symbolic system is computationally manipulated by the cognitive architecture so as enable the agent to successfully interact, in thought or action, with the external system. To borrow some helpful terminology from Swoyer (1991), the intuitive idea underlying this explanatory strategy is that, just as we use maps and scale models as ‘surrogates’ for reasoning about the real-world systems that those artifacts are structurally isomorphic with, the brain uses sub-personal cognitive models, encoded in symbolic systems, as surrogates for reasoning about the behaviorally-relevant distal systems that they’re isomorphic with. Following Ramsey, let’s call the mechanisms that are hypothesized to play this

---

5Note that I am using ‘isomorphism’ loosely here, to refer to the kind of resemblance relations that structural representations purportedly participate in, since that term is so familiar in this context. However, I will go on to argue that the resemblance relations at issue here are probably best understood as homomorphisms rather than isomorphisms.
The general idea that the mind reflects the abstract structure of external reality is prefigured in Aristotle’s image of a signet ring impressing its form on a blob of wax. However, it arguably finds its most influential contemporary expression in Kenneth Craik’s (1943) seminal book, *The Nature of Explanation*. Craik suggests that,

If [an] organism carries a ‘small-scale model’ of external reality and of its possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

(Craik, 1943, p.51)

Craik’s suggestion is compelling, and seems to pick out an explanatory role which, if filled, would help explain a range of interesting cognitive phenomena. However, at the time he was writing it was difficult for psychologists to see how this role could be filled. The situation changed with the advent of cognitive science and its attendant notion of computation, for, as Jerry Fodor (e.g. 1985) is fond of pointing out, the notion of computation provided a way of understanding how an internal symbolic system might be manipulated, via purely mechanistic processes, so as to produce effects that are relevant to the semantic properties of the symbols. Add to this the idea that the internal symbolic system embodies the same abstract structure as some external system, and we begin to see how Craik’s suggestion might be incorporated into a broadly mechanistic explanation of how the mind works.

Once cognitive scientists glimpsed the mechanistic bona fides of mental models, they invoked them freely to explain a wide range of cognitive phenomena. Perhaps the most prominent proponent of mental models was Philip Johnson-Laird (1983), who drew direct inspiration from Craik and appealed to mental models to explain such sophisticated cognitive capacities as problem solving and deductive reasoning.
However, the influence of the general idea that the mind manipulates internal structural isomorphs of external systems went far deeper, and some theorists suggested that all of the various representations posited by cognitive scientists ought to be unified under the aegis of the notion of structural representation; for example, Palmer (1978) defined cognitive representation in terms of a “correspondence (mapping) from objects in the represented world to objects in the representing world such that at least some relations in the represented world are structurally preserved in the representing world” (p.266-7). Similarly, many philosophers of cognitive science emphasized the importance of structural representations to explain cognition, especially the kind of ‘sophisticated’, stimulus-independent capacities discussed by Craik and Johnson-Laird; for example, Cummins (1994) claims that “what makes sophisticated cognition possible is the fact that the mind can operate on something that has the same structure as the domain it is said to cognize” (pp. 297-298).

Ramsey holds that mental models or structural representations are ubiquitous throughout classical cognitive science. He claims that they play an essential role within such quintessentially classicist theories as “Newell’s production-based SOAR architecture (1990), Winograd’s SHRDLU model (1972), Anderson’s various ACT theories (1983), Collins and Quillian’s semantic networks (1972), [and] Gallistel’s computational accounts of insect cognition (1998)” (Ramsey, 2007, p.79; citations provided therein). However, Ramsey claims that structural representations are rarely invoked by theories in connectionist cognitive science. As I mentioned earlier, I think this claim is false. States that serve as models or internal isomorphs play just as important an explanatory role in connectionism as they do in classicism. One way to argue for this point is to show that many of the kinds of connectionist systems that Ramsey

---

6Where a cognitive theory counts as ‘connectionist’ in the present sense precisely when it invokes associative connectionist models of the kind that I described in Chapter 2.
explicitly considers, such as the multilayer feedforward perceptrons discussed in the PDP literature of the ‘80s, in fact employ structural representations, despite Ramsey’s claims to the contrary. Garzón & Rodriguez (2009) develop this line of objection to Ramsey by pointing out that techniques like cluster analysis show that the metric relations between hidden-layer activation patterns in a trained multilayer perceptron reflect relations between entities in the problem domain that the perceptron was trained to classify. Indeed, this is the standard interpretation of how perceptron classifiers work, and is widely endorsed by the most prominent proponents of connectionism. As the Churchlands (2002) put it, “The various distance relationships between the learned clusters... within the activation space of a given population of neurons, are collectively and literally isomorphic with the similarity relationships that objectively exist between the various categories in the external world” (p.907).

Ramsey doesn’t directly address the point that even canonical connectionist systems are standardly characterized as vehicles of structural representation, but some of his related comments suggest the following reply: any structural relations between activation patterns purportedly revealed by cluster analysis are merely dispositionally latent within the network, and therefore don’t reflect structural relations in the task domain. However, some have made similar claims about the structural relations purportedly encoded by classical systems. For example, quoting Cummins (1989), O’Brien & Opie (2001) suggest that conceptions of representation based on structural resemblance have no place in classicism since “nothing is more obvious than that [symbolic] data structures don’t resemble what they represent” (Cummins, 1989, pp.30-31). The problem with the general claim here, whether it’s made on behalf of connectionists or classicists, is that it rests on a mistaken ‘pictorialist’ conception of structural resemblance. When considering whether a dynamic physical mechanism serves as a model, we must look to the functional, dispositional properties of the mechanism, not to its static, categorical properties. As Craik (1943) insightfully put it, “a model need not
resemble the real object pictorially; Kelvin’s tide predictor, which consists of a number of pulleys on levers, does not resemble a ride in appearance, but it works in the same way in certain essential respects” (p.51). What matters is not how the mechanism appears to the human eye, but how the mechanism works, and cluster analysis seems to show that classifier perceptrons work by instantiating the categorical structure of the domain that they classify.7

A second way to show that structural representations play an important explanatory role in connectionism is to consider some less familiar connectionist models, and argue that they too invoke structural representations. This strategy is pursued by Shagrir (2012), who discusses recent models of oculomotor control in computational neuroscience, which posit stable states of reverberating activity in a recurrent neural network that can occupy various points along a line attractor. The current location of the attractor state serves to keep track of the current position of the eyes. According to Shagrir, the network functions as a structural representation since “the state-space of the network mirrors the space of eye positions” (ibid., pp.13-14). When evaluating the significance of Shagrir’s example, it’s important to keep in mind that Ramsey allows that some connectionist theories invoke structural representations; he cites Grush’s (2004) ‘emulation theory’ as an example. Ramsey’s claim is that such theories are recherché exceptions to general trend of explaining cognition without appealing to structural representations. Shagrir is perhaps insufficiently attentive to this point, so it’s important to note that the oculomotor model he discusses is an instance of a general class of attractor network models discussed in earlier Chapters that are widely invoked throughout contemporary computational neuroscience to explain a diverse

7Ramsey himself expresses the present point when he writes that “there need be no superficial visual resemblance between representation and what is represented” (p.81). However, he doesn’t notice that this seems to undermine his claims about the merely ‘tacit’ nature of structural representations in connectionist networks.
range of psychological phenomena, and the points he makes about the oculomotor model could be generalized to other models within the same class. Similarly, as Grush would be the first to admit, his ‘emulation theory’ did not spring fully formed from the head of Zeus, but rather offers a synthesis of existing work on the neuroscience of motor control; ‘emulator’ is simply Grush’s term for a kind of theoretical posit — a forward model — that is ubiquitous in mainstream motor neuroscience, and which has historical antecedents that can be traced back to pioneering work by von Helmholtz in the nineteenth century. The theories mentioned here are far from recherché; they’re about as central to contemporary neuroscience as can be.

The general line of response to Ramsey that I’ve been sketching here, according to which structural representations are not propriety to classicist theories, has been developed by several other authors in addition to those I’ve already mentioned, so I will not pursue it further here. Instead, I will pursue an alternative line of response, which to my knowledge has not been explored in the existing literature. My central goal in this paper is not to adjudicate whether classicism or connectionism has custody over structural representations, but to diagnose whether the notion of structural representation is legitimate; I will argue that there is simply no distinction between structural representations and the purportedly non-representational ‘receptors’ that Ramsey contrasts them with.

To develop this argument, I should first consider whether structural representations satisfy a representational job description. Ramsey holds that they clearly do, since their

---

8See Eliasmith (2005) for a review.
9See Miall & Wolpert (1996) for a review.
10See also Grush (2008); Sprevak (2011).
11Note that the boundary between these two kinds of replies to Ramsey is fuzzy. Some of those who have argued that connectionists invoke structural representations can be understood to be arguing, implicitly, that states that intuitively seem to be receptors in fact also count as structural representations. However, to my knowledge nobody has explicitly developed an argument to the effect that all and only structural representations are receptors.
explanatory function is to model external systems. Like the scale models used by, say, aeronautical engineers, they are intuitively about the systems they are structurally isomorphic with. Moreover, the fact that they are isomorphic with a given system is relevant to their causal role, since it’s in virtue of that structural resemblance that they can be manipulated so as to guide an agent’s successful interactions with that system. However, unlike the scale models used by engineers, they needn’t be interpreted by an intelligent agent to play the content-relevant, behavior-guiding causal roles that they do; insofar as they can be manipulated by purely formal, computational operations, there’s no metaphysical mystery about how they could play those roles in the absence of a homunculus.

To evaluate Ramsey’s sanguine attitude about the job prospects for structural representations, I think we need a much clearer view of what exactly structural representations are. This is especially urgent given that, as Ramsey rightly points out, structural representations seem to be a species of what Peirce called icons — entities that represent by virtue of resembling what they are about — and there are of course venerable objections to resemblance-based theories of representation. One of the most classic such objections is directed specifically at the view that mental representations resemble their intentional contents. Surely, the objection runs, my mental image of, say, an orange isn’t itself colored orange.

This objection seems decisive against a resemblance-based theory of representation that explicates resemblance in terms of the co-instantiation of ‘first-order’ monadic properties. However, structural views of representation sidestep this objection by cashing resemblance out in terms of the co-instantiation of ‘second-order’, relational properties (O’Brien & Opie, 2001; Shepard & Chipman, 1970). A canonical example of an iconic representation that represents in this sense is a cartographic map: the map instantiates the same relational (geometric or topological) structure as the terrain it represents. However, as I suggested earlier, the functional architecture of a mechanism
might also embody the same relational structure as an external system — and indeed, it’s arguably this kind of dynamic rather than pictorial structural resemblance that’s at issue when discussing models that function in the absence of the interpretative faculties of a homunculus.

Ramsey, like many other proponents of structural representation, attempts to make the notion of shared relational properties more precise by appealing to the mathematical notion of isomorphism. However, a notion of resemblance cashed in terms of isomorphism is still a notion of resemblance, and Goodman (1968) famously developed what many take to be decisive objections to any kind of resemblance-based theory of representation. His central argument is that the logical properties of resemblance relations seem utterly different from those of representation relations: resemblance is reflexive, symmetric, and transitive, whereas representation is none of these.

To avoid the brunt of this objection, some theorists have attempted to explicate structural representation in terms of the notion of homomorphism rather than isomorphism.\textsuperscript{12} Homomorphisms are a more permissive kind of structure-preserving mapping, which can obtain between systems with different cardinalities, and which needn’t be either symmetric or transitive.\textsuperscript{13} Cashing out a notion of structural representation in terms of homomorphisms has also seemed attractive because it allows that representing systems might be less than perfect simulacra of represented systems.

Appealing to homomorphisms appears to resolve some of the problems with an unvarnished resemblance-based theory of representation. However, it still leaves us

\textsuperscript{12}See, e.g., Bartels (2006), who develops this idea in the context of debates about scientific representation, a context in which many of the present issues about structural representation are recapitulated.

\textsuperscript{13}Recall from Chapter 2.2.2 that an isomorphism is a bijective (i.e. one-one and onto) function from one set-theoretic structure to another, which preserves the relations defined over the elements of each structure. A homomorphism is, like an isomorphism, a structure-preserving mapping from one set-theoretic structure to another, but unlike isomorphisms, homomorphisms needn’t be bijective. Thus, a homomorphic mapping can be many-one, and needn’t map onto all the of the elements in the represented structure.
with the problem that resemblance is reflexive. Intuitively, representations don’t represent themselves. There’s an arguably deeper problem here, too: abstract structural relations are ubiquitous. Slough off enough detail, and virtually anything might be homomorphic with virtually anything else. A map of the NYC subway system might reflect the topology of connections between actual subway stops in Manhattan, but it might also reflect the topology of connections between a population of neurons in your brain. Yet the map surely doesn’t represent your neurons. In the case of artifacts like subway maps, the relevant homomorphisms are plausibly constrained by interpretative intentions of agents; but how do we constrain the homomorphisms that cognitive representations participate in without appealing to homunculi?

Ramsey’s response to this problem is puzzling. He holds that the morphisms that constitute structural representation relations are constrained by the methodology of cognitive explanation: “the explanandum itself... determines what it is that is being modeled” (Ramsey, 2007, p.94). So Ramsey seems to think that if, for example, a cognitive theory posits an internal map to explain a rat’s ability to navigate home, the map represents spatial relations in the rat’s environment — as opposed to, say, some configuration of stars in Andromeda — because the theory is about the rat’s ability to navigate its environment (ibid., p.96). This seems to entail that the content of structural representations is radically observer-dependent; that, for example, rats had to wait until they were studied by human ethologists until their cognitive maps acquired determinate content. One might think this constitutes a reductio of Ramsey’s view.14

14 Ramo (2007, p.98-9) does respond to a kind of observer-dependency worry in this context, but not the specific worry that I’m raising here. He addresses the concern that the structural representations posited by classical cognitive scientists are merely useful fictions. Like Ramsey, I don’t think we should lose any sleep over that concern. The worry that I’m raising here is different: it’s that Ramsey’s explication of the kind of structural representations posited by classical cognitive scientists entails that the content of a structural representation is radically observer-dependent. One might be troubled by that worry without having any specific views about the scientific realism debate.
Ramsey does hint at an alternative view, though he doesn’t distinguish it from the one just criticized, namely that the relevant morphisms are constrained by causal relations between representing and represented structures. On this view, the rat’s cognitive map represents the geometry of its environment at least in part because the map is causally responsive to specific geometric properties of the environment. A second way in which causation might help constrain the relevant morphisms is via the causal role that the representing system plays within an agent’s cognitive architecture; the idea here is that the representing system is used to guide the agent’s interactions with the represented system. This sort of use-condition is widely invoked in discussions of cognitive representation, and is clearly related to the idea that a representation ought to play a causal role that is relevant to its content; but it might also help with the indeterminacy problem presently under consideration. To paraphrase Dretske (1988), it is only by using a representation in the production of movements whose successful outcome depends on what is being represented can indeterminacy about the target of the representation be overcome (p.70). So appealing to external causal relations, or internal causal roles, seems to make headway on the problem of the ubiquity of structural resemblance. It also seems to resolve the problem of the reflexivity of resemblance. For while it’s plausible that a rat’s cognitive map is causally responsive to the geometric properties of the rat’s environment, the reverse is surely not.

Gallistel detailed characterization of representation in terms of functioning homomorphisms, which I articulated in Chapter 2.2.2, incorporates the various advances over a simplistic resemblance theory that I’ve just surveyed. Gallistel’s account thus promises a precise articulation of the kind of structural view of representation that Ramsey gestures at. Recall that, according to Gallistel, a system $A$ counts as a representation of a system $B$ just in case it satisfies three conditions: first, $A$ is homomorphic
with $B$; second, this homomorphism is established and sustained by causal relations between the two systems, such that variations in $A$ are causally influenced by variations in $B$; and third, $A$ causally interfaces with motor control systems such that it can guide the agent’s behavior with respect to $B$ in ways that reflect the relevance of $B$ for the agent. Thus, for Gallistel, representation relations are *functioning homomorphisms*: abstract structural similarities, sustained by causal relations, which serve to inform an agent’s behavior.

Gallistel’s account is clearly an instance of a structural view of representation in Ramsey’s sense, and Ramsey himself characterizes Gallistel’s account as an exemplar of a structural view (recall the quote on page 315). It’s true that there are some differences of emphasis between the two authors’ accounts — for example, Ramsey tends to focus on the role of structural representations in surrogative reasoning, whereas Gallistel emphasizes that they are used to “control and direct appropriate behavior” (Gallistel & King, 2010, p.55) — however, these differences are superficial. For example, Gallistel sometimes expresses his view in terms of surrogative reasoning. In a (1990b) paper, he writes that “In order for a representation to exist, the neural or mental representatives of environmental variables must enter into combinatorial neural or mental processes that generate *valid inferences* about the represented variables” (p.4, my emphasis). Conversely, Ramsey’s conception of surrogative reasoning is very liberal, and seems more or less coextensive with the kinds of behavioral control processes emphasized by Gallistel. As I’ll discuss in more detail later, one of Ramsey’s central examples of a structural representation is the mechanism depicted in Figure 5.1 below: an S-shaped groove inside a toy car, which interfaces with the car’s driveshift and steering wheel in such a way that when the car moves forward, a rudder traces the course of the groove, thereby steering the car and enabling it to navigate an S-shaped road. Ramsey holds that the car is able to successfully ‘reason’ about the road using the groove as a surrogate, by virtue of the structural resemblance between the groove and the road.
Now, one might quibble about whether a process like this is properly called ‘reasoning’, but what’s important to note here is simply that both Ramsey and Gallistel, a crucial condition for something’s being a structural representation is that it’s *used* in the right way.

![Figure 5.1: Ramsey’s example of a structural representation: an S-shaped groove inside a toy car interfaces with the car’s driveshift and steering wheel in such a way that when the car moves forward, a rudder traces the course of the groove, thereby turning the wheel and enabling the car to navigate an S-shaped road. The S-shaped groove embodies the same abstract structure as the road, and serves to guide the car along the road, thus it satisfies Ramsey’s criteria for structural representation (from Ramsey, 2007, p.199).](image)

Despite these basic similarities, we’ve seen in this section that there are some lacunae in Ramsey’s attempt to articulate a notion of structural representation that avoids the traditional problems with resemblance-based theories of representation. Gallistel’s notion of a functioning homomorphism seems to fill these lacunae, and hence appears to provide the strongest version of a structural conception of representation on offer.\textsuperscript{16}

\textsuperscript{16}Note that I don’t take myself to have provided a strong defense of a Gallistelian view of structural representation against the traditional objections to resemblance-based theories. For such a defense, see Isaac (2013). My claim here is simply the conditional one that if any account of structural representation is viable, Gallistel’s is the most plausible candidate.
Thus in what follows I’ll take ‘structural representation’ to be coextensive with whatever satisfies Gallistel’s conditions for participating in a functioning homomorphism.

5.2.2 Receptor Representations

Let’s now turn to an apparently very different kind of state that has been invoked to explain cognition, and ask whether it satisfies a job description for representations. Ramsey argues that when we look at the family of theories under the broad banner of ‘connectionism’, we find that such theories routinely appeal to states that are said to represent events of a certain type by virtue of being reliably elicited by, or correlated with, such events. We see this most clearly in neurophysiology, when individual neurons are characterized as ‘detectors’ of the class of entities that reliably elicit maximal activation within that neuron. An example that’s perhaps most familiar to philosophers is the ‘bug detector’ in the retina of the frog described by Lettvin et al. (1959), but similar examples abound within neurophysiology; one finds edge detectors (Hubel & Wiesel, 1962), face detectors (Desimone, 1991), even particular spatial location detectors (O’Keefe & Dostrovsky, 1971). Ramsey argues that we find essentially the same idea in the PDP literature of the ‘80s, where stable patterns of activity in trained multilayer perceptrons are said to represent the inputs that they were elicited by. Whether the focus is on activity in single neurons, or in neural networks, the explanatory pattern seems to be the same: internal states are said to represent distal events by reliably ‘detecting’ them.

Ramsey calls the states that play this kind of explanatory role ‘receptor representations’. However, he argues that they are not representations properly so-called, for the role they play is not a genuinely representational one. Simply functioning as a reliable receptor cannot be sufficient for something to count as a representation, claims Ramsey, since all sorts of mechanisms function as receptors without us having any inclination to think of them as representing anything. We already saw one example
when discussing the profligacy of Millikan’s view: adrenaline flow is reliably elicited by stressful events, but there’s arguably no strong sense in which it represents such events. Further examples abound: antibodies in the immune system, bimetallic strips in thermostats, infrared sensors in automatic faucets, and so-on. All of these things work by entering into particular states in response to certain distal conditions, but none of them function as representations by doing so. They simply function as triggers, or relays, or causal mediators.

The last example here provides a stark illustration of Ramsey’s point. While automatic faucets might be preferred over manually-controlled ones for various reasons, the role that infrared sensors play in automatic faucets isn’t any different from the role that handles play in manual ones; both mechanisms just serve to turn the water on. We might adopt the intentional stance with respect to automatic faucets and interpret them as ‘recognizing’ the hands that trigger them, but this gloss is superfluous when it comes to understanding how automatic faucets actually work. Ramsey holds that the same is true of the purportedly representational states posited by connectionists. Unlike structural representations, we can understand the explanatory role that these so-called receptor representations play without understanding them as inner stand-ins for external entities. To put the point another way, any content we might ascribe to receptors is irrelevant to explaining the role that the receptor plays within the system of which it is a part.

At this point one might worry that the preliminary characterization of receptors that we’ve been working with so far is a bit too thin to capture the notion of representation that connectionists have in mind when they describe feature detectors or states of networks as representations. For it seems that virtually any state that’s causally dependent on another counts as a receptor representation according to this initial characterization. Yet surely the notion of representation that connectionists employ when they claim, for example, that place cells represent a particular spatial location is some-
what more constrained than the idea that, say, a shattered window pane ‘represents’ the rock that broke it. The question that presents itself, then, is whether we can discipline the notion of receptor representation in such a way that it doesn’t encompass just *any* causal relation, yet retains the character of ‘detection’, while also satisfying the job description challenge.

One might have thought that the literature on indicator semantics could provide a promising source of constraints, since the central project in that literature is to explain intentional content in terms of reliable causal dependencies, while avoiding the obvious problem that not all causal dependencies involve intentional content. However, as I mentioned earlier, proponents of indicator semantics tend to ignore questions about what representations are and how they function. One notable exception is Dretske (1988), whose account of content is largely motivated by concerns about the explanatory role of representations. So, following Ramsey, let’s consider whether Dretske’s account might help to discipline the notion of receptor representation so that it passes the job description challenge.

Dretske’s account of intentional content is driven by two central ideas. The first, common to all versions of indicator semantics, is that the relation between intentional states and their contents is fundamentally a kind of *causal* relation. In addressing the aforementioned problem that not just any causal relation involves intentional content, Dretske also attempts to do justice to a second idea about the nature of intentional content, which we encountered earlier when discussing the job description challenge: that intentional states play causal roles that are somehow relevant to their content. As Dretske (1988) puts it, “The fact that [intentional states] have a content... must be relevant to the kind of effects they produce” (p.80). Dretske brings these ideas together by holding that a given state *R* that is causally dependent on some type of external event *C* comes to have the intentional content that *C* when *R* is selected, via natural selection or individual learning, to play a specific causal role within the system
of which it is a part, by virtue of the fact that it is causally dependent on C. That is, R is about C not only when it responds to C, but when it has the function of so responding, by virtue of the fact that its responsiveness to C is exploited by the system so as to guide appropriate behavior with respect to C.

Dretske's account seems to provide a characterization of receptor representations that doesn't massively over-generalize in the way that the original characterization did. To begin to see why, consider the 'bug detectors' in frogs described by Lettvin et al. (1959). The researchers described these cells as bug detectors, rather than small-dark-moving-dot detectors, even though they showed that the cells are maximally responsive to small dark moving dots. This suggests that the researchers were employing a more robust notion of representation than one according to which a mechanism represents whatever it is causally responsive to. Given the presumably uncontroversial assumptions that the frog's bug detectors interface with mechanisms that mediate bug-directed tongue strikes, and that they have been selected to do so because they reliably respond to bugs, Dretske's account of content — taken as an account of representation — seems to capture this more robust notion; it seems to provide a sense in which bug detectors genuinely detect bugs as such. Moreover, Dretske's account would presumably encompass many of the other kinds of states claimed by connectionists to be representations, yet would exclude shattered window-panes and the like. So Dretske's account seems to provide a robust explication of the notion of receptor representation.

But do receptors in Dretske's sense satisfy a representational job description? At first glance it would seem that they do. Indeed, Dretske expressly develops his account in order to capture the three aspects of our commonsense conception of representation discussed in Section 5.2. However, Ramsey argues that appearances here are illusory.

---

17 The distinction between an indicator's being selected over the course of phylogeny, and its being selected over the course of ontogeny, plays an important theoretical role for Dretske. However the distinction is largely orthogonal to present debates, so, like Ramsey, I'll elide over it in what follows.
He points out that the fact that a structure is responsive to certain conditions, and that it is enlisted to play a particular functional role within a system by virtue of that responsiveness, doesn’t entail that the structure plays a distinctively \textit{representational} role, for “a structure can be employed... \textit{qua} reliable respondent without being employed... \textit{qua} representation” (2007, p.138). For example, he points out that one might plant a tree at a particular location to provide shade at a particular time of the day, exploiting the fact that the shadow of the tree depends on the position of the sun. However, this dependency is used to provide shade, not to provide information about the position of the sun. Of course, causal dependencies might \textit{sometimes} be exploited for representational purposes. You might use the tree as a sundial, taking the shadow it casts to inform you about the sun’s position, and hence the time of day. Ramsey’s point, however, is that not all cases of exploiting a causal dependency are cases of exploiting a causal dependency \textit{for representational purposes}. So, he argues, merely satisfying Dretske’s conditions isn’t sufficient for something to count as a representation. Even if we make receptors more robust by appealing to Dretske’s conditions, it seems that they still fail to meet the job description challenge.

\subsection*{5.3 Problems in the Workplace}

I find Ramsey’s anti-representationalist argument unconvincing. To begin to see why, first notice that even if we endorse all the premises of the argument, the conclusion simply doesn’t follow. Ramsey wants to argue that the ‘receptor’ mechanisms typically characterized as representations by connectionists — such as feature detectors, states of networks, and the like — aren’t really representations at all. But notice that all that the argument at the end of the preceding section entitles us to conclude is that such mechanisms aren’t representations \textit{simply by virtue of being indicators that are exploited for behavioral control}. This doesn’t entail that receptor mechanisms aren’t
representations; it just entails that insofar as they are representations, they must be so by virtue of satisfying certain additional conditions.\textsuperscript{18}

To illustrate, consider a case that parallels Ramsey’s example of using a tree for shade. The brains of fruit flies contain neurons that are responsive to temperature, and these so-called ‘thermosensors’ allow fruit flies to avoid extreme heat (Hamada et al., 2008). Like the shadow cast by Ramsey’s tree, the causal dependency of thermosensor activity on temperature is exploited for the purpose of staying cool — but it doesn’t follow from this that the dependency is not also exploited for informing the fruit fly about the location of heat sources. Similarly, I might read the weather report to decide whether to carry my umbrella, but it doesn’t follow that the weather report isn’t also informing me about the likelihood of rain. Now, I haven’t yet provided a strong reason to think that receptors like thermosensors are representations; my point is simply that Ramsey’s argument, taken on its own terms, doesn’t provide a strong reason to think that receptors are not representations.

5.3.1 Structural Receptors?

However, I think there are more fundamental reasons why Ramsey’s argument fails. We can begin to uncover these reasons by first noting that the length of the tree’s shadow in Ramsey’s example is homomorphic to the angle of elevation of the sun above the horizon. That’s why we could use the length of the shadow to measure the sun’s elevation; according to standard ways of thinking about measurement, the possibility of measurement presupposes the existence of a homomorphism from the measured into the measuring structures (Krantz et al., 1971). Now, if we were to use the length of the shadow to measure the sun’s elevation, would we be using the shadow as a structural

\textsuperscript{18}Sprevak (2011) makes essentially the same point when he writes that “what satisfies the receptor notion, by itself, may not fulfill the job description of a representation, but the wider explanatory role that it plays in explaining successful behaviour may justify its labelling as a representation.”
representation or as a receptor? There’s a sense in which the shadow’s length would be serving as a kind of *model* of the sun’s elevation for us, yet the shadow would also provide a way for us to *detect* the sun’s elevation; indeed, it’s precisely because the shadow ‘detects’, or covaries with, the elevation of the sun, that it could be used to measure the sun’s elevation. Further, supposing that there *is* a robust sense in which the shadow in this case would be serving as a structural representation, why wouldn’t Ramsey’s objection to the representational bona fides of a Dretskean conception of receptors carry over to it? That is, why couldn’t we object, by parity with Ramsey’s objection to receptor representations, that if one were to use the tree’s shadow to stay cool, one would be exploiting a causally-mediated homomorphism for a specific purpose — albeit a *non-representational* purpose — and hence that merely satisfying the conditions for being a functioning homomorphism isn’t sufficient for something to play a distinctively representational role?

Ramsey might reply that this question betrays a misunderstanding of the kind of explanatory role that is constitutive of structural representationhood; structural representations are supposed to be homomorphisms between a representing system and a represented system that can be exploited to successfully guide an agent’s inferences or behavior *with respect to the represented system*. Arguably, that’s not happening in the case of using the tree’s shadow for shade — you’re not using the shadow to inform your behavior with respect to the sun’s position, you’re just using it to stay cool. So perhaps this case doesn’t satisfy the conditions for structural representationhood after all. But why couldn’t a proponent of receptor representations reply in a precisely parallel fashion? Indeed, in criticizing the Dretskean reconstruction of the receptor notion of representation, Ramsey seems to overlook a central tenet of Dretske’s theory. According to Dretske, indicators of $C$ become representations when they are enlisted to interface with certain motor commands *by virtue of the fact* that their being so enlisted leads to successful behavior with respect to $C$. As Dretske (1988) puts it, “Only by
using an indicator in the production of movements whose successful outcome depends on what is being indicated can [indeterminacy about the target of a representation] be overcome” (p.70).

If proponents of structural representation can appeal to this sort of use-condition to deflect the shadow-used-as-shade counterexample, why can’t proponents of receptor representations? It certainly seems plausible that neuroscientists have this sort of condition in mind when, for example, they characterize neurons that respond to small, dark moving dots as bug detectors; plausibly, such neurons have been selected to control tongue-snapping because they enable successful bug-catching. Perhaps one might be dissatisfied with the present line of response to the shadow-used-as-shade counterexample; however, it’s not clear why one would be any more satisfied with the response in the case of structural representations than in the case of receptors — or vice-versa. So it’s beginning to appear that structural representations and receptors are on all fours when it comes to meeting the job description challenge.

Now that I’ve exposed some raw concerns about Ramsey’s arguments, I’ll refine them into a more substantive objection. This will involve two main steps. First, note that many of Ramsey’s canonical examples of receptors involve mechanisms whose states vary in proportion to the magnitude of some distal quantity. For example, the conformation of the bimetallic strip in a thermostat varies in proportion to the ambient temperature. Many of the receptor-like mechanisms posited by neuroscientists are similar in this respect. For example, the activity of thermosensors in fruit flies also varies in proportion to the ambient temperature. Call such mechanisms analog receptors. As I suggested above, it seems clear that analog receptors participate in

---

19 As I’m using the term here, ‘analog’ is not synonymous with ‘continuous’. Rather, I’m using ‘analog’ in roughly Lewis’s (1971) sense, according to which magnitudes in the representing system are directly proportional to magnitudes in the represented system. However, when I go on to distinguish between ‘analog’ and ‘binary’ receptors, the distinction I wish to mark is not Lewis’ distinction between analog and digital representation. Binary receptors are arguably still cases of analog representation in Lewis’
homomorphisms, from the various possible magnitudes that the receptor is responsive to, to the various possible states of the receptor mechanism; that is, relations between different different states of the mechanism model relations between different magnitudes. I see no reason to deny this; indeed, it strikes me as an uncontroversial application of conventional ideas about the nature of measurement. It’s difficult to see how one could deny that analog receptors participate in homomorphisms without also denying that, say, thermometers are measuring instruments. Moreover, the sense in which analog receptors participate in homomorphisms seems precisely the same sense in which structural representations do; indeed, the psychologists who have done most to explicate the theoretical underpinnings of the notion of structural representation emphasize the essential similarity between that notion and the notion of representation found in measurement theory (Gallistel, 1990b; Palmer, 1978).  

Now, it’s true that not all of the mechanisms characterized as receptors by Ramsey obviously count as analog receptors. For example, as their name suggests, feature detectors have traditionally been characterized as having the function of detecting the presence of a given property, rather than of covarying with the magnitude of some quantity. It’s worth noting that this interpretation of feature detectors is in fact highly controversial. Problems with this interpretation were revealed by a neural network model that was trained to extract information about the shape of an object from patterns of shading (Lehky & Sejnowski, 1988). The network contained units that came to have response properties that were remarkably similar to the ‘edge’ and

\[^{20}\text{Gallistel (1990b) writes that “representation should have the same meaning in psychology as it has in mathematics” (p.1), and that “those familiar with the theory of measurement... will recognize the parallel between this use of representation and its use in measurement theory” (p.2).}\]

\[^{21}\text{The idea, of course, is not that a feature detector fires when and only when the feature that it is tuned to is in its receptive field; neurons are noisy critters, and are constantly firing even in the absence of external stimulation. The idea is that although a feature detector might fire across a range of frequencies, it’s only firing above a certain threshold frequency that is functionally relevant to its role as a detector.}\]
‘line’ detectors in the visual cortex, yet an analysis of the role that the units played within the network suggested that they served not to detect contours, but to extract curvature information. The authors drew the general lesson that to understand the function of a feature detector, we must look beyond its response profile to the entire role it plays within the network in which it is embedded. Consonant with this suggestion, an emerging consensus in computational neuroscience suggests that feature detectors serve to communicate the degree of error between feedforward sensory information and feedback predictions. From this perspective, feature detectors do serve to covary with magnitude of some quantity, hence count as analog receptors, so all the points made in the previous paragraph apply to them.

However, even if we interpret feature detectors traditionally, as what we might call binary receptors, there’s still a robust sense in which they participate in homomorphisms: relations between states of the mechanism reflect relations between states of the system that the mechanism is responsive to. Such homomorphisms might be very simple, but they’re homomorphisms all the same. Similarly, mechanisms like oil lights on car dashboards, smoke alarms, and pregnancy tests might participate in very simple homomorphisms with the systems they measure, but they nevertheless do participate in homomorphisms with those systems; that’s how they’re capable of measuring them. An oil light is no less a measuring instrument than a fuel gauge, simply because it can occupy fewer states; it’s just a less discriminating measuring instrument.

The essential point underlying this discussion is that homomorphisms are flexible.

---

22A seminal application of this idea to ‘edge’ and ‘line’ detectors is Rao & Ballard (1999).

23This might be a bit quick. Some, such as van Fraassen (2008), argue that there’s simply no sense to be made of the idea that a homomorphism might hold from one concrete, physical system to another, since the technical notion of a homomorphism is only well-defined in the domain of abstract, mathematical systems. This issue deserves further discussion, but it’s orthogonal to my present concerns, since it doesn’t provide any grist for the mill of someone who wants to claim that there’s a substantive theoretical distinction between what I’m calling ‘analog’ and ‘binary’ receptors, or between receptors and structural representations. Insofar as the notion of homomorphism applies to any of these mechanisms, it applies to all of them.
Earlier we saw one dimension along which they’re flexible: ‘horizontally’, they might apply between virtually any two systems. We saw that this dimension of flexibility poses a prima facie problem for theories that attempt to explain representation in terms of homomorphisms, but that the problem can arguably be assuaged by appealing to external causal relations and internal causal roles. However, the present discussion has revealed that even if we restrict our attention to functioning, causally-mediated homomorphisms, homomorphisms are still flexible along a second dimension: ‘vertically’, they might hold at many different levels of granularity. This is because homomorphisms exist wherever there is a relation of reliable causal covariation that might be exploited for measurement purposes, and such relations exist at many different levels of granularity — a mechanism that covaries with some system might be highly articulated and capable of embodying fine-grained distinctions, or very simple and only capable of making binary distinctions. While this flexibility needn’t pose a problem for a structural theory of representation, it does begin to suggest that such a theory will be far more liberal than its proponents might have initially thought.

That was the first step in my objection to Ramsey’s anti-representationalist argument. The second is to observe that, aside the emphasis on homomorphisms, Gallistel’s ‘structural’ theory of representation is essentially just a notational variant of Dretske’s ‘receptor’ theory. Both authors emphasize that cognitive representation is fundamentally a matter of exploiting a causal relation between representing and represented systems so as to guide successful interactions with the represented system. Recall from the discussion at the end of Section 5.2.1 that although both Ramsey and Gallistel sometimes characterize structural representations in terms of surrogative reasoning, neither of them thinks that such ‘reasoning’ need be especially sophisticated. For both authors, what matters is that the represented system is used in the right way; namely, to guide successful interactions with the represented system. But we’ve seen repeatedly now that Dretske emphasizes essentially the same point. If there’s a substantive
difference between Gallistel’s view of representation and Dretske’s, it’s that the former requires that representing systems be homomorphic with represented systems. Yet we just saw that this isn’t a substantive difference at all; insofar as changes in representing systems reliably covary with changes in represented systems, the homomorphism requirement will be satisfied. Homomorphisms are vertically flexible. It’s true that the kinds of mechanisms we intuitively associate with a Dretske-style, receptor-based notion of representation are relatively simple analog or binary receptors. However, there’s no reason such a notion shouldn’t encompass richer mechanisms that are intuitively more model-like. And in any case, when we focus on the theoretical content of the receptor- and structural-based notions of representation, rather than the connotations of words like ‘detector’ and ‘model’, it seems that whatever satisfies the conditions of one notion will satisfy the conditions of the other.

Let’s see how these points play out in the context of a specific example. The physiology and behavior of almost all animals on Earth is modulated by a circadian rhythm that reflects the period of the Earth’s axial rotation. These rhythms are mediated by circadian clocks: biochemical mechanisms that produce an endogenous oscillation of roughly 24 hours, which is entrained to the phase of the Earth’s day-night cycle by zeitgebers, or daylight cues. Circadian clocks play a crucial role in regulating a range of metabolic processes and behaviors that depend for their success on being synchronized with the day-night cycle. Thus circadian clocks seem to satisfy Gallistel’s criteria for representationhood: they enter into homomorphisms with external systems; these homomorphisms are established and sustained by causal relations; and they are exploited in the service of behavioral success. Indeed, Gallistel (1998) agrees: “The circadian clock... is perhaps the simplest of all the well-documented functioning [homomorphisms] between brain processes and the world” (p.28). However, it also seems

---

24The word I replaced here is ‘isomorphism’. In earlier work, Gallistel tended to express his view
clear that circadian clocks satisfy the conditions of Drestke’s criteria for representationhood: they reliably respond to the period and phase of the day-night cycle, and it’s presumably in virtue of that responsiveness that circadian clocks have been naturally selected to guide behaviors that depend for their success on being coordinated with the day-night cycle. And again, Dretske (1988, p.89) himself seems to agree that circadian clocks qualify as representations in his sense.

It’s worth noting that in a few animals, circadian clocks do not participate in functioning homomorphisms. For example, the Somalian cavefish has evolved for millions of years in the absence of sunlight, and possesses a vestigial circadian ‘clock’ with a period of 47 hours (Cavallari et al., 2011). This mechanism plays an important role in mediating internal metabolic processes, but it is not causally responsive to an external system that it guides behavior with respect to. Thus it doesn’t count as a representation in either Gallistel’s or Dretske’s sense. This example underscores that a ‘receptor’ condition is crucial for the cogency of a structural notion of representation, just as a ‘homomorphism’ condition is an essential component of a receptor notion of representation. Again, it seems that whatever satisfies the conditions of one notion satisfies the conditions of the other. There’s simply no distinction between receptors and structural representations.

One might object at this point that although there’s no categorical distinction between receptors and structural representations, there’s nevertheless a graded distinction. Just as there are genuinely bald people, even though the category of baldness is fuzzy at the boundaries, there are genuine receptors, even though that category is fuzzy; receptors are just, as it were, degenerate functioning homomorphisms. This is true, but irrelevant. While we might distinguish between very simple and more

in terms of isomorphisms, but in more recent work he expresses it in terms of homomorphisms, presumably due to a recognition of the problems with an isomorphism-based view of representation of the kind we discussed earlier.
complex functioning homomorphisms, the question at issue is whether this marks a distinction between representations and non-representations. And I see no theoretical or intuitive reason to think that it does. Ramsey (2007) writes that he’s “willing to be fairly unrestrictive about what qualifies as a map, model or simulation, as long as there is a clear explanatory benefit in claiming the system uses the structure in question as such” (p.82). I quite agree; I just think that the notion of a functioning homomorphism captures a clear sense in which a structure is used as a model even when the homomorphism is very simple. To reiterate an earlier example, a car’s oil light is no less a measuring instrument than a fuel gauge simply because it makes fewer distinctions; the richness of the homomorphism it participates in pertains to its representational expressiveness, not to its representational status.

Ramsey thinks that analogies like these pump our intuitions in misleading directions. He thinks that receptor-like artifacts such as oil lights or pregnancy tests can only function as representations when they’re interpreted by intelligent agents, and that when we find it intuitive that sub-personal receptors might function as representations in a similar way, “we overlook the fact that when the cognitive agent is removed from the picture... the process becomes just like other causal processes that we intuitively judged to be non-representational in nature” (p.218). On the other hand, he thinks that structural representations can function as representations without being interpreted by intelligent agents: “[a] mindless system can... take advantage of the structural isomorphism between internal structures and the world, and in so doing, employ elements of those internal structures as representations-qua-stand-ins” (p.200). Note that this objection presupposes the distinction I’m objecting to; it’s not clear why the quote just mentioned wouldn’t hold true of very simple ‘structural isomorphisms’ such as those mediated by binary receptors, as well as the more complex ones that Ramsey seems to have in mind. But merely replying that there is no distinction between intuitively receptor-like and intuitively map-like systems with respect
to Ramsey’s concerns about expunging the homunculus might be a pyrrhic victory for someone who wants to defend representationalism from Ramsey’s arguments, for one might worry that Ramsey’s concerns about the automatization of receptor-like surrogates applies to all surrogates; as Garzón & Rodriguez (2009) point out, one might worry that the homomorphisms that purportedly underlie a surrogative function can only be fixed by the interpretative faculties of an intelligent agent.

I think this concern is mistargeted; although both Ramsey (2007) and Garzón & Rodriguez (2009) are right to raise concerns about how surrogates could be automated, they fail to identify when and why those concerns arise. As I mentioned earlier, when we focus only on structural resemblances that are grounded in the static, ‘pictorial’ properties of a system, it’s difficult to see how those resemblances could be exploited without someone to perceive or interpret them. However, as Craik (1943) insightfully pointed out, if we shift our attention to resemblances that are grounded in the dynamic, functional properties of a system, it becomes clearer how resemblances could be exploited mechanistically; mechanisms can resemble one another not just in how they look, but in how they work. But that’s true regardless of whether the operation of a mechanism sustains simple resemblances or more complex ones. I take it that there’s no deep mystery about how, for example, the ‘resemblance’ between the activity in a fly’s thermosensors and the ambient temperature can be exploited to enable the fly to avoid aversive temperatures.

To repeat a familiar refrain, the key question about whether a system counts as a representation concerns not how complex it is, but whether it is used as a surrogate.

---

25To relate the point here to my earlier characterization of functioning homomorphisms in terms of measurement theory, measurement needn’t involve the assignment of numerals to a given system, which must be interpreted by an intelligent agent; measurement procedures can be automated within a system — think, for example, of how the measurement of temperature is automated within a thermostat. An automated measurement procedure just is a procedure that mediates a functioning homomorphism in Gallistel’s sense.
for another system. This is the central idea underlying Gallistel’s ‘structural’ view of representation, but it’s also the central idea underlying Dretske’s ‘receptor’ view. Following Frank Ramsey, Dretske frequently employs a metaphor that William Ramsey would surely find congenial: representations are maps by means of which we steer (ibid., p.79). So Dretske doesn’t differ from the latter Ramsey in his conception of what representations essentially do, he just has a broader conception of the kinds of mechanisms that can function as internal, behavior-guiding ‘maps’. Of course, nobody thinks that maps in the relevant sense need be very much like cartographic maps, with contour lines and a compass rose. The arguments in the present section suggest that to exclude receptors from the ambit of a structural conception of representation is to take the metaphor of a cartographic map too literally. Receptors can participate in homomorphisms, and can function as behavior-guiding ‘maps’ by doing so. Receptors just are structural representations.

5.3.2 Mental Representations

At this point, a proponent of structural representations might complain that I’ve been less than charitable. Surely, the complaint might go, there’s some distinction to be made between mechanisms like feature detectors and the kinds of mechanisms that psychologists have had in mind when they’ve talked of ‘mental models’, which goes beyond a mere difference in the relative richness of the homomorphisms that those mechanisms participate in. After all, mental models are supposed to be mental, whereas feature detectors are arguably not that different from simple mechanical devices that might be found in all sorts of non-intentional systems such as thermostats. The picture we get of mental models from pioneering discussions by Craik (1943), and later elaborations by classical cognitive scientists like Johnson-Laird (1983), is that they’re introspectively accessible states that agents can manipulate in thought, independently of the flux of real-time perception, which enable agents to achieve relatively sophisticated cognitive
feats, such as drawing lessons from past experiences, planning for the future, reasoning about counterfactual possibilities, and so forth. This certainly seems to be what Craik (1943) had in mind when he characterized mental models as states that enable an agent to try out “possible actions within its head” (p.51), and what Johnson-Laird (1983) had in mind when he invoked mental models to explain capacities like counterfactual reasoning. It also seems to be what Cummins (1994) has in mind when he writes that “sophisticated cognition” is enabled by the mind’s capacity to “operate on something that has the same structure as the domain it is said to cognize” (pp.297-298).

The ‘sophisticated’, stimulus-independent cognitive capacities associated with the picture of mental models I’ve just sketched have historically engendered the most venerable and compelling arguments for thinking that there are such things as mental representations in the first place. For example, the best argument that Locke provides for believing in *ideas* appeals to our capacity to think about things in memory and imagination. Since those things aren’t literally in our minds, Locke claims, “it is necessary that something else, as a sign or representation of the thing it considers, should be present to [the mind]; and these are ideas” (Locke, 1689 [1975], IV.xxi.4, my emphasis). Clark & Toribio (1994) update this argument for the information-processing age by pointing out that certain cognitive capacities are ‘representation-hungry’. As Clark (2001) later puts the point, “Internal representations look prima facie essential for such... activities as dreaming of Paris, mulling over U.S. gun control policy, planning next year’s vacation, counting the windows of your New York apartment while on holiday in Rome, and so on... All these cases, on the face of it, require the brain to use internal stand-ins for potentially absent, abstract, or non-existent states of affairs” (p.129).

It seems to me that considerations like these provide a quick argument against Ramsey’s anti-representationalist and pro-eliminativist conclusions. It’s surely uncontroversial that we *have* stimulus-independent cognitive capacities, such as the capacity
to recall episodic memories about specific events in our past, prospectively ‘project’ ourselves into future episodes, entertain pseudo-perceptual mental imagery, and so forth, and that such capacities stand in need of explanation. Moreover, it seems reasonable to call the mechanisms that explain those capacities, whatever they turn out to be, ‘representations’, simply by virtue of the capacity they explain — in the same way that it would have been reasonable to call the mechanisms of photosynthesis, whatever they had turned out to be, ‘sunlight-to-food converters’.

However, I will not pursue this line further since my primary goal is not to undermine Ramsey’s conclusions, but rather to scrutinize the premises that purportedly lead to those conclusions, so as to better understand the explanatory ambit of the widely invoked notion of structural representation. The preceding discussion has suggested that ‘sophisticated’, stimulus-independent cognitive capacities are mediated by mental representations, and that there’s an intuitive distinction between such representations and simple mechanisms like feature detectors, whether or not there might be some attenuated sense in which the latter count as representations. But the question at issue is whether that intuitive distinction maps on to the distinction between structural representations and receptors. We just saw that when Clark discusses stimulus-independent cognitive capacities, he emphasizes the importance of appealing to internal stand-ins. This seems to accord with Ramsey’s emphasis on the surrogative, standing-in role that in his view is constitutive of representationhood, however, the two authors mean quite different things by ‘standing-in’. For Clark, a stand-in is a system that can be manipulated offline, in the absence of a direct causal connection with the system it stands in for, whereas for Ramsey, a stand-in is a system that enables successful interactions with the system it stands in for, by virtue of the fact that both systems embody the same abstract structure. While the former author doesn’t emphasize the structural resemblance aspect, and the latter doesn’t emphasize the offline manipulation aspect, nevertheless both aspects seem complementary; a natural thought is that it’s precisely
because system $A$ embodies the same abstract structure as system $B$ that $A$ can be used as an effective offline surrogate for $B$.\textsuperscript{26} So perhaps we might explore the idea that structural representations and receptors are not to be distinguished on the basis of representational status, but rather on the basis of the basis of representational genus; the idea would be that what demarcates structural representations from receptors is that only the former can be decoupled from the systems they’re functionally homomorphic with, and moreover that such decoupleable functioning homomorphs are mental models, the vehicles of stimulus-independent cognition. Receptors, qua functioning homomorphs, might be representations, but they’re not, on this view, mental representations.

Now, Ramsey, Gallistel, and many other proponents of structural representations do not in fact hold that decoupleability is a necessary condition for something to count as a structural representation, or that structural representations are distinctively mental. Recall one of Gallistel’s examples of a structural representation, a circadian clock. Circadian clocks are found not just in animals, but in almost all organisms on Earth, including plants. And like the circadian clocks in animals, those in plants exist for a reason, namely to enable the organism to engage in behaviors (or ‘behaviors’, if you prefer) that depend for their success on being coordinated with the day-night cycle. So circadian clocks in plants seem to satisfy the conditions for being structural representations; indeed, they function in essentially the same way as those in animals, and involve many of the same evolutionarily-conserved biochemical mechanisms (Kay,

\textsuperscript{26}This line of thought seems to underlie many discussions of structural representation, which employ terms like ‘standing-in’ and ‘surrogative reasoning’. However, it’s often unclear whether these terms are being used in Clark’s sense or Ramsey’s. My impression is that there’s a tendency in the literature to assume that for a representation to be used as a surrogate just is for it to be manipulated offline, but as our discussion of Ramsean surrogates has shown, many surrogates are causally coupled to the systems they’re surrogates for. Indeed, this is arguably the normal function of surrogates — think of a map being used in conjunction with landmark recognition to navigate through an environment.
Yet arguably they can’t generally be manipulated offline, and they’re certainly not distinctively mental. In personal communication, Ramsey has agreed that circadian clocks in plants might count as structural representations; he emphasizes that what’s essential is whether a system is used as an internal surrogate, not whether it’s distinctively mental. But in any case, to see the point at issue, we needn’t look to such exotica as circadian clocks in plants. Recall that one of Ramsey’s central examples of a structural representation is the S-shaped groove in the toy car depicted in Figure 5.1. This cannot in any sense be manipulated ‘offline’; it works by being directly coupled to the car’s motion. Moreover Ramsey expressly discusses this example to make the point that structural representations aren’t distinctively mental; he argues that structural representations can perform their surrogative function in mindless systems like toy cars, and hence can function as representations without being interpreted by an intelligent homunculus (see pp.193-203).

Nevertheless, we might take the proposal that structural representations are necessarily decoupleable as a friendly amendment to Ramsey’s view, which strengthens his anti-representationalist and pro-eliminativist argument. For it seems to provide a way for Ramsey to distinguish structural representations from receptors, and hence to avoid the skeptical argument I developed in the Section 5.3.1. Moreover it does so in a way that is congenial to the eliminativist strand of his argument. Ramsey’s eliminativist argument, as it stands, seems strangely at odds with his view of representation. Recall the contours of the argument: while classicism once attempted to explain cognition by invoking genuinely representational ‘structural’ states, connectionism is eclipsing classicism and promises to explain cognition in terms of non-representational ‘receptors’. But what if connectionist explanations were to prove inadequate, and classicism experi-

\[\text{\textsuperscript{27}}\text{To underscore the point made in the previous section, note also that circadian clocks in plants satisfy the conditions for being Dretskean receptors, for essentially the same reasons.}\]
enced a resurgence? If, as Ramsey seems to think, there’s nothing distinctively *mental* about the structural representations posited by classical explanations, how would the explanatory success of classicism vindicate intentional psychology and dispel the bogey of eliminativism? Whether or not a cognitive theory threatens the existence of mental states seems to turn on whether it posits mental representations, not representations *simpliciter.*

But while it might be charitable to stipulate that decoupleability is a necessary condition for something to be a structural representation, I see no corresponding reason to hold that the *lack* of decoupleability is a necessary condition for something to be a receptor. Many of the most representative models of the neo-connectionist literature of the ‘80s were simulations of stimulus-independent cognitive capacities, and didn’t assume that the representations they modeled could be elicited only by immediate perceptual stimulation. Indeed, in the ‘bible’ of neo-connectionism, Rumelhart et al. (1986c) explicitly address the question of how activity in neural networks might be manipulated offline in the course of cognitive processes like planning or counterfactual reasoning, and suggest that this is achieved by a “mental model” instantiated by a “relaxation network which takes as input some specification of the actions we intend to carry out and produces an interpretation of ‘what would happen if we did that’” (p.41), thereby generating predictions about events in the world. In accordance with this general picture, recent evidence suggests that even feature detectors can be elicited offline; for example, neurons in early visual cortex seem to be activated during visual mental

---

28 Of course, if one could show that a cognitive theory doesn’t posit representations in *any* sense, one would have shown that it doesn’t posit *mental* representations. But as I pointed out at the beginning of Section 5.3, Ramsey’s argument doesn’t show that; even granting all the premises, the most the argument could show is that connectionist mechanisms aren’t representations by *virtue of functioning as receptors*, not that such mechanisms fail to be representations *simpliciter*. Indeed, the naturalistic methodology that Ramsey adopts precludes him from developing the kind of argument at issue; recall that Ramsey rightly holds that it is forlorn to seek a general analysis of representation that will encompass all and only representations.
imagery (Kosslyn & Thompson, 2003). Of course, one might stipulate that receptors cannot be manipulated offline so as to preserve the distinction between structural representations and receptors, but the resulting notion of a receptor wouldn’t be the one that could feature in Ramsey’s anti-representationalist argument, for there’s no reason to think that connectionists employ such an etiolated notion.

Putting Ramsey’s argument aside, haven’t we at least found that by cashing the notion of structural representation in terms of decoupleable functioning homomorphs, we’ve helped to elucidate the nature of ‘mental models’, the vehicles of stimulus-independent cognition? Not at all. For despite what I suggested earlier, circadian clocks in plants can be manipulated offline in the service of anticipatory behavior. In a fascinating discussion of adaptive behavior in plants, Garzón & Keijzer (2011) point out that some plants, notably *Lavatera cretica*, are capable of reorienting their leaves overnight so as to face the ‘anticipated’ location of the sun in the morning, thereby maximizing daily light intake. The circadian clocks in *Lavatera* can sustain this behavior over several days in the absence of sunlight (Schwartz & Koller, 1986). Similarly, the circadian clocks in plants of the genus *Arabidopsis* help to muster an anticipatory chemical defense against herbivorous insects overnight, even in the absence of sunlight (Goodspeed et al., 2012). The circadian clocks in these plants count as functioning homomorphs, as I argued earlier, and there seems to be a robust and interesting sense in which they can be used offline, so they seem to count as structural representations in the revised sense under consideration; nevertheless, they surely don’t count as mental representations, i.e. the vehicles of cognitive processes like episodic memory, planning, and mental imagery. Although Garzón & Keijzer (2011) take the evidence of offline anticipation in plants to show that plants are “cognitive in a minimal, embodied sense” (p.166), they don’t hold that plants have minds in the full-blooded sense that’s presently at issue. They rightly point out that for the purpose of understanding the continuity between life and mind, it is important to explore the oft-neglected com-
monalities between plant and animal behavior, but that “it goes without saying that for other purposes the emphasis may be rightly placed upon the [differences]” (p.156). Just so; for the purpose of elucidating the nature of mental representations, the notion of structural representation — even if supplemented with a decoupleability condition — simply won’t do. The claim that structural representations “make sophisticated cognition possible” might be true, but only in the sense in which atoms make sophisticated cognition possible; it can’t be the manipulation of structural representations as such that makes sophisticated cognition possible, since structural representations are manipulated in all sorts of mindless systems, such as plants.

29I should note that I don’t advocate an anthropocentric view of the domain of organisms that possess mental representations. I think that many animals, and probably even insects, have mental representations. For example, Clayton et al. (2001) have provided compelling evidence that scrub jays have episodic memory. But the ingenuity of this experiment, and the controversy surrounding its interpretation, highlights the fact that the capacities that mental representations mediate are highly non-trivial to demonstrate, and are substantially different from the ‘cognitive’ capacities of plants.
Chapter 6

Conclusion: Steps Toward a New Job Description

I’m caught in between future’s fantasies and memories I played back.

— Vast Aire, *The Cold Vein*

In attempting to articulate a notion of representation that will play a role in cognitive explanations, philosophers and psychologists have frequently appealed to notions like isomorphism and structural resemblance. Proponents of these ‘structural’ notions of representation often develop their views against the backdrop of the picture of mental models that I sketched at the beginning of the last chapter: states that can be manipulated in thought, during cognitive processes like episodic memory, counterfactual reasoning, and mental imagery. However, when these theorists attempt to articulate what exactly structural representations are, the background picture tends to fall away, unnoticed. We see this clearly in Ramsey’s (2007) discussion of structural representations. Although Ramsey often characterizes structural representations as ‘mental models’ of the kind discussed by Johnson-Laird (1983), we’ve seen that the substance of his view is best captured by Gallistel’s (1998) notion of a *functioning homomorphism*,
and that there’s nothing distinctively mental about representations so understood — even if we add the further condition that functioning homomorphs must be decoupleable from the systems that they’re homomorphic with. Functioning homomorphs, whether decoupleable or not, can be found in all sorts of non-mental systems, such as plants. Indeed, Ramsey explicitly endorses the claim that structural representations can be found in non-mental systems, as a premise in his argument that such representations needn’t be interpreted by a homunculus to function as representations.¹

The widespread tendency to conflate structural representations with mental models I think helps to explain some of the intuitive appeal of Ramsey’s argument against the representational bona fides of connectionist mechanisms. There seems to be a clear difference between the mental models mediating our thoughts about the past and future, and, say, single neurons that respond to oriented lines. So if one identified this difference with the purported distinction between structural representations and receptors, one might find Ramsey’s argument compelling. But this would involve a threefold error: First, it would rest on an etiolated caricature of the explanatory repertoire of connectionists; mainstream connectionists appeal to a range of complex, hierarchically structured systems to explain ‘sophisticated’ cognition. Second, it would presuppose a distinction between structural representations and receptors that has no theoretical content. And third, it would assume, falsely, that structural representations are distinctively mental.

In closing, I’d like to look at this third assumption more closely, since I think it can

¹It’s worth noting that this strategy for showing that structural representations can function as representations in the absence of a homunculus — what Ramsey calls the ‘mindless strategy’ of showing that structural representations can play a role within non-mentalistic systems, and hence aren’t distinctively mental — seems to conflate two distinct questions. It’s one thing to ask whether a type of representation is distinctively mental, and it’s quite another to ask whether a type of representation can function as a representation within a purely mechanistic system. While the mindless strategy might be sufficient for expunging the homunculus, it’s surely not necessary; to suppose otherwise would seem to assume a kind of dualism according to which minds cannot be explained mechanistically.
help diagnose where Ramsey’s reasoning went awry. Although Ramsey officially holds that a representation is any internal mechanism that functions as a behavior-guiding map, his thinking sometimes seems to be guided by a tacit conception of representation as something distinctively mentalistic. When he evaluates the representational credentials of structural and receptor representations, he seems to vacillate between these two conceptions. He seems to evaluate receptor representations with respect to a mentalistic conception of representation, and rightly concludes that they’re not representations in this sense, yet he seems to evaluate structural representations with respect to a non-mentalistic conception of representation, and rightly concludes that they are representations in this sense. But because of the tacit equivocation, he fails to notice that structural representations and receptors are the same in both respects: both are representations in that they function as behavior-guiding maps, but neither plays a distinctively mentalistic role in doing so.

This tacit equivocation sometimes becomes explicit. Recall that Ramsey sometimes characterizes the role of an internal, behavior-guiding map in terms of what Swoyer (1991) calls ‘surrogative reasoning’. Consonant with a strict interpretation of ‘surrogative reasoning’, as something that only agents are capable of, Ramsey holds that the central difference between receptors and ‘real’ representations is that the latter, but not the former, participate in “a process that is properly or naturally viewed as something like learning about or making inferences about some state of affairs with which the representation stands in some nomic relation” (2007, p.141). However, we’ve also seen that Ramsey often thinks of surrogative reasoning far more liberally, as a process that, for instance, might be carried out by the mechanically-driven rudder in the toy car depicted in Figure 5.1. It’s this liberal conception that underwrites Ramsey’s claim that the “surrogative, ‘standing-in-for’ property is not dependent upon any inference or learning process” (ibid., p.201, my emphasis). So Ramsey seems to think of surrogative reasoning as a distinctively mentalistic process when evaluating the representational
status of receptors, and denies that receptors are representations on that basis, yet he thinks of it far more liberally when evaluating structural representations.

I think this reveals two important lessons for discussions about the role of representation in cognitive explanations. The first is that it’s crucial to be clear about whether or not ‘representation’ is being used as shorthand for ‘mental representation’. Just as there’s little reason to think that all the things we call representations, ranging from traffic lights to thermoreceptors, fall under the same explanatory generalizations, we shouldn’t assume that sub-personal representations and mental representations have much in common beside the label ‘representation’. The second is that, contrary to widespread assumptions, the notion of a structural representation, qua an offline, behavior-guiding map, doesn’t help to explicate the distinctive nature of mental representation. One might complain that in describing my purported counterexample of a circadian clock that is used offline to control anticipatory leaf reorientation behavior in plants, I’m using terms like ‘offline control’, ‘anticipation’ and ‘behavior’ in ways that are far more liberal than proponents of mentalistic construals of structural representations would like. But that’s precisely my point; simply invoking these terms doesn’t explicate a robust notion of distinctively mental representation. One must imbue those terms with content.

How do we imbue a notion of mental representation with content? That’s a story for another time. However, I’d like to close this dissertation with a couple of suggestive observations. Contrary to some caricatures of connectionism as the heir to behaviorism, research into the mechanisms of ‘representation hungry’, stimulus-independent cognitive capacities is one of the most active and productive areas of contemporary neuroscience. The emerging picture, supported by a wealth of neuroimaging, neuropsychological, and neurophysiological evidence, is that capacities like episodic memory, prospective ‘mental time travel’, and mental imagery involve overlapping functional networks with a common core, which subserves a simulative function much like that
envisioned by Rumelhart et al. (1986c): top-down signals elicit a ‘re-enactment’ of activity in cortical areas involved in bottom-up perception.\textsuperscript{2} These areas contain feature detectors of the kind that Ramsey disparages as mere ‘receptors’. So, far from undermining our self-conception as intentional agents with rich mental lives, contemporary connectionism seems well on its way to revealing the mechanisms of our mentality. And even the humble feature detector seems to be playing a crucial role in that endeavor.

\footnotesize\textsuperscript{2}See Danker & Anderson (2010), Kent & Lamberts (2008), and Schacter et al. (2008) for reviews.
Bibliography


Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cam-
bridge, MA: MIT Press.


Hall.


Tolman, E. (1938). The determiners of behavior at a choice point. *Psychological


