From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Neuroscience

2014

Beyond simple model-free reinforcement learning in human decision making

Solway, Alec

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01fb494856q

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Beyond simple model-free reinforcement learning in human decision making

Alec Solway

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the
Princeton Neuroscience Institute
Adviser: Matthew M. Botvinick

June 2014

© Copyright by Alec Solway, 2014.

All rights reserved.

Abstract

Over the last two decades, there has been a large scale effort in cognitive neuroscience to understand learning and decision making from the perspective of simple model-free reinforcement learning algorithms. This interest was invigorated in the mid 1990's, when it was realized that the phasic activity of midbrain dopaminergic neurons resembles reward prediction errors. The algorithms studied formalize the notion of learning from past experiences through trial and error. Although important, there are many aspects of behavior they cannot explain. More recent work has begun to fill in some of these gaps by borrowing yet additional ideas from computational reinforcement learning. One line of inquiry has concentrated on aligning goal-directed behavior, which resembles the common sense notion of "planning", with *model-based* reinforcement learning. This work has aimed to understand how the brain is able to learn the world model prescribed by the model-based framework, and to characterize the neural correlates of the value functions it predicts. This thesis adds to this work by offering two separate, but related, algorithmic accounts of how the brain may be able to actually map the world model into a decision. Existing data are examined and new experiments are performed. A second line of inquiry has concentrated on understanding behavior from the perspective of *hierarchical* reinforcement learning. The thesis makes two contributions to this area as well. First, it is shown that the brain codes pseudo-reward prediction errors, a prediction error in response to a faux reward signal that is used to train skills that are not in themselves useful, but that may be used to achieve other means. Second, an optimality framework is provided for understanding which skills are most beneficial to have when confronted with an ensemble of tasks.

To my parents,

Marina and Jim

Acknowledgements

I am eternally grateful to a number of people without whom this work would be substantially worse, or not exist at all. My parents and I emigrated from Ukraine when I was six, with only four suitcases in tow. It is a somewhat typical story of Soviet émigrés of the time, but not one without challenges. The ingenuity and stubbornness they displayed in making something out of nothing has been a deep source of inspiration for me. For their support, I am always grateful. I dedicate this thesis to them.

My most cherished memories from Ukraine are from the time spent with my grandparents, Aleksey and Yelena. My grandfather brought home a seemingly infinite supply of what I can best describe as small "spare parts" from the engineering facility where he worked. These were the tinkertoys of the Soviet era. He taught me to play chess. I can't help but think that those experiences during my formative years helped structure my thinking, leading me to engineering and then science years later. My grandmother has been a tremendous source of support and inspiration, from her stories about fleeing Moscow during the war, to her learning to use a computer for the first time at the age of eighty-two. My sister Faye has provided love and friendship, aptly timed comic relief, and most importantly, a great deal of indulgence in my suggestions of watching vast amounts of Star Trek during her recent summer visits.

My longtime partner Elizabeth provided day-to-day support throughout the entire process. I am thankful for every minute of it, from her lending an ear, to making a great meal or cup of tea when I had to work late into the night. She is truly one of the most kind and generous people I know.

A number of people are responsible for my professional development. Before starting my Ph.D., I had the fortune of working with and being mentored by Michael Kahana. Mike taught me how to think like a computational cognitive

neuroscientist, how to be rigorous but bold with my ideas, and how to navigate the many non-scientific aspects of a scientific career. He has always been generous with his ideas (which seem to come out of nowhere and in endless supply), advice, and time. Many members of the Kahana lab had a large influence on me as well. I will not name all of them, but I want to single out Christoph Weidemann, with whom I worked on my very first research project looking for grid cells in human epilepsy patients, and Sean Polyn, who did not mind me constantly barging into his office to ask questions about computational modeling. Marieke van Vugt and Per Sederberg, although also members of the Kahana lab, I worked closely with only when the three of us were at Princeton. Both provided valuable guidance during my rotation projects the first year.

My Ph.D. adviser, Matthew Botvinick, has been my mentor and guide during the last four and a half years. One thing to know about Matt is that the scope of his curiosity truly has no bounds. A Neuroscience Institute poster from a few years back had a listing of the faculty and their interests. It listed the specialty of each, and for Matt it said "the neural basis of cognition." I think this is an apt description — Matt tackles problems from a wide spectrum — and does so with incredible skill and insight. Working with him has expanded my scientific horizon in a multitude of completely unexpected ways. In addition, Matt taught me how to break down and present highly technical material in an accessible fashion. Not only has this improved my writing and oral presentations, but it has given me a new way of thinking about problems.

Many others at Princeton were also influential in my development. I want to further single out José Fernandes and Carlos Diuk; our fun and successful collaborations are reflected in the work of Chapters 4 and 5. In addition to the training I received in the process of doing my dissertation work, I also learned a great deal in the classroom. In particular, David Tank and Alan Gelperin put a lot

of effort into constructing and supervising an enormously fun and useful lecture and lab course that all trainees take in their first semester. Sofia Machado Dibble was my fearless lab partner during this time, providing company and laughs as we debugged our rig late into the evening.

A great deal of gratitude goes to Debbie Yee and Jin Hyun Cheong, who helped run many of the behavioral experiments.

Finally, I would like to thank the rest of my defense committee, who in addition to Matt, consisted of Carlos Brody, Jonathan Cohen, Yael Niv, and Jordan Taylor.

Contents

Abstract iii

Acknowledgements v

List of Tables xiii

List of Figures xiv

1 Introduction **1**

1.1 Habits, learning, and basic model-free reinforcement learning algorithms 1

1.2 The need for model-based reinforcement learning 7

1.3 Hierarchical reinforcement learning 11

2 Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates **15**

2.1 Abstract 15

2.2 Introduction 16

2.3 Goal-directed decision making: definition and manifestations 17

2.3.1 Toward a computational account 21

2.4 Reframing the computational problem 24

2.4.1 A probabilistic representation of reward 26

2.4.2 A generative model for reward 28

2.4.3 Abductive inference 31

2.5 Algorithmic framework 33
2.5.1 Evidence integration 35
2.5.2 Goal-directed decision via iterative inference 37
2.6 Simulations 39
2.6.1 Simulation 1: instrumental choice 40
2.6.2 Simulation 2: sequential decision 47
2.6.3 Predictions 50
2.7 Neural implementation 53
2.7.1 Four interlocking neural systems 54
2.8 Neural network model 58
2.8.1 Simulations 61
2.8.2 Predictions 70
2.9 General discussion 72
2.9.1 Related work in machine learning and theoretical neuroscience 73
2.9.2 Spreading activation models of spatial navigation 74
2.9.3 Evidence-integration models of decision making 75
2.9.4 Departures from rationality 77
2.9.5 Learning 79
2.9.6 Capacity limitations, heuristics, and problem representation . 81
2.9.7 Relations with habitual action selection 84
2.9.8 Conclusion 85
2.10 Appendix A: Formal analysis 86
2.10.1 Monotonicity 86
2.10.2 Convergence 89
2.10.3 Convergence under random utility 90
2.10.4 Relation to sequential-sampling models 91
2.11 Appendix B: Simulation procedures 93

2.11.1 Graphical model . . . 93
2.11.2 Neural network . . . 94
2.12 Addendum: Notes on belief propagation . . . 97
2.12.1 Messages for one time slice . . . 97
2.12.2 Messages for two time slices . . . 99

3 Evidence integration in model-based tree search 101
3.1 Abstract . . . 101
3.2 Introduction . . . 102
3.3 Results . . . 105
3.3.1 Experiment 1 . . . 105
3.3.2 Model . . . 110
3.3.3 Experiment 2 . . . 114
3.4 Discussion . . . 116
3.4.1 More complex decision trees . . . 117
3.4.2 Attention . . . 119
3.4.3 Probabilistic transitions . . . 120
3.4.4 Response variability . . . 120
3.4.5 Implementation and neural correlates . . . 121
3.4.6 Relation to previous work . . . 123
3.4.7 Conclusion . . . 123
3.5 Materials and methods . . . 124
3.5.1 Participants . . . 124
3.5.2 Task . . . 124
3.5.3 Model . . . 127

4 A neural signature of hierarchical reinforcement learning 129
4.1 Summary . . . 129

4.2 Introduction . . . 130
4.2.1 EEG experiment . . . 134
4.2.2 fMRI experiments . . . 136
4.2.3 Behavioral experiment . . . 138
4.3 Discussion . . . 141
4.4 Experimental procedures . . . 144
4.4.1 An HRL model of the delivery task . . . 144
4.4.2 EEG experiment . . . 145
4.4.3 fMRI experiments . . . 149
4.4.4 Behavioral experiment . . . 154
4.5 Supplemental information . . . 156
4.5.1 EEG experiment . . . 156
4.5.2 fMRI experiment 1 . . . 157
4.5.3 Supplemental figures . . . 161

5 Optimal behavioral hierarchy **165**
5.1 Abstract . . . 165
5.2 Author summary . . . 166
5.3 Introduction . . . 166
5.4 Results and discussion . . . 170
5.4.1 Formal approach . . . 170
5.4.2 Behavioral experiments . . . 176
5.5 Methods . . . 182
5.5.1 Computational analysis . . . 182
5.5.2 Ancillary procedures . . . 187
5.6 Supplementary methods and results . . . 195
5.6.1 Computational analysis . . . 195
5.6.2 Supplementary experimental results . . . 200

6 Future directions **204**

References **210**

List of Tables

2.1 Specification of belief propagation messages employed in Simulation 3 95

3.1 Hand tuned model parameters 116

List of Figures

2.1 Examples of goal-directed decision making 20
2.2 Model-based reinforcement learning as Bayesian inference 27
2.3 Different ways of searching for optimal policies 30
2.4 Evidence integration in the Bayesian network 36
2.5 Example of evidence integration 38
2.6 Results of Simulations 1-2 43
2.7 Variance in choice and reaction time in the Bayesian network 45
2.8 Neural network model 62
2.9 Neural network simulation of Padoa-Schioppa and Assad (2006) . . 64
2.10 Neural network simulation of Lau and Glimcher (2008) 65
2.11 Simulation of Mushiake et al. (2006) 67
2.12 Evidence integration in simple incentive choice 69

3.1 Decision trees representing decision problems in previous and current work 105
3.2 The sequence of events in each experimental trial. 106
3.3 Results of Experiment 1 107
3.4 A cartoon depiction of the model 111
3.5 Results of Experiment 2 115

4.1 Illustration of HRL Dynamics 132

4.2 Task and Predictions from HRL and RL 134
4.3 Results of EEG Experiment 135
4.4 Results of fMRI Experiment 1 136
4.5 Results of Behavioral Experiment 139
4.6 Task and relationship between joystick and truck movement 162
4.7 Accuracy, reaction times, and evoked potentials conditioned on these variables 163
4.8 ERPs for conditions involving a shift of attention 164
4.9 BOLD activity contrasted for conditions involving a shift of attention 164

5.1 Rooms domains 168
5.2 Results of behavioral experiments 177
5.3 Calculating model evidence example 195

Chapter 1

Introduction

1.1 Habits, learning, and basic model-free reinforcement learning algorithms

Imagine driving to work each day. If you have worked in the same location for even a short but reasonable amount of time, a large part of this activity is likely to be effortless. Although you're receiving a barrage of rich sensory cues, your *responses* to these cues are well established. When arriving at Harrison Street, make a right. At Nassau Street, make a left. If another car appears within a threshold distance, begin to brake. Although your actual behavior is admittedly much more complex (and as will be revealed shortly, is the topic of this thesis), much of it fits well with the common sense notion of "habit". Based on the information you receive from the world (the stimulus), you execute a pre-programmed response. But how did you learn this response? And why this response? Out of all of the possible things you could do (smack your lips, pound your chest), you learned to *turn right* at Harrison Street.

Formal work studying conditioning in the laboratory can be traced back almost a century to the seminal work of Pavlov (Pavlov, 1927, but see also Thorndike,

1911). Stumbling on his initial result by accident, Pavlov went on to map what we now call classical or Pavlovian conditioning. In this setting, a conditional stimulus (e.g. the ringing of a bell) is paired with an unconditional stimulus (e.g. food). The unconditional stimulus is usually understood to signal primary reward, which determines individuals' fitness on evolutionary timescales. Obvious examples are food and sex. The supposition is thus that first-order (or perhaps more accurately, zeroth-order) learning has already been "solved" by natural selection, and there is nothing to do be done. Organisms will prefer stimuli which favor the survival of their genetic lineage over those that don't. The question then is, how higher order, or more complicated, contingencies are learned. In the simplest case, repeated trials with the conditional stimulus preceding the unconditional stimulus allows the former to build up value over time. This build up is measured by the conditional response, a new response to the previously neutral stimulus. For example, after many pairings of a bell with food, the bell itself begins to elicit a salivary response, presumably because it comes to *predict* the food.

While classical or Pavlovian conditioning addresses the situation in which an organism learns to predict exogenous events controlled by the environment, operant or instrumental conditioning concerns how such stimuli may be combined with actions, with the goal being to influence the environment to produce the greatest reward (Skinner, 1935; Thorndike, 1911). In general, food does not magically appear after the ringing of a bell; an animal has to instead go out and look for it. In the laboratory, this is operationalized by simple actions, such as pulling a lever. Although the instrumental learning problem appears in some sense harder than the classical conditioning problem, both involve the transfer of information between a rewarding stimulus and a previously neutral entity (another stimulus, or the conjunction of a stimulus and action).

The data generated by work on conditioning proved to be fertile ground for theorists, who aimed to build compact computational descriptions (models) that could simultaneously capture the many nuances of learning. Although many models were built, a few deserve special mention because they are now widely considered direct progenitors of the modern cognitive neuroscientific view of learning.

Bush and Mosteller (1951) provided a model of reinforcement which is summarized by the simple equation:

$$\Delta p = a(1 - p) - bp. \tag{1.1}$$

Here, p is the probability of taking an action, a is a parameter controlling reward, and b is a parameter controlling cost. The parameters a and b may themselves be complex functions of the environment, although the model is silent about their form. As one might intuitively expect, reward drives up the probability of taking an action, and cost drives the probability down. Critically, learning depends on how good a predictor of reward the action already is, reaching asymptote at $a/(a + b)$. At this value, the rewards and costs are balanced, and no learning occurs (i.e. Δp is zero).

Building on this work, Rescorla and Wagner (1972) described a model for learning in situations where more than one (conditional) stimulus was present. In their model, the update rule for stimulus i is:

$$\Delta V_i = \alpha\beta(\lambda - \Sigma_j V_j). \tag{1.2}$$

Here, α and β are learning rates related to the conditional and unconditional stimuli, λ is the asymptotic strength or value of the unconditional stimulus, and V_i is the value of conditional stimulus i. Although the model is spelled out in terms of stimulus values rather than directly in terms of actions probabilities, all that is

needed to remedy the situation is a function to map from one to the other. We can ignore this detail for now. The critical part here, again, is that learning depends on the discrepancy between the asymptotic value of the unconditional stimulus, λ, and the current prediction of that value, taken as a sum across all of the conditional stimuli present on the current trial, $\Sigma_j V_j$.

Still missing, however, was a description of the *temporal evolution* of events. Both Bush and Mosteller (1951) and Rescorla and Wagner (1972) assumed that learning happens at the trial level, leaving unspecified what actually happens *inside* a trial. However, such level of detail is necessary for explaining behavior in the real world, where "trials" are not necessarily delineated. Instead, stimuli are presented one after another, with the identity of each dependent (perhaps probabilistically) on the previous one. Further, actions also influence stimulus transitions, and stimuli and actions jointly determine reward. Learning which actions lead to (possibly delayed) reward is known as the temporal credit assignment problem (Sutton, 1984). A simple, yet powerful, technique for solving it is called temporal-difference learning (e.g. Samuel, 1959; Sutton & Barto, 1981). The basic idea has many variants. A formulation for Pavlovian conditioning known as TD(0), with table based state lookups, is as follows (Sutton & Barto, 1998):

$$V(s_t) = V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]. \tag{1.3}$$

Here $V(s_t)$ is the *value* of state (stimulus) s at time t, r_t is the *immediate reward* at time t, and γ is a discount factor on future reward. Recursive application of Equation 1.3 throughout experience converges (under reasonable assumptions, Dayan, 1992) to the true underlying expected value for each state. State *value* refers to the cumulative reward expected over an extended (possibly infinite) time horizon when starting from the state. An instrumental formulation, known as

SARSA (Rummery & Niranjan, 1994), has a similar form:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]. \tag{1.4}$$

Equation 1.4 is almost identical to Equation 1.3, except that it computes values for state–action pairs rather than just states. Actions can be selected based on the learned values in a variety of ways. A popular choice is to use the softmax function:

$$Pr(a_t|s_t) = \frac{\exp\{\tau Q(s_t, a_t)\}}{\sum_b \exp\{\tau Q(s_t, b_t)\}}. \tag{1.5}$$

The softmax function attempts to balance exploitation and exploration, preferring actions that have the highest value, but also allowing actions that are deemed suboptimal to compete. This provides an opportunity for revising incorrect value estimates.

A wide variety of reinforcement learning algorithms now exist (Sutton & Barto, 1998). Central to all of them, however, is the same basic principle that appeared in all of the models discussed above: the reward prediction error. The organism learns to *predict* the values associated with states and/or actions, and acts in accordance with these predictions. The predictions are adjusted based on observed outcomes. If everything goes according to plan, no changes are made. On the other hand, if things get better or worse, the value function is adjusted up or down accordingly. Consider what happens in Equation 1.3. If $V(s_t)$ is an accurate predictor of future reward, it must equal the immediate reward received after leaving state s_t, plus all of the reward expected after that:

$$V(s_t) = r_{t+1} + \gamma V(s_{t+1}). \tag{1.6}$$

If the organism now leaves state s_t and really does receive reward r_{t+1}, the net change in $V(s_t)$ (the part in the bracket on the left hand side of Equation 1.3) is zero. If r_{t+1} or $V(s_{t+1})$ (from other interactions with state s_{t+1}) are greater than or less than expected, $V(s_t)$ is adjusted accordingly, scaled by α.[1]

Although formulation of the above and related algorithms were inspired by animal learning, they apply equally well to any agent, biological or artificial. Computer scientists have developed these ideas in earnest, and a large body of theoretical and empirical work now exists detailing how artificial agents can learn in a wide variety of environments. The link back to animal (and human) learning originated in the mid 1990's (Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997), when it was realized that the phasic firing of midbrain dopaminergic neurons represents a signal very similar to the reward prediction error described above.

This initiated a large scale effort to further characterize the properties of dopamine neurons, and learning more generally, within the computational reinforcement learning framework (Glimcher, 2011; Maia, 2009; Niv, 2009; Schultz, 2010). Although the dopamine reward prediction error hypothesis is not without detractors (Berridge, 2007), many researchers believe that the brain operates using the principles of computational reinforcement learning, even if the details are disagreed on. Simple learning rules like (1.3) and (1.4) can learn a lot. Nevertheless, they are insufficient to explain much of real life behavior.

[1] α is the learning rate and acts to "smooth" learning in noisy environments. Consider what happens in a state where rewards are Gaussian, and the organism receives a large rare reward from the tail of the distribution. Without α (or setting $\alpha = 1$), the value would immediately be updated to reflect this fact. Setting $\alpha < 1$ discounts this evidence, allowing past experiences to influence the value as well.

1.2 The need for model-based reinforcement learning

Let's return to the driving example. Assume for simplicity an encoding of the world where each block of a street is a state, and each action corresponds to driving in one of four directions. You can keep going straight, make a left or right turn, or if you're really stuck, you can back up. Furthermore, each step you take is costly: You're using gasoline and putting wear on your car, and you're not engaging in other activities that may otherwise be rewarding. We'll encode this cost (or negative reward) as -1 on each step, although the scale here is arbitrary. We set γ to one for simplicity. Learning the appropriate path is now a simple matter of applying Equation 1.4 and adjusting your policy (what you should do in each state) accordingly. Initialize $Q(s, a) = 0$ for all s and a, start at your house (or in fact anywhere), and make a random series of turns. Upon arriving at work, $Q(s_{work}, \cdot)$ will be zero (there is no further cost incurred once you've arrived at your goal), and every other state–action combination on the path will have value $0 + \alpha[-1 + 0 - 0] = -\alpha$. If you were to take the same path again, each action value except the last turn immediately before work would similarly be updated to $-\alpha + \alpha[-1 + (-\alpha) - (-\alpha)] = -2\alpha$. The last action is special, because the subsequent (goal) state incurs no cost. Here the updated value would be $-\alpha + \alpha[-1 + 0 - (-\alpha)] = -2\alpha + \alpha^2$, slightly less than -2α. On the next iteration, the last step will continue to be associated with a lower cost compared to the previous steps, but the step before it will itself also be associated with a lower cost than the steps before *it*. Of course the same exact path doesn't have to (and shouldn't) be taken each time; the policy should be adjusted to match the value function. That is, you should begin to avoid more costly states and adjust your path to incur the least cost.

The example so far is meant to further illustrate that the type of reinforcement learning algorithms discussed so far learn value functions and policies directly through trial and error. You begin with a potentially blank slate of the world, and

proceed to try different things out. As you gain experience, you learn (e.g. using Equation 1.4) which actions are better (more rewarding or less costly in the long run) than others.

Consider now the following modification. Rather than incurring a fixed cost for each turn, you incur a variable cost depending on the amount of traffic on the next street. You proceed as before, using Equation 1.4 to learn a policy that minimizes the average cost (the expected amount of traffic) of your journey. You learn that the best path from your home to work is to take US-1, turn right on Harrison Street, and finally turn left on Nassau Street. One morning, you hear on the radio that Harrison Street is flooded. How can you use this information to help you plan an alternate route? The value function you've learned through previous experience is of little use; Harrison Street was open when you learned it, and it does not take into account the new information.[2] The only way Equation 1.4 would help is if you now took additional trips with the barricade in place, learning, over time, a new value function and policy. People do not generally behave this way, however. Rather, they incorporate the new information rather quickly into their decision making.

The modified driving example is an instance of *detour behavior* (Tolman & Honzik, 1930). Two further examples serve to illustrate the problem with the above algorithms. The first, *latent learning* (Blodgett, 1929), involves learning about the transition structure of the environment separately from the reward structure. Within the context of the ongoing driving example, imagine you've now moved to a new (small) town, and you're trying to get the lay of the land. It's Sunday afternoon, and you have no particular place to be, so your reward function has no information about any particular goal. Applying Equation 1.4 throughout your drive would be of little value: You would learn how to get to the street with the minimum amount of traffic (an alley probably), but not how to get to any particu-

[2]Of course, it is not *entirely* useless, since *parts* of it are still correct.

larly useful place. In particular, when it comes time to drive to work on Monday morning, the value function and policy you learned will not get you there. Yet, most people would be able to use the information learned the previous day to piece together a route.

The second example, *outcome revaluation* (Adams & Dickinson, 1981), similarly involves a changing reward function. Imagine you learned to drive to work using Equation 1.4 as in the original example, but you now hear on the radio that there is construction on Harrison Street. You can still get through the street: The transition structure of the environment hasn't changed. However, the cost has gone up significantly, as you will now be stuck in traffic a lot longer. As before, your old value function and policy do not reflect this information, but most people can easily plan an alternate route.

Computational reinforcement learning has something to say about these scenarios as well. The ideas here are in fact older than many of the ones described above (Bellman, 1957), but their application to human learning and decision making is more recent (Dolan & Dayan, 2013). This area of reinforcement learning is called *model-based*, while the ideas described above belong to the class of *model-free* algorithms. The *model* in the name refers to the two pieces of information already alluded to: A transition function that dictates movements between states in response to actions, and a reward function that dictates how good or bad those moves are. This information is inherent in the world. In the examples above, a certain relationship between street blocks (states) is imposed by the layout of the town, while the cost is influenced by gas prices and the actions of other drivers. Model-free agents bypass learning this information, and instead estimate the optimal value function and policy directly from experience. A model-based agent, by contrast, first learns the model, and then estimates the value function and policy from the model itself.

How does this help remedy our understanding of behavior in the above situations? In detour behavior, one can incorporate the knowledge that Harrison Street is closed locally into the learned transition function, by changing where taking a right from US-1 leads (it leads back to US-1). The rest of the model is still correct, and the modified model can be used to construct a new route. Outcome revaluation also involves a local change to the model, but to the reward rather than to the transition function. Finally, latent learning simply requires learning the transition and reward functions at separate times.

Most of the work on model-based reinforcement learning in humans has concentrated on two broad questions. First, how the model is learned (Bornstein & Daw, 2012, 2013; Glascher, Daw, Dayan, & O'Doherty, 2010), and second, whether correlates of the value function predicted by the framework can be found in the brain (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; D. A. Simon & Daw, 2011b; Wunderlich, Dayan, & Dolan, 2012). Much less attention has so far been given to algorithmic questions – how the brain actually maps the model into a decision. Chapters 2 and 3 approach this topic from what at first appears to be very different perspectives.

Chapter 2 builds on the work of Botvinick and An (2009), assuming that the brain learns and maintains a generative Bayesian model of the transition and reward functions. Decision making proceeds by conditioning on reward (in Bayesian terms, treating it as *evidence*, or something that has already been obtained), and performing reverse inference to reason about the policies that would have brought this about. The new work considers the neural implications of this framework, using it to explain a wealth of existing data. This allows us to build a bridge from model-based reinforcement learning to evidence integration models that have become popular in other areas of cognitive science, especially perceptual decision making (Gold & Shadlen, 2007).

Chapter 3 begins from the evidence accumulation perspective, building in particular on the work of Rangel and colleagues on simple choice (Milosavljevic, Malmaud, Huth, Koch, & Rangel, 2010; Krajbich, Armel, & Rangel, 2010; Krajbich & Rangel, 2011). Simple choice is a form of model-based decision making where there is only a single step of action and a handful (2-4) of choices. Two new behavioral experiments are presented that extend the simple choice paradigm to multiple steps of action. Then, a new accumulator model is developed to explain the relationship between choice and reaction time data in these tasks. The chapter ends with a broad but detailed outline of the future work necessary to fully understand how the brain performs model-based reinforcement learning. Chapters 2 and 3 are connected through the evidence accumulation perspective. Although a formal understanding of the relationship between the models presented in these two chapters is yet to be developed, broad qualitative similarities can be seen.

1.3 Hierarchical reinforcement learning

A very different type of problem with basic model-free (and model-based) reinforcement learning involves its ability to scale to large state and action spaces. In the driving example, we chose a rather abstract representation for convenience: States were blocks and actions were turns. However, this representation encapsulates a large part of the decision problem one would actually face. When you chose to turn onto US-1, you first had to decide to get into your car. That decision itself involved further subcomponents, such as getting out of bed and getting dressed. The finest level of detail perhaps involves deciding between individual muscle movements. Now, imagine using Equation 1.4 to learn the muscle movements necessary to get to work through trial and error. The possible combination of states and actions is enormous; it would take a long time to hit upon the correct combination once,

much less multiple times. Model-based reinforcement learning fares no better, as the organism still has to manipulate an enormous model of the environment.[3]

A second but related problem involves the transfer of skills between different tasks. Let's say you're able to somehow learn to go south on US-1 by using model-free methods and trying out different muscle movements. Part of this sequence involves getting into your car. If you're now tasked with going *north* on US-1, you still have to first get into your car, but you have no way of knowing which part of your policy is still correct (the part that gets you to your car) and which isn't (the part that gets you to US-1 south, because you now have to go north). Not all is lost and perhaps starting with your old policy can still shorten the overall time it takes to learn, but there are better ways. A similar argument can be made for the model-based case.

The framework of hierarchical reinforcement learning (Barto & Mahadevan, 2003) can deal with both of these problems. There are several variants; this thesis focuses on the options framework (Sutton, Precup, & Singh, 1999). The basic idea is simple: In addition to primitive actions, it is assumed the agent has access to temporally extended actions called options. Each option can be initiated and terminated in a subset of states. Moreover, each option has its own policy. When the agent is executing an option, that option's policy controls behavior. This aids learning, exploration, and planning by allowing the agent to take (and reason about) much larger steps through the environment than it otherwise could. Furthermore, the set of actions (options) at each level of the hierarchy may be reduced. In our example, the top level, which may have options like *go-to-car*, *turn-steering-wheel-left*, *turn-steering-wheel-right*, etc., has no need to reason about individual muscle movements. The problem of skill transfer is solved in the obvious way: The agent

[3]There are strong connections between the algorithms for model-based and model-free reinforcement learning. This can particularly be seen in the Dyna architecture (Sutton, 1990). Here, model-free methods are used to learn model-based policies by using the model to generate simulated experiences, and then treating the outcomes of those experiences as if they were real.

can learn an option specific policy in one task, and then invoke the same option in another task.

How do you learn option policies? The answer turns out to be simple: you can apply the same learning mechanisms that are used in standard (flat) reinforcement learning. Going back to our driving example again, let's say you know that *go-to-car* is a useful abstraction to have, but you don't know how to actually go about doing it. One way to learn is to introduce a reward in the state corresponding to you being in the car, and then use standard model-free or model-based methods (augmented with perhaps lower level options) to obtain the appropriate policy. The reward that would have to be introduced is called *pseudo-reward*, to differentiate it from the standard external reward provided by the environment. That is, going to your car is not rewarding in itself, but because it is a useful skill to have, a faux reward is introduced to aid learning.

Pseudo-rewards drive pseudo-reward prediction errors in the same way external rewards drive standard prediction errors. Chapter 4 describes work looking at the neural correlates of pseudo-reward prediction errors. My contribution shows that the feedback related negativity, an event-related potential seen in the electroencephalograph, relates to pseudo-reward prediction errors in the same way it relates to standard reward prediction errors. The work of my colleague José Fernandes, looking at behavioral measures and fMRI BOLD correlates, is also reported.

Chapter 5 takes a step back to address the deeper question of how useful options are identified in the first place. This is a more general problem in hierarchical reinforcement learning, and is not limited to understanding how organisms implement an already existing theory. In the example above, we assumed a particular hierarchy (that *go-to-car*, *turn-steering-wheel-left*, etc., are the useful subcomponents), and a similar assumption is made in Chapter 4. Chapter 5 outlines a framework, based on Bayesian model comparison, for understanding which decompositions

should be learned. It suggests that useful decompositions maximally aid learning *on average* across an ensemble of tasks, and simultaneously compress the problem, allowing the optimal policies to be stored with the minimal number of bits. The predictions of the model are tested in four behavioral experiments, performed by my colleagues Carlos Diuk, Natalia Córdova, and Debbie Yee, and the chapter reports on these results as well.

Chapter 6 discusses additional directions for future research not mentioned in the previous chapters, and provides concluding remarks.

Chapter 2

Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates

Solway, A., & Botvinick, M. M. (2012). *Psychological Review*, 119(1), 120–154.

2.1 Abstract

Recent work has given rise to the view that reward-based decision making is governed by two key controllers: a habit system, which stores stimulus–response associations shaped by past reward, and a goal-oriented system that selects actions based on their anticipated outcomes. The current literature provides a rich body of computational theory addressing habit formation, centering on temporal-difference learning mechanisms. Less progress has been made toward formalizing the processes involved in goal-directed decision making. We draw on recent work

in cognitive neuroscience, animal conditioning, cognitive and developmental psychology, and machine learning to outline a new theory of goal-directed decision making. Our basic proposal is that the brain, within an identifiable network of cortical and subcortical structures, implements a probabilistic generative model of reward, and that goal-directed decision making is effected through Bayesian inversion of this model. We present a set of simulations implementing the account, which address benchmark behavioral and neuroscientific findings, and give rise to a set of testable predictions. We also discuss the relationship between the proposed framework and other models of decision making, including recent models of perceptual choice, to which our theory bears a direct connection.

2.2 Introduction

Since the earliest days of both psychology and neuroscience, investigators interested in decision making and the control of behavior have recognized a fundamental distinction between *habitual* action and *goal-directed* or *purposive* action. Although this opposition has obvious roots in commonsense notions from folk psychology, its first rigorous expression emerged in a classic debate in the behaviorist era. On one side of this debate, Hull (1943), Spence (1956), and others characterized action selection as driven primarily by immediate associations from internal and environmental states to responses. On the other, Tolman (1932), McDougall (1923), and others portrayed action as arising from a process of prospective planning, involving the anticipation, evaluation, and comparison of action outcomes. Over time, this early view of habit and goal directedness as mutually exclusive accounts of behavior has given way to a more inclusive multiple-systems account, under which habitual and goal-directed control coexist as complementary mechanisms for action selection (Daw, Niv, & Dayan, 2005; Dayan, 2009; Dickinson, 1985; Bullock &

Rhodes, 1993; Doya, 1999; Glascher et al., 2010; Platt et al., 2008; Rangel, Camerer, & Montague, 2008; Rangel & Hare, 2010; Samejima & Doya, 2007). This more recent perspective licenses the study of each form of action control in its own right, and sizable literatures have developed concerning both habitual stimulus–response based action selection and planning-based control (see, e.g., Bargh, Green, & Fitzsimons, 2008; Bekkering, Wohlschlager, & Gattis, 2000; Gergely & Csibra, 2003; Wood & Neal, 2007; Yin & Knowlton, 2006).

Despite exciting progress in both arenas, however, a nagging imbalance has gradually arisen: Over the past decade, research on habitual, stimulus–response behavior has crystallized around an increasingly explicit set of computational ideas, originating from the field of reinforcement learning (Sutton & Barto, 1998). These ideas have not only provided a context for interpreting and predicting patterns of behavior (Sutton & Barto, 1981, 1990; Wickens, Kotter, & Houk, 1995); they have also enabled new and detailed insights into the functional contributions of specific brain structures, including the striatum and the midbrain dopaminergic system (Barto, 1995; Houk, Adams, & Barto, 1995; Joel, Niv, & Ruppin, 2002; Montague et al., 1996; Ribas-Fernandes et al., 2011; Schultz et al., 1997). In contrast, research on goal-directed behavior, for all its sophistication, has not developed a similarly mature computational core.

In the present work, we contribute toward closing this gap in psychological and neuroscientific theory by proposing a neuro-computational account of goal-directed decision making.

2.3 Goal-directed decision making: definition and manifestations

It is important, from the outset, to be precise about what the expression *goal-directed decision making* is intended to denote. Here, as in the animal conditioning literature, we use the term to describe decision making based directly on predictions concerning action outcomes and their attendant incentive values. As implied by this definition, goal-directed decision making requires the agent to have access to two distinct forms of knowledge. First, it requires access to stored information about action-outcome contingencies, a body of knowledge that Tolman (1932, 1948) famously referred to as a "cognitive map." Second, as Tolman (1932, 1949) also observed, in order for preferences to emerge over prospective outcomes, action-outcome knowledge must be integrated with incentive knowledge, knowledge of the reward values associated with individual world states. Integration of these two forms of knowledge allows the selection of actions judged most likely to bring about preferred outcomes (Balleine & Dickinson, 1998a).

Working from this conception of goal-directed decision making, animal conditioning research has generated a number of experimental paradigms that operationalize the construct, making it possible to diagnose goal directedness in observed behavior. One particularly important experimental manipulation is known as *outcome revaluation* (Adams & Dickinson, 1981; Balleine, 2005; Balleine & Dickinson, 1998c; Colwill & Rescorla, 1985b; Klossek, Russell, & Dickinson, 2008). Here, an animal first learns to perform actions that yield specific rewards (e.g., learning to pull a chain that yields one kind of food and to press a lever that yields another). The appeal or reward value of one of the outcomes is then altered, for example, by allowing the animal to eat its fill of a particular food (the *specific satiety* procedure; Balleine & Dickinson, 1998c; Colwill & Rescorla, 1985a, by pairing that food

with an aversive event such as toxin-induced illness (*conditioned aversion*; Adams, 1982; Adams & Dickinson, 1981; Colwill & Rescorla, 1985a, 1988) or by inducing a change in motivational state (Balleine, 1992; Balleine & Dickinson, 1994; Dickinson & Dawson, 1989). Under appropriate circumstances, this intervention results in a rapid shift in behavior either away from or toward the actions associated with the relevant outcome. Such a shift is interpreted as reflecting goal-directed behavior because it implies an integration of action-outcome knowledge with representations of outcome reward value.

Another key experimental manipulation involves breaking the causal contingency between a specific action and outcome. Here, typically, the animal first learns to associate delivery of a certain food with a particular action but later begins to receive the food independently of the action. The upshot of this "contingency degradation" is that the animal less frequently produces the action in question (Colwill & Rescorla, 1986; Dickinson & Mulatero, 1989; Williams, 1989). Such behavior provides evidence that actions are being selected based on (appropriately updated) internal representations of action-outcome contingencies, thus meeting the criteria for goal directedness.

The same definition for goal directedness extends to decisions involving sequences of action (Daw et al., 2011; Ostlund, Winterbauer, & Balleine, 2009; D. A. Simon & Daw, 2011b). An illustrative example, introduced by Niv, Joel, and Dayan (2006), involves a rat navigating through a two-step T maze, as shown in Figure 2.1 (lower right). The animal in this scenario must make a sequence of two left-right decisions, arriving by these at a terminus containing an item with a particular incentive value. A goal-directed decision at S_1 would require retrieval of a sequence of action-outcome associations – linking a left turn at S_1 with arrival at S_2 and a left turn at S_2 with cheese – as well as access to stored information about the incentive value of the available outcomes. Building on this simple example, Niv et al. (2006)

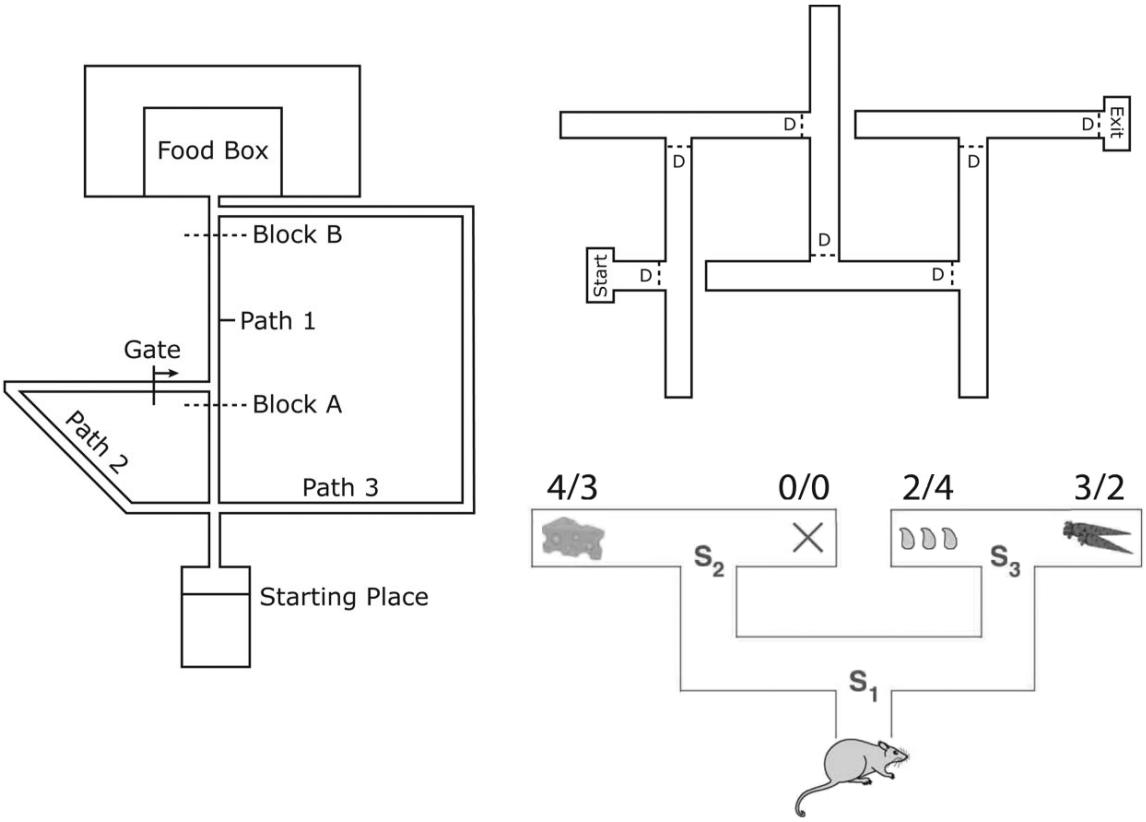

Figure 2.1: Left: Maze used to demonstrate detour behavior, redrawn from Tolman and Honzik (1930, p. 223) Upper right: Maze used to demonstrate latent learning, redrawn from Blodgett (1929, p. 117). D = door. Lower right: T-maze scenario from Niv et al. (2006) Outcome values relate to hungry (left) and thirsty (right) states. Reprinted from "A Normative Perspective on Motivation," by Y. Niv, D. Joel, and P. Dayan, 2006, *Trends in Cognitive Sciences*, 10, p. 376. Copyright 2006 by Elsevier.

provided an illustration of how revaluation plays out in the multistep decision context. They considered a scenario in which fluid deprivation is used to make the rat thirsty, inducing a change in the reward values associated with the four outcomes (see Figure 2.1). This change in the animal's internal representations of incentive value, when integrated into the prospective operations involved in goal-directed decision making, results in a different action at S_1.

Although this T-maze example represents only a thought experiment, some of the issues it addresses were engaged in recent experiments by Ostlund et al. (2009). Here, rats were trained to execute two-step sequences in order to obtain food

rewards. The rats had access to two levers. When a rat pressed the right lever and then the left, a bit of sucrose was delivered. When the levers were pressed in the opposite order, the rat received polycose. The sequences left-left and right-right, meanwhile, yielded no reward. Following training, one of the food rewards was devalued through satiety. When presented with the two levers in this setting, rats tended to execute the sequence yielding the nondevalued food more frequently than the opposite sequence. Ostlund et al. (2009) also showed analogous changes in sequence production following contingency degradation.

Two further standard operationalizations of goal-directed decision making derive from the classic research championed by Tolman. In the *latent learning* paradigm (Blodgett, 1929), rats run a compound T maze as shown in Figure 2.1 (upper right), until they reach the box labeled "exit." After several sessions, a food reward is placed in the exit box. After the animals discover this change, there is an immediate reduction in the frequency of entrances into blind alleys. Animals suddenly take a much more direct path to the exit box than they had previously. In *detour* behavior, as described by Tolman and Honzik (1930), rats run a maze configured as in Figure 2.1 (left). When the most direct route (Path 1) is blocked by a barrier at location A, the animals tend to opt for the shortest of the remaining paths (Path 2). However, when the block is placed at location B, animals take the third path. In each of these cases, a change in action-outcome contingencies triggers immediate adjustments in behavior, providing a hallmark of goal-directed decision making.

2.3.1 Toward a computational account

Our interest in the present work is in understanding the computations and mechanisms that underlie goal-directed decision making, as it manifests in behaviors like the ones just described. Given the recent success of temporal-difference models in

research on habit formation, one approach might be to draw from the same well, surveying the wide range of algorithms that have developed in artificial intelligence, machine learning, and operations research for solving multistep decision problems based on preestablished contingency and incentive knowledge (see Bertsekas & Tsitsiklis, 1996; Puterman, 2005; Russell & Norvig, 2002; Sutton & Barto, 1998). We do believe that it is important to consider such procedures for their potential biological relevance,[1] and later we will circle back in order to do so. However, the theory we present draws its inspiration from a rather different source, looking to previous research in neuroscience, psychology, and computer science that has invoked the notion of a *probabilistic generative model*. In order to set the scene for what follows, we will briefly unpack this construct and highlight previous work in which it has been applied.

Generative models in psychology and neuroscience. Over recent years, a broad formal perspective has taken root within both cognitive and neural research, in which probabilistic inference plays a central organizing role. A recurring motif, across numerous applications of this perspective, is that of inverse inference within a generative model. The basic idea emerged first in research on visual perception. Early on, Helmholtz (1860/1962) characterized vision as a process of unconscious inference, whose function is to diagnose the environmental conditions responsible for generating the retinal image. In recent years, this perspective has found expression in the idea that the visual system embodies a generative model of retinal images, that is, an internal model of how the ambient scene (objects, textures, lighting, and so forth) gives rise to patterns of retinal stimulation. More specifically, this generative model encodes a conditional probability distribution, $p(image|scene)$. The inference of which Helmholtz spoke is made by inverting this

[1] As detailed in the General Discussion, the idea that we pursue also has precedents in machine learning, although it does not yet figure among the standard approaches to solving sequential decision problems.

generative model using Bayes' rule, in order to compute the posterior probability $p(scene|image)$ (Dayan, Hinton, Neal, & Zemel, 1995; Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996; Yuille & Kersten, 2006).

The influence of this generative perspective has gradually spread from perception research to other fields. In particular, it has played an important role in recent work on motor control. Here, the generative (or forward) model maps from motor commands to their postural and environmental results, and this model is inverted in order to establish a mapping from desired effects to motor commands (Carpenter & Williams, 1995; Jordan & Rumelhart, 1992; Kilner, Friston, & Frith, 2007; Körding & Wolpert, 2006; Rao, Shon, & Meltzoff, 2007; Wolpert, Doya, & Kawato, 2003; Wolpert, Ghahramani, & Jordan, 1995). Beyond motor control and perception, theories centering on probabilistic inference over generative models have figured in numerous other realms, including language (Chater & Manning, 2006; Xu & Tenenbaum, 2007), memory (Hemmer & Steyvers, 2009), conceptual knowledge (Chater & Oaksford, 2008; Griffiths, Steyvers, & Tenenbaum, 2007), perceptual categorization (Yu, Dayan, & Cohen, 2009), and–significantly–causal learning and the learning of action-outcome contingencies (Blaisdell, Sawa, Leising, & Waldmann, 2006; Glymour, 2001; Gopnik et al., 2004; Gopnik & Schulz, 2007; Green, Benson, Kersten, & Schrater, 2010; Sloman, 2005; Tenenbaum, Griffiths, & Niyogi, 2007).

One exciting aspect of the generative approach in psychology is that its terms can be transposed, in very much the same mathematical form, into accounts of the underlying neural computations. The notion of inverse inference within a generative model has played a central role in numerous recent theories of brain function, both in visual neuroscience (Ballard, Hinton, & Sejnowski, 1983; Barlow, 1969; T. S. Lee & Mumford, 2003; Rao & Ballard, 1999) and elsewhere (Dayan et al., 1995; Friston, 2005; Knill & Pouget, 2004; Mumford, 1992, 1994).

Goal-directed decision making as inverse inference. Our central proposal in the present work is that goal-directed decision making, like so many other forms of human and animal information processing, can be fruitfully understood in terms of probabilistic inference. In particular, we propose that goal-directed decisions arise out of an internal generative model, which captures how situations, plans, actions, and outcomes interact to generate reward. Decision making, as we characterize it, involves inverse inference within this generative model: The decision process takes the occurrence of reward as a premise and leverages the generative model to determine which course of action best explains the observation of reward.

Although this specific idea is new to psychology and neuroscience, it has a number of direct and indirect precedents in machine learning, as we later detail (Attias, 2003; Botvinick & An, 2009; G. F. Cooper, 1988; Dayan & Hinton, 1997; Hoffman, Freitas, Doucet, & Peters, 2009; Shachter & Peot, 1992; Toussaint & Storkey, 2006; Verma & Rao, 2006b). In what follows, we draw many of our raw materials from such work, but also reshape them to yield an account that makes maximal contact with existing psychological and neuroscientific theory.

Overview. The ensuing presentation is divided into three main sections, corresponding to the three levels of theoretical analysis famously proposed by Marr (1982; see also Jones & Love, 2011). We begin in the next section by considering the computational problem underlying goal-directed control. The succeeding section moves on to consider the algorithm or procedure involved in solving that computational problem. Finally, in a third section, we consider the level of neural implementation. Following these three core sections of the paper, we discuss the relationship between the present ideas and earlier work, and consider directions for further development.

2.4 Reframing the computational problem

In building a formal theory, we take as our point of departure an insight recently expressed by Daw et al. (2005; see also Dayan & Niv, 2008), which is that goal-directed decision making can be viewed as a version of *model-based reinforcement learning*. The "model" referred to in this term comes in two parts: a *state-transition function*, which maps from situation–action pairs to outcomes, and a *reward function*, which attaches a reward value to each world state. Model-based reinforcement learning refers to the project of discovering an optimal (reward-maximizing) *policy*, or mapping from states to actions, given this two-part model (Sutton & Barto, 1998).

To state this more formally: Model-based reinforcement learning begins with a set of givens, which include a set of states, S; a set of actions, A; a state-transition function $T(s \in S, a \in A, s' \in S)$, which specifies the probability of arriving in state s' after having performed action a in state s; and a reward function $R(s)$, which assigns a scalar reward value to each state. The computational problem is then to choose a policy $\pi(s, a, t) = p(a|s, t)$ that maximizes expected cumulative reward over steps of action t up to some planning horizon T:

$$\underset{\pi}{\operatorname{argmax}} \; E\left[\sum_{t=1}^{T} p_t(s|\pi)R(s)\right]. \tag{2.1}$$

Our objective is to reframe this problem in terms of probabilistic inference. As a first step in that direction, the problem's ingredients, as well as their interrelations, can be represented in the form of a probabilistic graphical model (see Bishop, 2006; Koller & Friedman, 2009; Pearl, 1988). Figure 2.2A begins construction of this model with an initial set of three nodes. The node S represents a variable indicating the decision maker's current situation or state.[2] This node is shaded to indicate

[2]Representing state as a multinomial variable is obviously a massive simplification. However, the graphical model formalism can accommodate richer representations of state, including factored

that its value is known or observed by the decision maker; the initial state is a "given" in the action-selection problem. The node A represents a variable whose values correspond to available actions, and Π represents a set of state-specific policy variables, with values corresponding to state–action pairs. The two arrows converging on A indicate that the current action a depends on both the current state s and the policy π for that state. More specifically, node A is associated with the conditional probability distribution $p(A = a|S = s, \Pi = \pi)$ or, for brevity, $p(a|s, \pi)$.[3]

Figure 2.2B expands the model to incorporate a representation of the transition function. As above, the latter is defined as a probability distribution $p(s'|s, a)$, where s' is the value of a variable representing action outcomes or successor states. This variable is represented by node S' in the figure, with incoming arrows indicating its joint dependence on S and A.

Figure 2.2C completes the structure by incorporating a representation of the reward function. Here, we add a node $\hat{R}$ representing reward value, with an afferent arrow to indicate that the value $\hat{r}$ depends on the outcome state s'. (The reason for the change in notation from R to $\hat{R}$ will be disclosed in a moment.)

The architecture developed so far addresses only a single step of action. However, it is readily extended to sequences. As shown in Figure 2.2D, this extension is accomplished by duplicating part of the existing structure, providing a series of state, action, policy, and reward nodes, one for each step of the action sequence. In

---

or distributed representations and representations involving continuously valued features. The same comment applies to the action representations discussed below.

[3] In the present case, where only a single step of action is planned and the initial state is known, there is in fact no need to distinguish between action and policy variables. However, we include policy variables for two reasons. First, they allow the model to accommodate situations where the initial state is uncertain at the time of planning. This is often the case, for example, in behavioral experiments where a participant must prepare to respond to an impending stimulus, without yet knowing the exact identity of the stimulus. Indeed, this is precisely the scenario involved in most experiments that have demonstrated coding for specific tasks in prefrontal cortex (see the Neural Implementation section). Second, we include policy variables for parallelism with the multistep case, where they are in fact computationally necessary.

extending the architecture in this way, we also introduce one final new element: a variable representing the cumulative reward accrued over an action sequence ($\hat{R}_c$).

2.4.1 A probabilistic representation of reward

To this point, our model has been built from materials directly provided by traditional reinforcement learning. At the present juncture, however, we make our first move toward reframing the goal-directed decision making problem by choosing a special form for the representation of reward. In reinforcement learning, as well as in many quarters of economics and psychology, reward magnitude is generally formalized as a scalar value. In view of this, the most intuitive approach in fleshing out our graphical model might be to treat $\hat{R}$ as a continuous variable, whose value directly corresponds to reward magnitude or utility (see, e.g., Attias, 2003). However, we find it fruitful to represent reward in a different way. Specifically, we cast $\hat{R}$ as a binary variable, with discrete values of one and zero. Reward magnitude is then encoded as the probability $p(\hat{r} = 1)$, for which we use the shorthand $p(\hat{r}|s')$. Under this encoding, a state s' associated with large positive reward would give $p(\hat{r}|s')$ close to one. If the state were associated with large negative reward (punishment), $p(\hat{r}|s')$ would fall near zero. In the sequential setting (see Figure 2.2D), the cumulative reward variable $\hat{R}_c$ is also be treated as binary, with

$$p(\hat{r}_c = 1) = \frac{1}{T}\left(\sum_{t=1}^{T} \hat{r}_t\right), \tag{2.2}$$

where $\hat{r}_t$ is the $\hat{R}$ node associated with step t of the plan (Tatman & Shachter, 1990).

To prevent misapprehension, it is worth emphasizing that what is represented using this approach is reward *magnitude*, not reward probability. Although the value $p(\hat{r})$ is a probability, it is being used as the vehicle for representing the size of a deterministic reward. On first blush, this approach to representing reward

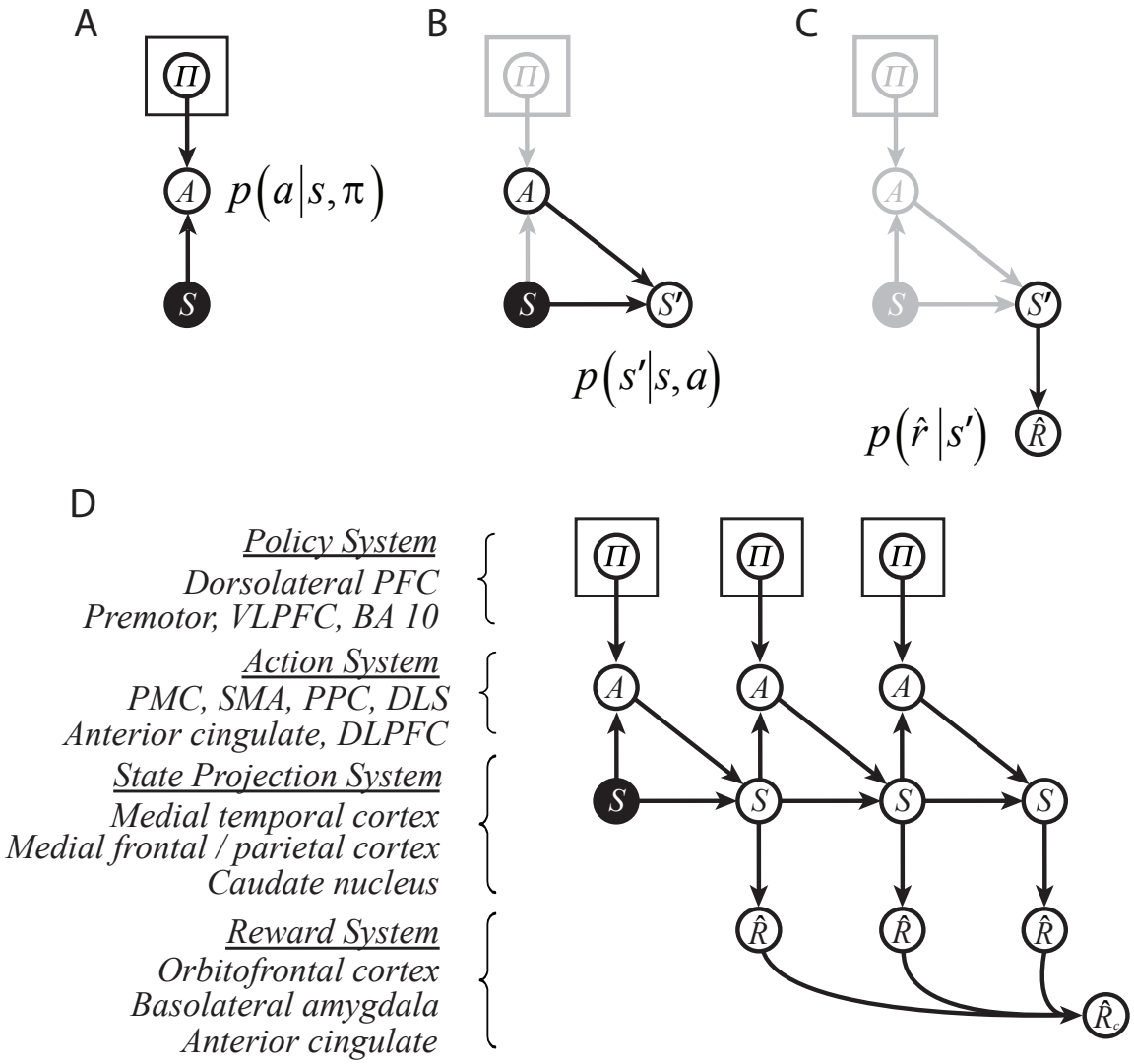

Figure 2.2: Elements of the computational account. Rectangular plates surrounding policy nodes indicate the inclusion of one such node per state (see Appendices). PFC = prefrontal cortex; VLPFC = ventrolateral prefrontal cortex; BA = Brodmann area; PMC = premotor cortex; SMA = supplementary motor area; PPC = posterior parietal cortex; DLS = dorsolateral striatum; DLPFC = dorsolateral prefrontal cortex.

may seem rather perverse. However, as we later discuss in detail, it has precedents in economics, psychology, and neuroscience, as well as in decision theory and machine learning (Shachter & Peot, 1992; Toussaint & Storkey, 2006). For example, in the psychology literature, N. Stewart, Chater, and Brown (2006) have proposed that the utility of a choice item is quantified as the probability that this item would be judged preferable to a randomly selected comparison item (see also Kornienko, 2010). And in neuroscience, data suggest that utility is encoded in part through the firing rates of neurons in orbitofrontal cortex, i.e., the probability that these neurons will fire within a small time window (see, e.g., Padoa-Schioppa & Assad, 2006). In both cases, as in our model, utility is encoded through the probability of a binary event.

By adopting this binary format for reward representation, we bring about a subtle but important change in how the goal-directed decision problem is framed. In the conventional case, where reward is represented as an ordinary real number (which we shall continue to denote by r), the problem is to find the policy that maximizes expected reward magnitude (see Equation 2.1). In the scenario we are considering, the problem is instead to maximize the *probability* of a discrete event, $p(\hat{r} = 1|\pi)$. Goal-directed decision making thus assumes the form of a likelihood maximization problem. This seemingly incidental point has far-reaching ramifications, which we unpack in what follows.

2.4.2 A generative model for reward

As we have noted, the graphical model in Figure 2.2 can be seen as simply one way of representing the standard ingredients of a model-based reinforcement learning problem. However, another way of viewing it is as a generative model for reward. That is, the model represents the interrelated factors–initial states, policies, actions, and outcomes–that together give rise to reward events.

To illustrate, we can "query" the variable $\hat{R}$, asking for the marginal probability $p(\hat{r}|s)$. In the one-step model, this probability depends on the remaining variables in the following way:

$$p(\hat{r}|s) = \sum_{s',a,\pi} p(\hat{r}|s')p(s'|s,a)p(a|s,\pi)p(\pi). \tag{2.3}$$

Note that the first factor in this sum is simply the reward function. The second term is the transition function, and the third expresses the effect of policies on action selection. The final term represents the decision maker's prior bias toward specific policies, expressed as a probability distribution. Each of these factors corresponds to the conditional probability distribution (CPD) at a specific node in the graph.

An important aspect of probabilistic graphical models is that they provide a substrate for *conditional inference*. Given an observed or known value for one or more variables, one can query the conditional distribution for any other set of variables (see Bishop, 2006; Koller & Friedman, 2009). Indeed, Equation 2.3 already provides an illustration of this, as here the value of the initial state s was an observed quantity. The same approach could be used to obtain the marginal probability of $p(\hat{r} = 1)$, given a commitment to a specific policy. This is obtained by treating Π as an observed variable ($\Pi = \pi$), as illustrated in Figure 2.3 (top), and computing

$$p(\hat{r}|s,\pi) = \sum_{s',a} p(\hat{r}|s')p(s'|s,a)p(a|s,\pi). \tag{2.4}$$

Given the definition of $\hat{r}$, the conditional probability computed here corresponds to the expected reward under the designated policy π. As indicated in Figure 2.3 (top), in the multistep setting, the expected cumulative reward for a specific set of policy choices can be inferred by computing the conditional probability of $\hat{r}_c$.

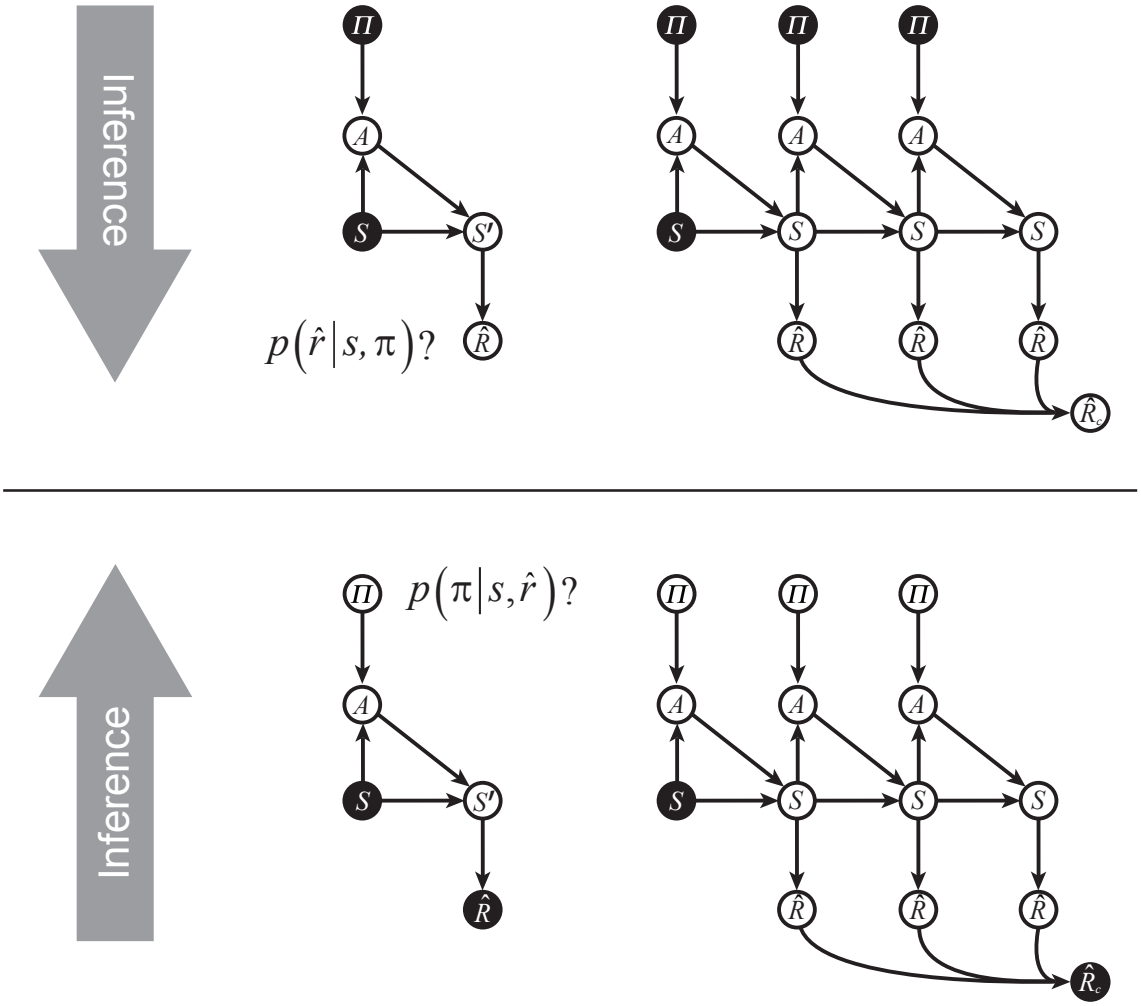

Figure 2.3: Top: Conditioning on a policy. Bottom: Conditioning on reward. Filled nodes indicate variables with stipulated values.

Note that conditioning on a policy and querying the reward variable in this way offers one potential method for solving the computational problem we have laid out. The decision maker could iterate through all available policies, keeping a record of the expected reward $p(\hat{r}|\pi, s)$ for each, and then choose the policy that maximizes that quantity. As we discuss later, this procedure may be relevant to decision making in the biological case, in some instances. However, there is also another, more interesting route to solving the computational problem.

2.4.3 Abductive inference

As discussed in the introduction, the notion of a generative model has been applied extensively in work on vision. There, the proposal has been that perception seeks an explanation for retinal inputs, based on a generative model capturing the way that environmental situations give rise to those inputs. Note that the observed data in this case (i.e., the pattern of retinal stimulation) is at the "output" end of the generative model. The model is not used to reason from causes to effects but is rather inverted to reason *abductively* (i.e., from effects to causes).

The same logic can be applied within our generative model of reward. Rather than conditioning on policies and computing rewards, it is possible to invert the model in order to reason from rewards to policies (see Figure 2.3, bottom). Specifically, leveraging our binary representation of reward, we can condition on $\hat{r} = 1$ and apply Bayes' law to compute

$$p(\pi|s, \hat{r}) \propto p(\hat{r}|s, \pi)p(\pi) = \sum_{s',a} p(\hat{r}|s')p(s'|s, a)p(a|s, \pi)p(\pi). \tag{2.5}$$

As illustrated in Figure 2.3, the same approach can be applied in the multistep case by conditioning on $\hat{r}_c = 1$.

Notice that if there is no initial bias toward any specific policy (the priors $p(\pi)$ are uniform across all values of π), the right-hand side of Equation 2.5 is identical to that of Equation 2.4. That is,

$$p(\pi|s, \hat{r}) = p(\hat{r}|s, \pi). \tag{2.6}$$

This suggests an alternative way of framing the computational problem involved in goal-directed decision making. According to our earlier formulation, the objective was to find a policy to maximize $p(\hat{r}|\pi)$. It is now evident that an equally valid objective is to find a policy to maximize $p(\pi|\hat{r})$. Conditioning on $\hat{r} = 1$, the task is to identify the policy that best explains that "observation." In what ensues, we refer to this procedure as *policy abduction*, considering that it involves reasoning from effects (reward) to their explanations or causes (policies for action).

It should be noted that our ability to make this important turn derives specifically from our having adopted a binary representation of reward, choosing to work with $p(\hat{r} = 1|s)$ rather than $R(s)$. To see this, consider what happens if we attempt to condition on a scalar representation of reward. The most obvious approach here would be to replace the $\hat{R}$ node in Figure 2.2 with a node R representing $p(r|s')$, a probability density function over the real numbers. One might then (naively) set up to find $\text{argmax}_\pi\, p(\pi|r)$. However, what specific value of r would one condition on here? If the range of R were bounded, one might be tempted to condition on its maximum: $\text{argmax}_\pi\, p(\pi|r = r_{max})$. However, this will not answer. What if the outcome state s' affording that maximum is not reachable–or not reachable with certainty–given the current situation, as will generally be the case? In the end, there is no tractable way of conditioning on a traditional scalar reward representation. The shift to a binary representation of reward is a critical step in reframing goal-directed decision making as abductive inference.

To recap, we have moved in this section through three interrelated ways of characterizing the computational problem involved in goal-directed decision making: (a) the conventional framing, which centers on the maximization of expected reward; (b) an alternative, maximum-likelihood view; and (c) a final transformation of the problem, which calls for the inversion of a generative model of reward. In the next section, we retain a focus on the last of these problem formulations, turning to a consideration of the procedures by which the problem might be solved.

2.5 Algorithmic framework

Given the preceding discussion, the appropriate procedure for goal-directed decision making may appear self-evident: In order to find $\text{argmax}_\pi\, p(\hat{r}|s,\pi)$, condition on $\hat{r} = 1$ and evaluate $\text{argmax}_\pi\, p(\pi|s,\hat{r})$. It is true that this approach will yield the optimal policy under certain restricted circumstances. However, under others it would backfire. For one thing, the procedure requires that the decision maker begin with no bias toward any specific policy, because as indicated by Equation 2.5, such prior biases enter into computing the posterior distribution $p(\pi|s,\hat{r})$. Another, more daunting problem arises in the multistep setting. Here, taking $\text{argmax}_\pi\, p(\pi|s,\hat{r}_c)$ at each policy variable (see Figure 2.3, lower right) can lead to incorrect decisions. This is because, in the setting of sequential decision making, the optimal decision at any step depends on what actions are planned for later steps.

As an illustration of this important point, consider the decision faced by the rat in the two-step T maze discussed earlier and shown in Figure 2.1. The numbers at the top of that figure (ahead of each slash) indicate the reward values associated with items contained at the maze termini. Obviously, the optimal choice at the first decision point is to head left. However, this is only true if the animal's plan at the next juncture, S_2, is to head left again. If the animal plans instead to head right

if faced with decision point S_2, then the best choice at S_1 is actually to go right. The same is true if the animal has not yet made any decision about what to do at S_2 or S_3; if the animal is equally likely to head left or right at these points, then the best plan at S_1 is to go right. Given this kind of interdependence, a procedure that makes independent decisions at each stage of the plan would yield unreliable results.

Before considering how a biological decision-making algorithm might cope with these issues, let us introduce one further circumstance in which simple policy abduction might fail to yield a reward-maximizing response. This is suggested by so-called random utility models of economic decision making. In such models, the value associated with any particular outcome is not a fixed quantity: Each time the decision maker retrieves a value for an outcome, the result is drawn from a probability distribution (see Gul & Pesendorfer, 2006; Manski, 1977). According to one standard version of this idea, the goal of decision making is to maximize expected reward given such "noisy" readings of outcome value (Busemeyer, 1985; Busemeyer & Townsend, 1993; Glimcher, 2008; Platt et al., 2008; Rustichini, 2008; Shadlen, 2008).

In order to incorporate random utility into our graphical-model framework, we can simply add a stochastic component to the CPD at $\hat{R}$. Thus, rather than $p(\hat{r}|s')$, we have $p(\hat{r}|s', z)$, where Z is a random variable (see Figure 2.4). Although this changes the reward model available to the decision maker, the decision problem–to maximize $p(\hat{r}|s, \pi)$, now equal to the expectation $E_Z[p(\hat{r}|s, \pi, Z)]$–remains unchanged. Note that in this setting, as in the others we have enumerated, policy abduction is not assured to deliver the policy with the highest expected return; even a policy that maximizes $p(\hat{r}|s, \pi, z)$ may not maximize $E_Z[p(\hat{r}|s, \pi, Z)]$.

Notice that decision making under random utility, as we have just characterized it, bears a close resemblance to perceptual decision-making problems involving

ambiguous or noisy stimuli. A highly studied example is the dot-motion task introduced by Newsome, Britten, and Movshon (1989). Here, the subject is required to identify the predominant direction of motion in a dynamic display (see Figure 2.4, top). Formally, the challenge is to decide between competing hypotheses (i.e., true directions of motion), given observations that provide information that is both incomplete and potentially equivocal: incomplete in the sense that $p(x|y) < 1.0$ for all available hypotheses x and any single observation y, and equivocal in the sense that for two observations y_1 and y_2 and hypotheses x_1 (the true hypothesis) and x_2 (false), it might occur that both $p(x_1|y_1) > p(x_2|y_1)$ and $p(x_2|y_2) > p(x_1|y_2)$.

In fact, this decision-making situation is isomorphic to our random utility scenario, where the single "observation" $\hat{r} = 1$ provides information about candidate policies that is potentially both partial and equivocal. In both scenarios, it is hazardous to commit to an answer based on only a single observation. Given this parallel, in order to make progress in understanding goal-directed decision making, it may be fruitful to consider current models of perceptual decision making. As discussed next, these center on the theme of evidence integration.

2.5.1 Evidence integration

An abundance of research suggests that, in the case of perceptual decision making, human and animal decision makers mitigate uncertainty by pooling across a series of observations. According to current evidence-integration models (see Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006), in any interval during the decision process, having made the series of observations $\mathbf{y}$ and a new observation y_{new}, the decision maker updates a representation of the posterior probabilities $p(x|\mathbf{y})$ by combining them with the likelihoods $p(y_{new}|x)$: $p(x|\mathbf{y}, y_{new}) \propto p(y_{new}|x)p(x|\mathbf{y})$. In so-called random walk or drift-diffusion models of two-alternative forced choice decision (see Figure 2.4, top), accumulated evidence is represented in the form of

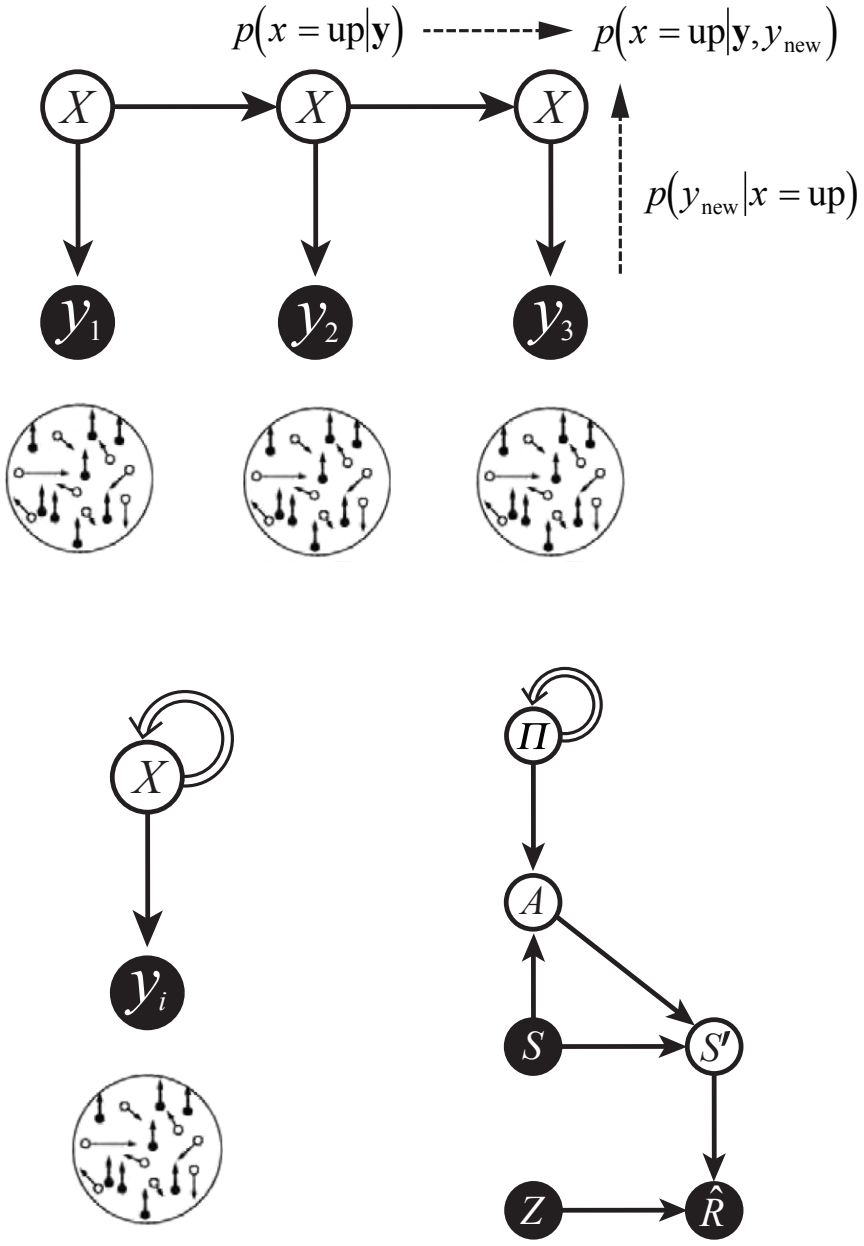

Figure 2.4: Top: Evidence integration in the dot-motion task, focusing on the hypothesis that the underlying stimulus motion is in the upward direction. Bottom left: The graph in the top panel can also be diagrammed as a dynamic Bayesian network, with a recurrent connection running from and to the variable X. Bottom right: An architecture for evidence integration, based on the graph from Figure 2.3.

a log posterior ratio, to which is added a log-likelihood ratio representing the evidence from each new observation (see Beck & Pouget, 2007; Bogacz et al., 2006; Gold & Shadlen, 2007; Rao, 2007; Ratcliff & McKoon, 2008). Given an unlimited number of observations, this procedure is guaranteed to converge to the correct hypothesis. Moreover, when a response threshold is introduced (see Figure 2.5), the procedure becomes equivalent to the so-called sequential probability ratio test (Wald & Wolfowitz, 1948), which guarantees the minimum attainable reaction time for any given error rate.

Evidence-integration models have shown striking success in accounting for behavioral data not only in perceptual tasks but also in memory retrieval (Ratcliff, 1978), lexical decision (e.g., Wagenmakers et al., 2004), cognitive control (Liu, Holmes, & Cohen, 2008), and other contexts. Indeed, as reviewed later, efforts have been made to adapt the framework to reward-based decision making (Rangel & Hare, 2010; Rustichini, 2008; Usher, Elhalal, & McClelland, 2008). The apparent ubiquity of evidence-integration procedures in human and animal decision making, along with the particular parallels we have noted, makes it inviting to consider the potential relevance of these procedures to the framework we have developed for goal-directed decision making.

2.5.2 Goal-directed decision via iterative inference

In our framework, under policy abduction, the "observation" $\hat{r} = 1$ is adopted, and the posteriors $p(\pi|s, \hat{r})$ are computed.[4] As we have seen, this approach is not robust and can go awry in the presence of nonuniform priors, random utility, or sequential problem structure. However, by analogy to evidence-integration models of perceptual choice, the inference procedure can be *repeated*. On each iteration n, the observation $\hat{r} = 1$ is reinstated, and the policy posteriors are updated using

[4]From here forward, to avoid clutter, we suppress the noise variable Z.

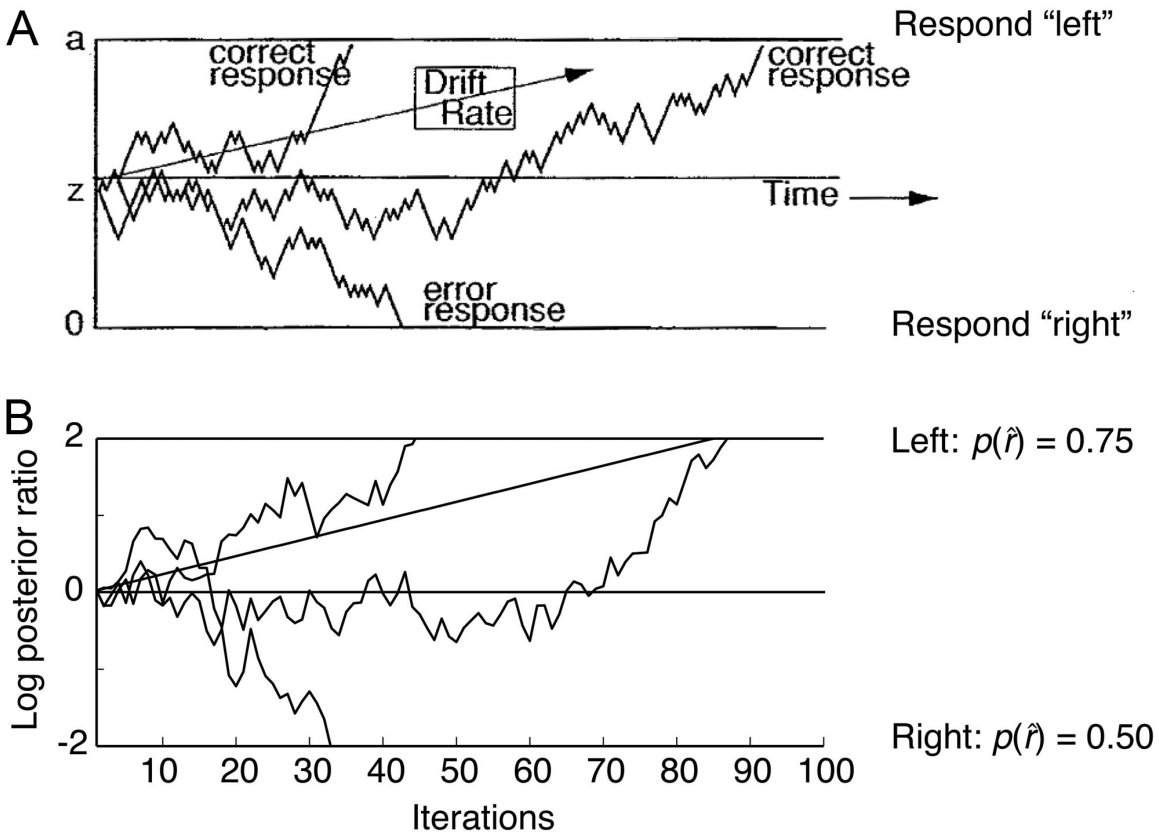

Figure 2.5: A. Evolution of the decision variable in a sequential-sampling model of a left-right visual motion judgment, both in the absence of noise (straight trajectory labeled "Drift Rate") and with the addition of noise (remaining trajectories). From "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks," by R. Ratcliff and G. McKoon, 2008, *Neural Computation*, 20, p. 876. Copyright 2008 by the Massachusetts Institute of Technology. **B**. Evolution of the log posterior ratio in the present model, as applied to a forced choice between outcomes with values as shown at right, both with and without noise (random utility).

Bayes' rule:

$$p_n(\pi|s,\hat{r}) \leftarrow \alpha p(\hat{r}|s,\pi)p_{n-1}(\pi|s,\hat{r}), \tag{2.7}$$

where α is a normalization coefficient that ensures the left-hand side term sums to one across all values of π.

Both mathematically and conceptually, this iterative procedure directly parallels the standard evidence-integration model as applied to the dot-motion task (see

Figure 2.4). Rather than noisy perceptual observations, we have stochastic observations of reward.[5] In both cases, observations are translated into likelihoods–respectively, $p(y|x)$ and $p(\hat{r}|s, \pi)$–which are used to update an evolving posterior distribution. Indeed, as in other evidence-integration models, the iterative procedure in Equation 2.7 is guaranteed to converge to the correct decision, that is, to find the optimal policy, as shown formally in Appendix A. Furthermore, in the single-step case, if a response threshold is imposed as in Figure 2.5, the procedure is guaranteed to yield the lowest error rate for a given expected decision time, just as in the sequential probability ratio test (see Appendix A).

Although it was random utility that led us to consider an evidence-integration approach, it turns out that the iterative procedure we have obtained also overcomes the other hazards enumerated at the outset of this section. Specifically, the procedure is guaranteed to converge to the optimal policy even in the presence of an initial bias toward a nonoptimal policy, and as demonstrated in Appendix A it will also find the optimal sequential policy in the multistep decision-making case. Indeed, in the multistep setting, our procedure shares structure with iterative procedures found in reinforcement learning and dynamic programming, where repeated updates allow a diffusion of information across temporally distributed events (see Sutton & Barto, 1998; Toussaint & Storkey, 2006).

2.6 Simulations

Having arrived at an algorithmic account, we turn now to a set of simulations that show the procedure in action, illustrating its applicability to hallmark patterns of behavior in goal-directed decision making. Technical details, sufficient to repli-

[5] In the evidence-integration framework one has a fixed likelihood function and stochastic observations. In the present model, one has instead a fixed observation and a stochastic likelihood function. Mathematically, these two cases are notational variants of one another.

cate these simulations, are presented in Appendix B (relevant code is available at `http://www.princeton.edu/~matthewb`).

2.6.1 Simulation 1: instrumental choice

Simulation 1.1. Simple binary choice. We begin with the simplest possible case: two-alternative forced choice with deterministic outcomes. For concreteness, and to prepare the ground for later simulations, consider a laboratory scenario in which a rat has access to two levers, positioned to its left and right. Pressing the left lever yields one kind of food and pressing the right lever another (see, e.g., Balleine & Dickinson, 1998c). Let us assume that the rat prefers the food associated with the left lever, at baseline, and assign a scalar reward value $r = 2$ to this food and a reward value $r = 1$ to the other.

The situation is modeled by defining three states, *no-food* (the initial state, $r = 0$), *food1*, and *food2*; and two policies, *press-left* and *press-right*, matched with corresponding actions. Our framework requires that reward values be represented as probabilities $p(\hat{r}|s')$. In order to map from traditional, unbounded scalar reward values (r) to probabilities between zero and one, we employ the following simple linear transformation (with alternatives discussed later):

$$p(\hat{r}|s') = 0.5\left(\frac{R(s')}{r_{max}} + 1\right), r_{max} := \max_{s'} |R(s')|. \tag{2.8}$$

For the present scenario, this yields $p(\hat{r}|food1) = 1.00$ and $p(\hat{r}|food2) = 0.75$.

The question is how the rat decides, based on its knowledge of the causal structure of the environment and its preferences over outcomes, which lever to press. One way of reaching a decision would involve the procedure shown in Figure 2.3 (top). Here, the policy variable is treated as observed, first set to press-left, then separately to press-right. In both cases, forward inference yields specific

posterior probabilities at the reward node. The probability $p(\hat{r}|s,\pi)$ turns out to be larger under the press-left policy (1.00) than under press-right (0.75), providing a sufficient basis for choice.

The potential relevance of serial policy evaluation, along the lines just described, has been recognized in recent theoretical work on animal decision making (see, e.g., Daw et al., 2005; Smith, Li, Becker, & Kapur, 2004), and recent single-unit recording data in rodents provide apparent evidence for serial consideration of future actions and outcomes at behavioral choice points (Johnson & Redish, 2007; Johnson, van der Meer, & Redish, 2007). However, our theory focuses on a different, more parallelized decision procedure. Here, the reward variable is treated as observed ($\hat{r} = 1$), and inference yields posterior probabilities for the two available policies. Figure 2.6A shows the evolution of these posteriors, over iterations of inference within a single decision-making "trial." Also displayed is the expected value of the current mixture of policies: the average of $p(\hat{r}|s,\pi)$, weighted by the posterior probability of π on the current iteration; i.e., the marginal probability $p(\hat{r}|s)$. As the figure shows, as time elapses within the decision-making episode, the model converges to the optimal deterministic policy.

To make clear what is going on "under the hood" in this simulation, let us step through the computations performed during its first three iterations. At the outset, the initial or prior probabilities $p(\pi)$ for the policies press-left and press-right are both equal to 0.5. Labeling these policies π_L and π_R the first iteration uses Equation 2.5:

$$p_1(\pi_L|s,\hat{r}) \propto p(\hat{r}|s,\pi_L)p(\pi_L) = 1 \times 0.5 = 0.5,$$
$$p_1(\pi_R|s,\hat{r}) \propto p(\hat{r}|s,\pi_R)p(\pi_R) = 0.75 \times 0.5 = 0.375.$$

Dividing each of these values by their sum, to normalize, yields $p_1(\pi_L|s,\hat{r}) = 0.57$ and $p_1(\pi_R|s,\hat{r}) = 0.43$. On the second iteration, the results of iteration 1 are fed into Equation 2.7:

$$p_2(\pi_L|s,\hat{r}) \propto p(\hat{r}|s,\pi_L)p_1(\pi_L|s,\hat{r}) = 1 \times 0.57 = 0.57,$$
$$p_2(\pi_R|s,\hat{r}) \propto p(\hat{r}|s,\pi_R)p_1(\pi_R|s,\hat{r}) = 0.75 \times 0.43 = 0.3225.$$

Normalizing, again by dividing both values by their sum, yields $p_2(\pi_L|s,\hat{r}) \approx 0.64$ and $p_2(\pi_R|s,\hat{r}) \approx 0.36$. On the third iteration, the results of iteration 2 are fed back into Equation 2.7:

$$p_3(\pi_L|s,\hat{r}) \propto p(\hat{r}|s,\pi_L)p_2(\pi_L|s,\hat{r}) = 1 \times 0.64 = 0.64,$$
$$p_3(\pi_R|s,\hat{r}) \propto p(\hat{r}|s,\pi_R)p_2(\pi_R|s,\hat{r}) = 0.75 \times 0.36 = 0.27.$$

Normalization yields $p_3(\pi_L|s,\hat{r}) = 0.70$ and $p_3(\pi_R|s,\hat{r}) = 0.30$. On the fourth iteration, these results are fed back into Equation 2.7, and the process continues in that fashion. One way of summarizing the whole procedure in this simple case, where only a single step of action considered and no noise is involved, is to note that the policy posterior $p_n(\pi|s,\hat{r})$ on each iteration n is proportional to $p(\hat{r}|s,\pi)^n p(\pi)$.

Simulation 1.2. Stochastic choice. In implementing random utility above, we introduced a random variable Z, which parameterized the reward function $p(\hat{r}|s',z)$. For simplicity, this aspect of the model was set aside in Simulation 1.1, as it shall be in subsequent simulations. In the present simulation, however, we examine its impact on the decision-making process.

To this end, we assigned the variable Z a multivariate normal distribution with zero covariance (see Appendix B). Under these conditions, the decision dynamics take the form of a drift-diffusion process, isomorphic to those purported to underlie perceptual decision making (see Appendix A). The model's behavior is

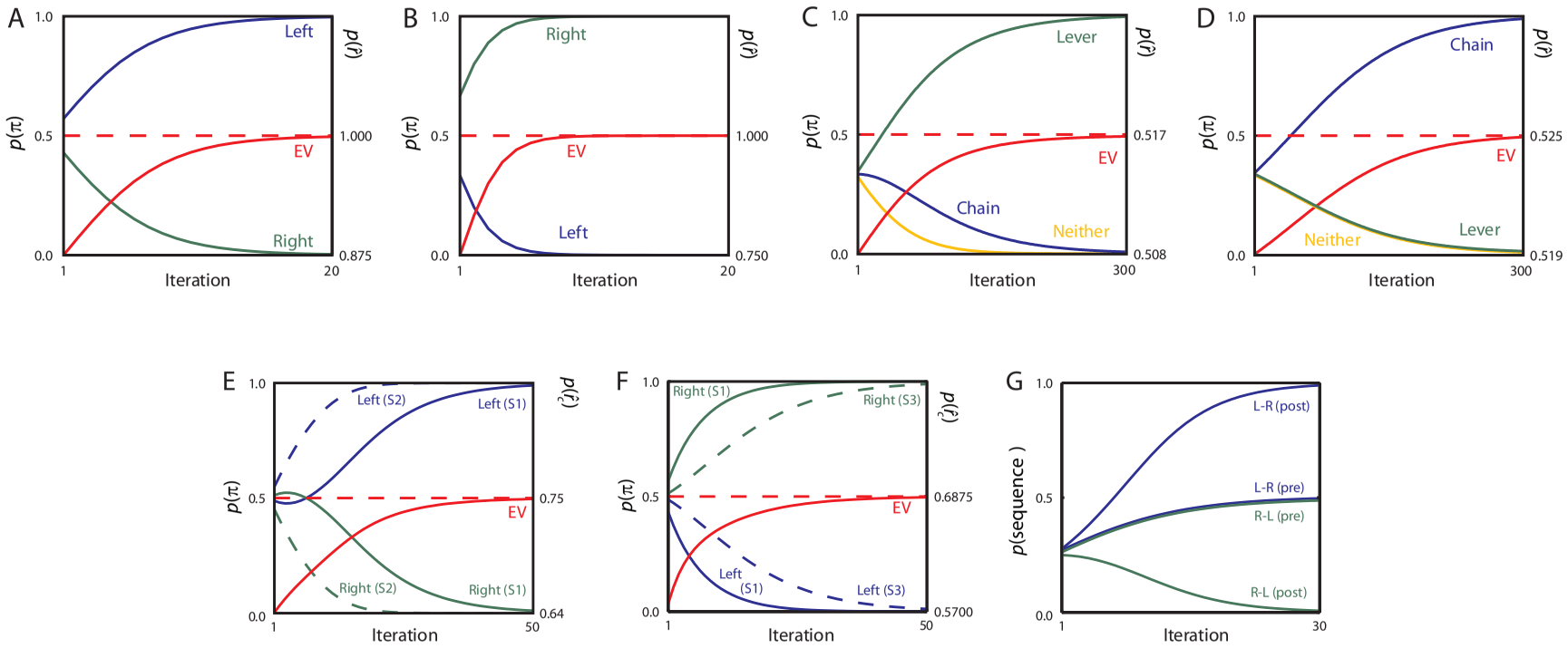

Figure 2.6: Results of Simulations 1.1 (**A**), 1.3 (**B**), 1.4 (**C–D**), 2.1 (**E**), 2.2 (**F**), and 2.3 (**G**). Blue, green, and yellow traces indicate the posterior probability of indicated actions/policies at each processing iteration. Red traces indicate the probability $p(\hat{r} = 1)$ given the mixture of policies at each iteration, proportional to the expected reward for that mixture. Dashed red lines indicate $p(\hat{r} = 1)$ for the optimal policy. In Panel G, the two most central data series are offset for legibility; the values were in fact precisely equal across the two. pre = pre-devaluation; post = post-devaluation.

illustrated in Figure 2.5B, in the same lever-choice scenario considered in Simulation 1.1. For comparison with Figure 2.5A, the figure shows the log posterior ratio, $\log[p(left)/p(right)]$, rather than the individual posteriors. In the absence of noise, this quantity follows a straight-line course, mirroring the constant drift rate of the drift-diffusion model (see Bogacz et al., 2006; Gold & Shadlen, 2007; Ratcliff & McKoon, 2008). With Z active, the log posterior ratio follows a serpentine course, tending toward the optimal policy but sometimes deviating in the other direction.

If a response threshold is introduced, as shown in Figure 2.5B, the match to the drift-diffusion model is complete. This formal link allows the present model to account for some important behavioral data concerning choice proportions and reaction times in reward-based decision making. Figure 2.7A shows data from an experiment by Padoa-Schioppa and Assad (2006), in which monkeys chose between two juice offers. A central finding in this study concerned choice variability. When one of the alternatives presented was much more valuable than the other, that option was always selected, but as the alternatives came closer together in value, the animals showed a graded increase in choice variability. When our model is faced with decisions between rewards with varying degrees of separation, it shows precisely the same kind of behavior, as illustrated in Figure 2.7B.

In a related study, Padoa-Schioppa, Jandolo, and Visalberghi (2006) showed that incentive disparity can also affect reaction time, with decisions taking longer when options are closely matched in value (see Figure 2.7C; see also Rangel, 2008; Rustichini, Dickhaut, Ghirardato, Smith, & Pardo, 2005). This finding is also captured by our model under random utility, as shown in Figure 2.7D.

An important realm of data addressed by standard evidence-integration models centers on reaction-time distributions. As shown in Figure 2.7E, in many decision-making settings such distributions assume a characteristic skewed shape, with the distribution becoming broader under conditions leading to greater choice vari-

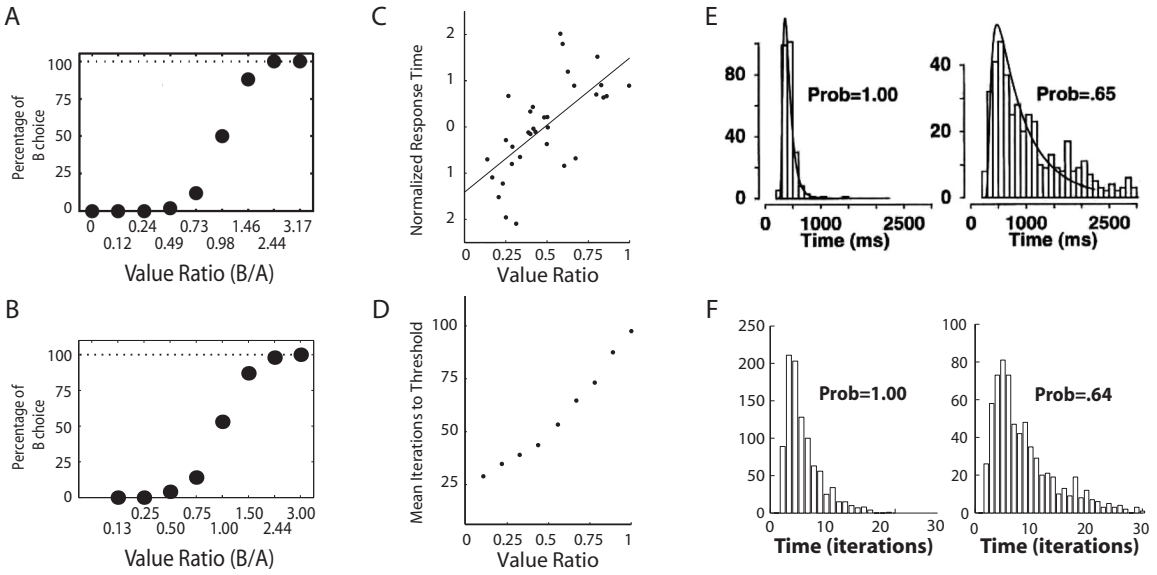

Figure 2.7: **A**. Choice data from Padoa-Schioppa and Assad (2006, Figure 1B, p. 223). Value ratio indicates the subjective value of choice option B relative to option A as inferred from choice behavior. Copyright 2006 by Macmillan Publishers Ltd: Nature. **B**. Choice data from Simulation 1.2, including random utility and a response threshold on the log posterior ratio of 2.0. Reward values for choice options were selected so as to yield the ratios shown on the x-axis. Each point reflects the choice proportion over a sample of 1,000 trials. **C**. Response time data from Padoa-Schioppa et al. (2006). Reprinted from *Cognition*, 99, C. Padoa-Schioppa, L. Jandolo, and E. Visalberghi, "Multi-Stage Mental Process for Economic Choice in Capuchins," p. B6, Copyright 2006 by Elsevier. **D**. Response times in the simulation associated with Panel B. **E**. Response-time distributions in a two-alternative perceptual judgment, under stimulus conditions yielding uniform judgments (Prob = 1.00) and more variable judgments (Prob = .65). The superimposed curve shows the fit of a drift-diffusion model. Adapted from "Modeling Response Times for Two-Choice Decisions," by R. Ratcliff and J. N. Rouder, 1998, *Psychological Science*, 9, Figure 5, p. 352. Copyright 1998 by Sage. **F**. Response-time distributions from the simulation associated with Panels B and D, with outcome value ratios chosen so as to yield choice variabilities close to those in the Ratcliff and Rouder (1998) experiment.

ability. As shown in Figure 2.7F, our model generates reaction time distributions showing the same characteristics. The model thus predicts that reaction-time distributions in goal-directed choice should resemble those observed in other settings, including perceptual decision making (Ratcliff & Rouder, 1998) and memory retrieval (Ratcliff, 1978). To our knowledge, reaction-time distributions in reward-based decision making have not yet been studied experimentally.

Simulation 1.3. Outcome devaluation. As a further proof of concept, our paradigm can be used to simulate incentive devaluation. For this purpose, we return to the two-lever scenario and the model introduced in Simulation 1.1. In Balleine and Dickinson (1998c), to take a representative study, the incentive value of one of two action outcomes was devalued by specific satiety, leading to an immediate reduction in performance of the associated action. This devaluation effect can be captured in our model by simply changing the reward value associated with one food outcome. To simulate the effect of this, we reduced the reward value associated with the formerly preferred food from $r = 2$ to $r = 0$. Note that this change directly affects only the CPD of the reward variable; $p(\hat{r}|s')$ is reduced, for the case where s' corresponds to the devalued food. When inference is performed, however, the impact of this local change propagates to the level of the policy node, yielding a reversal in choice (see Figure 2.6B).

Simulation 1.4. Contingency degradation. As discussed in the introduction, changes in goal-directed decisions can be induced not only by revaluation of outcomes but also by changes in patterns of causal contingency (Dickinson & Mulatero, 1989; Williams, 1989). A representative demonstration was reported in Colwill and Rescorla (1986). Here, rats were given access to a lever and a chain. If the lever was pressed, a preferred food was delivered with probability 0.05. Pulling the chain yielded a less preferred food, again with probability 0.05. Under these conditions, not surprisingly, animals came to favor the lever. However, in the next phase of the

experiment, the causal link between the lever and the preferred food was broken by delivering the preferred food with probability 0.05 regardless of the animal's action (or inaction). Following this change, animals shifted their efforts toward the chain response.

To simulate this effect, we adapted the model from Simulation 1.1 to include three policies, corresponding to the actions *chain*, *lever*, and *neither* and the four states *no-food* ($r = 0$), *food1* ($r = 1$), *food2* ($r = 2$), and *both-foods* ($r = 3$). When the model was parameterized to reflect the initial contingencies in the experiment, evidence integration led to selection of the lever action (see Figure 2.6C). When the CPD $p(s'|s, a)$ was updated to predict the later contingencies, where *food2* could occur without any action, and *both-foods* could occur following *chain* (but not following *lever*), preference shifted to the *chain* response (see Figure 2.6D).

In addition to illustrating contingency degradation, this simulation demonstrates the ability of the present framework to cope with probabilistic outcomes. Given an accurate representation of outcome contingencies, evidence integration will yield the response with the highest expected utility. Indeed, the likelihood $p(\hat{r}|s, \pi)$, which marginalizes over outcomes s', can be viewed as a direct representation of expected utility (see Equation 2.4).

2.6.2 Simulation 2: sequential decision

Here we apply the iterated architecture from Figure 2.2D to simulate benchmark phenomena in multistep decision making.

Simulation 2.1. A two-stage decision problem. As an initial illustration of sequential choice, we focus in this simulation on the two-step T-maze scenario from Niv et al. (2006), described in the introduction and illustrated in Figure 2.1. The states included in our model of this situation include the terminal reward items (*cheese*, *carrot*, *water*, and *null*), as well as the three preceding choice points.

Following Niv et al. (2006), we assume the baseline reward values $R(\textit{cheese}) = 4$, $R(\textit{carrots}) = 3$, $R(\textit{water}) = 2$, $R(\textit{null}) = 0$. Figure 2.6E shows the decision trajectory produced by evidence integration in this problem setting. The model converges on the sequence *left-left*, a policy that takes it to the preferred cheese reward.

If we were to "look under the hood," tracing the computations on successive iterations at each stage of the plan, the story would be identical to that in Simulation 1.1, with the following important caveat: The calculations bearing on the first stage of the plan (i.e., the policy at S_1) are impacted by the current policy posteriors at stage 2 (i.e., at S_2 and S_3). For example, the first iteration computes the posterior probability of adopting the left and right policies at S_1. Calling these $\pi_L^{S_1}$ and $\pi_R^{S_1}$, Equation 2.5 gives

$$p_1(\pi_L^{S_1}|s,\hat{r}) \propto p(\hat{r}|s,\pi_L^{S_1})p(\pi_L^{S_1}),$$
$$p_1(\pi_R^{S_1}|s,\hat{r}) \propto p(\hat{r}|s,\pi_R^{S_1})p(\pi_R^{S_1}).$$

The likelihood terms here–$p(\hat{r}|s,\pi_L^{S_1})$ and $p(\hat{r}|s,\pi_R^{S_1})$–depend implicitly on what is planned for S_2 and S_3, that is, on $p(\pi_L^{S_2})$, $p(\pi_R^{S_2})$, $p(\pi_L^{S_3})$, and $p(\pi_R^{S_3})$.

This dependence manifests in the time courses plotted in Figure 2.6E. Note the trajectory of the solid blue and green traces in the figure, which relate to the decision at S_1. Although the decision ultimately tips toward *left*, early on there is transient movement toward *right*. This effect stems directly from the fact that the optimal first-step choice depends on what is planned for later steps. As discussed earlier, if the animal is equally likely to go left or right upon reaching either S_2 or S_3, the expected reward for a left turn at S_1 is $(4 + 0)/2$, and that for a right turn is $(2 + 3)/2$. The (locally) optimal choice at S_1 is thus to turn right. Eventually, as better plans emerge for S_2 and S_3, the expected reward for left and right turns at S_1 move toward 4 and 3, respectively, making it preferable to turn left at S_1.

As discussed below, the kind of dynamics reflected in this simulation, arising from the interdependence of decision-making operations across plan steps, gives rise to testable model predictions.

Simulation 2.2. Cumulative reward and cost-benefit analysis. A key feature of multistep decision problems is the need to compute cumulative rewards when rewards are distributed across steps of action. A simple and ubiquitous case arises in effort-based decision making, where a cost-benefit analysis must take into account both distal rewards and the cost of proximal effort. A number of rodent studies have examined this cost-benefit analysis by placing an animal inside a T maze where both arms contain food, but where one also contains a scalable barrier that the animal must surmount to access the food reward. The common finding is that, unless the reward on the barrier side is larger by a sufficient degree, animals will forgo it, avoiding the effort required (Salamone, Correa, Farrar, & Mingote, 2007; Walton, Kennerley, Bannerman, Phillips, & Rushworth, 2006).

This sort of cost-benefit analysis can be modeled very naturally within the present framework. For simplicity, we do so using the two-step T-maze scenario already established. Here, we reimpose the original reward values on the outcome states, but also imagine that there is now a scalable barrier placed at S_2. The cost of traversing this barrier is inserted into the model by reducing $R(S_2)$ to -2. Evidence integration under these circumstances yields the decision trajectory in Figure 2.6F, which reflects the inference that the value of the most preferred reward is not worth the associated cost in effort.

Simulation 2.3. Outcome revaluation and contingency degradation. As discussed in the introduction, outcome revaluation can affect decisions in multistep settings, just as in simpler decision tasks. To recap one relevant study, Ostlund et al. (2009) trained rats to execute two two-step lever-press sequences (*left-right*, *right-left*), which yielded sucrose and polycose, respectively. When one of these

outcomes was devalued through satiety, the animals tended to favor the sequence yielding the nondevalued food.

Note that, although the Ostlund et al. (2009) experiment involves lever pressing rather than maze navigation, the form of the decision problem aligns precisely with the two-step T maze from Niv et al. (2006). State S_1 in Figure 2.1 now corresponds to the rat's initial situation, facing the two levers, with available actions *press-left* and *press-right*. State S_2 corresponds to the rat's situation after having pressed the left lever once; state S_3 to the situation after pressing the right lever once.[6] The outcomes for *press-left* and *press-right* are, respectively, *null* ($r = 0$) and *polycose* ($r = 11$) at S_2; and *sucrose* ($r = 1$) and *null* ($r = 0$) at S_3. Using the same model architecture that we used to simulate the two-step T maze, these initial conditions lead to selection of the sequences *left-right* and *right-left* (with equal probability) over *left-left* and *right-right* (see Figure 2.6G). Simulating devaluation by reducing R(*sucrose*) to 0.5 leads to a preference for *left-right* over all other sequences, in line with the empirical observation (see Figure 2.6G).[7]

Ostlund et al. (2009) also showed analogous changes in sequence production following contingency degradation. Simulating contingency degradation in the present model, using the approach established in Simulation 1.4, yields parallel results (data not shown). Using a similar logic, the model can be applied in a straightforward way to account for the classic latent learning and detour effects described in the introduction (see Botvinick & An, 2009).

[6]Note that the rat's "state" at S_2 and S_3 might thus be understood as factoring in an internal representation of past actions. However, as Ostlund et al. (2009) noted, this is not strictly necessary, as visual, tactile, and proprioceptive information might suffice to discriminate among the relevant situations.

[7]It is worth remarking that the computational account we are offering here for the findings of Ostlund et al. (2009) differs from those authors' own interpretation. Ostlund et al. considered the observed pattern of behavior to indicate the involvement of "chunked" representations of action sequences. The present simulation illustrates that chunking is not in fact necessary. Having noted this, however, we hasten to add that chunked or hierarchical representations are nonetheless likely to play a role in goal-directed decision making, a point to which we return in the General Discussion.

2.6.3 Predictions

In addition to demonstrating the ability of our framework to account for benchmark phenomena in goal-directed behavior, the foregoing simulations give rise to several testable predictions.

One of these arises from Simulation 2.1 and pertains to decision time course. As shown in Figure 2.6E, the model in this simulation displayed a sort of decision-making reversal, traveling toward one policy early on and then, later, toward another. The origins of this effect, as discussed earlier, lie in the recursive structure of the planning problem: The optimal policy for any stage of the plan depends on what is planned for later stages. In Simulation 2.1, this general principle combined with a specific set of conditions, according to which the outcome with the *maximum* value lay in one direction, and the outcomes in the other direction had a larger *mean* value. Our model predicts that this mean–max conflict situation should trigger a similar reversal at the level of neural response representations in human or animal subjects. One way of testing this prediction behaviorally would be to impose response deadlines in order to elicit speeded choice reactions. Under these circumstances, the model predicts that short-latency responses in mean–max conflict conditions should show below-chance accuracy.

Another prediction arises from Simulation 1.4. As illustrated there and demonstrated formally in Appendix A, our evidence-integration algorithm yields mathematically sound decisions in the face of probabilistic outcomes. However, an interesting and somewhat surprising effect arises during this process. Recall that with each iteration of the decision-making process, for each planned action, our model computes a posterior probability distribution over outcomes (s'). It turns out that this posterior distribution is optimistic. That is, it is weighted toward high-utility outcomes. For example, in Simulation 1.4, selection of the *lever* action prior to contingency degradation led to the outcome *food2* with probability 0.05.

However, at asymptote, the model attaches to this outcome a posterior probability of 0.08.

To see the origins of this optimism effect, recall that decision making begins with an assumption of reward (i.e., the premise $\hat{r} = 1$). This assumption feeds into the calculation of outcome probabilities, with the natural consequence that they are weighted toward states with higher utility. It is important to emphasize that this aspect of the model does not affect the model's actual decisions; as we have noted, the model's choices of action conform to sound calculations of expected utility. Nevertheless, even as the model chooses rationally, it gives rise to optimistic estimates of outcome probability. This translates into a further testable prediction of the present theoretical account.

The predicted optimism effect bears an interesting relationship to what previous work has labeled the "illusion of control." Here, individuals make more optimistic outcome predictions when their actions are freely chosen than when their actions are dictated (Presson & Benassi, 1996). For example, Langer (1975) found that experimental participants expressed greater confidence in their chances of winning a drawing when they were permitted to select a ticket from among a set of objectively equivalent tickets than when a random ticket was simply given to them. A standard explanation for this effect has been that choice serves as a cue falsely implying outcome controllability (Langer, 1975; Presson & Benassi, 1996). The present work suggests a different, though perhaps not incompatible, explanation, which is that choice gives rise to optimism as a natural consequence of the computations involved in goal-directed decision making.

How do these predictions compare with those of competing theories? This is not a straightforward question to answer, given the dearth of psychological and neuroscientific theory concerning the processes underlying goal-directed decision making, particularly in sequential domains. However, it is perhaps useful

to consider whether different machine-learning algorithms for model-based reinforcement learning might give rise to comparable predictions. In this respect, the above predictions concerning choice dynamics appear not to arise from algorithmic approaches in which depth-first tree search is employed (see, e.g., Smith et al., 2004), or where choice depends on backward induction (starting at the goal and working backward, in the spirit of successive subgoaling). On the other hand, the same predictions might obtain in more parallel procedures, such as the classical value iteration algorithm (see Sutton & Barto, 1998). In contrast, our model's prediction concerning optimistic state representation appears problematic even for the latter planning procedure and thus stands as a particularly distinctive prediction of the present framework.

2.7 Neural implementation

To this point, we have considered goal-directed decision making in abstract cognitive or information-processing terms. However, ultimately what is needed is an account that makes direct contact with neuroscientific data, pinpointing the neural structures and processes that give rise to goal-directed decisions. One of the most exciting aspects of recent empirical research on goal-directed decision making is that it has begun to shed some light on the relevant functional anatomy, identifying critical brain regions and, in some cases, characterizing the response properties of the neurons they contain. Despite such progress, we still lack a working model of how these brain structures interface and interact in order to support goal-directed decision making.

In this section we leverage the present theory to sketch out such a functional neural model. More specifically, we translate the theory into neural terms at two distinct levels of description. First, at a structural level, we map the elements

of our model to specific gross brain regions, as discussed in the next subsection. Then, at a finer grain, we cash out the proposed information-processing operations within a neural network model, yielding a coarse account of how neurons within the relevant brain regions may collaborate in generating goal-directed behavior.

2.7.1 Four interlocking neural systems

The graphical architecture we have been considering contains variables of four types, which represent, respectively, (a) *policies*, (b) *actions*, (c) current and projected situations or *states*, and (d) *reward* or *utility*. As noted previously, these four domains of representation, along with the transition and reward functions that link them, constitute the givens of the model-based reinforcement learning problem. However, each of the four representational domains can also be mapped to distinct sets of neuroanatomic regions. Making this mapping ties the four strata of our model to specific brain systems, opening the door to a consideration of the model's potential neuroscientific implications.

1. The policy system. Recall that the policy nodes in our model represent mappings from situations to responses. In the brain, representations of this kind have been shown to reside within the dorsolateral prefrontal cortex (DLPFC). Single-unit recording studies in primates and complementary functional neuroimaging studies in humans have indicated that one important function of the DLPFC may be to represent task sets or "rules" (Asaad, Rainer, & Miller, 2000; Bunge & Wallis, 2007; Sakai, 2008; Wallis, Anderson, & Miller, 2001; White & Wise, 1999). The content of such rules is typically understood to establish a set of relationships between stimuli and responses (Bunge, 2004). According to the guided activation theory of E. K. Miller and Cohen (2001), a critical function of the DLPFC is to bias the flow of neural activation in pathways between stimulus and response representations, supporting transmission along task-relevant pathways. Given this role, it is not

surprising that the DLPFC has been heavily implicated in planning and goal direction (Anderson, Albert, & Fincham, 2005; Duncan, Emslie, Williams, Johnson, & Freer, 1996; Goel & Grafman, 1995; Lengfelder & Gollwitzer, 2001; E. K. Miller & Cohen, 2001; Shallice, 1982; Shallice & Burgess, 1991; Tanji & Hoshi, 2008; Tanji, Shima, & Mushiake, 2007; Unterrainer & Owen, 2006). Furthermore, studies on outcome devaluation in rodents (Balleine & Dickinson, 1998b; Corbit & Balleine, 2003; Killcross & Coutureau, 2003; although see Ostlund & Balleine, 2005) suggest that it depends critically on prelimbic cortex, a structure judged by some to represent a homologue to the primate DLPFC (Fuster, 1997; Kesner, 2000; Uylings, Groenewegen, & Kolb, 2003).

Although the DLPFC is the area most heavily implicated in policy representation, there are data suggesting that policy, task set, or rule representations may also reside in other portions of the frontal lobe, including premotor cortex (Wallis & Miller, 2003), ventrolateral prefrontal cortex (Bunge, 2004; Bunge et al., 2005), presupplementary area (Dosenbach et al., 2006; Rushworth, Walton, Kennerley, & Bannerman, 2004), and the frontal pole (Sakai & Passingham, 2003). The policy stratum in our model thus summarizes a role that is carried out in the brain by a densely interconnected network of cortical regions, with the DLPFC as an important hub.

2. The action system. Within our model, policy nodes interface with nodes representing actions. If the pertinent actions are understood as bodily movements, then the set of relevant brain areas is relatively straightforward to identify, and would include premotor and supplementary motor cortices, portions of cingulate and parietal cortex, and associated sectors within the dorsal striatum. However, goal-directed decision making can involve more abstract forms of action, including actions defined in terms of ends rather than motoric means, implicating intraparietal and inferior frontoparietal cortex (Hamilton & Grafton, 2006, 2008), or

temporally extended behaviors, currently speculated to be represented in portions of prefrontal cortex (see Badre, 2008; Botvinick, 2008). The action variables in our model thus, once again, summarize the role of a specific network of areas.

3. The state projection system. Within our model's architecture, action nodes project to and receive projections from nodes representing current and projected situations or states.[8] On the neuroscientific side, it is clear that the brains of higher animals must contain representations of anticipated states, as well as their dependencies on earlier states and actions (Atance & O'Neill, 2001; Gopnik & Schulz, 2007; Schütz-Bosbach & Prinz, 2007). However, despite considerable research, the neuroanatomical site of such representations is only beginning to emerge. Early studies of spatial navigation in rodents led to the idea that cognitive map representations might reside in the hippocampus (O'Keefe & Nadel, 1978), and recent research suggests that hippocampal place cells may represent projected future locations (Diba & Buzsáki, 2007; Johnson & Redish, 2007; Johnson et al., 2007). Lesion studies have also provided evidence for the involvement of medial temporal lobe structures (entorhinal cortex, if not hippocampus) in the representation of action-outcome contingencies during instrumental learning (Corbit, Ostlund, & Balleine, 2002). Convergent neuropsychological research in humans indicates that medial temporal lobe structures may play a critical role in allowing visualization of future events, including action outcomes and goals (Buckner & Carroll, 2007; Hassabis, Kumaran, Vann, & Maguire, 2007; Schacter, Addis, & Buckner, 2007), possibly as part of a larger network including regions within medial and lateral parietal cortex (see also Hamilton & Grafton, 2008), lateral temporal cortex, and medial frontal cortex (see also K. Matsumoto, Suzuki, & Tanaka, 2003; K. Matsumoto & Tanaka, 2004; Tanaka, Balleine, & O'Doherty, 2008). Still other work has suggested that

[8] It should be noted that our model intends "state" to encompass not only ambient environmental circumstances but also internal state, including the state of working memory (implicating relevant DLPFC and parietal areas), affective state (amygdala, insula, and other affect-related structures), and homeostatic conditions (hypothalamus).

the DLPFC may play a role in representing projected action outcomes, including both final "goal" states and intermediate "means" states (Fuster, 1997; Mushiake et al., 2006; Saito, Mushiake, Sakamoto, Itoyama, & Tanji, 2005), and a recent study by Hamilton and Grafton (2008) suggests that the right inferior frontal cortex may also be involved in representing action outcomes.

At the subcortical level, there is strong evidence for the involvement of specific basal ganglia structures in the representation of action-outcome contingencies. Research in rats has shown that damage to or inactivation of the dorsomedial striatum impairs sensitivity to outcome devaluation and changes in instrumental contingency (Balleine, 2005; Yin, Knowlton, & Balleine, 2005; Yin, Ostlund, Knowlton, & Balleine, 2005). This fits well with research implicating the caudate nucleus, the primate homologue of the dorsomedial striatum, in action-outcome contingency detection (Tanaka et al., 2008) and planning (Monchi, Petrides, Strafella, Worsley, & Doyon, 2006; Unterrainer & Owen, 2006) in humans. A potential role for the striatum in representing action-outcome contingencies is particularly interesting given evidence for overlapping inputs from dorsal and orbital prefrontal areas within anterior striatum (Cavada, Compañy, Tejedor, Cruz-Rizzolo, & Reinoso-Suárez, 2000; Haber, Kim, Mailly, & Calzavara, 2006), a convergence that fits well with the structure of our graphical model.

4. The reward system. The final set of elements in our model are nodes representing reward. Here again, the variables in question can be understood as summarizing the representational role of a specific set of brain regions. In this case, the relevant regions include, most prominently, the orbitofrontal cortex and the basolateral amygdala. The orbitofrontal cortex (OFC) has been extensively implicated, across species, in the representation of the incentive value of stimuli, including anticipatory coding for the value of predicted and even imagined outcomes (Arana et al., 2003; Bray, Shimojo, & O'Doherty, 2010; Kringelbach, 2005;

Montague & Berns, 2002; Padoa-Schioppa & Assad, 2006; Plassmann, O'Doherty, & Rangel, 2007; Rolls, 2004, 2006). This function has been linked to a role in goal-directed decision making (Frank & Claus, 2006; Roberts, 2006; Rolls, Everitt, & Roberts, 1996; Schoenbaum & Setlow, 2001; Schultz, Tremblay, & Hollerman, 2000; Wallis, 2007), based in part on studies demonstrating OFC involvement in revaluation phenomena (De Araujo, Kringelbach, Rolls, & McGlone, 2003; Gottfried, O'Doherty, & Dolan, 2003; Izquierdo, Suda, & Murray, 2004; LaBar et al., 2001; Pickens, Saddoris, Gallagher, & Holland, 2005; Valentin, Dickinson, & O'Doherty, 2007; however, see Ostlund & Balleine, 2007).

Despite important differences in function, the basolateral amygdala (BLA) has also been extensively implicated in the representation of incentive value of stimuli, including action outcomes, and in the guidance of goal-directed behavior (Arana et al., 2003; Balleine, 2005; Baxter & Murray, 2002; Holland & Gallagher, 2004). Like OFC, BLA has been directly implicated in revaluation phenomena (Balleine, Killcross, & Dickinson, 2003; Corbit & Balleine, 2005; Gottfried et al., 2003; LaBar et al., 2001). Indeed, there is evidence that revaluation sensitivity may depend on a functional interaction between BLA and OFC (Baxter, Parker, Lindner, Izquierdo, & Murray, 2000), suggesting that these structures might be most fruitfully regarded as two components within an integrated system for reward representation (Cavada et al., 2000; Schoenbaum, Setlow, Saddoris, & Gallagher, 2003).[9]

Figure 2.2 summarizes the proposed correspondences between elements of our model and functional neural structures. At one level, these parallels simply recapitulate existing ideas concerning the functional roles of the implicated brain areas. However, because we have drawn the parallels in the context of an explicit algorithmic model, what results is a proposal concerning the way that the relevant

[9] Although we have focused on OFC and BLA as substrates for the representation of utility, it should be noted that there is evidence that the costs of effort, as studied in Simulation 2.2, may be represented in different structures, in particular the dorsal anterior cingulate cortex (Botvinick, Huffstetler, & McGuire, 2009; Rudebeck, Walton, Smyth, Bannerman, & Rushworth, 2006).

neural structures interact to support goal-directed decision making. Of course, this account is specified at a very high level of abstraction. What we ultimately need is an account of the computations carried out by the neurons residing in each of these anatomical regions. In the next section, we extend the present account to make contact with this level of description.

2.8 Neural network model

The pivotal operation in our graphical framework (as in many applications of probabilistic graphical models) involves computing a marginal distribution for each variable the graph contains. What is required, in order to translate our account into neural terms, is an account of how this marginalization operation might be carried out in a neural network. Fortunately, a number of recent theoretical papers have addressed just this problem (Beck & Pouget, 2007; Deneve, 2008; T. S. Lee & Mumford, 2003; Litvak & Ullman, 2009; Ma, Beck, Latham, & Pouget, 2006; Pouget, Dayan, & Zemel, 2003; Rao, 2007). One approach that is particularly well suited to the present application was proposed by Rao (2005). Rao focused on a classic algorithm for marginalization in graphical models, known as *belief propagation* (Pearl, 1988). Belief propagation operates through *message passing*: Each variable node in the network sends to each of its neighbors a vector-valued message, the components of which encode specific marginal probabilities. The outgoing messages at each node are computed by combining incoming messages with information stored locally at the node. After the information from each node propagates throughout the network, the messages converging at each node can be combined to compute the marginal distribution for the pertinent variable (for full details of the algorithm, see Koller & Friedman, 2009; Pearl, 1988).

The propagation of messages in belief propagation is manifestly similar to the propagation of activation within a neural network; indeed, the algorithm was originally inspired by neural network research (Weiss & Pearl, 2010). Making good on this similarity, Rao (2005) suggested how networks of biological neurons might directly implement the belief propagation algorithm, applying the resulting approach to several specific problems, including evidence integration in perceptual decision making. Briefly, Rao's (2005) proposal was that each variable in the underlying graph is represented by a group of neurons, each coding for a particular message component in its instantaneous firing rate. The passage of messages between neighboring variables translates to synaptic transmission of firing-rate information, with synaptic weights and dendritic operations helping to transform the set of incoming messages into new outgoing messages.[10]

We applied the proposal from Rao (2005) in order to transpose our theory into the format of a neural network (for implementational details and simulation procedures, see Appendix B; simulation code is available at `http://www.princeton.edu/~matthewb`). Starting from the two-alternative forced-choice model introduced in Simulation 1.1, the resulting recurrent neural network is shown in Figure 2.9. Each disk in the figure corresponds to a single neuron-like unit, which carries a scalar activation value between zero and one, representing its instantaneous firing rate. This activation value corresponds to a specific message component prescribed by belief propagation, and each group of color-matched units together represents a particular set of probabilities, as spelled out in the table in Figure 2.8. For example, the red units at the top of the network diagram together encode the "message" $p(\pi|\hat{r}, s)$. As such, their activation values should evolve like the pol-

[10]The scheme from Rao (2005) carries with it two particularly speculative assumptions, which are important to acknowledge. First, it requires multiplicative interactions between presynaptic neurons. Although both modeling and empirical work have begun to shed light on how this might be accomplished (Mel, 1992, 1993; Polsky, Mel, & Schiller, 2004), further work is necessary to elucidate the details of these mechanisms. Second, this approach assumes that dendrites are able to approximate a logarithmic transformation (see Rao, 2005, for discussion).

icy posteriors in the graphical model, as diagrammed in Figure 2.6A. Figure 2.8A confirms that this is indeed the case.

The neural network in Figure 2.8 may seem rather elaborate for such a simple task (i.e., two-alternative forced choice with deterministic outcomes). However, it should be borne in mind that, by design, the architecture accommodates more complex scenarios, including problems with stochastic action-outcome contingencies, and problems where the initial state is not uniquely known at the time of planning (see Footnote 3). Furthermore, the neural network, like the probabilistic graphical model on which it is based, does more than map from rewards to policies: It also projects outcome states and expected rewards, as detailed in what follows. Perhaps most important, it is straightforward to apply the same implementational approach to multistep decisions. As an illustration, we converted to neural network form the two-step T-maze model described in Simulation 2.1. Figure 2.8B shows unit activations over iterations of processing for units coding for the policy at the first and second stages of the network, analogous to Figure 2.6E.[11]

A critical aspect of information processing in biological neural networks is its stochasticity, apparent in the random variability in the interspike interval (Shadlen & Newsome, 1998). The impact of this variability can be captured in the present implementation by relating the activity of each unit to the variable number of spikes that might be fired by a biological neuron during a small time interval (see Appendix B). Figure 2.12B shows the behavior of the policy units in the two-alternative forced-choice network when variability is introduced in this way. The dynamics of the decision-making process here resemble those arising in Simulation 1.2, under random utility, and the network shows the same dependence of choice proportion on incentive disparity (see Figure 2.8C). In the present case, however,

[11]The minor differences between Figures 2.8B and 2.6E arise from the fact that an exact algorithm was used in Simulation 2 (see Appendix B). Belief propagation is, technically speaking, an approximate inference algorithm in graphs that contain loops, and so is not guaranteed to yield marginals precisely equivalent to those arising from exact algorithms (see Koller & Friedman, 2009).

the model's behavior arises not from randomness isolated to the utility function but instead from randomness in neural firing throughout the entire network. This feature of the neural network implementation fits well with recent neuroscientific analyses of economic decision making, which have asserted that the variability traditionally ascribed to random utility should indeed be seen as simply reflecting variability in neural activity (see Shadlen, 2008).

2.8.1 Simulations

Our neural network implementation presents a further opportunity to test the present theoretical framework against empirical data. If the model is valid, then, despite its simplicity, it seems reasonable that the response profiles of the units within it should correspond to those of actual neurons in the relevant brain systems. The following simulations document several such parallels.

Simulation 3.1. State value. Recent neuroscientific studies have distinguished sharply between two forms of value representation. Studies of OFC suggest that many neurons in this region code for *state value*, the reward value associated with specific states, outcomes, or goods (Padoa-Schioppa, 2011; Tremblay & Schultz, 1999). Studies in several other areas, including dorsal striatum (Hori, Minamimoto, & Kimura, 2009; Kim, Sul, Huh, Lee, & Jung, 2009; Lau & Glimcher, 2008; Lauwereyns, Watanabe, Coe, & Hikosaka, 2002; Pasquereau et al., 2007; Samejima, Ueda, Doya, & Kimura, 2005) and parietal cortex (Dorris & Glimcher, 2004; Platt & Glimcher, 1999; Sugrue, Corrado, & Newsome, 2004), have identified neurons that code for *action value*. During decision making, these neurons code for specific actions, but in a way that depends on the expected reward for the relevant action.

Our neural network implementation contains units coding for state value and for action value. Units coding for state value lie at the bottom of the diagram in Figure 2.8 (shown in purple), in a sector of the model we earlier related to OFC.

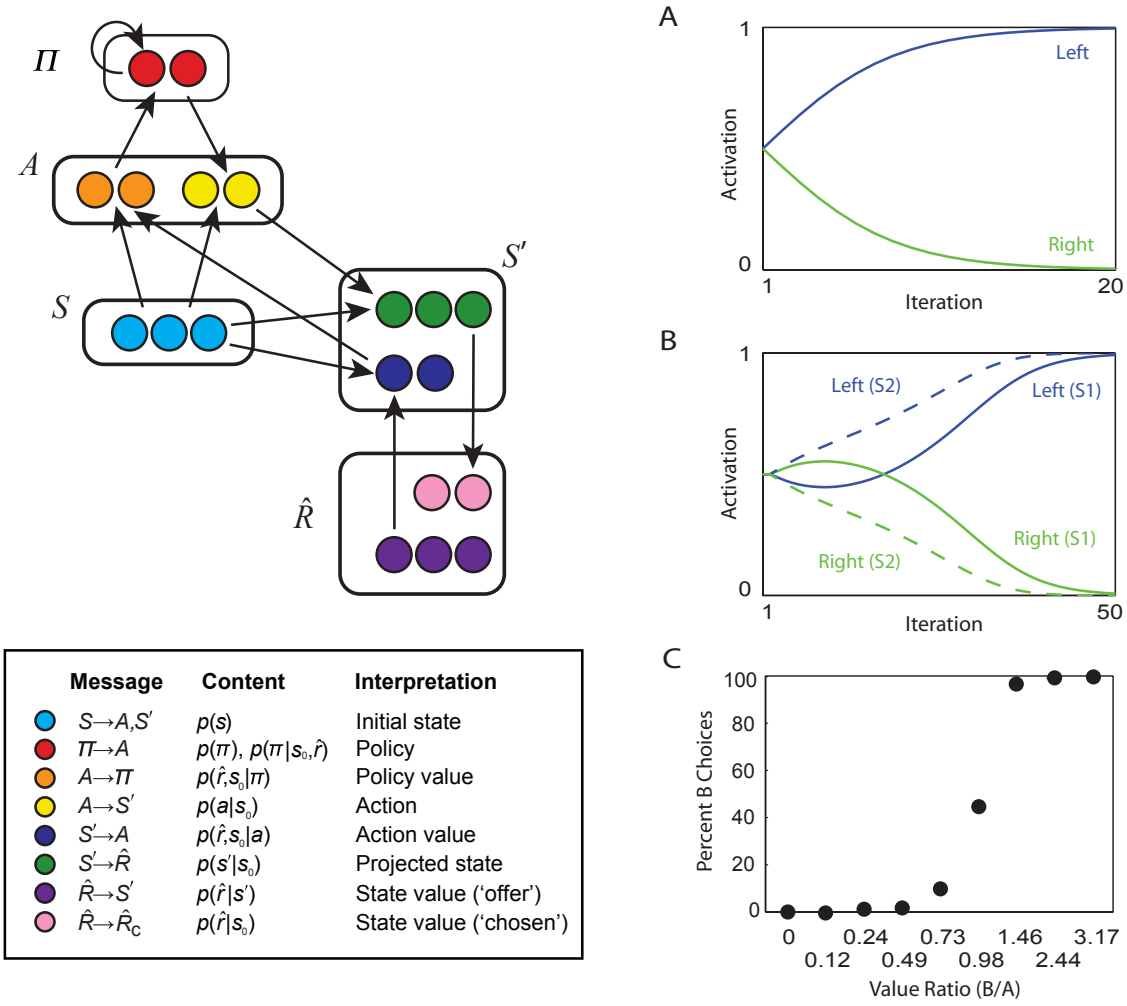

Figure 2.8: Left: Neural network implementation for two-alternative forced choice decision, with unit colors keyed to the table below. Arrows indicate all-to-all connections between the indicated unit groups. The group shown in pink derives from the multistep model and is included for Simulation 3.1. **A**. Replication of Simulation 1.1 (cf. Figure 2.6A). **B**. Replication of Simulation 2.1 (cf. Figure 2.6E). **C**. Replication of Simulation 1.2. (cf. Figure 2.7B).

To illustrate the correspondence, we used the network to simulate a neurophysiological study by Padoa-Schioppa and Assad (2006). Here, monkeys chose between different quantities and types of juice by making a saccade to one of two locations. Single-unit recordings in OFC revealed that a subset of neurons were sensitive to the offers made on each trial, independent of the monkey's subsequent choice. Figures 2.9A and 2.9B show the firing rates of two neurons, each encoding the value of a particular juice offer. We modeled this task using three states (*decision*, *juice-A*, and *juice-B*) and two actions (*saccade-left* and *saccade-right*). The values of the messages $\hat{R} \rightarrow S'$ for the series of decisions in Figures 2.9A and 2.9B are shown in Figures 2.9D and 2.9E, respectively.

In addition to discovering neurons coding for "offer value," Padoa-Schioppa and Assad (2006) discovered OFC neurons coding for "chosen value," the value of the option ultimately selected by the animal (Figure 2.9C). In our model, chosen value corresponds to the marginal probability $p(\hat{r}|s)$, which appears as a message in the multistep version of our model (pink units in Figure 2.8). The activation of the relevant unit, across the series of decisions denoted in Figure 2.9C, is shown in Figure 2.9F.

Simulation 3.2. Action value. Representations of action value are borne by different units within our model, specifically the units labeled $S' \rightarrow A$ and shown in blue in Figure 2.8. To illustrate, we used the model to simulate another single-unit recording study, by Lau and Glimcher (2008). Here, monkeys chose between visual targets yielding different quantities of juice. The study revealed that neurons within the dorsal striatum coded for specific eye movements, but in a way that reflected the reward to be expected for executing them (see Figures 2.10A-B). We modeled this task using the same approach as in Simulation 3.1, with three states and two actions. Figures 2.10C-D shows the effect of action values (quantified as in Lau & Glimcher, 2008; see Appendix B) on the activity of one $S' \rightarrow A$ unit in our neural

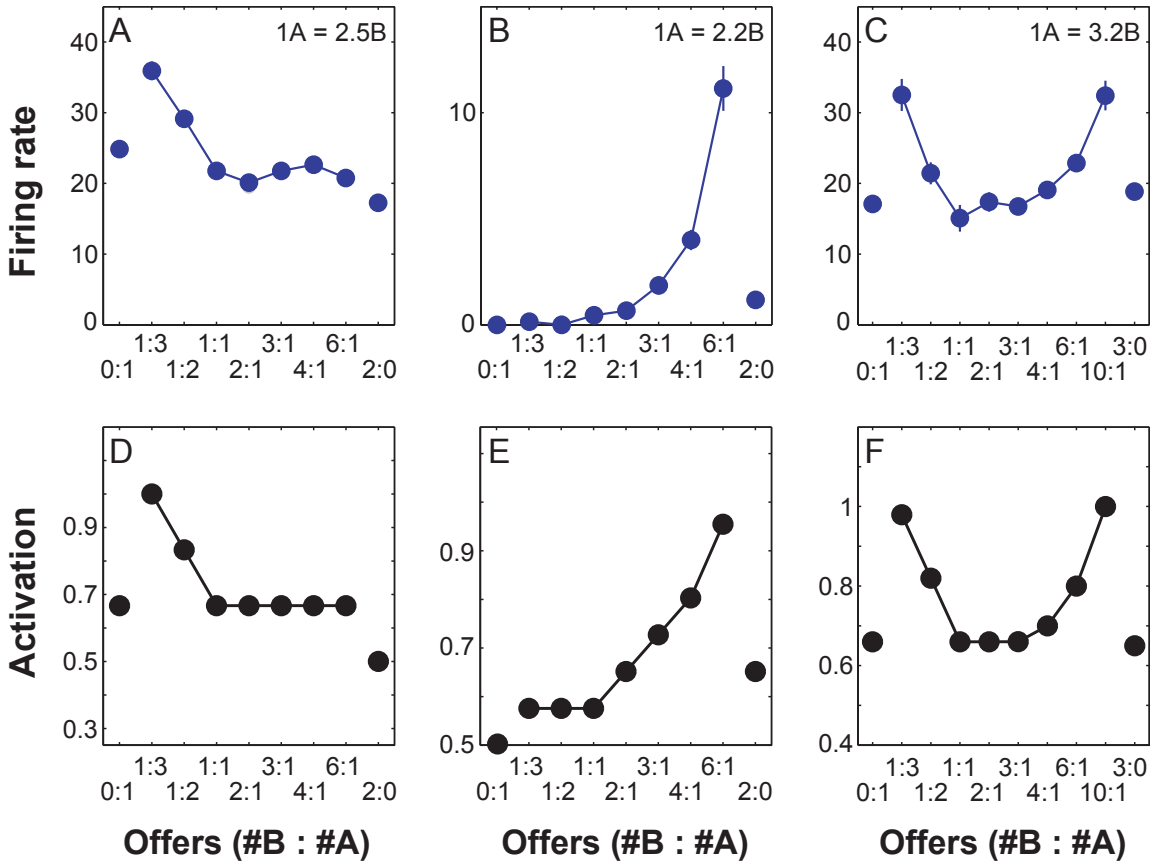

Figure 2.9: **A–C** Data from Padoa-Schioppa and Assad (2006, p. 225). Copyright 2006 by Macmillan Publishers Ltd: Nature. **D–E**. Results from Simulation 3.1.

network model. Like that of the neurons in the empirical study, this unit's activity varies with the expected value of one action but is insensitive to the value of the opposing action.

Simulation 3.3. Sequence planning. In addressing multistep decision making, our model posits separate policy, action, state, and reward representations for each plan step (see Figure 2.2). If this is a valid picture of the mechanisms underlying goal-directed decision making, step-specific representations should be evident in the relevant neural structures. Evidence in support of this comes from a number of studies focusing on action representations, in which neurons have been reported to code conjunctively for specific actions and their positions within a planned sequence (Barone & Joseph, 1989; Botvinick & Plaut, 2009; Inoue & Mikami, 2006;

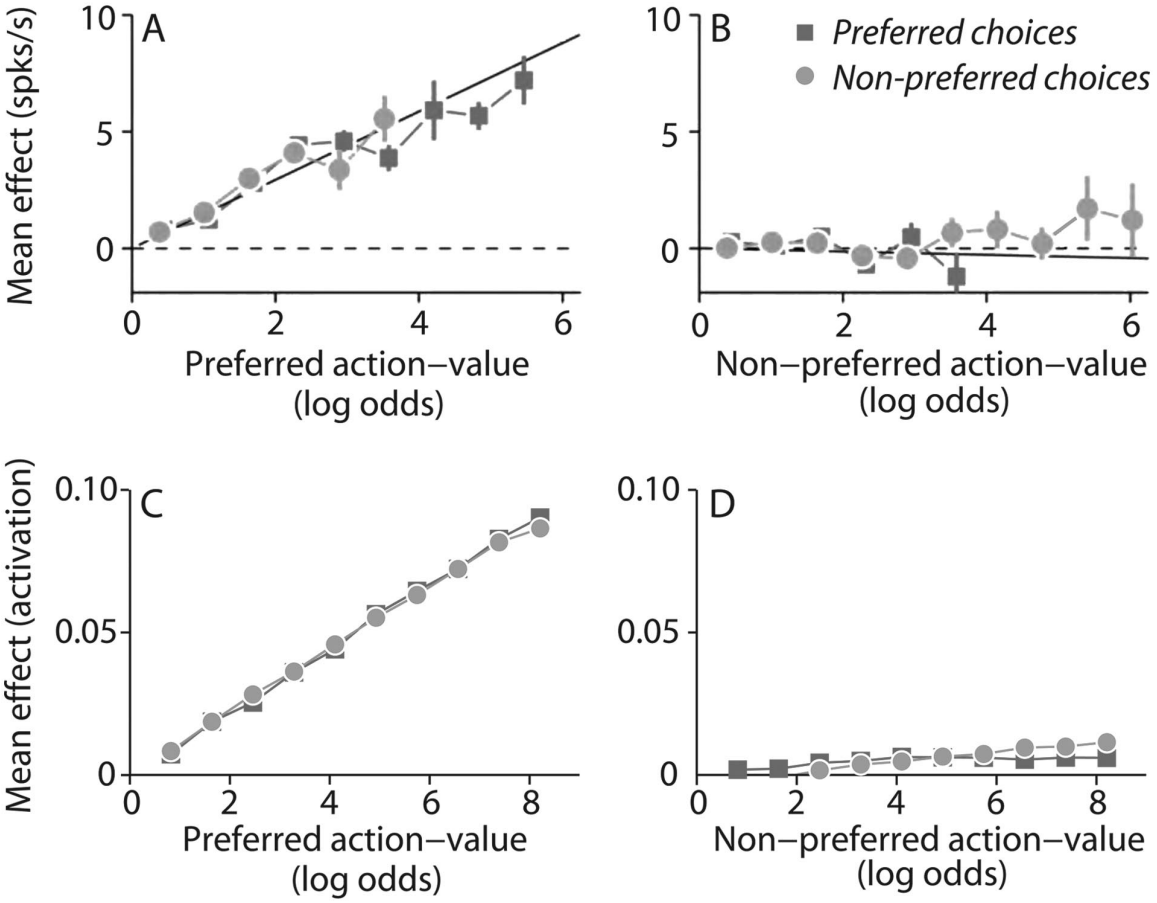

Figure 2.10: A–B. Data from Lau and Glimcher (2008). Reprinted from *Neuron*, 58, B. Lau and P. W. Glimcher, "Value Representations in the Primate Striatum During Matching Behavior," p. 457, Copyright 2008 by Elsevier. Preferred action refers to the action (saccade) whose execution preferentially excites the index neuron. Action value was quantified in terms of the impact of objective reward quantities (volume of water) on choice probability. **C–D**. Results of Simulation 3.2.

Ninokura, Mushiake, & Tanji, 2004). Such studies have also revealed important information about the timing of activation in such neurons, which may be important for evaluating the validity of our model of sequential decision making.

To focus on one particularly rich example, Mushiake et al. (2006) reported an experiment in which monkeys were presented with a maze display, indicating a goal location, as shown in Figure 2.11A. Shortly thereafter, a set of additional barriers was added to the maze, as shown in the figure. The animal's task was to navigate from the center of the maze to the goal location. The researchers found, recording in DLPFC, that many neurons coded for specific directions of movement within the maze, showing selectivity also for the ordinal position of the movement (first, second, or third in the solution sequence; see Figure 2.11B). These neurons became active before the onset of the first action, consistent with a role in planning. A critical additional finding was that neurons coding for successive actions became active at around the same time (see Figure 2.11C), suggesting that planning of the three required movements occurred more or less in parallel.

In order to simulate these results, we implemented a three-stage model. Considering that the action units in our neural network model convey the probability $p(a|s)$, and thus carry the same information as the A variables in our graphical model, we performed this simulation using the graphical model implementation for convenience. The state space included the set of occupiable positions in the maze, with available actions including movement in the four cardinal directions. The transition function dictated that movement into a barrier yielded no change in position, and reward was associated with the single goal location ($p(\hat{r}) = 0.7$; elsewhere 0.05).

The central result of the simulation is shown in Figure 2.11D. This shows the evolution of the action posteriors for the first, second, and third steps in the plan, as they converge to the correct plan $up \rightarrow left \rightarrow up$. Of course, our model includes

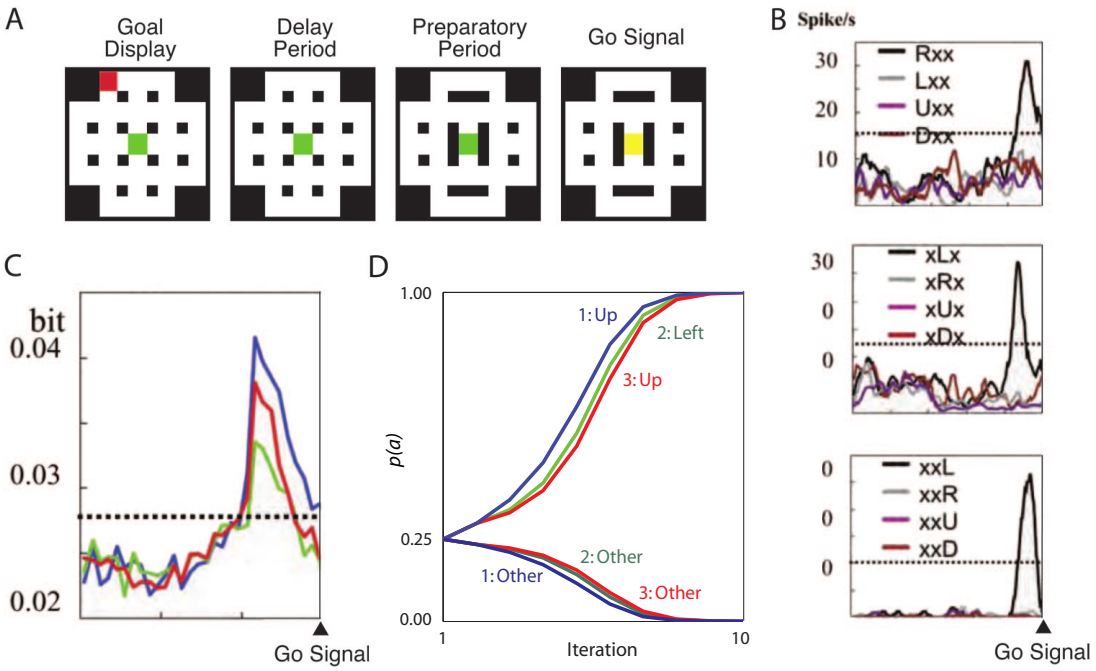

Figure 2.11: **A**. Example displays from Mushiake et al. (2006, p. 663), showing the sequential presentation of goal and barrier locations. Reprinted from *Neuron*, 50, H. Mushiake, N. Saito, K. Sakamoto, Y. Itoyama, and J. Tanji, "Activity in Lateral Prefrontal Cortex Reflects Multiple Steps of Future Events in Action Plans," p. 633, Copyright 2006 by Elsevier. **B**. Response profiles of three dorsolateral prefrontal neurons studied by Mushiake et al. (2006, p. 635). The arrowhead on the x-axis indicates the onset of the visual signal cuing the animal to begin navigating the maze. The top panel shows a neuron selective for rightward movement on the first step; the middle panel shows a neuron selective for leftward movement on the second step; and the lower panel shows a neuron selective for leftward movement on the third step. **C**. Data from Mushiake et al. (2006, p. 636) showing simultaneous emergence, over a population of prefrontal neurons, of information concerning first (blue), second (green), and third (red) actions during planning. **D**. Results of Simulation 3.3. Numbers indicate the relevant action variable (as though moving from left to right in the architecture shown in Figure 2.2D). "Other" indicates actions down, right, and left on Steps 1 and 3 and actions up, down, and right on Step 2. Note that Mushiake et al. (2006) also presented data relating to plan execution, which are omitted here.

conjunctive representations of action and ordinal position and thus matches this aspect of the empirical data by design. What the figure shows, additionally, is that the decision processes at the three steps follow highly overlapping time courses, very much in line with the parallel activation observed in the Mushiake et al. (2006) study.

It is revealing to compare these results with those from Simulations 2.1 and 2.2 (see Figures 2.6E and 2.6F). The present simulation shows that within our model, as in Mushiake et al.'s (2006) study, planning at successive steps can be highly parallel in time. Figure 2.6, in contrast, shows cases where planning is more asynchronous. In Figure 2.6F, the decision at the first step of a two-step plan emerges first. In Figure 2.6E the order is reversed, with the decision at the second step evolving faster. As this contrast indicates, although our model can be fit to the findings from Mushiake et al. (2006), the model more generally predicts that the relative timing of decision making across stages of a multistep plan will vary systematically with the specific set of outcome contingencies involved in the decision task.

Simulation 3.4. Evidence integration in simple incentive choice. Earlier, we compared our graphical model account with evidence integration models of perceptual decision making. We are now in a position to consider this parallel from a neuroscientific point of view. A range of studies have mapped the elements of the evidence-integration framework onto specific neural regions, in the context of specific perceptual tasks. The most extensive research has focused on the dot-motion paradigm reviewed earlier and diagrammed in Figure 2.4. Here, neurophysiological research has focused on localizing two critical functions. The first is the "integrator" itself, the area or areas in which information about visual motion accumulates over time, leading neural activity to approach or retreat from decision thresholds. Activity fitting with this description has been identified in

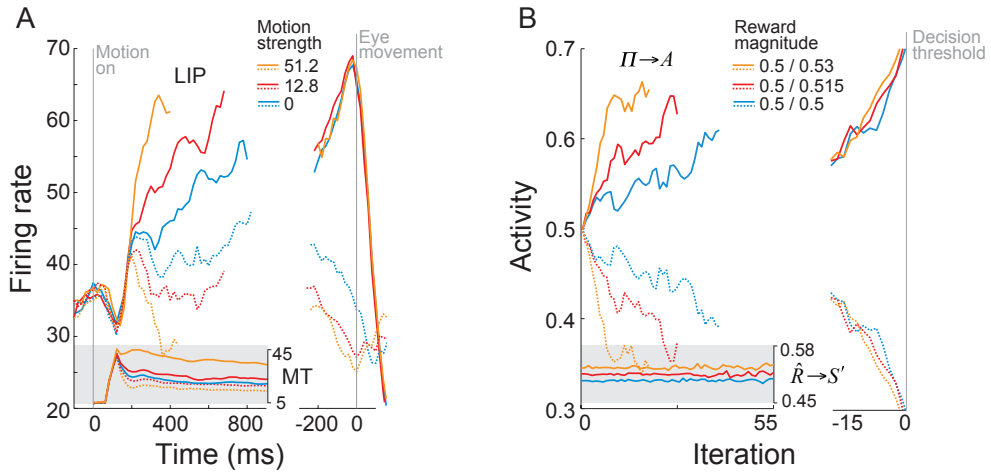

Figure 2.12: A: Representative findings from lateral intraparietal area (LIP) and middle temporal area (MT) during motion discrimination. From "The Neural Basis of Decision Making," by J. I. Gold & M. N. Shadlen, 2007, *Annual Review of Neuroscience*, 30, p. 548. Copyright 2007 by Annual Reviews, Inc. B: Results from Simulation 3.4. Upgoing data series in the main panel are for the unit representing the chosen policy; downgoing time series are for the unchosen policy. As in Panel A, red and yellow data series are based only on trials involving correct (reward-maximizing) responses.

lateral intraparietal area (LIP) as illustrated in Figure 2.12A, based on work by Gold and Shadlen (2007).

The second focus of neuroscientific work has been to identify the source of input to the integrator, that is, the source of the evidence feeding into the evidence-integration mechanism. Not surprisingly, in the dots task this has been tracked to cortical area MT, which has long been known to encode information concerning visual motion (see Gold & Shadlen, 2001). Unlike neurons in LIP, MT neurons show relatively stable tonic activity during viewing of dot-motion stimuli, consistent with the idea that they are coding for instantaneous information in the display, rather than integrating this information over time (see Figure 2.12A, inset).

Earlier, we highlighted the fact that the policy variable in our model behaves like an integrator. In this regard, our theory draws a direct analogy between the role of LIP in perceptual decision tasks and the role of DLPFC in goal-directed decision

making. The analogy is reinforced in Figure 2.12B, which shows the activity in the units coding for policy marginals in our neural network model (red in Figure 2.8) over a set of two-alternative decision problems varying in incentive disparity (see Appendix B for simulation methods).

If the role of DLPFC is analogous to that of LIP in perceptual decision making, then what area is analogous to MT? That is, what area provides the "evidence" that is integrated over time within DLPFC? In formal terms, we earlier identified this evidence with the likelihood $p(\hat{r}|\pi, s)$. In the setting of simple binary choice, where there is a one-to-one correspondence between actions and outcomes, note that this value is exactly equal to $p(\hat{r}|s')$. As a consequence, in simple choice, the "evidence" entering into the integration process corresponds to the activation of the units labeled $\hat{R} \rightarrow S'$ in our neural network model (purple in Figure 2.8). In Simulation 3.1, we compared the function of this set of units with that of neurons representing state value in OFC. The analogue to MT, according to our model, is therefore OFC. The analogy is elaborated in Figure 2.12B (inset), which shows activity in state-value units for the same choice problems used to generate the policy time courses above. Like MT, these units show stable tonic activity indicating the "strength of evidence" for one choice over the other.

2.8.2 Predictions

In these simulations, we have focused on cases where signals within the neural network have readily identifiable correlates in the current neuroscientific literature. Other aspects of the neural network model lead to further testable predictions. For example, the units labeled $S \rightarrow \hat{R}$ and shown in green in Figure 2.8 represent the probabilities of outcome states.[12] The model predicts that such representations

[12]Interestingly, these messages represent the probability of states conditional on the current policy distribution but not on $\hat{r} = 1$. As a result, the predictions represented here do not show the same "optimistic" bias as the marginal state probabilities discussed under Predictions following

should be identifiable within the brain and (less obviously) that sequence planning should activate neural representations of sequences of future states, with order-specific coding as has been demonstrated for actions (see Simulation 3.3). Some neuroscientific evidence consistent with prospective state coding was discussed earlier. With regard to representation of multiple future states during planning, suggestive evidence is provided by Saito et al. (2005), who showed that neurons in prefrontal cortex encode both immediate and final goal locations in parallel during planning in a maze navigation task. Having noted this, we acknowledge that other studies have uncovered representations of state that fit less tidily into the present account. Johnson and Redish (2007) observed hippocampal activation apparently coding for projected future positions during path planning. However, in contrast to the activation reported by Saito et al. (2005), these activations were activated serially in time rather than concurrently. In other recent work, Stalnaker, Calhoon, Ogawa, Roesch, and Schoenbaum (2010) reported neurons in dorsal striatum coding for action-outcome conjunctions. Such representations do not figure in our neural network model and therefore present a challenge to be examined in future work.[13]

A further prediction stems from the fact that our model posits separate representations of expected reward for each stage in a multistage plan. Given the parallels we have drawn to OFC and amygdala, this predicts that similar, step-specific reward representations should be identifiable in one or both of these regions during the planning of sequential actions. To our knowledge, neural activity

Simulations 1-2. Another set of messages, present when the model is expanded to encompass more than one step of action, does show the "optimism" effect. Thus, the framework predicts that it should be possible to find multiple representations of outcome probability, some of which are, and some of which are not, optimistic. Some evidence in favor of this kind of multiple coding is reported by Kool, Getz, and Botvinick (2013).

[13]One interesting possibility, which was intimated by the Stalnaker et al. (2010), is that these neurons are involved in representing the transition function. This is consistent with the finding, from this same study, that many striatal neurons coded in a tonic fashion for the specific response–outcome associations active during the current block of trials.

in these regions has not been studied in the setting of sequence planning (though see D. A. Simon & Daw, 2011b).

Finally, Simulation 3.2 leads to specific predictions concerning neural action-value representations. In previous work, such representations have generally been assumed to support model-free or habitual action selection (see, e.g., Samejima et al., 2005). Our model shows how action-value representations might arise during goal-directed decision making. Furthermore, our model suggests a close link between action-value and state-value representations, with the latter providing part of the basis for computing the former during the course of single decision-making episodes (for related proposals, see Hasselmo, 2005). This leads to the novel prediction that disruptions of neural state-value representations (e.g., in OFC) should disrupt action-value coding (e.g., in parietal cortex or striatum).

2.9 General discussion

In the present paper, we have advanced an account of goal-directed decision making. With a nod to David Marr, we have specified the theory at computational, algorithmic, and implementational levels. At the computational level, the proposal aligns with contemporary theories in vision, motor control, and other domains, which center on inverse inference within a generative model. In the present work, the generative model in question captures the way in which policies, actions, and states work together to generate rewards, and model inversion reveals the policy that best explains the occurrence of reward. The procedures involved in carrying out this inversion link the present account with current theories of perceptual decision making, which center on iterative evidence integration. Like such theories, the present one can be translated into neural terms, providing an account of how populations of neurons spanning relevant brain areas may work together to yield

goal-directed decisions. Across the algorithmic and implementational levels, the theory we have presented accounts for a range of behavioral and neurophysiological observations and gives rise to testable predictions. In this final section, we pan back to consider the relationship between the ideas we have presented and previous work, and enumerate some areas for further development.

2.9.1 Related work in machine learning and theoretical neuroscience

As intimated earlier, although the notion of reward-based decision making as inference has been little explored in psychology or neuroscience, versions of the idea have been in play for several decades within decision theory and machine learning. Initial proposals for how to solve decision problems through probabilistic inference in graphical models, including the idea of encoding reward as the posterior probability of a random utility variable, were put forth by G. F. Cooper (1988). Related ideas were presented by Shachter and Peot (1992), including the use of nodes that integrate information from multiple utility nodes. More recently, Attias (2003) and Verma and Rao (2006b) have used graphical models to solve shortest path problems, leveraging probabilistic representations of rewards, although not in a way that guarantees convergence to reward-maximizing plans. More closely related to the present research is work by Toussaint and Storkey (2006) employing the expectation-maximization algorithm, a technique with interesting but insufficiently explored relations to evidence-integration procedures (see also Dayan & Hinton, 1997; Furmston & Barber, 2009; Hoffman et al., 2009).

Although close in spirit, our framework does not fully parallel any of this previous work. Perhaps the most important difference is at the level of the research objective: Our aim in the present work has been to maximize not computational power but rather explanatory power, by engaging wherever possible with estab-

lished principles and findings in psychology and neuroscience. Our efforts to relate the present theory to accounts of perceptual decision making and to available functional neuroanatomic and neurophysiologic data are emblematic of this objective.

Within neuroscience, one recent line of work that has explored reward-based decision making from an inference-centered point of view is by Friston, Daunizeau, and Kiebel (2009). This work adopts the generative perspective, proposing that the brain is shaped through learning to minimize its own "surprise" by maximizing the accuracy of its predictions about external inputs. Action selection is then modeled by introducing the additional assumption that the brain is configured to predict the perceptual feedback that would be produced by adaptive actions. The objective of minimizing surprise is then met by selecting actions that assure the predicted inputs. Beyond its shared focus on inference within a generative model, this approach is somewhat different from the one we have taken. In the theory of Friston et al., the role of the central generative model is to predict observations (perceptual inputs), and the role of action is to realize those observations. The generative model at the center of our work is itself the substrate for action selection, accomplished through inverse inference from the fixed initial "observation" of reward. Interestingly, the model of Friston et al. (2009) deliberately eschews any explicit representation of reward; reward is encoded implicitly through the distributions that express the agent's predictions. Although such an implicit encoding may be computationally feasible (see Furmston & Barber, 2009), it does not square well with the neurophysiological data reviewed earlier (e.g., Padoa-Schioppa & Assad, 2006), which provide strong evidence for explicit neural representations of reward.

2.9.2 Spreading activation models of spatial navigation

One other area in which some work has been done on the neurocomputational basis of goal-directed decision making is spatial navigation. The predominant approach in such work is represented in studies by Schmajuk and colleagues (Schmajuk & Thieme, 1992; Voicu & Schmajuk, 2002) and subsequent simulations by Hasselmo and colleagues (Hasselmo, 2005; Koene & Hasselmo, 2005). Both sets of models assume a network of simple neuron-like processing elements representing environmental states or locations, which plays the role of the cognitive map. In Schmajuk's models, inputs representing incentive value activate rewarded locations, and activation spreads from these locations to adjacent ones until the frontier of activation reaches the agent's current location. This results in an activation map, from which actions can be selected through a hill-climbing procedure (for related work, see Bugmann, Taylor, & Denham, 1995; Gaussier, Revel, Banquet, & Babeau, 2002; Girard, Filliat, Meyer, Berthoz, & Guillot, 2005; Martinet, Passot, Fouque, Meyer, & Arleo, 2008; Muller, Stead, & Pach, 1996; Reid & Staddon, 1998). Hasselmo's models (Hasselmo, 2005; Koene & Hasselmo, 2005) follow the same general approach, but allow activation to spread "forward" from the agent's initial state (see also Smith et al., 2004). These models also explicitly represent actions and action-outcome relationships, permitting the models, at least in principle, to be applied beyond the domain of spatial navigation.

The framework we have put forth shares a definite family resemblance with such spreading-activation models. In particular, one can relate the propagation of activation within these networks to the message-passing operations within our neural network implementation. A relative strength of our model, once again, is that it offers an explicit formal characterization of the computations involved,[14]

[14]Hasselmo (2005) discussed parallels between his model and the policy iteration procedure in reinforcement learning. Some underlying links to policy iteration have also been considered

establishing a link between these computations and inference-based operations in other information-processing domains, as well as to normative and empirical accounts of perceptual decision making. Furthermore, by implementing goal-directed decision making in probabilistic terms, our models also naturally extend to settings involving uncertain outcomes and multiple sources of reward or cost, settings not generally addressed by spreading activation models.

2.9.3 Evidence-integration models of decision making

A key feature of our account is its incorporation of an iterative procedure transparently related to the sequential probability ratio test, an optimal procedure for sequential hypothesis testing. As we have emphasized, this aspect of our model links it closely with current theories of perceptual decision making, in particular those leveraging the drift-diffusion formalism. Our neural-network implementation reinforces this connection, based as it is on a recent effort to translate such decision-making theories into neural terms.

The success of drift-diffusion models in perceptual decision making and other domains has inspired several researchers to apply the same framework to reward-based decisions. In several cases, the proposal has been to import the drift-diffusion model en bloc, simply relabeling the inputs to the process as the utilities of choice objects (Krajbich et al., 2010; Rangel, 2008; Rustichini, 2008; Shadlen, 2008). The present work complements and extends such efforts in two ways. First, it furnishes an explicit statistical interpretation for evidence integration in the context of reward-based decision making. In the case of perceptual decision making, such an interpretation is ready to hand: The evidence-integration process is understood as an implementation of the sequential probability ratio test, with perceptual inputs

in planning-as-inference (see Toussaint & Storkey, 2006). Further exploration of this underlying formal connection would be of interest.

playing the role of the data, representations of stimulus identity playing the role of hypotheses, and a well-characterized likelihood function $p(data|hypothesis)$ linking the two (see Figure 2.4). In contrast, prior applications of the evidence-integration framework to reward-based decision making have not, to our knowledge, been associated with a corresponding statistical interpretation. The present work bridges this gap. In our framework, the fictive observation $\hat{r} = 1$ plays the role of the data; each hypothesis corresponds to the belief that the observation $\hat{r} = 1$ is explained by a particular policy; and the likelihood function is $p(\hat{r}|\pi, s)$.

In addition to providing this formal interpretation for evidence-integration models of reward-based decision, the present work generalizes the approach. Indeed, the standard drift-diffusion model can be seen as a limiting case of the present framework, which obtains in the setting of two-alternative forced choice with one-to-one, deterministic action-outcome contingencies (see Appendix A). The present account widens the scope of the evidence-integration paradigm to accommodate stochastic action-outcome contingencies and multistep planning.[15]

Alongside direct applications of the drift-diffusion model, several models have adapted the evidence-integration framework to reward-based decision in more elaborate and specialized ways. Such work includes the leaky competitive accumulator (LCA) model of Usher and colleagues (Bogacz, Usher, Zhang, & McClelland, 2007; Tsetsos, Usher, & Chater, 2010; Usher et al., 2008), the decision-by-sampling (DBS) framework of Stewart and colleagues (N. Stewart, 2009; N. Stewart et al., 2006), and decision field theory (DFT), as proposed by Busemeyer and colleagues (Busemeyer & Townsend, 1993; Busemeyer & Diederich, 2002). Our model has features in common with all three of these, given their shared use of sequen-

[15]The statistical interpretation offered above for the one-step scenario transfers to the multistep case: The likelihood in this instance, for the set of policy variables π_t at each stage t of the plan, is $p(\hat{r}_c|\pi_t; \pi_{\tau \neq t})$. The optimality property that obtains in the one-step case does not transfer. To our knowledge, optimal decision making (in the sense involved in the sequential probability ratio test) has not been studied in the setting of multistep planning. This strikes us as a fascinating area for future study, into which the present work may provide a portal.

tial sampling, along with integrator-like mechanisms. One important difference is that the LCA, DBS, and DFT models all focus heavily on multi-attribute decision making, where choice options are characterized along multiple feature dimensions. Extending the present framework to engage the multi-attribute case is an important area for future development. Elaborating the computational architecture to accommodate multiple feature dimensions is, in itself, quite straightforward, as demonstrated by related work in machine learning using factored state representations (see, e.g., Toussaint & Storkey, 2006). The key question for future work is whether introducing factored representations into the present framework gives rise to patterns seen in human multi-attribute choice (see Busemeyer & Diederich, 2002; Tsetsos et al., 2010).

2.9.4 Departures from rationality

A central preoccupation in work with the LCA, DBS, DFT, and related models has been with putative departures from rationality, as defined by classical expected-value theory. The ability of such models to account for biases and heuristic use in decision making may at first appear to reflect a fundamental difference in approach from the one we have pursued. It is, after all, true that our framing of the goal-directed decision making problem is normative in form, taking the maximization of expected reward (or subjective utility) as its objective. In this respect, the present framework aligns with a wide range of other work that adopts a normative approach to decision making (e.g. Anderson, 1990; Bogacz et al., 2006; Geisler, 2003; Niv et al., 2006). A particularly strong resonance is with work taking a normative perspective on action understanding (Bekkering et al., 2000; Csibra & Gergely, 2007; Gergely & Csibra, 2003), some of which has also adopted an explicitly probabilistic approach (C. L. Baker, Saxe, & Tenenbaum, 2009; Rao et al., 2007; Verma & Rao, 2006a).

Having said this, it is also important to note that our account presumes that decision making is rational only relative to the decision maker's internal model of the problem (see H. A. Simon, 1987). Throughout the present work we have assumed, for simplicity, that this model accurately captures the objective probabilities associated with action-outcome contingencies, and represents reward values in a simple linear fashion (see Equation 2.8). However, the framework naturally accommodates representations of contingency and reward that depart from this default case. In particular, the distribution $p(\hat{r}|s')$ could be assumed to have the asymmetric sigmoid form of the utility function posited by prospect theory (Kahneman & Tversky, 1979), and the distribution $p(s'|s,a)$ could be assumed to distort objective outcome probabilities as occurs in prospect theory's weighting function. Under these assumptions, the present model would inherit the ability of prospect theory to account for such phenomena as loss aversion and interactions between outcome probability and valence in determining risk attitude.

This approach of simply "plugging in" functions from prospect theory has precedents in the decision modeling literature (see, e.g., Tsetsos et al., 2010; Usher et al., 2008) and could arguably be justified in our model–independently of the behavioral phenomena to be explained– based on neurophysiological data identifying neural response profiles resembling those functions (Fox & Poldrack, 2008; Hsu, Krajbich, Zhao, & Camerer, 2009). However, it would perhaps be more satisfying if the relevant functional forms could be understood as emerging naturally through learning rather than simply stipulated. The psychology, neuroscience, and economics literatures suggest some interesting possibilities in this regard, which may have further relevance to departures from strict rationality, as we discuss next.

2.9.5 Learning

The work we have presented, like most work on goal-directed decision making, has focused on the question of how decisions are made in the presence of an established internal model of the task domain. A truly comprehensive theory would have to include an account of how that internal model arises (Glascher et al., 2010; Green et al., 2010). The theory we have presented is, we believe, quite amenable to such an extension. Indeed, formal methods for learning in graphical models are well developed (Jordan, 1998), and analogies have already been made between the relevant algorithms and learning processes in humans (Chater, Tenenbaum, & Yuille, 2006; Gopnik & Schulz, 2007).

From a purely formal perspective, the most obvious approach to learning in our graphical model would be to base the CPD at each node on event counts, as these provide maximum-likelihood estimates of the true distributions (see Koller & Friedman, 2009). Thus, for example, if an action a in situation s can lead to two outcomes s'_1 and s'_2, the transition probabilities could be estimated as the count ratios $N_{s,a,s'\,1}/N_{s,a}$ and $N_{s,a,s'\,2}/N_{s,a}$. Attias (2003) has demonstrated the feasibility of combining this form of learning with concurrent inference-based decision making.[16]

However, an appealing alternative approach to learning is suggested by recent work in psychology and economics. As mentioned earlier, in the DBS model of Stewart and colleagues (N. Stewart, 2009; N. Stewart et al., 2006), continuous

[16]One interesting issue that arises when learning and action selection are interleaved is that action choices can affect what is learned. The learner can thus engage in active learning, in which actions are taken to maximize information gain (Castro et al., 2008; Kruschke, 2008; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Another setting where action can be motivated by the value of information is partial observability, where the state of the environment is not entirely available to immediate perception (Behrens, Woolrich, Walton, & Rushworth, 2007; Howard, 1966). The models we presented assumed full state observability. However, Toussaint, Charlin, and Poupart (2008) and Furmston and Barber (2009) have described how similar principles can be applied to partially observable problems. Evaluating the fit between the resulting account and human behavior in analogous task contexts presents an interesting challenge.

quantities such as utilities and outcome probabilities arise out of a tournament-like process. To compute the utility of a particular item, for example, one compares the item against a series of reference items, sampled from memory based on their frequency of occurrence in past experience. The proportion of comparisons in which the index item is judged preferable to the reference item becomes the scalar representation of the index item's utility (for a related proposal in economics, see Kornienko, 2010; van Praag, 1968).

It seems inviting to consider how this tournament-based approach could be integrated into our framework, for several reasons. First, the approach provides a natural interpretation for our binary representation of reward: $p(\hat{r}|s')$ could be interpreted as the proportion of "victories" enjoyed by s' in the relevant tournament. Note that here, because $p(\hat{r}|s')$ depends on the set of states against which s' is compared, the value $p(\hat{r}|s')$ acquires the property of range adaptation (see Kornienko, 2010; N. Stewart et al., 2006). This is appealing from a neuroscientific perspective, because recent studies have demonstrated range adaptation in neural representations of reward (Kobayashi, de Carvalho, & Schultz, 2010; Padoa-Schioppa, 2009). Range adaptation is also appealing from the perspective of behavioral economics; as Stewart and colleagues (N. Stewart, 2009; N. Stewart et al., 2006) have detailed, adaptive coding provides an explanation for the emergence of both the utility and weighting functions from prospect theory. Furthermore, because adaptive coding makes the representation of utility (and other quantities) context dependent, it gives rise to a number of phenomena that present a challenge for standard expected utility (e.g., similarity, compromise, and attraction effects; see Busemeyer & Diederich, 2002; Kornienko, 2010; N. Stewart, 2009; Tsetsos et al., 2010). Evaluating the potential role for these considerations within the present theory is an important target for future research.

One further germane aspect of learning arises from cognitive research on planning and problem solving. Such work highlights the importance not only of learning about state transitions but also of learning to represent states themselves (Chase & Simon, 1973). The principles underlying such representational or state-space learning (see Gershman & Niv, 2010) are still poorly understood, and incorporating this aspect of learning into models of goal-directed behavior stands as an important long-range challenge.

2.9.6 Capacity limitations, heuristics, and problem representation

Cognitive research on planning highlights another characteristic of human goal-directed decision making that we have not considered thus far: its rather strict capacity limitations. A central take-home message from prior research is that human planners are incapable of reasoning precisely about complex problems, due largely to limitations on working memory capacity, and thus resort to a number of simpler problem-solving heuristics (Newell & Simon, 1972; Novick & Bassok, 2005; Unterrainer & Owen, 2006). One way of understanding such capacity limitations within the present framework would be in terms of a limit on the number of future steps of action that can be concurrently represented. This limit could be a property of the underlying processing architecture (i.e., an inherent limit on the number of segments within the structure posited in Figure 2.2, bottom). A structural limit of this flavor has been independently proposed in work on multitasking (Koechlin & Hyafil, 2007), and inherent limits on depth of search have also been heavily discussed in work on decision making in economic games (see Camerer, 2003). One reason such a depth limit might make functional sense in the problem settings we have considered relates to the impact of noise. In any model of planning where random variability plays a role, adding a stage to the planning depth will inject more noise into the planning process. Given that decision making at each stage of

a sequential plan is dependent on other stages, it seems likely that the impact of noise will grow in a nonlinear fashion as planning depth increases, making deep search intractable (see Daw et al., 2005).

Another perspective on capacity limitations that may have relevance within the present account comes from work suggesting that human cognition does not leverage probability distributions in their entirety, but rather only samples from such distributions. Under this approach, capacity limitations in information processing are understood to arise from limitations on the number of samples that can be made during a single decision-making event. This general idea, which leverages machine learning algorithms for approximate inference, has been applied to magnitude estimation (Vul & Pashler, 2008) and sentence processing (Levy, Reali, & Griffiths, 2008).[17] The notion of sampling has already entered into the present work, both in connection with random utility and in our neural network implementation. Evaluating the more general relevance of the sampling hypothesis to goal-directed decision making is an inviting area for further theory development.

As noted earlier, the cognitive planning literature not only documents capacity limitations but goes on to characterize the strategies used by human planners to mitigate or cope with those limitations (Newell & Simon, 1972). Some of the relevant ideas are readily transposed into the present theory. For example, one method of coping with limited or costly processing capacity is to simplify problem representations. This has been proposed, in particular, to explain intransitivities in multi-attribute choice (Kalenscher, Tobler, Huijbers, Daselaar, & Pennartz, 2010; Shah & Oppenheimer, 2008; Tversky, 1969, 1972). Such a strategy would enter into the present theory at the level of the underlying generative model, as this

[17] One interesting aspect of sampling-based techniques for approximate inference in graphical models is that they are inherently serial in operation, in some cases involving "particles" that traverse the graphical structure in a wavelike fashion (see Koller & Friedman, 2009). Exploring the application of such procedures in the present modeling context may thus allow contact with evidence that planning in challenging circumstances can take a serial form, often involving serial subgoaling.

model is in essence a representation of the decision problem. Strategic selection of this model might thus be considered part of an adaptive procedure for goal-directed decision making. Accounting for this model-specification stage presents an important challenge for development of the present theory, as for any theory of goal-directed decision making or planning.

Another planning strategy that helps in overcoming capacity limitations is referred to as hill climbing. Here, a goal is pursued by selecting actions that reduce the discrepancy between the present state and the goal state (Newell & Simon, 1972). Within the present model, this strategy would correspond to imposing a special or auxiliary reward function, which values states in proportion to their similarity to a goal state. Of course, to make good on this proposal, it would be necessary to supplement the present theory with an account of how reward functions might be strategically chosen. This is an issue that comes up in the field of hierarchical reinforcement learning, a field whose relevance to psychology and neuroscience we have recently considered elsewhere (Botvinick, Niv, & Barto, 2009; Ribas-Fernandes et al., 2011).

Indeed, one further strategy for mitigating the impact of limited capacity on goal-directed decision making, both in machine learning and in human cognition, is through hierarchical representation. Hierarchical action representations simplify the planning problem, allowing plans to reach deeper into the future through efficient coding of action subsequences (see Botvinick, Niv, & Barto, 2009). As discussed in Simulation 2.3, Ostlund et al. (2009) reported devaluation behavior that they interpreted as direct evidence for "chunked" action representations in goal-directed behavior. Although, in our earlier discussion, we suggested the relevant data might be explained without chunking, it seems certain that, in the general case, hierarchical action representations do play a role in goal-directed decision making. In recent work, Toussaint et al. (2008) have provided an initial demon-

stration of how hierarchical representation can be integrated with inference-based planning. It would be interesting to consider how the relevant computational issues relate to recent findings suggesting that prefrontal cortex houses a topographically organized hierarchy of action representations (Badre, 2008).

Human capacity limitations in planning, as well as the strategies and heuristics used to cope with them, have of course been a central concern in production-system models including ACT-R and SOAR (Anderson et al., 2004; Laird, Newell, & Rosenbloom, 1987). Such models stand in a complex relationship to models that approach goal-directed decision making from a reinforcement learning perspective, as recently discussed by Dayan (2009). One difference is in the way the underlying problem is typically framed. Production system models, following the tradition in problem-solving research, have tended to focus on tasks defined by an explicit a priori goal. In work inspired by reinforcement learning, including the work we have presented here, specific goal states do not figure at all in the formulation of the computational problem, which focuses instead on the generic goal of reward maximization. Recent versions of SOAR and ACT-R have begun to incorporate representations of reward into their accounts of action selection (see Anderson et al., 2004; Nason & Laird, 2005). However, in both cases the role of such representations appears to align more with the action values found in model-free reinforcement learning than with the freestanding reward function that is central to model-based or goal-directed action. Of course, this is not to say that production system models could not implement goal-directed choice procedures. Indeed, many ACT-R models contain action-outcome information in declarative memory, and recent work has also used declarative memory for rewards to guide action selection (see T. C. Stewart, West, & Lebiere, 2009). The challenge for production system models lies not in any restriction on their representational capacities but instead in their very flexibility. Such models could, in principle, implement any of

a range of procedures for goal-directed decision making; the architectures, in and of themselves, do not furnish a specific theory. Nonetheless, because production system models and, in particular, ACT-R, take detailed account of basic cognitive faculties (perhaps most importantly the dynamics of memory), we believe they may offer a useful context in which to compare theories of goal-directed decision making, including the one we have advanced here.

2.9.7 Relations with habitual action selection

In the present work, we have modeled goal-directed decision making in isolation, but as recent work has emphasized, human and animal behavior also rests upon habitual action selection, supported by different computational and neural mechanisms. A final important area for further development of the current account involves the question of how goal-directed decision making mechanisms interface with the habit system (Botvinick & Plaut, 2006; R. P. Cooper & Shallice, 2006; Coutureau & Killcross, 2003; Daw et al., 2005; Killcross & Coutureau, 2003). One way to model the role of habits in the present framework might be as additional inputs to policy variables, biasing policy selection toward habitual configurations. Another potential point of contact between goal-directed and habit mechanisms might also be at the planning horizon: Rather than encoding immediate reward at the final step of a multistep plan, it might make more sense to represent a cached "reward-to-go" value, a central element in model-free temporal-difference learning algorithms (see Sutton & Barto, 1998). Capping off explicit prospective "rollouts" with value representations of this kind has become standard in recent machine learning models of forward planning in partially observable domains (see Ross, Pineau, Paquet, & Chaib-Draa, 2008). Whether an application of these ideas within the present framework would align with available behavioral and neural evidence concerning the goal/habit interface will be an interesting question to pursue.

2.9.8 Conclusion

Despite a veritable explosion in computational work addressing habitual action selection, inspired largely by theories linking dopamine with temporal-difference learning, relatively little work has been done to specify the computational principles involved in goal-directed decision making. The present work contributes toward rectifying this imbalance. In addition to adopting the view that goal-directed decision making can be viewed in the terms provided by model-based reinforcement learning, our proposal seeks to account for such decision making in terms that figure equally in other domains of neural information processing, including other types of decision making, motor control, perception, and beyond. By portraying goal-directed decision making as probabilistic inference, the work we have presented fits into a broad movement within both psychology and neuroscience, which sees inference as providing a lingua franca, applicable across content domains as well as across computational, algorithmic, and implementational levels of description (Chater & Oaksford, 2008; Doya, Ishii, Pouget, & Rao, 2006; Jones & Love, 2011).

Given the early stage of computational research on goal-directed decision making, the most important contribution of the present work is simply to chart out one sector in the space of possible computational approaches. By performing this role, we hope, the work will, at the very least, provide a useful stepping-stone toward further computational and empirical research in this important domain.

2.10 Appendix A: Formal analysis

The main text introduced an iterative procedure for solving finite horizon Markov decision problems within graphical models of the kind displayed in Figure 2.2. Here we provide formal proofs of monotonicity and convergence (based on

Botvinick & An, 2009), which guarantee that the algorithm will converge to an optimal policy. To recap, the procedure is as follows: (a) Initialize the policy nodes with any set of nondeterministic priors. (b) Treating the initial state and $\hat{R}_c$ as observed variables, with $\hat{r}_c = 1$, use the junction tree algorithm or a comparable algorithm to infer the posterior distributions over all policy nodes. (c) Set the prior distributions over the policy nodes to the values (posteriors) obtained in step b. (d) Go to step b. The proofs follow:

2.10.1 Monotonicity

We show first that, at each policy node, the probability associated with the optimal policy will rise on every iteration. Define $\pi^\star$ as follows:

$$p(\hat{r}_c|\pi^\star, \pi^+) > p(\hat{r}_c|\pi', \pi^+) \forall \pi' \neq \pi^\star, \tag{2.9}$$

where π^+ is the current set of probability distributions at all policy nodes at all subsequent steps within the plan (i.e., to the right within the model architecture. Note that we assume here, for simplicity, that there is a unique optimal policy at each step.) The objective is to establish that

$$p(\pi_n^\star) > p(\pi_{n-1}^\star), \tag{2.10}$$

where n indexes processing iterations. The evidence integration procedure stipulates that

$$p(\pi_n) > p(\pi_{n-1}|\hat{r}_c), \tag{2.11}$$

where π represents any value (i.e., policy) of the decision node being considered. Substituting this into 2.10 gives

$$p(\pi^{\star}_{n-1}|\hat{r}_c) > p(\pi^{\star}_{n-1}). \tag{2.12}$$

From this point on the focus is on a single iteration, which permits us to omit the relevant subscripts. Applying Bayes' law to 2.12 yields

$$\frac{p(\hat{r}_c|\pi^{\star})p(\pi^{\star})}{\sum_{\pi} p(\hat{r}_c|\pi)p(\pi)} > p(\pi^{\star}). \tag{2.13}$$

Canceling and bringing the denominator up, this becomes

$$p(\hat{r}_c|\pi^{\star}) > \sum_{\pi} p(\hat{r}_c|\pi)p(\pi). \tag{2.14}$$

Rewriting the left-hand side, we obtain

$$\sum_{\pi} p(\hat{r}_c|\pi^{\star})p(\pi) > \sum_{\pi} p(\hat{r}_c|\pi)p(\pi). \tag{2.15}$$

Subtracting and further rearranging,

$$\sum_{\pi} [p(\hat{r}_c|\pi^{\star}) - p(\hat{r}_c|\pi)]p(\pi) > 0, \tag{2.16}$$

$$[p(\hat{r}_c|\pi^{\star}) - p(\hat{r}_c|\pi^{\star})]p(\pi^{\star}) + \sum_{\pi' \neq \pi^{\star}} [p(\hat{r}_c|\pi^{\star}) - p(\hat{r}_c|\pi')]p(\pi') > 0, \tag{2.17}$$

$$\sum_{\pi' \neq \pi^{\star}} [p(\hat{r}_c|\pi^{\star}) - p(\hat{r}_c|\pi')]p(\pi') > 0. \tag{2.18}$$

Note that this last inequality (2.18) follows from the definition of $\pi^{\star}$.

Remark. Of course, the identity of $\pi^\star$ depends on π^+. In particular, the policy $\pi^\star$ will only be part of a globally optimal plan if the set of choices π^+ is optimal. Fortunately, this requirement is guaranteed to be satisfied, as long as no upper bound is placed on the number of processing cycles. Recalling that we are considering only finite-horizon problems, note that for policies leading to states with no successors, π^+ is empty. Thus, $\pi^\star$ at the relevant policy nodes is fixed and is guaranteed to be part of the optimal policy. The proof above shows that $\pi^\star$ will continuously rise. Once it reaches a maximum, $\pi^\star$ at immediately preceding decisions will perforce fit with the globally optimal policy. The process works backward, in the fashion of backward induction.

2.10.2 Convergence

Continuing with the same notation, we show now that

$$\lim_{n\to\infty} p_n(\pi^\star|\hat{r}_c) = 1. \tag{2.19}$$

Note that, if we apply Bayes' law recursively,

$$p_n(\pi^\star|\hat{r}_c) = \frac{p(\hat{r}_c|\pi^\star)p_n(\pi^\star)}{p_n(\hat{r}_c)} = \frac{p(\hat{r}_c|\pi^\star)^2 p_{n-1}(\pi^\star)}{p_n(\hat{r}_c)p_{n-1}(\hat{r}_c)} = \frac{p(\hat{r}_c|\pi^\star)^3 p_{n-2}(\pi^\star)}{p_n(\hat{r}_c)p_{n-1}(\hat{r}_c)p_{n-2}(\hat{r}_c)} \cdots \tag{2.20}$$

Thus,

$$p_n(\pi^\star|\hat{r}_c) = \frac{p(\hat{r}_c|\pi^\star)^n p_1(\pi^\star)}{\prod_{m=1}^{n} p_m(\hat{r}_c)}. \tag{2.21}$$

Therefore, what we wish to prove is

$$\frac{p(\hat{r}_c|\pi^\star)^\infty p_1(\pi^\star)}{\prod_{n=1}^{\infty} p_n(\hat{r}_c)} = 1, \tag{2.22}$$

or, rearranging,

$$\prod_{n=1}^{\infty} \frac{p_n(\hat{r}_c)}{p_n(\hat{r}_c|\pi^\star)} = p_1(\pi^\star). \tag{2.23}$$

Note that, given the stipulated relationship between $p(\pi)$ on each processing iteration and $p(\pi|\hat{r}_c)$ on the previous iteration,

$$\begin{aligned} p_n(\hat{r}_c) &= \sum_{\pi} p(\hat{r}_c|\pi) p_n(\pi) = \sum_{\pi} p(\hat{r}_c|\pi) p_{n-1}(\pi|\hat{r}_c) \\ &= \frac{\sum_\pi p(\hat{r}_c|\pi)^2 p_{n-1}(\pi)}{p_{n-1}(\hat{r}_c)} = \frac{\sum_\pi p(\hat{r}_c|\pi)^3 p_{n-1}(\pi)}{p_{n-1}(\hat{r}_c) p_{n-2}(\hat{r}_c)} = \frac{\sum_\pi p(\hat{r}_c|\pi)^4 p_{n-1}(\pi)}{p_{n-1}(\hat{r}_c) p_{n-2}(\hat{r}_c) p_{n-3}(\hat{r}_c)} \cdots \end{aligned} \tag{2.24}$$

With this in mind, we can rewrite the left-hand side product in 2.23 as follows:

$$\frac{p_1(\hat{r}_c)}{p(\hat{r}_c|\pi^\star)} \cdot \frac{\sum_\pi p(\hat{r}_c|\pi)^2 p_1(\pi)}{p(\hat{r}_c|\pi^\star) p_1(\hat{r}_c)} \cdot \frac{\sum_\pi p(\hat{r}_c|\pi)^3 p_1(\pi)}{p(\hat{r}_c|\pi^\star) p_1(\hat{r}_c) p_2(\hat{r}_c)} \cdot \frac{\sum_\pi p(\hat{r}_c|\pi)^4 p_1(\pi)}{p(\hat{r}_c|\pi^\star) p_1(\hat{r}_c) p_2(\hat{r}_c) p_3(\hat{r}_c)} \cdots \tag{2.25}$$

Note that, given A16, the numerator in each factor of 2.25 cancels with the denominator in the subsequent factor, leaving only $p(\hat{r}_c|\pi^\star)$ in that denominator. The expression can thus be rewritten as

$$\frac{1}{p(\hat{r}_c|\pi^\star)} \cdot \frac{1}{p(\hat{r}_c|\pi^\star)} \cdot \frac{1}{p(\hat{r}_c|\pi^\star)} \cdot \frac{\sum_\pi p(\hat{r}_c|\pi)^4 p_1(\pi)}{p(\hat{r}_c|\pi^\star)} \cdots = \sum_{\pi} \frac{p(\hat{r}_c|\pi)^\infty}{p(\hat{r}_c|\pi^\star)} p_1(\pi). \tag{2.26}$$

The objective is then to show that the above equals $p_1(\pi^\star)$. It proceeds directly from the definition of $\pi^\star$ that, for all π other than $\pi^\star$,

$$\frac{p(\hat{r}_c|\pi)}{p(\hat{r}_c|\pi^\star)} < 1. \tag{2.27}$$

Thus, all but one of the terms in the sum above approach zero, and the remaining term equals $p_1(\pi^\star)$. Thus,

$$\lim_{n\to\infty} \sum_{\pi} \frac{p(\hat{r}_c|\pi)^n}{p(\hat{r}_c|\pi^\star)^n} p_1(\pi) = p_1(\pi^\star) \blacksquare \tag{2.28}$$

2.10.3 Convergence under random utility

We show here that the algorithm will also converge to the optimal policy under random utility. We focus on the single-step model, but the proof can be extended to the multistep case. As in the main text, we assume that the distribution of the variable R depends jointly on S' and on a vector-valued random variable Z, whose elements are independent and identically distributed. Z is assumed to be sampled upon each iteration of the evidence-integration procedure described above and in the main text. Define

$$\pi^\star := \text{argmax}\, E_Z[p(\hat{r}|\pi, Z)], \tag{2.29}$$

where E indicates expectation, and conditioning on the initial state s is implicit. Adopting this definition, the last expression in 2.26 becomes (replacing $\hat{r}_c$ with $\hat{r}$)

$$\sum_{\pi} \left(\prod_{m=1}^{\infty} \frac{p(\hat{r}|\pi, z_m) p_1(\pi)}{p(\hat{r}|\pi^\star, z_m)} \right) = p_1(\pi^\star). \tag{2.30}$$

Given the present definition of $\pi^\star$, 2.27 translates to

$$E_Z \left[\frac{p_n(\hat{r}|\pi \neq \pi^\star, z_n)}{p_n(\hat{r}|\pi^\star, z_n)} \right] < 1. \tag{2.31}$$

The expected value for the product in 2.30 is equal to the product of the expected values for the individual factors (iterations) indexed by m. Given 2.31, the latter product goes to zero as m goes to infinity for every $\pi \neq \pi^\star$. Thus, the expected value

of the left-hand side in A22 must converge to $p_1(\pi^\star)$. It can be easily shown that the variance of that same expression goes to zero as m goes to infinity, guaranteeing that $p_1(\pi^\star)$ will converge to one.

2.10.4 Relation to sequential-sampling models

The main text asserted a link between the present model and evidence-integration or sequential-sampling models of perceptual decision making, including random walk and drift-diffusion models, which in the case of binary choice are known to implement the sequential probability ratio test (for reviews, see Bogacz et al., 2006; Gold & Shadlen, 2007). We show here that in the same setting of simple binary choice, the model we have proposed displays precisely the same dynamics. The analog to the decision variable in the standard random walk model is the log policy posterior ratio

$$\log\left(\frac{p_n(\pi_A|\hat{r})}{p_n(\pi_B|\hat{r})}\right), \tag{2.32}$$

where π_A and π_B are the two response options (policy values); as before, n is the iteration; and $\hat{r}$ is shorthand for $\hat{r} = 1$. It is easily shown that in the absence of noise, this value grows linearly with a step size equal to the log likelihood ratio given the evidence $\hat{r} = 1$. The increment in the decision variable on each time step is

$$\log\left(\frac{p_n(\pi_A|\hat{r})}{p_n(\pi_B|\hat{r})}\right) - \log\left(\frac{p_{n-1}(\pi_A|\hat{r})}{p_{n-1}(\pi_B|\hat{r})}\right). \tag{2.33}$$

Absorbing the second term into the first, and applying Bayes' law along with the stipulation that $p_n(\pi) = p_{n-1}(\pi|\hat{r})$, this becomes

$$\log\left(\frac{p(\hat{r}|\pi_A)p_{n-1}(\pi_A|\hat{r})}{p(\hat{r}|\pi_B)p_{n-1}(\pi_B|\hat{r})} \cdot \frac{p_{n-1}(\pi_B|\hat{r})}{p_{n-1}(\pi_A|\hat{r})}\right), \tag{2.34}$$

which reduces to

$$\log\left(\frac{p(\hat{r}|\pi_A)}{p(\hat{r}|\pi_B)}\right). \tag{2.35}$$

This last expression is a constant, confirming that the decision variable grows linearly with a step size equal to the log likelihood ratio.

In our model, drift rate variability derives purely from internal sources of noise. In our algorithmic account, the source of noise is understood as deriving from intrinsic variability in the reward function, modeled using the noise variable Z. If the distribution of Z is chosen as in our simulations (see Simulation Procedures below), then drift-rate variability assumes a uniform Gaussian form, as in the drift diffusion model.

2.11 Appendix B: Simulation procedures

2.11.1 Graphical model

All simulations were run using the Matlab Bayes Net Toolbox (Murphy, 2001), combined with custom Matlab (MathWorks, Natick, MA) code (available for download from `http://www.princeton.edu/~matthewb`).

Simulations addressing single-step decisions employed the architecture from Figure 2.2C. Multistep tasks were modeled using the architecture from Figure 2.2D, extended to include the minimum number of actions required for the task simulated. States, actions, and policies were represented by discrete, multinomial variables. Policies were modeled using a set of nodes connected to each action variable, with each node representing the policy for a single state. Each policy-node value corresponded to a unique, deterministic policy for the relevant state.

As described earlier, reward was modeled using a binary variable connected to each state variable, as described in the main text.

For each task modeled, a scalar reward value $R(s')$ was assigned to each state s'. The resulting set of reward values was then scaled to fall between zero and one and was used to define the CPD for the reward variable, using the linear transformation specified in Equation 2.8. For simplicity, temporal discounting was not applied, but the framework could accommodate it through appropriate changes to the reward-variable CPD.

Each simulation involved imposing a set of values on one variable or set of variables and computing the posterior distribution over another variable or variables. In all cases, posterior probabilities were computed using the junction tree algorithm (see Jensen, 2001). Iterative inference was conducted as described in the main text and Appendix A. In all simulations, distributions for all policy variables were initialized as uniform.

As shown in Figure 2.4 (bottom right), Simulation 1.2 included an additional multivariate normal variable Z, with the same dimensionality as S' and covariance $0.3I$. On each iteration of inference, a value of this variable was sampled and treated as observed. The probability $p(\hat{r}|s', z)$ was then determined as $P(logit(\rho) + z)$, where P is the standard logistic function, z is the value of the element of Z with the same index as s', and ρ is a parameter $p(\hat{r}|s', 0)$, denoted $p(\hat{r}|s')$ in the main text.

2.11.2 Neural network

In translating our generative model into neural network form, we followed the approach outlined by Rao (2005). As noted in the main text, that work proposes how belief propagation might be implemented in biological neural networks, with message components encoded in the proportional firing rates of individual neurons. Following this idea, our neural network models simply implement standard belief

Table 2.1: Specification of belief propagation messages employed in Simulation 3

Message	Specification
$m(S \rightarrow A, S')$	$p(s)^a = \langle 1, 0, 0 \rangle$
$m(\Pi \rightarrow A)$	$p_n(\Pi) = p_{n-1}(\Pi \vert \hat{r}, s_0)$
$m(\hat{R} \rightarrow S')$	$p(\hat{r} \vert S')$
$m(S' \rightarrow A)$	$p(S' \vert s_0, A) m(\hat{R} \rightarrow S')$
$m(A \rightarrow \Pi)$	$p(A \vert \Pi, S_0) m(S' \rightarrow A)$
$m(A \rightarrow S')$	$\alpha p(A \vert s_0, \Pi)^T m(\Pi \rightarrow A)^c$
$m(S' \rightarrow \hat{R})$	$\alpha p(S' \vert s_0, A)^T m(A \rightarrow S')$
$m(\hat{R} \rightarrow \hat{R}_c)^d$	$\alpha p(\hat{R} \vert S')^T m(S' \rightarrow \hat{R})$

[a] Here and in subsequent entries, s_0 indicates the observed initial state, and the notation $p(X)$ denotes a probability vector with one component for each discrete value of X. [b] Here and in subsequent entries, $p(Y|X)$ and $p(Y|X, z)$ indicate a matrix with a row for each value of X and a column for each value of Y. [c] Here and elsewhere, α denotes a normalization factor. [d] As discussed in the main text (Simulation 3.1), this message derives from the multistep model.

propagation, with a unit for each message component. For a detailed introduction to the operations underlying belief propagation, see Pearl (1988). In what follows, we provide simulation details that cannot be gleaned from this source or from Rao (2005).

The network depicted in Figure 2.8 was tailored to the two-alternative forced choice task scenario. The messages transmitted within the model were computed as indicated in Table 2.1. The message $m(\Pi \rightarrow A)$ was initialized as $\langle 0.5, 0.5 \rangle$, and updated as

$$m(\Pi \rightarrow A) \leftarrow \alpha m(\Pi \rightarrow A) \odot m(A \rightarrow \Pi), \tag{2.36}$$

where $\odot$ denotes component-wise multiplication. The messages used in the multistep model can similarly be derived from the general purpose equations prescribed by belief propagation. See Pearl (1988) for details.

Rao (2005) presented an account of how stochasticity in neural firing might enter into a biological implementation of belief propagation. We took a simpler

approach, which gives rise to similar network behavior (as confirmed in head-to-head comparison simulations). In our modified approach, rather than treating the marginal probability (p) carried by each message component as an instantaneous firing rate and transmitting its exact value to down-stream units, we drew a sample from $Binomial(N, p)$, normalized its value by N (a free parameter, set to 200 in our simulations except where otherwise noted), and transmitted the result. The resulting quantities can be interpreted in two ways. First, they can be interpreted as the proportion of N time bins within a fixed interval during which an index neuron fired. Alternatively, they can be interpreted as representing the proportion of N neurons, with identical receptive fields, firing within a fixed time window. Further details for several specific simulations follow:

Replication of Simulation 1.2 (Figure 2.8). Here, a threshold of 0.75 was used along with a value of 75 for the N parameter, and the data presented represent response proportions from a set of 1,000 trials.

Simulation 3.2. The approach taken in this simulation was based closely on the procedure followed in Lau and Glimcher (2008). First, 1,000 simulation runs were performed for every pairing $\langle p(\hat{r}|outcome\ action\ 1), p(\hat{r}|outcome\ action\ 2)\rangle$ in which each value fell between 0.5 and 0.6, inclusive, and constituted a multiple of 0.01 (threshold parameter = 0.8, N = 200). From each trial, the action chosen and the activation of one $S' \rightarrow A$ unit at the time of threshold traversal were recorded. The action for which the $S' \rightarrow A$ unit coded was treated as the "preferred" action in the remaining analysis steps. Following Lau and Glimcher (2008), a logistic regression was conducted to relate $p(\hat{r}|action\ 1)$ and $p(\hat{r}|action\ 2)$ to choice probability $[\log p(action\ 1)/p(action\ 2)]$. This yielded a regression coefficient of 0.82, that is,

$$\log\left(\frac{p(action\ 1)}{p(action\ 2)}\right) = 0.82p(\hat{r}|action\ 1) - 0.82p(\hat{r}|action\ 2). \tag{2.37}$$

Based on this result, the scale used to represent action value ($\overline{AV}$) on the x-axis in Figure 2.10C-D was

$$\overline{AV} = 0.82(p(\hat{r}|action) - 0.5), \tag{2.38}$$

with the quantity 0.5 intended to represent a reference or status quo reward value. Again following Lau and Glimcher (2008), the values plotted on the y-axis in Figure 2.10 represent the residuals ϵ from two linear regressions:

$$\begin{aligned} \epsilon_{\overline{AV}_{\text{preferred}}} &= S' \rightarrow A\ unit\ activity - \beta_1(action\ selected) - \beta_2(\overline{AV}_{\text{non-preferred}}), \\ \epsilon_{\overline{AV}_{\text{non-preferred}}} &= S' \rightarrow A\ unit\ activity - \beta_1(action\ selected) - \beta_2(\overline{AV}_{\text{preferred}}). \end{aligned} \tag{2.39}$$

Simulation 3.4. The data presented in Figure 2.12B are based on 50 simulation trials for each reward–value pairing, using a response threshold of 0.8 and $N = 200$.

2.12 Addendum: Notes on belief propagation

The following notes on loopy belief propagation did not appear in the published paper. The notation of Pearl (1988, Chapter 4) is used. The exception is that upward messages are represented by ρ instead of π because the latter refers to policies earlier in the chapter. The hat on $\hat{r}$ is suppressed throughout for clarity.

2.12.1 Messages for one time slice

$$\lambda(r) = [0, 1] \tag{2.40}$$

$$\lambda_r(s') = \lambda(s') = \sum_r \lambda(r) P(r|s') = P(r = 1|s') \tag{2.41}$$

$$\lambda_{s'}(a) = \lambda(a) = \beta \sum_{s'} \lambda(s') \sum_s P(s'|a, s) \rho_{s'}(s) \tag{2.42}$$

$$\lambda_a(\pi_i) = \lambda(\pi_i) = \beta \sum_a \lambda(a) \sum_s \sum_{\pi_j \neq \pi_i} P(a|s, \pi_0, \ldots, \pi_n) \rho_a(s) \prod_{j \neq i} \rho_a(\pi_j) \tag{2.43}$$

$$\rho_{s'}(a) = \alpha \sum_s \sum_{\pi_j} P(a|s, \pi_0, \ldots, \pi_n) \rho_a(s) \prod_j \rho_a(\pi_j) \tag{2.44}$$

$$\rho_a(\pi_i) = P(\pi) \tag{2.45}$$

$$\lambda_a(s) = \sum_a \lambda(a) \sum_{\pi_0, \ldots, \pi_n} P(a|s, \pi_0, \ldots, \pi_n) \prod_i \rho_a(\pi_i) \tag{2.46}$$

$$\lambda_{s'}(s) = \beta \sum_{s'} \lambda(s') \sum_a P(s'|a, s) \rho_{s'}(a) \tag{2.47}$$

$$\rho_{s'}(s) = \alpha P(s) \lambda_a(s) \tag{2.48}$$

$$\rho_a(s) = \alpha P(s) \lambda_{s'}(s) \tag{2.49}$$

In order to compute $P(r)$, we need to act as if r is not set and compute the message going down to r from s':

$$\rho_r(s') = \alpha \sum_a \sum_s P(s'|a, s) \rho_{s'}(a) \rho_{s'}(s) \tag{2.50}$$

Then $P(r) \approx \sum_{s'} P(r|s') \rho_r(s')$.

2.12.2 Messages for two time slices

$$\lambda(r_c) = [0,1] \tag{2.51}$$

$$\lambda_{r_c}(r') = \lambda(r') = \beta \sum_{r_c} \lambda(r_c) \sum_{r''} P(r_c|r', r'')\rho(r'') \tag{2.52}$$

$$\lambda_{r_c}(r'') = \lambda(r'') = \beta \sum_{r_c} \lambda(r_c) \sum_{r'} P(r_c|r', r'')\rho(r') \tag{2.53}$$

$$\lambda_{r'}(s') = \sum_{r'} \lambda(r')P(r'|s') \tag{2.54}$$

$$\lambda_{r''}(s'') = \lambda(s'') = \sum_{r''} \lambda(r'')P(r''|s'') \tag{2.55}$$

$$\lambda_{s'}(a) = \lambda(a) = \beta \sum_{s'} \lambda(s') \sum_{s} P(s'|a, s)\rho_{s'}(s) \tag{2.56}$$

$$= \sum_{s'} \lambda_a(s')\lambda_{s''}(s')\lambda_{r'}(s') \sum_{s} P(s'|a, s)\rho_{s'}(s) \tag{2.57}$$

$$\lambda_{s''}(s') = \beta \sum_{s''} \lambda(s'') \sum_{a'} P(s''|a', s')\rho_{s''}(a') \tag{2.58}$$

$$\lambda_{s''}(a') = \lambda(a') = \beta \sum_{s''} \lambda(s'') \sum_{s'} P(s''|a', s')\rho_{s''}(s') \tag{2.59}$$

$$\lambda_{a'}(s') = \sum_{a'} \lambda(a') \sum_{\pi'_j} P(a'|s', \pi'_0, \ldots, \pi'_n) \prod_j \rho(\pi'_j) \tag{2.60}$$

$$\lambda_a(\pi_i) = \lambda(\pi_i) = \beta \sum_{a} \lambda(a) \sum_{s} \sum_{\pi_j \neq \pi_i} P(a|s, \pi_0, \ldots, \pi_n)\rho_a(s) \prod_{j \neq i} \rho(\pi_j) \tag{2.61}$$

$$\lambda_{a'}(\pi'_i) = \lambda(\pi'_i) = \beta \sum_{a'} \lambda(a') \sum_{s'} \sum_{\pi'_j \neq \pi'_i} P(a'|s', \pi'_0, \ldots, \pi'_n)\rho_{a'}(s') \prod_{j \neq i} \rho(\pi'_j) \tag{2.62}$$

$$\rho_a(\pi_i) = \rho(\pi_i) = P(\pi) \tag{2.63}$$

$$\rho_{a'}(\pi'_i) = \rho(\pi'_i) = P(\pi') \tag{2.64}$$

$$\rho_{s'}(a) = \alpha \sum_{s} \sum_{\pi_j} P(a|s, \pi_0, \ldots, \pi_n)\rho_a(s) \prod_j \rho(\pi_j) \tag{2.65}$$

$$\rho_{s''}(a') = \alpha \sum_{s'} \sum_{\pi'_j} P(a'|s', \pi'_0, \ldots, \pi'_n)\rho_{a'}(s') \prod_j \rho(\pi'_j) \tag{2.66}$$

$$\rho_{a'}(s') = \alpha\lambda_{s''}(s')\lambda_{r'}(s')\sum_a\sum_s P(s'|a,s)\rho_{s'}(a)\rho_{s'}(s) \tag{2.67}$$

$$\rho_{s''}(s') = \alpha\lambda_{a'}(s')\lambda_{r'}(s')\sum_a\sum_s P(s'|a,s)\rho_{s'}(a)\rho_{s'}(s) \tag{2.68}$$

$$\rho_{r'}(s') = \alpha\lambda_{a'}(s')\lambda_{s''}(s')\sum_a\sum_s P(s'|a,s)\rho_{s'}(a)\rho_{s'}(s) \tag{2.69}$$

$$\rho_{r''}(s'') = \alpha\sum_{a'}\sum_{s'} P(s''|a',s')\rho_{s''}(a')\rho_{s''}(s') \tag{2.70}$$

$$\lambda_a(s) = \sum_a \lambda(a)\sum_{\pi_0,\ldots,\pi_n} P(a|s,\pi_0,\ldots,\pi_n)\prod_i \rho_a(\pi_i) \tag{2.71}$$

$$\lambda_{s'}(s) = \beta\sum_{s'}\lambda(s')\sum_a P(s'|a,s)\rho_{s'}(a) \tag{2.72}$$

$$\rho_{s'}(s) = \alpha P(s)\lambda_a(s) \tag{2.73}$$

$$\rho_a(s) = \alpha P(s)\lambda_{s'}(s) \tag{2.74}$$

In order to compute $P(r_c)$, we need to act as if r_c is not set and compute the message going down to r_c from r' and r'':

$$\rho_{r_c}(r') = \alpha\sum_{s'} P(r'|s')\rho_{r'}(s') \tag{2.75}$$

$$\rho_{r_c}(r'') = \alpha\sum_{s''} P(r''|s'')\rho_{r''}(s'') \tag{2.76}$$

Then, $P(r_c) \approx \sum_{r'}\sum_{r''} P(r_c|r',r'')\rho_{r_c}(r')\rho_{r_c}(r'')$.

Chapter 3

Evidence integration in model-based tree search

3.1 Abstract

Over the last decade, a number of researchers have related goal-directed decision making to model-based reinforcement learning. The majority of both theoretical and empirical work has been at the computational level, detailing what the brain is computing, but remaining largely agnostic about the particular mechanisms. Work on simple choice has been the exception. However, it is limited to a single step of action, whereas a hallmark of both goal-directed decision making and model-based reinforcement learning is the ability to plan several steps ahead. In this paper, we extend the evidence integration approach previously applied to data on simple choice to the multi-step case. We present two behavioral experiments, each with two steps of action, and show that a simple modified model is able to explain the observed relationship between choice and reaction time data. Our results suggest that the basic mechanisms central to simple choice are at play for more complex decision problems as well.

3.2 Introduction

Recent proposals have suggested that reward based decision making is driven by at least two distinct subsystems (e.g. Daw et al., 2005). A habitual system supports simple stimulus–response mappings, and a goal-directed system supports more complex and flexible behaviors, such as the ability to generalize existing knowledge to novel situations. A large body of recent work on the habitual system (e.g., Bayer & Glimcher, 2005; Berns, McClure, Pagnoni, & Montague, 2001; D'Ardenne, McClure, Nystrom, & Cohen, 2008; Daw et al., 2005, 2011; McClure, Berns, & Montague, 2003; Montague et al., 1996; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Schultz et al., 1997; Seymour et al., 2004; Zaghloul et al., 2009) has focused on its connection to model-free reinforcement learning, a suite of algorithms for learning stimulus–response associations through trial-and-error (Sutton & Barto, 1998). Although the details of these algorithms differ, they are based on the same basic underlying principle: Actions that lead to better than expected outcomes are reinforced, and actions that lead to worse outcomes are weakened.

In addition to providing a computationally precise theory of the behavioral data, an advantage of the modeling approach is that it provides a natural bridge to neural mechanisms. An ever increasing number of studies have implicated the midbrain dopaminergic system in coding reward prediction errors like the ones used by model-free reinforcement learning (see above, comprehensive recent reviews include Glimcher, 2011; Maia, 2009; Niv, 2009; Schultz, 2010). Subsequent work has focused on refining this original hypothesis and arbitrating between specific algorithms. On one side are valued-based algorithms, such as SARSA (Rummery & Niranjan, 1994) and Q-learning (Watkins, 1989), which make predictions about long-term average returns, and select actions associated with the highest returns. On the other side are policy gradient methods (Dayan & Abbott, 2001), which search for the best set of actions directly based on experience. Other algorithms,

like Actor-Critic (Barto, 1995), fall in-between. Although it is not entirely clear which of these algorithms the brain uses and when (Li & Daw, 2011; Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006; Roesch, Calu, & Schoenbaum, 2007), significant progress has been made.

Study of the goal-directed system has been following a similar path, although this work is more recent (Dolan & Dayan, 2013). Here, the goal-directed system has been equated with model-based reinforcement learning. In this framework, the decision maker has access to a model of the environment in the form of: 1) a transition function, specifying the (possibly probabilistic) transitions between states that result from each action, and 2) a reward function, specifying the reward expected after each transition. The information contained in the model is then converted into a decision, again using one of several different algorithms. The model may either be given to the decision maker, or it itself must be learned. Even in the latter case, the model-based framework can be advantageous compared to model-free learning. Changes in one part of the model can immediately be propagated to and influence decisions in all states. One line of work has focused on how the model is learned (Bornstein & Daw, 2012, 2013; Glascher et al., 2010).

A second line of work has focused on mapping intermediate (latent) quantities prescribed by the framework, namely, model-based values, to the brain (Daw et al., 2011; D. A. Simon & Daw, 2011b; Wunderlich et al., 2012). Similar to the model-free case, some model-based algorithms first map the model to the long-term returns expected after taking each action, and then select the actions associated with the highest returns. This second line of work has focused on understanding where in the brain these values are represented, and how they interact with the values prescribed by the model-free system. Missing, however, is an account of how these values are computed in the first place. The time seems ripe to begin to explore more

nuanced algorithmic questions having to do with how the brain actually performs the mapping from the model into a decision.

Some work along these lines has already been conducted under the more general rubric of goal-directed decision making, rather than model-based reinforcement learning, in the case of simple binary choice (Rangel, 2009). In these experiments, subjective values for a number of items are obtained by asking participants to rate each one individually. Then, on each trial, participants are asked to choose which of two items they would prefer to receive. The structure of a single trial, and its most parsimonious representation as a reinforcement learning problem, is displayed schematically in Figure 3.1A. The participant begins in a state represented by the open circle and has two actions available. One leads (deterministically) to a second state, and the other leads to a third state. Each transition is associated with reward equivalent to the subjective value assigned to each item.

Milosavljevic et al. (2010) examined the relationship between the two dependent behavioral measures captured in this task, accuracy (whether the higher rated item was selected) and reaction time, and the ratings of the items. They found a monotonically increasing relationship between mean accuracy and the difference in ratings, and a monotonically decreasing relationship between mean reaction time and the difference in ratings. When the ratings were further apart, decisions were faster and more accurate. Milosavljevic et al. (2010) explained these effects within the context of the drift-diffusion model (DDM; Ratcliff, 1978; Ratcliff & McKoon, 2008). The DDM stipulates that the evidence for each option comes in the form of noisy samples. These samples are then integrated over time, in order to increase the signal-to-noise ratio. A decision is made when the evidence for one option significantly exceeds the evidence for the other. Application of the DDM is limited to binary decisions, but multi-alternative extensions are available, and it

has been shown that one such extension is able to explain data on trinary choice as well (Krajbich & Rangel, 2011).

Real life decisions are more complicated in a different way, however. In addition to having a large breadth (the number of available actions in each state), they also have a large depth (the number of time steps across which planning takes place). Consider planning a route to the airport. A simple way to encode this as a reinforcement learning problem is to consider each block (street) a state, each direction an action, and the reward (or cost) proportional to the amount of traffic. Even if you had only three actions available at each step (turning left, turning right, and going straight), there are many steps to consider. In this paper, we begin to expand previous work on simple choice to deal with such situations. We present two behavioral experiments in which decisions are made across two time steps, and detail a computational model that can explain the relationship between accuracy and reaction time in these data. Of note is that the same evidence integration process key to explaining data on simple choice applies to more complicated problems as well. The model itself is not limited to two steps of action and applies to arbitrary decision trees. However, it is likely that yet additional mechanisms are at play in even more complex problems. We return to this point in detail in the Discussion.

3.3 Results

3.3.1 Experiment 1

Participants in both experiments started by rating a set of 270 products on a scale of 0-4 (Experiment 1) or 1-5 (Experiment 2). This was followed by a series of decision trials in which they were asked to choose the products they would most prefer to receive given the structure of the decision problem. The structure used in Experiment 1 is illustrated schematically in Figure 3.1B, and represents what is

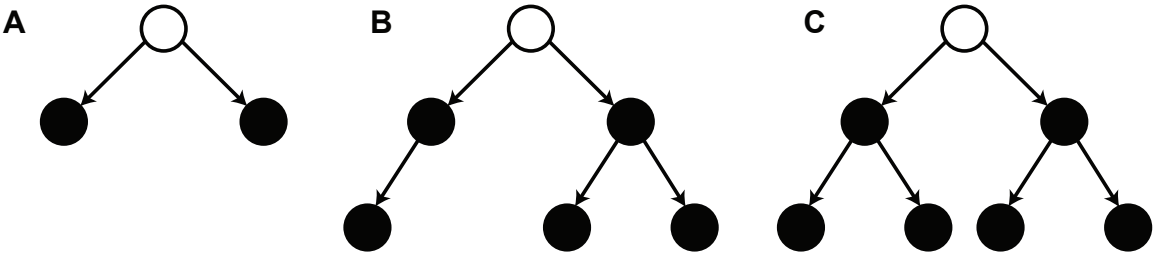

Figure 3.1: Decision trees representing decision problems in previous and current work. Each node is a state, and each arrow an action. Participants start in the state marked by the open circle. Each state marked with a closed circle contains reward. **A**. Simple choice (e.g., Krajbich et al., 2010; Milosavljevic et al., 2010). **B**. Experiment 1. **C**. Experiment 2.

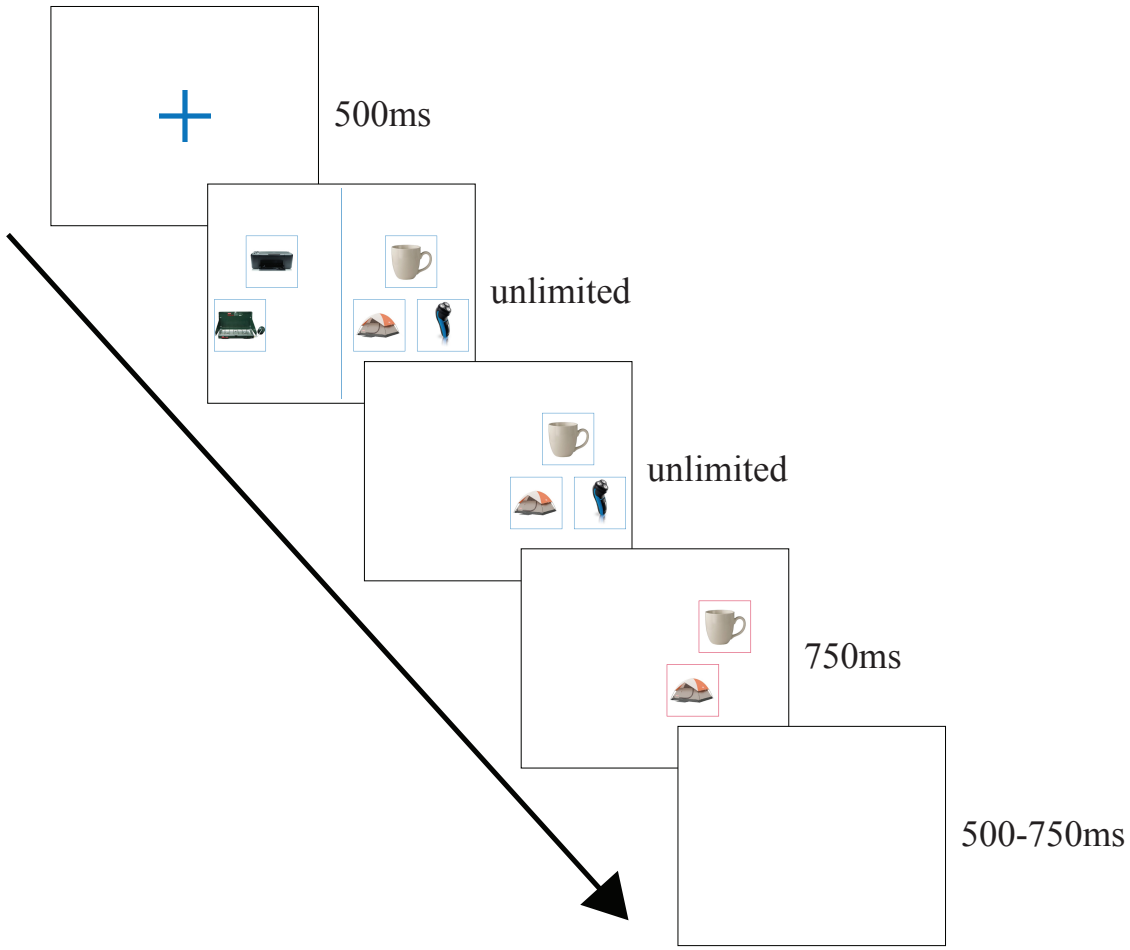

Figure 3.2: The sequence of events in each experimental trial.

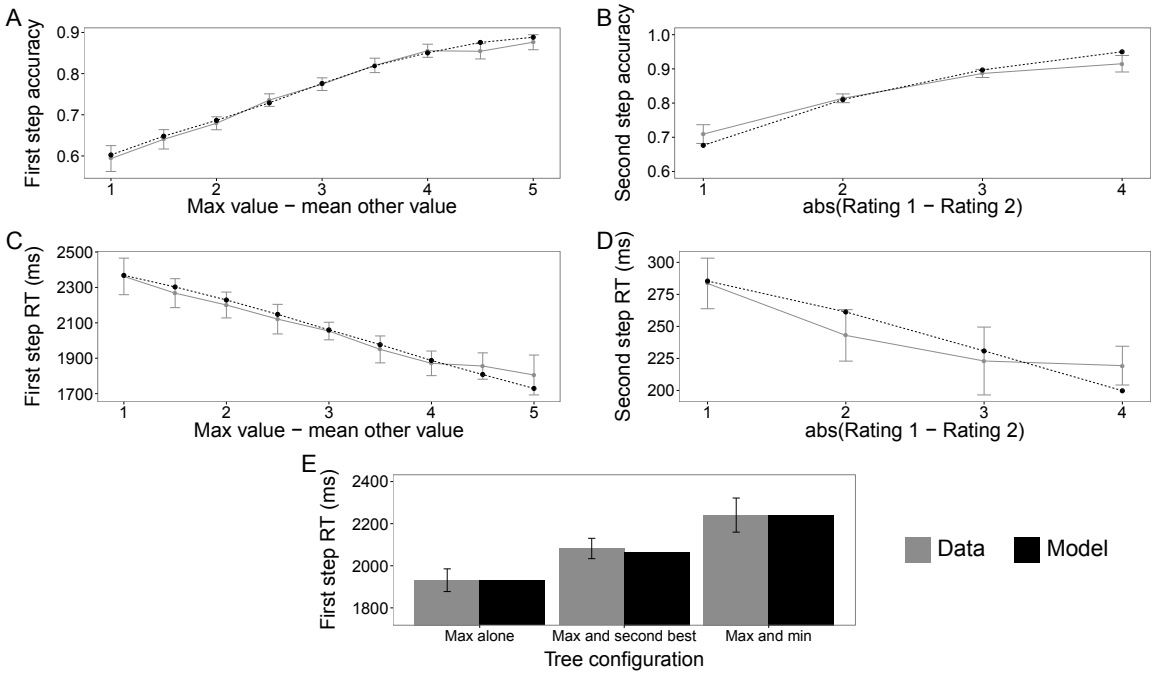

Figure 3.3: Results of Experiment 1. Empirical data appear in gray with solid lines, and the model in black with dashed lines. Bars represent within-subject confidence intervals. The model was simulated 100 times for each experimental trial using hand tuned parameters (see Table 3.1). In the figure and in the following description, *value* refers to the sum of the ratings in one arm of the decision tree. **A**. First step accuracy as a function of the difference between the maximum value and average of the other two values. A trial is considered correct if the participant selects the side of tree with the best arm. **B**. Second step accuracy as a function of the absolute difference between the ratings of the second step products. A trial is considered correct if the higher rated product is selected. Only trials where the participant selected the side of the tree with a second step decision on the first step are included. **C**. First step reaction time for correct trials. A trial is considered correct if the best arm was selected overall, taking both steps into account. Note that this is a more stringent definition than in A. **D**. Second step reaction time for correct trials. A trial is considered correct if the best arm was selected overall, taking both steps into account. **E**. First step reaction time for correct trials, as defined in C–D, as a function of the arms appearing together in the tree. For example, "Max and second best" means that the two arms corresponding to the two largest values appeared on one side, and the smaller valued arm appeared on the other side by itself.

perhaps the simplest possible extension of one step binary choice to the multi-step domain. Participants made either one or two binary (left/right) decisions on each trial. The first decision committed them to at least one product on the corresponding side of the screen. One side had a further left/right decision allowing participants to select a second product from amongst two others. The other side had a forced left/right "decision" committing participants to a particular product. The sequence of events within a single trial are displayed in Figure 3.2. Importantly, participants had access to the full structure of the problem, including reward at both steps, even when making their first step decision. The first step decision should thus reflect all of this information. This is the first set of results to which we now turn.

The dependent variables we wish to study are accuracy and reaction time. However, it is not immediately clear what the independent variable should be. The first impulse is to use the full structure of the decision tree. However, if one notes that it is possible for five different ratings to appear in each of the five positions, even when symmetry at both steps is taken into account, we are left with very few trials for each tree. Work on one step binary choice has focused on using the difference between the two ratings as the independent variable. This represents a sort of measure of decision difficulty, and it has been shown that larger differences are easier in that they result in faster and more accurate decisions (Milosavljevic et al., 2010; Krajbich et al., 2010). It would seem sensible to derive a similar measure for our two step problem. We do this by collapsing across the two steps, and treating the problem, for a moment, as a comparison between the three pairs of products. The value of each pair is simply the sum of the ratings of the products in the pair. One measure of decision difficulty then, similar to what has previously been used in one step multi-alternative choice (Krajbich & Rangel, 2011), is the difference between the maximum value and the mean of the other two values. The solid lines in Figures 3.3A and 3.3C plot accuracy and reaction time at the first step

as a function of this measure of difficulty. As one would expect if this measure were useful, accuracy increases and reaction time decreases as decisions get easier.

Similarly, we can look at accuracy and reaction time at the second step for those trials on which a second step decision had to be made, as a function of the difference between the ratings of the remaining two products. This is plotted in Figures 3.3B and 3.3D, which show the same pattern of results seen in one step binary choice.

The final aspect of the data that is of present interest is shown in Figure 3.3E. Here, we plot the first step reaction time as a function of the two values that are grouped together (i.e. on the same side) in the tree. There are three possibilities: the maximum and second best are together, the maximum and minimum are together, or the maximum is alone (the second best and minimum are together). Of particular note is the difference between the maximum and second best and the maximum and minimum groups. When the maximum value is together with the second best, the decision at the first step is faster compared to when the maximum is together with the minimum. This might suggest that participants prune the tree and discard the minimum before the second level decision on the other side is complete.

In order to confirm whether or not this is the case, we must also take into account differences in the distribution of decision difficulties between these three groupings. Because of the structure of the tree, these distributions are different, and this could contribute to the results in Figure 3.3E. This can be seen by the fact that the mean reaction time for the "Max alone" group is also faster than the other two. In order to perform a more stringent test of the pruning hypothesis, we performed a second analysis concentrated on the difference between the "Max and second best" and "Max and min" groups. We took as our index of difficulty the difference between the maximum value and average of the other two values on each trial, as discussed above, and discarded trials corresponding to difficulty levels that did not appear in both groups. Next, we computed mean overall accuracy (whether the best arm

was selected, taking both steps of the decision into account) and mean reaction time (for correct trials) separately for each difficulty level and subject. Finally, we computed the means within subject, across difficulty levels, in effect giving equal weighting to the trials at each level. Reaction times were still faster in the "Max and second best" condition compared to the "Max and min" condition (one-tailed t-test, $p = 0.015$), but accuracy was not higher (one tailed t-test, $p = 0.17$).

3.3.2 Model

We begin modeling this task by borrowing some of the assumptions made by drift-diffusion model. In particular, we assume that the reward function is represented by a set of noisy samples, and that these samples are accumulated over time. Rather than a single integrator representing the difference in evidence between two options, each of the five states (products) is associated with a separate integrator indicating the build up of preference for that product. On each iteration, we sample the entire reward function (all five products), and update the corresponding evidence:

$$\begin{aligned}
E_L^{t+1} &= E_L^t \quad + d_1 \cdot R_L \quad + \epsilon_L, \\
E_{L,L}^{t+1} &= E_{L,L}^t + d_1 \cdot R_{L,L} + \epsilon_{L,L}, \\
E_R^{t+1} &= E_R^t \quad + d_1 \cdot R_R \quad + \epsilon_R, \\
E_{R,L}^{t+1} &= E_{R,L}^t + d_1 \cdot R_{R,L} + \epsilon_{R,L}, \\
E_{R,R}^{t+1} &= E_{R,R}^t + d_1 \cdot R_{R,R} + \epsilon_{R,R}.
\end{aligned} \tag{3.1}$$

Here, E_L is the amount of evidence associated with the top left product (going left once), $E_{L,L}$ is the amount of evidence associated with the bottom leftmost product (going left and then left again), E_R is the amount of evidence associated with the top right product, and so on. Similarly, R_* represents the mean reward (the rating)

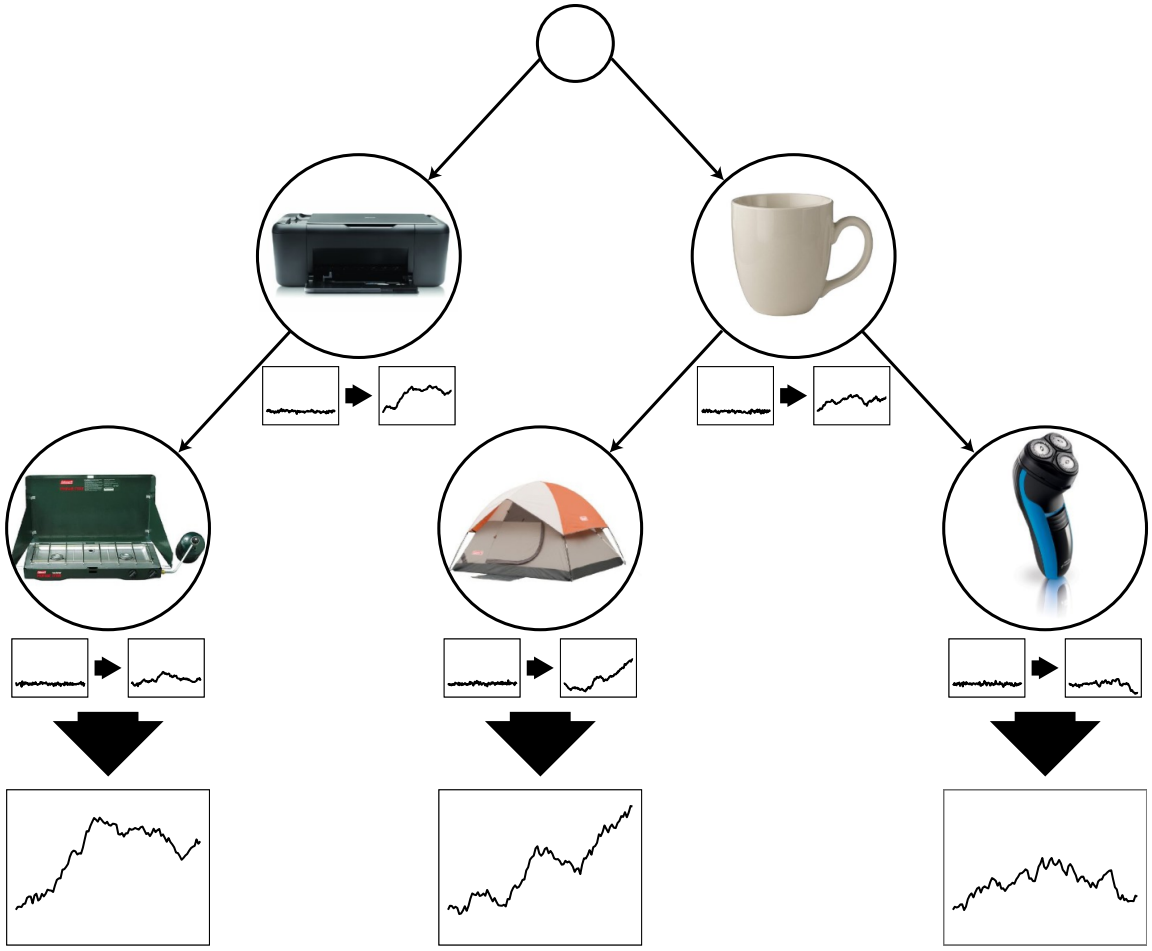

Figure 3.4: A cartoon depiction of the model. A separate integrator accrues evidence in favor of each state, and integrated values within each arm of the decision tree are combined. A decision is made when the largest integrator is significantly greater than the second largest integrator, or when the smallest integrator on one side of the tree is significantly greater than the largest integrator on the other side of the tree.

at each position. Each ϵ is an independent draw from a Gaussian distribution with mean 0 and standard deviation fixed to 0.01. The parameter d acts as a scaling factor on the drift rate, set by the underlying mean reward. We present Equation 3.1 in terms of a decision problem where the second step decision is on the right. The same set of updates, with an analogous set of labels, apply to a decision problem with the second step decision on the left. We take up the validity of sampling the entire reward function on each iteration, together with other assumptions, in the Discussion.

In the drift-diffusion model, a decision is made when the difference in evidence for the two options exceeds a threshold. When multiple integrators are involved, it is less clear what the decision rule should be. Previous work (Krajbich & Rangel, 2011) has shown that a decision rule whereby the difference between largest integrator and the second largest integrator is used fits well with data on one-step multi-alternative choice. We make a similar assumption here, but posit that the competition is between the arms of the decision tree rather than between individual items. That is, a decision is made when one of the following occurs:

$$\begin{aligned} E_L + E_{L,L} - max(E_R + E_{R,L}, E_R + E_{R,R}) > \theta_1, \\ E_R + E_{R,L} - max(E_L + E_{L,L}, E_R + E_{R,R}) > \theta_1, \\ E_R + E_{R,R} - max(E_L + E_{L,L}, E_R + E_{R,L}) > \theta_1. \end{aligned} \tag{3.2}$$

As before, the decision rule is written in terms of a decision problem where the second step decision is on the right, but the same rule with an analogous set of labels applies to decision problems where the second step decision is on the left.

When one of the above quantities crosses threshold, a decision is rendered for both steps of action. However, as we saw, the data suggest that participants also sometimes prune the tree and make a first step decision before the second step

decision is complete. This is captured by a second decision rule, which states that a first step decision can also be rendered if the smallest integrator on one side of the tree is significantly greater than the largest integrator on the other side of the tree. In Experiment 1, because one side has only a single arm, this is simply:

$$min(E_R + E_{R,L}, E_R + E_{R,R}) - (E_L + E_{L,L}) > \theta_2, \tag{3.3}$$

assuming the single arm is on the left side, and similarly with the labels transposed if it is on the right side. When the decision rule in 3.3 applies, the evidence integration process continues where it left off for products at the bottom level:

$$\begin{aligned} E_{R,L}^{t+1} &= E_{R,L}^{t} + d_2 \cdot R_{R,L} + \epsilon_{R,L} \\ E_{R,R}^{t+1} &= E_{R,R}^{t} + d_2 \cdot R_{R,R} + \epsilon_{R,R}. \end{aligned} \tag{3.4}$$

A second step decision is rendered when the difference in evidence for the two remaining products exceeds threshold:

$$\begin{aligned} E_{R,L} - E_{R,R} &> \theta_3 \\ E_{R,R} - E_{R,L} &> \theta_3. \end{aligned} \tag{3.5}$$

The model is displayed schematically in Figure 3.4. We simulated each trial in the data 100 times using Equations 3.1-3.5, and subjected the simulated trials to the same analysis as above. The results are plotted alongside the data using dashed lines in Figure 3.3. As can be seen, the model captures the general features of all five aspects of the data discussed above. It should be noted, however, that the free parameters were fit by hand, and the results shown in Figure 3.3 may thus not represent the very best fit that may be achieved.

The parameters were fit by hand for two reasons. First, because the model does not have known closed form solutions, we had to resort to simulation. Although each simulation of a single tree configuration is not prohibitively expensive in itself, the data contain many tree configurations, and each configuration has to be simulated multiple times. All together, this rendered any objective function that could be given to an automated optimization algorithm relatively expensive. Second, it is likely that some of the model's assumptions are too strong, and some details are left out. Thus, it does not seem sensible to look for parameters that can perfectly fit the data. Instead, we aim to capture its general features: Decision making proceeds by a process of evidence integration, with competition between the arms of the decision tree. The overall pattern of accuracy and reaction time effects seen is Figures 3.3A-D and 3.5A-D is a general feature of the model, and results from a wide range of parameter settings.

3.3.3 Experiment 2

The second experiment was similar to the first, except that both sides of the decision tree had a second step decision (Figure 3.1C). The goal was to replicate the first set of findings and modeling results with a new set of participants, and to test the model's ability to account for an additional arm in the tree. The data and simulation results are shown in Figure 3.5, which parallels what is shown in Figure 3.3 for Experiment 1.

We repeated our more stringent pruning test on data from the second experiment as well. Rather than performing pairwise tests (since all three groupings are now relevant to the question), we pooled the data in the "Max and third best" and "Max and min" groups together and compared it to the "Max and second best" group. We used the same strategy as in the first experiment, discarding trials corresponding to difficulty levels not appearing in both (new) groups, and computing

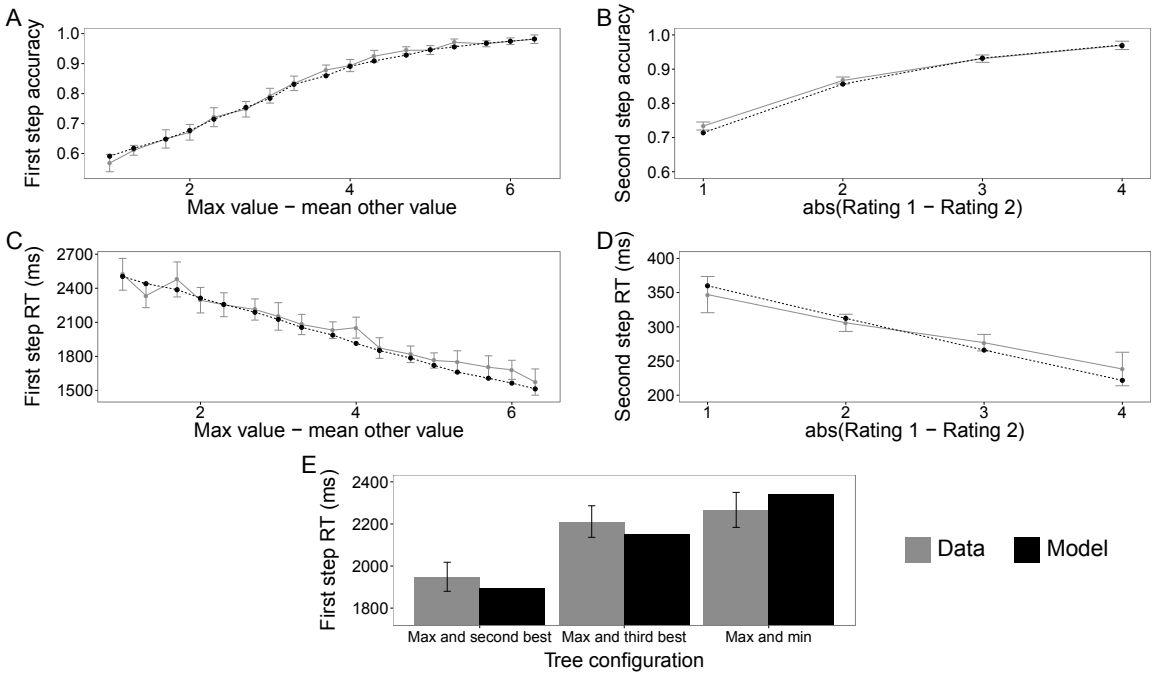

Figure 3.5: Results of Experiment 2. The panels parallel those of Figure 3.3 for Experiment 1. Empirical data appear in gray with solid lines, and the model in black with dashed lines. Bars represent within-subject confidence intervals. The model was simulated 100 times for each experimental trial using hand tuned parameters (see Table 3.1). In the figure and in the following description, *value* refers to the sum of the ratings in one arm of the decision tree. **A**. First step accuracy as a function of the difference between the maximum value and average of the other three values. A trial is considered correct if the participant selects the side of tree with the best arm. **B**. Second step accuracy as a function of the absolute difference between the ratings of the second step products. A trial is considered correct if the higher rated product is selected. **C**. First step reaction time for correct trials. A trial is considered correct if the best arm was selected overall, taking both steps into account. Note that this is a more stringent definition than in A. **D**. Second step reaction time for correct trials. A trial is considered correct if the best arm was selected overall, taking both steps into account. **E**. First step reaction time for correct trials, as defined in C–D, as a function of the arms appearing together in the tree. For example, "Max and second best" means that the two arms corresponding to the two largest values appeared on one side, and the two smaller valued arms appeared on the other side.

Table 3.1: Hand tuned model parameters

Parameter	Exp. 1	Exp. 2
d_1	0.000105 / ms	0.000135 / ms
θ_1	0.73	0.68
T_1	330 ms	550 ms
θ_2	1.08	0.74
d_2	0.00015 / ms	0.00019 / ms
θ_3	0.34	0.37
T_2	150 ms	150 ms

mean accuracy and reaction time first within and then across difficulty levels. Reaction time in the "Max and second best" group was faster compared to the the other group (one-tailed t-test, $p = 0.046$), but accuracy was not higher (one-tailed t-test, $p = 0.99$).

3.4 Discussion

Previous empirical and theoretical work has suggested that decision making is driven by at least two distinct subsystems: a habitual system that forms simple stimulus-response associations, and a goal-directed system that has access to a more informed model of the environment. Moreover, some theorists have suggested that the former accords with model-free reinforcement learning, and the latter, with model-based reinforcement learning (for a review, see Dolan & Dayan, 2013). The behavioral and neural correlates of model-free reinforcement learning have been under intense scrutiny since the suggestion that dopamine neurons code reward prediction errors (Barto, 1995; Montague et al., 1996; Schultz et al., 1997). Investigation of model-based reinforcement learning, by contrast, has begun in earnest only more recently. Most work thus far has focused on two key issues. First, if the brain supports model-based reinforcement learning, then it is likely that neural correlates of the state and action values predicted by the model-based

framework can be found. The first line of work has focused on determining whether or not this is the case, and on identifying the network of brain regions that support this encoding (Daw et al., 2011; D. A. Simon & Daw, 2011b; Wunderlich et al., 2012). A second line of work has focused on understanding how the brain learns the world model (Bornstein & Daw, 2012, 2013; Glascher et al., 2010), a necessary step to then using the model during decision making.

Both areas of research are sound first steps towards understanding how the brain implements model-based reinforcement learning, but the story they tell is incomplete. In particular, missing is an algorithmic account of the deliberation process: how the brain is able to convert the world model into a decision. Although not previously identified as belonging to the realm of reinforcement learning, some preliminary work on this question has been done. This work has focused on simple choice, with a single step of action and two to four choices (Gluth, Rieskamp, & Büchel, 2012; Milosavljevic et al., 2010; Krajbich et al., 2010; Krajbich & Rangel, 2011; Towal, Mormann, & Koch, 2013). However, a hallmark of both model-based reinforcement learning and goal-directed decision making is the ability to plan across multiple time steps. In this paper, we have begun extending the class of evidence integration models, previously applied to data on simple choice, to the multi-step domain. We have shown that more complex decisions also involve integrating noisy information about reward in parallel, but now not only across multiple states, but also across time. The decision competition is between the arms of a decision tree rather than between individual states. Furthermore, the evidence integration process continues not only across prospective time, but also across physical time. Actions that prune the tree are rendered when the reward structure allows for it, and the deliberation process continues where it left off. The model is able to explain the relationship between choice and reaction time in two new experiments. Nevertheless, substantial work remains to be done.

3.4.1 More complex decision trees

First, many real world decisions are even more complex than the two step problems studied here. They involve situations that have both more actions available at each step (a larger tree breadth) and more time steps (larger depth). We believe that the basic principles at play for the problems studied here to carry over to more complex problems as well. However, larger trees carry a higher tax on information processing systems, and likely rely on additional machinery to simplify the problem. Such machinery is important because not only does processing carry both intrinsic (Kool, McGuire, Wang, & Botvinick, 2013) and opportunity costs, but large enough trees may simply hit the capacity constraints of the system (Newell & Simon, 1972).

How the brain is able to simplify the problem is speculative, but a few intriguing possibilities exist. One approach, which we have considered to some degree in the present work, is pruning. The decision maker may start with the intention of exploring the entire tree, but then use a heuristic function during exploration to remove branches from further consideration. A recent study (Huys et al., 2012) explored decision making in structured trees with discrete rather than noisy rewards. It was found that people prune both at random, with a set probability of terminating search at each node, and also strategically, when encountering large losses. The latter form was hypothesized to be supported by a Pavlovian system. Translating these findings into the current framework, a branch may be pruned when the magnitude of a single sample of reward exceeds a critical lower boundary. Alternatively, rather than relying on a single sample, a similar process may operate on the evidence integrators, not unlike some versions of decision field theory (Roe, Busemeyer, & Townsend, 2001). That is, a branch is pruned when the evidence for it falls below threshold. Similar to the suggestion made by Roe et al. (2001) in the context of one step multi-attribute decision problems, the location

of this threshold may depend on tree complexity, with a higher (i.e. less stringent) boundary for more complex trees. This would allow many poor branches to quickly be discarded. Pruning of this type is different from the more informed type of pruning we consider in this paper. The former depends on a heuristic function that is local to the branches under consideration, while our models prunes only when the best a branch has to offer is already known to be a lot worse than the worst in an alternative branch (Equation 3.3).

Another approach, rather than pruning branches deemed likely to be bad, may be to sample the tree until a good (or *good enough*) path or set of paths is found. The ways in which this idea can be incorporated into the present framework are similar to the above. The arms of the tree can be sampled until n arms with rewards that exceed a threshold are found. Only these arms would then compete. Alternatively, n competitors could be chosen at random, and if the best integrator fails to reach an intermediate threshold within a set amount of time, additional competitors allowed in (perhaps in place of the current worst competitors).

Yet another approach to simplify the problem is to incorporate ideas from hierarchical reinforcement learning (Barto & Mahadevan, 2003; Botvinick, Niv, & Barto, 2009; Diuk, Tsai, Wallis, Botvinick, & Niv, 2013; Ribas-Fernandes et al., 2011; Solway et al., n.d.). How this may be combined with the evidence integration framework is less clear, but presents an exciting opportunity for future research.

3.4.2 Attention

The second issue which the present work does not address is the role of attention. It has been shown that visual fixations can bias the evidence integration process in simple choice, promoting the currently fixated option at the expense of the others (Krajbich et al., 2010; Krajbich & Rangel, 2011). The pattern of fixations is thought to be a function of both value and visual salience (Towal et al., 2013). Such

factors are no doubt at play here, although we cannot speak to them because we did not perform eye tracking. Integrating these two theories would open the door for interesting new predictions. For example, fixating on a first level product would promote both arms of the decision tree in which it participates. This would lead to excess pruning of the alternative side, but no advantage for disambiguating the arms in which the product is embedded. Likewise, fixating on a leaf node (i.e. a second level product) would promote only a single arm.

Although incorporating fluctuations in visual attention would help constrain theory, a long term goal should also involve incorporating *internal* fluctuations in attention. Understanding internal fluctuations is perhaps even more important for multi-step decision making, where the transition and reward functions are seldom visible, but rather, are learned piecewise. Consider again the route planning example from the introduction, where you are planning a route to the airport. The transition function consists of valid turns, from one street to another, and the reward (or cost) function encodes the amount of traffic usually present on each street. You likely learned this information across many trips through the city in which you live, many of which were not to the airport. Furthermore, the layout and current traffic situation are usually not visible, but have to be retrieved from memory. The dynamics by which this process unfolds are currently unknown, although some work has begun to address the issue (Bornstein & Daw, 2013; Wimmer & Shohamy, 2012).

3.4.3 Probabilistic transitions

Third, we have only addressed tree structured problems with deterministic transitions. The model-based reinforcement learning framework allows for additional complexity, namely, probabilistic transitions between arbitrary states, and the ability to visit the same state multiple times. Many real life decision problems have

this complexity as well, and understanding how the brain solves more general model-based reinforcement learning problems is an important challenge for future work.

3.4.4 Response variability

Both the data we present here and the data on simple choice suggest that responses are highly variable. However, the source of this noise is unclear. There are at least three possibilities. First, part of the noise may be endogenous, resulting from the way in which the reward function is coded by neurons (Solway & Botvinick, 2012). A second possibility is that, rather than relying purely on statistical information about reward, reward information may be retrieved from episodic memory (Bornstein & Daw, 2013; Wimmer & Shohamy, 2012). The dynamics of this retrieval process could contribute additional noise. Third, there is likely external noise contributed by the stimulus as well. We discussed the role that visual attention and object salience play in biasing choice. It follows that the randomness inherit in fixating on each item, and on the subcomponents of each item, must also contribute to the variability seen in choice and reaction time data. These three possibilities are not necessarily mutually exclusive.

Regardless where the noise comes from, an additional question concerns its utility. It would seem beneficial if the brain could retrieve discrete (mean) reward information, allowing decision making to proceed at a faster pace. Moreover, the relationship between mean reaction time and difference in value seems particularly perverse, with decision time increasing when the difference in value decreases. It would seem more sensible to observe the opposite pattern, with reaction time decreasing when the difference in value decreases, because opportunity costs are lower. Although our model is able to explain the observed relationship given the

assumptions borrowed from previous work, why the brain evolved this way in the first place is unclear.

3.4.5 Implementation and neural correlates

The model we presented is an algorithmic description of the decision process. An obvious question concerns implementation, how the brain actually realizes this process at the neural level. A recent paper addressing this question for binary choice (Hare, Schultz, Camerer, O'Doherty, & Rangel, 2011, see also Basten, Biele, Heekeren, & Fiebach, 2010; Rolls, Grabenhorst, & Deco, 2010) opens the door for speculation. In line with previous work, the authors found fMRI BOLD correlates of reward in ventromedial prefrontal cortex (for a review, see Rangel & Clithero, 2013). Going a step further, they built a simple neural version of the drift-diffusion model in which the integrator for each option is driven by a separate pool of neurons, interacting through lateral inhibition. They used the output of the model to look for correlates of the integration process itself, finding positive results in dorsomedial prefrontal cortex and intraparietal sulcus. Moreover, coupling between these regions and ventromedial prefrontal cortex increased during deliberation, and coupling between these regions and the appropriate side of motor cortex (depending on whether the decision was to the left or to the right) increased during the response period. It seems sensible that the same neural structures also support decision making across multiple time steps. Our model predicts a separate integrator for each state, all working in parallel across time, together with an additional group of integrators coding the sum of the evidence in each arm. Further, it predicts that integrators for states lower in the tree (i.e. further in time) persist after actions that prune the tree are rendered, allowing decision making to continue where it left off in the selected branch. One avenue for future work may be to instantiate a neural

version of our model using an approach similar to that of Hare et al. (2011), and to look for correlates of the model in the brain.

A point of contention concerns how this relates to previous work looking at neural correlates of model-based values (Daw et al., 2011; D. A. Simon & Daw, 2011b; Wunderlich et al., 2012). We started by pointing out that this work has been silent about how values are actually computed, but then took as our point of departure the evidence accumulation perspective, which bypasses values in favor of computing evidence. Although we do not compute values directly, the *expected* amount of evidence in favor of each arm is proportional to value on each iteration of the algorithm. However, whereas the above mentioned work on simple choice predicts the integrators should be found in dorsomedial prefrontal cortex, work looking at model-based values has largely implicated striatum and ventromedial prefrontal cortex.[1] It is not clear at present how these discrepancies can be resolved.

3.4.6 Relation to previous work

We previously suggested (Solway & Botvinick, 2012) that the brain might solve the model-based reinforcement learning problem by treating it in turn as a Bayesian inference problem (Koller & Friedman, 2009). The components of the model (i.e. the transition and reward function) were encoded together with the policies as a joint probability distribution, and decision making amounted to computing the posterior of the policy variables conditional on (maximizing) the reward. We used the framework to explain a number of qualitative behavioral and neural findings in the literature. The goal of the present work was to begin a more detailed quantitative study of how people solve multi-step decision problems. We thus opted for an approach that more closely resembles what has already been done

[1] A further complication is that BOLD activity reflecting the integrators should be proportional to the *sum* of the evidence at each iteration (a sum of sums, Hare et al., 2011). Thus, correlates of values should not necessarily coincide with correlates of the integrators.

in the context of simple choice. However, the approach of Solway and Botvinick (2012) is not necessarily far removed from the present model. As we explain in that paper, the evolving posterior over the policy variables can be understood as integrating reward information over time. Future work will need to more formally address the connection between these models.

3.4.7 Conclusion

Borrowing from work on simple choice, we've taken the first steps towards providing an algorithmic account of goal-directed decision making across multiple time steps. A significant amount of work remains to be done. However, as we outlined in the discussion above, the nature of much of this work is not necessarily mysterious, and the problems are ripe for both experimental and theoretical attack.

3.5 Materials and methods

3.5.1 Participants

Thirty participants were targeted for each experiment. As described below, each experiment consisted of two parts. The first involved rating products, and the second, making decisions between them. In order to progress to the second part, participants had to rate a set number of products at each rating level (although they were not told this, see below). Recruitment continued until there were thirty valid participants for each experiment. Overall, thirty-one participants were recruited for Experiment 1, and all progressed to the second part. However, the data for one participant were discarded because the participant went more than 30 minutes over the allotted time. Thirty-six participants were recruited for Experiment 2, and thirty of them progressed to the second part. Both parts of each experiment took

approximately two hours combined. Participants were compensated either $12 per hour, or with course credit. All experimental procedures were approved by the institutional review board of Princeton University.

3.5.2 Task

Each experiment consisted of two parts. The first part was nearly identical for both experiments and involved rating a set of 270 products. Products consisted of electronics, clothing, books, non-perishable foods, kitchen items, jewelry, and various novelties. Participants were asked to first familiarize themselves with the general range of products they would later encounter, and to then rate each product within the context of all of the other products on the list. The preview phase consisted of a series of screens showing 25 products at a time, each lasting 20 seconds. After all of the products were shown in this way, each product was displayed again, this time separately, with the numbers 0-4 below it (in Experiment 1, a scale of 1-5 was used in Experiment 2). Participants clicked with the mouse to select a rating. The order of the products for the preview phase and for the actual rating phase differed, both selected pseudorandomly for each participant.

In order to proceed to the second part, participants had to rate ten products at each rating level for Experiment 1 and twelve for Experiment 2. Participants were not told of this specific requirement beforehand, but were instructed to give relative ratings by considering each product in the context of all of the other products on the list, and to try to use all of the numbers during the rating period. Participants that did not meet this requirement did not move on two second part and were dismissed with two hours worth of compensation. This requirement was enforced to reduce carry-over effects between consecutive trials. For example, because five products were displayed on each trial of the decision phase in Experiment 1, and each could be associated with the same rating level, having ten products at each

level ensured that the same product would not have to appear on two consecutive trials. The same reasoning applies to Experiment 2, which involves six products per trial.

For participants meeting the above requirement, the set of products was then trimmed to have the same number at each rating level. For example, if 90 products were rated 0, 50 were rated 1, 30 were rated 2, 50 were rated 3, and 50 were rated 4, the 30 products rated a 2 were all kept, as were a random subset of 30 products at each other rating level. This helped put the products at each level on more equal footing. If there were more products at one level compared to the rest, each such product would appear in fewer trials, and possibly require additional processing time during presentation.

The second part of both experiments consisted of a series of decision trials. The structure of the decision problem differed between experiments, but the order of events within each trial remained the same (Figure 3.2). Each trial started with a 500ms fixation cross, which was followed by a self paced decision phase, and ended with a 750ms "feedback" phase where only the selected products remained on the screen. The inter-trial interval was 500ms with a 0-250ms jitter.

Experiment 1 consisted of a two step decision. One product appeared on the top left of the screen, and another appeared on the top right. Below each of these products, either one or two additional products appeared in a horizontal orientation (see Figure 3.2). In a random half of the trials, two products appeared on the bottom left and one on the bottom right, and in the other half, the side with the two products was swapped. Participants made up to two left/right decisions on each trial. The first decision committed them to the top product on the corresponding side, and to possibly make a second decision between the two bottom products on that side. If the side with a single product on the bottom was chosen, participants were forced to select it by pressing the corresponding keyboard key. The single product on the

bottom was offset slightly to the left or to the right of the product above it, as if a second product was next to it. It was offset to the left in a random half of trials, and to the right in the other half. The same two left/right keyboard keys were used as input for both steps. Participants kept their left index finger on the left key and their right index finger on the right key continuously throughout the experiment.

The products on each trial were pseudorandomly chosen such that: 1) the same product did not appear on consecutive trials, and 2) the difference between the value of the best arm and the average of the values of the other two arms was between 1 and 5, in increments of 0.5. The latter is the primary measure of trial difficulty as described in the Results section, where value is defined as the sum of the ratings in a single arm of the tree. For example, if the rating of the product on the left top was 4, left bottom left 5, left bottom right 3, right top 1, and right bottom left 5, the values were computed to be: left top + left bottom left, 9, left top + left bottom right, 7, and right top + right bottom left, 6. The difficulty of the trial was thus $9 - mean(6, 7) = 2.5$. Each participant completed 900 trials, 100 trials at each difficulty level, with four unlimited breaks offered evenly throughout the experiment.

Experiment 2 was similar to Experiment 1 except that both sides of the screen had a second step decision. Products were similarly chosen on each trial, with the difficulty measured by comparing four pairs of products instead of three. Difficulty levels between 1 and 6 were included, with the difficulty for each trial rounded to the nearest integer. Seventeen participants completed 660 trials, 110 at each difficulty level, and the remaining thirteen participants completed 780 trials, 130 at each difficulty level. Both groups were offered four unlimited breaks evenly throughout the experiment. Note that in Figure 3.5, difficulty level is rounded to the nearest tenth decimal place rather than to the nearest integer, as the data were less noisy than expected.

Participants did not actually receive any products at the end of any of the experiments, nor did they receive any kind of performance bonus for selecting products that were more highly rated during the first part. Instead, they were told to simply choose the products they would most prefer to receive *as if* they would actually get them after each trial.

Trials faster than 500ms and slower than 10s were discarded from analysis both in the data and in the model.

3.5.3 Model

The details of the model are spelled out in the main text. We add two additional points here. First, there was an additional set of parameters not mentioned, $T_{i\in 1,2}$, controlling the amount of non-decision time at each step. Second, we provide additional clarification about the second decision rule. The second decision rule states that a first step decision is rendered and the tree pruned if the smallest integrator on one side of the tree is significantly greater than the largest integrator on the other side of the tree. That is, when either of the following is true:

$$\begin{aligned} &min(E_L + E_{L,L}, E_L + E_{L,R}) - max(E_R + E_{R,L}, E_R + E_{R,R}) > \theta_2, \\ &min(E_R + E_{R,L}, E_R + E_{R,R}) - max(E_L + E_{L,L}, E_L + E_{L,R}) > \theta_2. \end{aligned} \tag{3.6}$$

This is the general form of the rule and the one used in Experiment 2. Equation 3.3 is exactly this rule for the special case of Experiment 1, where one side of the tree has a single arm.

Chapter 4

A neural signature of hierarchical reinforcement learning

Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). *Neuron*, 71(2), 370–379.

4.1 Summary

Human behavior displays hierarchical structure: simple actions cohere into subtask sequences, which work together to accomplish overall task goals. Although the neural substrates of such hierarchy have been the target of increasing research, they remain poorly understood. We propose that the computations supporting hierarchical behavior may relate to those in hierarchical reinforcement learning (HRL), a machine-learning framework that extends reinforcement-learning mechanisms into hierarchical domains. To test this, we leveraged a distinctive prediction arising from HRL. In ordinary reinforcement learning, reward prediction errors are computed when there is an unanticipated change in the prospects for accomplishing overall task goals. HRL entails that prediction errors should also occur in relation

to task subgoals. In three neuroimaging studies we observed neural responses consistent with such subgoal-related reward prediction errors, within structures previously implicated in reinforcement learning. The results reported support the relevance of HRL to the neural processes underlying hierarchical behavior.

4.2 Introduction

In recent years computational reinforcement learning (RL) (Sutton & Barto, 1998) has provided an indispensable framework for understanding the neural substrates of learning and decision making (Niv, 2009) shedding light on the functions of dopaminergic and striatal nuclei, among other structures (Barto, 1995; Montague et al., 1996; Schultz et al., 1997). However, to date, ideas from RL have been applied mainly in very simple task settings, leaving it unclear whether related principles might pertain in cases of more complex behavior (for a discussion, see Daw & Frank, 2009; Dayan & Niv, 2008). Hierarchically structured behavior provides a particularly interesting test case, not only because hierarchy plays an important role in human action (R. P. Cooper & Shallice, 2000; Lashley, 1951), but also because there exist RL algorithms specifically designed to operate in a hierarchical context (Barto & Mahadevan, 2003; Dietterich, 1998; Parr & Russell, 1998; Sutton et al., 1999).

Several researchers have proposed that such hierarchical reinforcement learning (HRL) algorithms may be relevant to understanding brain function, and a number of intriguing parallels to existing neuroscientific findings have been noted (Botvinick, 2008; Botvinick, Niv, & Barto, 2009; Diuk, Botvinick, Barto, & Niv, 2010; Badre & Frank, 2012; Haruno & Kawato, 2006). However, the relevance of HRL to neural function stands in need of empirical test.

In traditional RL (Sutton & Barto, 1998), the agent selects among a set of elemental actions, typically interpreted as relatively simple motor behaviors. The key

innovation in HRL is to expand the set of available actions so that the agent may now opt to perform not only elemental actions, but also multiaction subroutines, containing sequences of lower-level actions, as illustrated in Figure 4.1 (for a fuller description, see Experimental Procedures and Botvinick, Niv, & Barto, 2009).

Learning in HRL occurs at two levels. At a global level, the agent learns to select actions and subroutines so as to efficiently accomplish overall task goals. A fundamental assumption of RL is that goals are defined by their association with reward, and thus, the objective at this level is to discover behavior that maximizes long-term cumulative reward. Progress toward this objective is driven by temporal-difference (TD) procedures drawn directly from ordinary RL: following each action or subroutine, a reward prediction error (RPE) is generated, indicating whether the behavior yielded an outcome better or worse than initially predicted (see Figure 4.1 and Experimental Procedures), and this prediction error signal is used to update the behavioral policy. Importantly, outcomes of actions are evaluated with respect to the global goal of maximizing long-term reward.

At a second level, the problem is to learn the subroutines themselves. Intuitively, useful subroutines are designed to accomplish internally defined subgoals (Singh, Barto, & Chentanez, 2004). For example, in the task of making coffee, one sensible subroutine would aim at adding cream. HRL makes the important assumption that the attainment of such subgoals is associated with a special form of reward, labeled pseudo-reward to distinguish it from "external" or primary reward. The distinction is critical because subgoals may not themselves be associated with primary reward. For example, adding cream to coffee may bring one closer to that rewarding first sip, but is not itself immediately rewarding. In an HRL context, accomplishment of this subgoal would yield pseudo-reward, but not primary reward.

In order to make these points concrete, consider the video game illustrated in Figure 4.2, which is based on a benchmark task from the computational HRL liter-

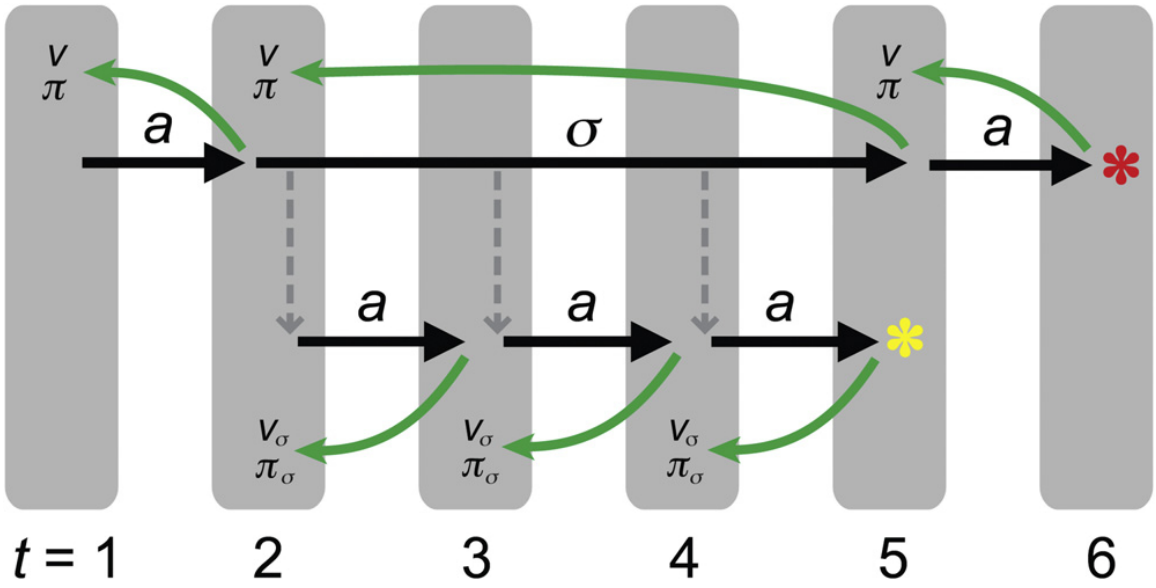

Figure 4.1: Illustration of HRL Dynamics. At $t = 1$, a primitive action (a) is selected. Based on the consequent state, an RPE is computed (green arrow from $t = 2$ to $t = 1$), and used to update the action policy (π) for the preceding state, as well as the value (V) of that state (an estimate of the expected future reward, when starting from that state). At $t = 2$ a subroutine (σ) is selected and remains active through $t = 5$. Until then, primitive actions are selected as dictated by σ (lower tier). A PPE is computed after each (lower green arrows from $t = 5$ to $t = 2$), and used to update the subroutine-specific action policy (π_σ) and state values (V_σ). These PPEs are computed with respect to pseudo-reward received at the end of the subroutine (yellow asterisk). Once the subgoal state of σ is reached, σ is terminated. An RPE is computed for the entire subroutine (upper green arrow from $t = 5$ to $t = 2$), and used to update the value and policy, V and π, associated with the state in which σ was initiated. A new action is then selected at the top level, yielding primary reward (red asterisk). Adapted from Botvinick, Niv, and Barto (2009), with permission from Elsevier.

ature (Dietterich, 1998). Only the colored elements in the figure appear in the task display. The overall objective of the game is to complete a "delivery" as quickly as possible, using joystick movements to guide the truck first to the package and from there to the house. It is self-evident how this task might be represented hierarchically, with delivery serving as the (externally rewarded) top-level goal and acquisition of the package as an obvious subgoal. For an HRL agent, delivery would be associated with primary reward and acquisition of the package with pseudo-reward. (This observation is not meant to suggest that the task must be represented hierarchically. Indeed, it is an established point in the HRL literature that any hierarchical policy has an equivalent nonhierarchical or flat policy, as long as the underlying decision problem satisfies the Markov property.) Our neuroimaging experiments proceeded on the assumption that participants would represent the delivery task hierarchically. However, as we discuss later, the neuroimaging results themselves, together with results from a behavioral experiment, provided convergent evidence for the validity of this assumption. See Supplemental Experimental Procedures, available online, for further discussion.

Consider now a version of the task in which the package sometimes unexpectedly jumps to a new location before the truck reaches it. According to RL, a jump to point A in the figure, or any location within the ellipse shown, should trigger a positive RPE because the total distance that must be covered in order to deliver the package has decreased. (Note that we assume temporal discounting, which implies that attaining the goal faster is more rewarding. We also assume that current subgoal and goal distances are always immediately known, as they were for our experimental participants from the task display.) By the same token, a jump to point B or any other exterior point should trigger a negative RPE. Cases C, D, and E are quite different. Here, there is no change in the overall distance to the goal, and so no RPE should be triggered, either in standard RL or in HRL. However, in case

C the distance to the subgoal has decreased. Thus, according to HRL, a jump to this location should trigger a positive PPE. Similarly, a jump to location D should trigger a negative PPE (note that location E is special, being the only location that should trigger neither an RPE nor a PPE). These points are illustrated in Figure 4.2 (right), which shows RPE and PPE time courses from simulations of the delivery task based on standard RL and HRL (for simulation methods, see Experimental Procedures).

These points translate directly into neuroscientific predictions. Previous research has revealed neural correlates of the RPE in numerous structures (Breiter, Aharon, Kahneman, Dale, & Shizgal, 2001; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; Holroyd & Coles, 2002; Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003; O'Doherty et al., 2003; Ullsperger & von Cramon, 2003; Yacubian et al., 2006). HRL predicts that neural correlates should also exist for the PPE. To test this, we had neurologically normal participants perform the delivery task from Figure 4.2 while undergoing EEG and, in two further experiments, fMRI.

4.2.1 EEG experiment

The EEG experiment included 9 participants, who performed the delivery task for a total of 60 min (190 delivery trials on average per participant). One-third of trials involved a jump event of type D from Figure 4.2; these events were intended to elicit a negative PPE. Earlier EEG research indicates that ordinary negative RPEs trigger a midline negativity typically centered on lead Cz, sometimes referred to as the feedback-related negativity or FRN (Holroyd & Coles, 2002; Holroyd et al., 2003; Miltner, Braun, & Coles, 1997). Based on HRL, we predicted that a similar negativity would occur following the critical jumps (type D) in our task. To provide a baseline for comparison, another third of the trials involved jump events of type E.

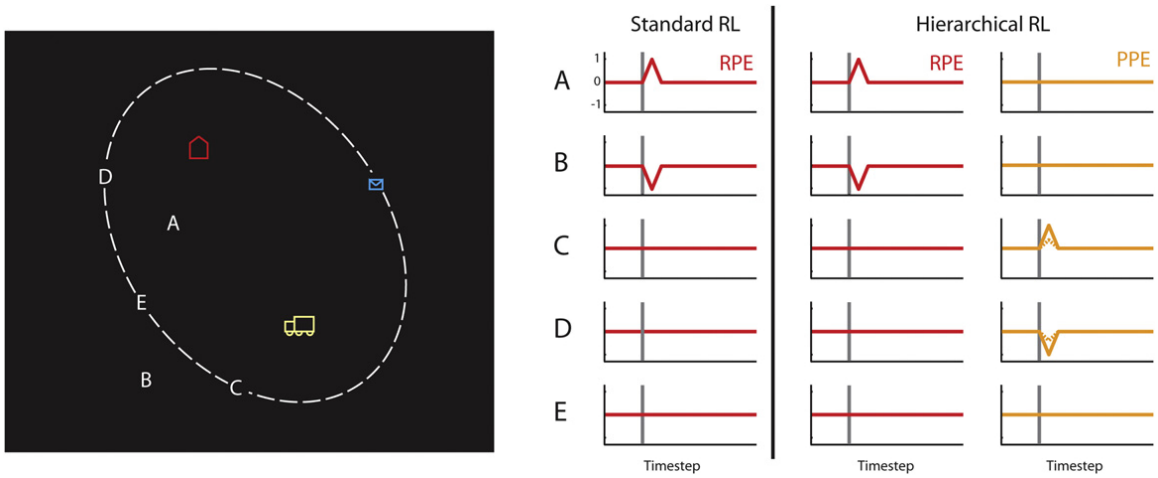

Figure 4.2: Task and Predictions from HRL and RL. Left view is task display and underlying geometry of the delivery task. Right view shows prediction-error signals generated by standard RL and by HRL in each category of jump event. Gray bars mark the time step immediately preceding a jump event. Dashed time courses indicate the PPE generated in C and D jumps that change the subgoal's distance by a smaller amount. For simulation methods, see Experimental Procedures.

Stimulus-aligned EEG averages indicated that class D-jump events triggered a phasic negativity in the EEG ($p < 0.01$ at Cz; Figure 4.3, left), relative to the E-jump control condition. (Like the ERP obtained in this study, the FRN sometimes takes the form of a relative negativity occupying the positive voltage domain, rather than absolute negativity. For germane examples, see Nieuwenhuis et al., 2005; Yeung, Holroyd, & Cohen, 2005. Like the FRN, this negativity was largest in the fronto-central midline leads (including Cz, see Figure 4.3, right), and although the observed negativity peaked later than the typical FRN, its timing is consistent with studies of equivalent complexity of feedback (T. E. Baker & Holroyd, 2011).

4.2.2 fMRI experiments

In our first fMRI experiment, a group of 30 new participants performed a slightly different version of the delivery task, again designed to elicit negative PPEs. As in the EEG experiment, one-third of trials included a jump of type D (as in Figure 4.2),

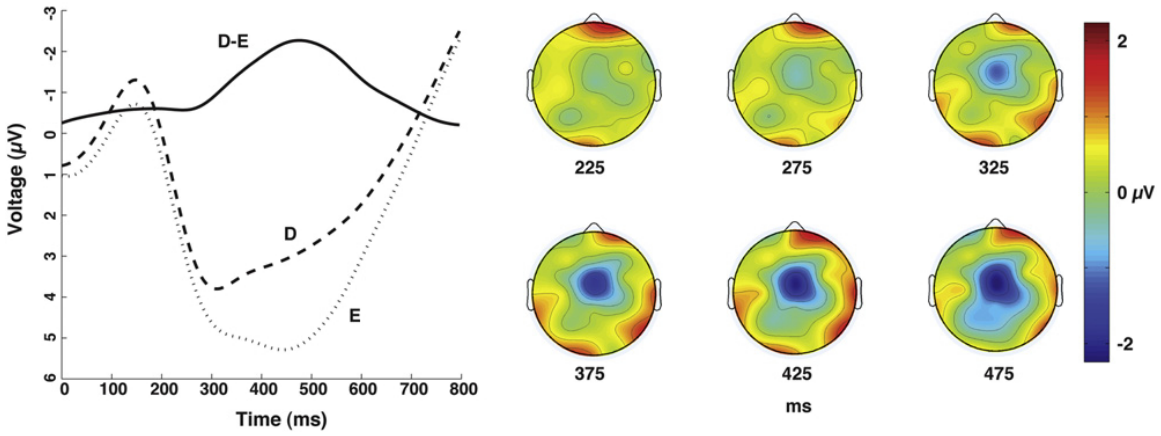

Figure 4.3: Results of EEG Experiment. Left view shows evoked potentials at electrode Cz, aligned to jump events, averaged across participants. D and E refer to jump destinations in Figure 4.2. The data series labeled D-E shows the difference between curves D and E, isolating the PPE effect. Right view is scalp topography for condition D, with baseline condition E subtracted (topography plotted on the same grid used in Yeung et al., 2005).

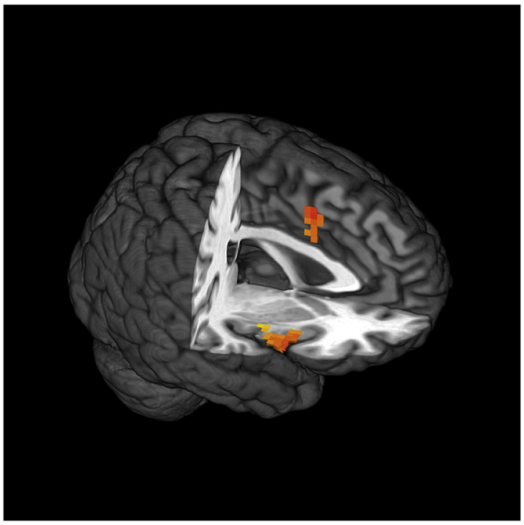

Figure 4.4: Results of fMRI Experiment 1. Shown are regions displaying a positive correlation with the PPE, independent of subgoal displacement. Talairach coordinates of peak are 0, 9, and 39 for the dorsal ACC, and 45, 12, and 0 for right anterior insula. Not shown are foci in left anterior insula (−45, 9, −3) and lingual gyrus (0, −66, 0). Color indicates general linear model parameter estimates, ranging from 3.0 x 10^{-4} (palest yellow) to 1.2 x 10^{-3} (darkest orange).

and another third included a jump of type E. Type D jumps, by increasing the distance to the subgoal, were again intended to trigger a PPE. However, in the fMRI version of the task, unlike the EEG version, the exact increase in subgoal distance varied across trials. Therefore, type D jumps were intended to induce PPEs that varied in magnitude (Figure 4.2). Our analyses took a model-based approach (O'Dohert, Hampton, & Kim, 2007), testing for regions that showed phasic activation correlating positively with predicted PPE size.

A whole-brain general linear model analysis, thresholded at $p < 0.01$ (cluster-size thresholded to correct for multiple comparisons), revealed such a correlation in the dorsal anterior cingulate cortex (ACC; Figure 4.4). This region has been proposed to contain the generator of the FRN (Holroyd & Coles, 2002, although see Nieuwenhuis et al., 2005 and Discussion below). In this regard the fMRI result is consistent with the result of our EEG experiment. The same parametric fMRI effect was also observed bilaterally in the anterior insula, a region often coactivated with the ACC in the setting of unanticipated negative events (Phan, Wager, Taylor, & Liberzon, 2004). The effect was also detected in right supramarginal gyrus, the medial part of lingual gyrus, and, with a negative coefficient, in the left inferior frontal gyrus. However, in a follow-up analysis we controlled for subgoal displacement (e.g., the distance between the original package location and point D in Figure 4.2), a nuisance variable moderately correlated, across trials, with the change in distance to subgoal. Within this analysis only the ACC ($p < 0.01$), bilateral anterior insula ($p < 0.01$ left, $p < 0.05$ right), and right lingual gyrus ($p < 0.01$) continued to show significant correlations with the PPE.

In a series of region-of-interest (ROI) analyses, we focused in on additional neural structures that, like the ACC, have been previously proposed to encode negative RPEs: the habenular complex (Salas, Baldwin, De Biasi, & Montague, 2010; Ullsperger & von Cramon, 2003), nucleus accumbens (NAcc) (Seymour, Daw,

Dayan, Singer, & Dolan, 2007), and amygdala (Breiter et al., 2001; Yacubian et al., 2006). (These analyses were intended to bring greater statistical power to bear on these regions, in part because their small size may have undermined our ability to detect activation in them in our whole-brain analysis, where a cluster-size threshold was employed.) The habenular complex was found to display greater activity following type D than type E jumps ($p < 0.05$), consistent with the idea that this structure is also engaged by negative PPEs. A comparable effect was also observed in the right, though not left, amygdala ($p < 0.05$).

In the NAcc, where some studies have observed deactivation accompanying negative RPEs (Knutson, Taylor, Kaufman, Peterson, & Glover, 2005), no significant PPE effect was observed. However, it should be noted that NAcc deactivation with negative RPEs has been an inconsistent finding in previous work (for example, see J. C. Cooper & Knutson, 2008; O'Doherty, Buchanan, Seymour, & Dolan, 2006). More robust is the association between NAcc activation and positive RPEs (Hare et al., 2008; Niv, 2009; Seymour et al., 2004). To test this directly, we ran a second, smaller fMRI study designed to elicit positive PPEs, specifically looking for activation within a NAcc ROI. A total of 14 participants performed the delivery task, with jumps of type C (in Figure 4.2) occurring on one-third of trials and jumps of type E on another third. As described earlier, a positive PPE is predicted to occur in association with type C jumps, and in this setting significant activation ($p < 0.05$) was observed in the right (though not left) NAcc, scaling with predicted PPE magnitude.

4.2.3 Behavioral experiment

We have characterized the results from our EEG and fMRI experiments as displaying a "signature" of HRL, in the sense that the PPE signal is predicted by HRL but not by standard RL algorithms (Figure 4.2). However, there is an important

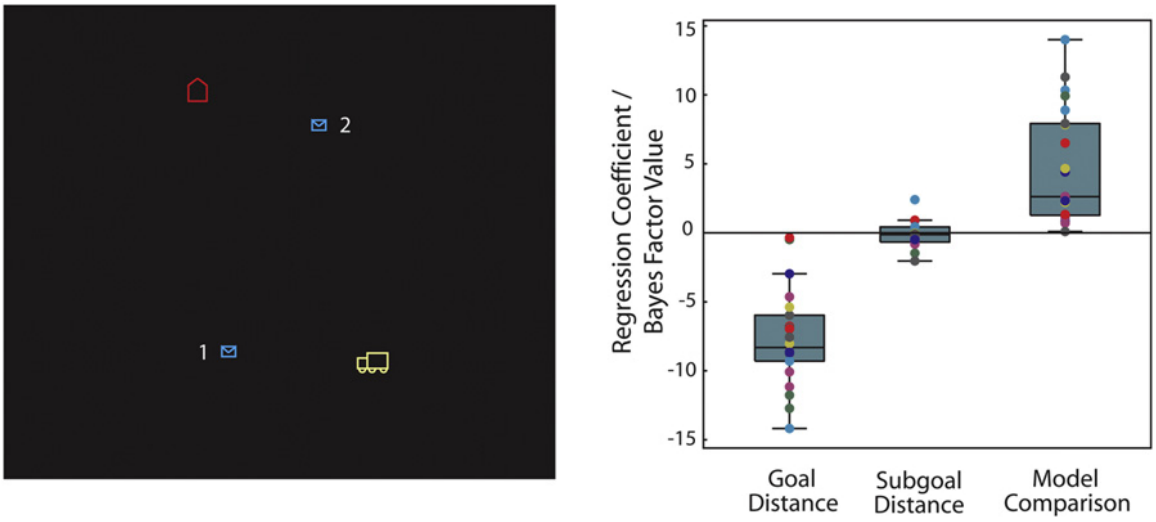

Figure 4.5: Results of Behavioral Experiment. Left view is an example of a choice display. Subgoal 1 would always be on an ellipse defined by the house and the truck. In this example subgoal 2 has smaller overall distance to the goal and larger distance to the truck relative to subgoal 1 (labels not shown to participants). Right view shows results of logistic regression on choices and of the comparison between two RL models. Choices were driven significantly by the ratio of distances of the goal of the two subgoals (left box, central mark is the median, edges correspond to 25th and 75th percentiles, whiskers to extreme values, outliers to individual dots outside box and whiskers; each colored dot represents a single participant's data), whereas the ratio of distances to subgoal did not significantly explain participant's choices (middle box). Bayes factors favored the model with only reward for goal attainment and no reward for subgoal against the one with reward for subgoal and goal attainment (right box).

caveat that we now consider. In our neuroimaging experiments we assumed that reaching the goal (the house) would be associated with primary reward. (The same points hold if "primary reward" is replaced with "secondary" or "conditioned reinforcement.") We also assumed that reaching the subgoal (the package) was not associated with primary reward but only with pseudo-reward. However, what if participants did attach primary reward to the subgoal? If this were the case, it would present a difficulty for the interpretation of our neuroimaging results because it would lead standard RL to predict an RPE in association with events that change only subgoal distance (including C and D jumps in our neuroimaging task).

In view of these points, it was necessary to establish whether participants performing the delivery task did or did not attach primary reward to subgoal attainment. In order to evaluate this, we devised a modified version of the task. Here, 22 participants delivered packages as before, though without jump events. However, at the beginning of each delivery trial, two packages were presented in the display, which defined paths that could differ both in terms of their subgoal distance and the overall distance to the goal (Figure 4.5, left). Participants indicated with a key press which package they preferred to deliver.

We reasoned that if goal attainment were associated with primary reward, then (assuming ordinary temporal discounting) the overall goal distance associated with each of the two packages should influence choice. More importantly, if we were correct in our assumption that subgoal attainment carried no primary reward, then choice should not be influenced by subgoal distance, i.e., the distance from the truck to each of the two packages.

Participants' choices strongly supported both of these predictions. Logistic regression analyses indicated that goal distance had a strong influence on package choice ($M = -7.6$, $p < 0.001$; Figure 4.5, right; larger negative coefficients indicate a larger penalty on distances). However, subgoal distance exerted no appreciable influence on choice ($p = 0.43$), and the average regression coefficient was near zero (-0.16). The latter observation held even in a subset of trials where the two delivery options were closely matched in terms of overall distance (with ratios of overall goal distance between 0.8 and 1.2).

These behavioral results strongly favor our HRL account of delivery task, over a standard RL account. (The behavioral data are consistent with a standard RL model that attaches no reward to subgoal attainment, but as noted earlier, such a model offers no explanation for our neuroimaging results.) To further establish the point, we fit two computational models to individual subjects' choice data: (1) an

HRL model, and (2) a standard RL model in which primary reward was attached to the subgoal (see Experimental Procedures). The mean Bayes factor across subjects–with values greater than one favoring the HRL model–was 4.31, and values across subjects differed significantly from one (two-tailed t test, $p < 0.001$; see Figure 4.5, right).

4.3 Discussion

We predicted, based on HRL, that neural structures previously proposed to encode TD RPEs should also respond to PPEs–prediction errors tied to behavioral subgoals. Across three experiments using a task designed to elicit PPEs, without eliciting RPEs, we observed evidence consistent with this prediction. Negative PPEs were found to engage three structures previously reported to show activation with negative RPEs: ACC, habenula, and amygdala; and activation scaling with positive PPEs was observed in right NAcc, a location frequently reported to be engaged by positive RPEs.

Of course the association of these neural responses with the relevant task events does not uniquely support an interpretation in terms of HRL (see Poldrack, 2006). However, aspects of either the task or the experimental results do militate against the most tempting alternative interpretations. Our behavioral study provided evidence against primary reward at subgoal attainment, closing off an interpretation of the neuroimaging data in terms of standard RL. Given previous findings pertaining to the ACC, the effect we observed in this structure might be conjectured to reflect response conflict or error detection (Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Krigolson & Holroyd, 2006; Yeung, Botvinick, & Cohen, 2004). However, additional analyses of the EEG data (see Figure 4.7 and Supplemental Experimental Procedures) indicated that the PPE effect persisted even after con-

trolling for response accuracy and for response latency, each commonly regarded as an index of response conflict.

Another alternative that must be addressed relates to spatial attention. Jump events in our neuroimaging experiments presumably triggered shifts in attention, often complete with eye movements, and it is important to consider the possibility that differences between conditions on this level may have contributed to our central findings. Although further experiments may be useful in pinning down the precise role of attention in our task, there are several aspects of the present results that argue against an interpretation based purely on attention. Note that, in previous EEG research, exogenous shifts of attention have been associated with a midline positivity, the amplitude of which grows with stimulus eccentricity (Yamaguchi, Tsuchiya, & Kobayashi, 1995). (A midline negativity has been reported in at least one study focusing on endogenous attention (Grent & Woldorff, 2007), but the timing of this potential differed dramatically from the difference wave in our EEG study, peaking at 1000-1200 ms poststimulus, hundreds of milliseconds after our effect ended.) In fact we observed such a positivity in our own data, in Cz, when we compared jump events (D and E) against occasions where the subgoal stayed put, an analysis specifically designed to uncover attentional effects (Figure 4.8). In contrast the PPE effect in our data took the form of a negative difference wave (Figure 4.3), consistent with the predictions of HRL and contrary to those proceeding from previous research on attention.

Our fMRI results also resist an interpretation based on spatial attention alone. As detailed in the Supplemental Experimental Procedures, we did find activation in or near the frontal eye fields and in the superior parietal cortexregions classically associated with shifts of attention (Corbetta, Patel, & Shulman, 2008)–in an analysis contrasting all jump events with trials where the subgoal remained in its original

location (Figure 4.9). However, as reported above, activity in these regions did not show any significant correlation with our PPE regressor (Figure 4.4).

If one does adopt an HRL-based interpretation of the present results, then several interesting questions follow. Given the prevailing view that TD RPEs are signaled by phasic changes in dopaminergic activity (Schultz et al., 1997), one obvious question is whether the PPE might be signaled via the same channel. ACC activity in association with negative RPEs has been proposed to reflect phasic reductions in dopaminergic input (Holroyd & Coles, 2002), and the habenula has been proposed to provide suppressive input to midbrain dopaminergic nuclei (Christoph, Leonzio, & Wilcox, 1986; M. Matsumoto & Hikosaka, 2007). Thus, the implication of the ACC and habenula in the present study, as well as the involvement of the NAcc – another structure that has been proposed to show activity related to dopaminergic input (Nicola, Surmeier, & Malenka, 2000) – provides tentative, indirect support for dopaminergic involvement in HRL. At the same time, it should be noted that some ambiguity surrounds the role of dopamine in driving reward-outcome responses, particularly within the ACC (for a detailed review, see Jocham & Ullsperger, 2009). Indeed, some disagreement still exists concerning whether the dorsal ACC is responsible for generating the FRN (compare Holroyd et al., 2004; Nieuwenhuis et al., 2005; van Veen, Holroyd, Cohen, Stenger, & Carter, 2004). Thus, the present findings must be interpreted with appropriate circumspection. Above all, it should be noted that our HRL-based interpretation does not necessarily require a role for dopamine in generating the observed neural events. Indeed, if the PPE were conveyed via phasic dopaminergic signaling, this would give rise to an interesting computational problem because proper credit assignment would require discrimination between PPE and RPE signals (for discussion, see Botvinick, Niv, & Barto, 2009).

Another important question for further research concerns the relation between the present findings and recent data concerning the representation of action hierarchies in the dorsolateral prefrontal cortex (Badre, 2008; Botvinick, 2008). Neuroimaging and neuropsychological studies have lately given rise to the idea that the prefrontal cortex may display a rostrocaudal functional topography, which separates out task representations based on some measure of abstractness (Badre, Hoffman, Cooney, & D'Esposito, 2009; Christoff, Keramatian, Gordon, Smith, & Mädler, 2009; Grafman, 2002; Kouneiher, Charron, & Koechlin, 2009). One speculation, which could be tested through further research, is that HRL-like mechanisms might be responsible for shaping such representations and gating them into working memory in an adaptive fashion (see Botvinick, Niv, & Barto, 2009; Reynolds & O'Reilly, 2009).

One final challenge for future research is to test predictions from HRL in settings involving learning-driven changes in action selection. As in many neuroscientific studies focusing on RL mechanisms, our task looked at prediction errors in a setting where behavioral policies were more or less stable. It may also prove useful to study the dynamics of learning in hierarchically structured tasks, as a further test of the relevance of HRL to neural function (see Badre & Frank, 2012; Diuk et al., 2010)

4.4 Experimental procedures

4.4.1 An HRL model of the delivery task

To make our computational predictions explicit, we implemented both a standard and a hierarchical RL model of the delivery task, based on the approach laid out in Botvinick, Niv, and Barto (2009). Simulations were performed in MATLAB (The MathWorks, Natick, MA); the relevant code is available for download from `http://www.princeton.edu/~matthewb`.

For the standard RL agent, the state on each step t, labeled s_t, was represented by the goal distance (gd), the distance from the truck to the house, via the package, in units of navigation steps. For the HRL agent the state was represented by two numbers: gd and the subgoal distance (sd), i.e., the distance between the truck and the package. Goal attainment yielded a reward (r) of one for both agents, and subgoal attainment a pseudo-reward (ρ) of one for the HRL agent. On each step of the task, the agent was assumed to act optimally, i.e., to take a single step directly toward the package or, later in the task, toward the house. The HRL agent was assumed to select a subroutine (σ) for attaining the package, which also resulted in direct steps toward this subgoal (for details of subtask specification and selection, see Figure 4.1 and Botvinick, Niv, & Barto, 2009; Sutton et al., 1999).

For the standard RL agent, the state value at time t, $V(t)$, was defined as γ^{gd}, using a discount factor $\gamma = 0.9$. Thus, the RPE on steps prior to goal attainment was:

$$RPE = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) = \gamma^{1+gd_{t+1}} - \gamma^{gd_t}. \tag{4.1}$$

The HRL agent calculated RPEs in the same manner but also calculated PPEs during execution of the subroutine σ. These were based on a subroutine-specific value function (see Botvinick, Niv, & Barto, 2009; Sutton et al., 1999), defined as $V_\sigma(s_t) = \gamma^{sd_t}$.

Thus, the PPE on each step prior to subgoal attainment was:

$$PPE = \rho_{t+1} + \gamma V_\sigma(s_{t+1}) - V_\sigma(s_t) = \gamma^{1+sd_{t+1}} - \gamma^{sd_t}. \tag{4.2}$$

To generate the data shown in Figure 4.2, we imposed initial distances (gd, sd) equaling 949 and 524. Following two task steps in the direction of the package, at a point with distances 849 and 424, in order to represent jump events distances were

changed to 599 and 424 for jump type A, 1449 and 424 for type B, 849 and 124 for type C, 849 and 724 for type D, and 849 and 424 for type E. Dashed data series in Figure 4.2 were generated with jumps to 849 and 236 for type C and 849 and 574 for type D.

4.4.2 EEG experiment

Participants

All experimental procedures were approved by the Institutional Review Board of Princeton University. Participants were recruited from the university community, and all gave their informed consent. Nine participants were recruited (ages 18-22 years, M = 19.7, 4 males, all right handed). All received course credit as compensation, and in addition received a monetary bonus based on their performance in the task.

Task and procedure

Participants sat at a comfortable distance from a shielded CRT display in a dimly lit, sound-attenuating, electrically shielded room. A joystick was held in the right hand (Logitech International, Romanel-sur-Morges, Switzerland).

The computerized task was coded using MATLAB (The MathWorks) and the MATLAB Psychophysics Toolbox, version 3 (Brainard, 1997). On each trial, three display elements appeared: a truck, a package, and a house (Figure 4.6A). These objects occupied the vertices of a virtual triangle with vertices at pixel coordinates 0 and 180, 150 and 30, and 0 and 180, relative to the center of the screen (resolution 1024 x 768) but assuming a random new rotation and reflection at the onset of each trial. The task was to move the truck first to the package and then to the house. Each joystick movement displaced the truck a fixed distance of 50 pixels. For

reasons given below the orientation of the truck was randomly chosen after every such translation, and participants were required to tailor their joystick responses to the truck's orientation, as if they were facing its steering wheel (Figure 4.6A). For example if the front of the truck were oriented toward the bottom of the screen, rightward movement of the joystick would move the truck to the left. This aspect of the task was intended to ensure that intensive spatial processing occurred at each step of the task, rather than only following subgoal displacements.

Responses were registered when the joystick was tilted beyond half its maximum displacement (Figure 4.6A). Between responses the participant was required to restore the joystick to a central position (Figures 4.6A and 4.6B). When the truck passed within 30 pixels of the package, the package moved inside the truck icon and remained there for subsequent moves. When the truck containing the package passed within 35 pixels of the house, the display cleared, and a message reading "10¢" appeared for a duration of 300 ms (participants were paid their cumulative earnings at the end of the experiment). A central fixation cross then appeared for 700 ms before the onset of the next trial.

On every trial, after the first, second, or third truck movement, a brief tone occurred, and the package flashed for an interval of 200 ms, during which any joystick inputs were ignored. On one-third of such occasions, the package remained in its original location. On the remaining trials, at the onset of the tone, the package jumped to a new location. In half of such cases, the distance between the package's new position and the truck position was unchanged by the jump (case E in Figure 4.2 of the main text). In the remaining cases the distance from the truck to the package was increased by the jump, although the total distance from the truck to the house (via the package) remained the same (case D in Figure 4.2). In these cases the jump always carried the package across an imaginary line connecting the truck and the house, and always resulted in a package-to-house distance of 160 pixels. In all

three conditions the package would be on an ellipse defined by the locations of the old subgoal, the house, and the position of the truck at the time of the jump. By the definition of an ellipse, overall distance to the house was preserved.

At the outset of the experiment, each participant completed a 15 min training session, which was followed by the hour-long EEG testing session. Participants completed 190 trials on average (range 128-231). Trials were grouped into blocks, each containing six trials: two trials in which the position of the package did not change, two involving type E jumps, and two type D jumps. The order in which trials of a particular type occurred was pseudorandom within a block. Participants were given an opportunity to rest for a brief period between task blocks.

Data acquisition

EEG data were recorded using Neuroscan (Charlotte, NC) caps with 128 electrodes and a Sensorium (Charlotte, VT) EPA-6 amplifier. The signal was sampled at 1000 Hz. All data were referenced online to a chin electrode, and after excluding bad channels were rereferenced to the average signal across all remaining channels (Hestvik, Maxfield, Schwartz, & Shafer, 2007). EOG data were recorded using a single electrode placed below the left eye. Ocular artifacts were detected by thresholding a slow-moving average of the activity in this channel, and trials with artifacts were not included in the analysis. Less than four trials per subject matched this criterion and were excluded from the analysis (less than two per condition).

Data analysis

Epochs of 1000 ms (200 ms baseline) were extracted from each trial, time locked to the package's change in position. The mean level of activity during the baseline interval was subtracted from each epoch. Trials containing type D jump were separated from trials containing jumps of type E, and ERPs were computed for

each condition and participant by averaging the corresponding epochs. The ERPs shown in Figure 4.3 (main text) were computed by averaging across participants.

The PPE effect was quantified in electrode Cz (following Holroyd & Coles, 2002). The PPE effect was quantified for each subject by taking the mean voltage during the time window from 200 to 600 ms following each jump, for the two jump types. A one-tailed paired t test was used to evaluate the hypothesis that type D jumps elicited a more negative potential than type E jumps. For comparability with previous studies, topographic plots are shown for electrodes FP1, FP2, AFz, F3, Fz, F4, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, O1, Oz, and O2 (as in Yeung et al., 2005, F7 and F8 were an exception, given that the used cap did not have these electrode locations).

4.4.3 fMRI experiments

Participants

Participants were recruited from the university community and all gave their informed consent. For the first fMRI experiment, 33 participants were recruited (ages 18-37 years, M = 21.2, 20 males, all right handed). Three participants were excluded: two because of technical problems and one who was unable to complete the task in the available time. For the second experiment, 15 participants were recruited (ages 18-25 years, M = 20.5, 11 males, all were right handed). One participant was excluded for failure to complete the task in the available time. All participants received monetary compensation at a departmental standard rate. Participants in the second experiment also received a small monetary bonus based on task performance.

Task and procedure

An MR-compatible joystick (MagConcept, Redwood City, CA) was used. The task was identical to the one used in the EEG experiment, with the following exceptions. For the first experiment initial positions of the icons were randomly assigned to the screen respecting a minimal distance of 150 pixels between icons. For the second experiment initial positions of the icons were rotations or reflections, varied randomly, of a preestablished arrangement of icons of a predetermined triangle with vertices truck (0, 200), package (151, −165), and house (0, −200) (coordinates are in pixels, referenced to the center of the screen). On type D jumps, the destination of the package was chosen randomly from all locations satisfying the conditions that they (1) increase truck-to-package distance, but (2) leave total path length to the goal (house) unchanged. The forced delay involved in the task interruption (tone, package flashing) totaled 900 ms. At the completion of each delivery, the message "Congratulations!" was displayed for 1000 ms (Figure 4.6D), followed by a fixation cross that remained on screen for 6000 ms.

The first fMRI experiment consisted of three parts: a 15 min behavioral practice outside the scanner, an 8 min practice inside the scanner during structural scan acquisition, and a third phase of approximately 45 min, where functional data were collected. During functional scanning, 90 trials were completed, in 6 runs of 15 trials each. At the beginning and end of each run, a central fixation cross was displayed for 10,000 ms. The average run length was 7.5 min (range 5.7-11).

The task and procedure in the second fMRI experiment were identical to those in the first, with the following exceptions. Type D jumps were replaced with type C jumps (see Figure 4.2 in the main text). In these cases, the distance between truck and package always decreased to 120 pixels. The message "10¢" appeared for 500 ms, indicating the bonus earned for that trial. Immediately following this, a

fixation cross appeared for 2500 ms, followed by onset of the next trial. The average run length was 6.8 min (range 4.7-10.7).

Image acquisition

Image acquisition protocols were the same for both experiments. Data were acquired with a 3 T Siemens Allegra (Malvern, PA) head-only MRI scanner, with a circularly polarized head volume coil. High-resolution (1 mm^3 voxels) T1-weighted structural images were acquired with an MP-RAGE pulse sequence at the beginning of the scanning session. Functional data were acquired using a high-resolution echo-planar imaging pulse sequence ($3 \times 3 \times 3$ mm voxels, 34 contiguous slices, 3 mm thick, interleaved acquisition, TR of 2000 ms, TE of 30 ms, flip angle 90°, field of view 192 mm, aligned with the anterior commissure-posterior commissure plane). The first five volumes of each run were ignored.

Data analysis

Data analysis was similar for both experiments. Data were analyzed using AFNI software (Cox, 1996). The T1-weighted anatomical images were aligned to the functional data. Functional data were corrected for interleaved acquisition using Fourier interpolation. Head motion parameters were estimated and corrected allowing six-parameter rigid body transformations, referenced to the initial image of the first functional run. A whole-brain mask for each participant was created using the union of a mask for the first and last functional images. Spikes in the data were removed and replaced with an interpolated data point. Data were spatially smoothed until spatial autocorrelation was approximated by a 6 mm FHWM Gaussian kernel. Each voxel's signal was converted to percent change by normalizing it based on intensity. The mean image for each volume was calculated and used later as baseline regressor in the general linear model, except in the

ROI analysis where the mean image of the whole brain was not subtracted from the data. Anatomical images were used to estimate normalization parameters to a template in Talairach space (Talairach & Tournoux, 1988), using SPM5 (`http://www.fil.ion.ucl.ac.uk/spm/`). These transformations were applied to parameter estimates from the general linear model.

General linear model analysis

For each participant we created a design matrix modeling experimental events and including events of no interest. At the time of an experimental event, we defined an impulse and convolved it with a hemodynamic response. The following regressors were included in the model: (a) an indicator variable marking the occurrence of all auditory tone/package flash events; (b) an indicator variable marking the occurrence of all jump events (spanning jump types E and D in Experiment 1 and types E and C in Experiment 2); (c) an indicator variable marking the occurrence of type D jumps (C jumps in Experiment 2); (d) a parametric regressor indicating the change in distance to subgoal induced by each D (or C) jumps, mean centered; (e and f) indicator variables marking subgoal and goal attainment; and (g) an indicator variable marking all periods of task performance, from the initial presentation of the icons to the end of the trial. Also included were head motion parameters, and first- to third-order polynomial regressors to regress out scanner drift effects. In Experiment 1, a global signal regressor was also included (comparable analyses omitting the global signal regressor yielded statistically significant PPE effects in the ACC, bilateral insula, and lingual gyrus, in locations highly overlapping with those reported in the main text).

Group analysis (Experiment 1)

For each regressor and for each voxel, we tested the sample of 30 subject-specific coefficients against zero in a two-tailed t test. We defined a threshold of $p = 0.01$ and applied correction for multiple comparison based on cluster size, using Monte Carlo simulations as implemented in AFNI's AlphaSim. We report results at a corrected $p < 0.01$.

Follow-up analysis (Experiment 1)

Our experimental prediction related to the change in distance between truck and package induced by type D-jump events, i.e., the change in distance to subgoal, or PPE effect. However, jump events also varied in the degree to which they displaced the package (i.e., the distance from its original position to its post-jump position), and this distance correlated moderately with the increase in subgoal distance. Therefore, it was necessary to evaluate whether the regions of activation identified in our primary GLM analysis might simply be responding to subgoal displacement (and possible attendant visuospatial or motor processes), rather than the increase in distance to subgoal. To this end, we looked at each area identified in the primary GLM, asking whether the area continued to a show significant PPE effect even after this regressor was made orthogonal to subgoal displacement. In order to avoid bias in this procedure, we employed a leave-one-out cross-validation approach, as follows. For every subgroup of 29 participants (from the total sample of 30), we reran the original GLM, identifying voxels that: (1) showed the PPE effect at significance threshold of $p = 0.05$ (cluster-size thresholded to compensate for multiple comparisons); and (2) fell within 33 mm of the peak-activation coordinates for one of the six clusters identified in our primary GLM (dorsal anterior cingulate, bilateral anterior insulae, left lingual gyrus, left inferior frontal gyrus, and right supramarginal gyrus). The resulting clusters were used as ROIs for the critical test.

Focusing on the one subject omitted from each 29 subject subsample, we calculated the mean coefficient within each ROI for the PPE effect, after orthogonalizing the PPE regressor to subgoal displacement (and including subgoal displacement in the GLM). This yielded 30 coefficients per ROI. Each set was tested for difference from zero, using a two-tailed t test.

ROI analysis

For the first fMRI experiment, we defined NAcc based on anatomical boundaries on a high-resolution T1-weighted image for each participant; habenula, using peak Talairach coordinates (5, 25, 8), guided by Ullsperger and von Cramon (2003), surrounded by a sphere with a radius of 6 mm (Salas et al., 2010); and amygdala, drawn using the Talairach atlas in AFNI. For the second experiment we defined NAcc in the same way as for the first experiment. Mean coefficients were extracted from these regions for each participant. Reported coefficients for all ROIs are from general linear model analyses without subtraction of global signal. The sample of 30 (or 14 for the second experiment) subject-specific coefficients was tested against zero in a two-tailed t test, with a threshold of $p < 0.05$.

4.4.4 Behavioral experiment

Participants

A total of 22 participants were recruited from the Princeton University community (ages 18–22 years, 11 male). All provided informed consent and received a nominal payment. Task and Procedure The experiment was composed of three phases. In the first phase, participants completed ten deliveries, with the procedure matching that used in our fMRI studies. However, no jump events occurred in this or later phases of the experiment. The second phase consisted of ten further delivery

trials. However, here, at the onset of each trial, the participant was required to choose between two packages (Figure 4.5). The location of the truck and the house was chosen randomly. The location of one package, designated subgoal one, was randomly positioned along an ellipse with the truck and house as its foci and a major-to-minor axis ratio of 5/3. The position of the other package, subgoal two, was randomly chosen, subject to the constraint that it fall at least 100 pixels from each of the other icons.

At the onset of each trial, each package would be highlighted with a change of color, twice (in alternation with the other package), for a period of 1.5 s. Highlighting order was counterbalanced across trials. During this period the participant was required to press a key to indicate his or her preferred package when that package was highlighted. After the key press, the chosen subgoal would change to a new color. At the end of the choice period, the unchosen subgoal was removed, and participants were expected to initiate the delivery task. The remainder of each trial proceeded as in phase one.

The third and main phase of the experiment included 100 trials. One-third of these, interleaved in random order with the rest, followed the profile of phase two trials. The remaining trials began as in phase two but terminated immediately following the package-choice period.

Data analysis

To determine the influence of goal and subgoal distance on package choice, we conducted a logistic regression on the choice data from phase three. Regressors included (1) the ratio of the distances from the truck to subgoal one and subgoal two, and (2) the ratio of the distances from the truck to the house through subgoal one and subgoal two. To test for significance across subjects, we carried out a two-tailed t test on the population of regression coefficients.

To further characterize the results, we fitted two RL models to each participant's phase three choice data. One model assigned primary reward only to goal attainment and so was indifferent to subgoal distance per se. A second model assigned primary reward to the subgoal as well to the goal.

Value in the first case was a discounted number of steps to the goal, and in the second case it was a sum of discounted number of steps to the subgoal and to the goal. Choice was modeled using a softmax function, including a free inverse temperature parameter. The fmincon function in MATLAB was employed to fit discount factor and inverse temperature parameters for both models and reward magnitude for subgoal attainment for the second model. We then compared the fits of the two models calculating Bayes factor for each participant and performing a two-tailed t test on the factors.

4.5 Supplemental information

4.5.1 EEG experiment

Controlling for response conflict. As noted in the main text, it was important to evaluate whether the ERP effect observed might reflect error or response conflict detection, factors that have been shown in previous studies to induce phasic midline negativities (Botvinick et al., 1999; Krigolson & Holroyd, 2006; Yeung et al., 2004). To rule out an explanation in terms of error-detection, we conducted an analysis that excluded trials where errors occurred. Although there is no discrete criterion for response corrections in the task, it is possible to distinguish between highly accurate and less accurate responses. We defined response accuracy in terms of the angle between the perfect joystick movement (the movement that would have taken the truck directly toward the package) and the actual movement, setting an upper bound of 22.5° for highly accurate responses, based on an inspection of

the response distribution (Figure 4.7A). For clarity, we also only considered trials where the package displacement required a change in the truck path spanning at least 45°. Repeating our original ERP analysis, focusing only on trials involving highly accurate responses, yielded the ERP data shown in Figure 4.7B. As in the original analysis, the difference between jumps of type D and E was significant ($p = 0.019$).

The other alternative explanation we wished to evaluate was related to conflict detection. It was possible that type D jumps caused greater response conflict than type E, perhaps because a greater time was needed to pin down the direction to the new package location (more distant in case D than E). In order to test this explanation, we adopted the common approach of treating reaction time (RT) as an index of conflict. Considering only data from trials with highly accurate responses, mean RT in condition D (1013 ms) did not differ significantly from mean RT for condition E (1049 ms, paired two-tailed t-test, $p = 0.39$). In fact, unconditioned on accuracy, mean RT following type D jumps (849 ms) was smaller than that following type E jumps (926 ms, paired two-tailed t-test, $p < 0.01$), further militating against an explanation based on conflict. RT distributions for responses immediately following type D and E jumps (collapsing across participants) are shown in Figure 4.7C. RTs in both conditions displayed a clear bimodal distribution, and the difference in mean RT could be largely attributed to a difference in the proportion of fast (and relatively inaccurate) responses versus that of slower (and more accurate) responses. To control for RT, we limited consideration to responses that fell within the slower component of the bimodal distribution in both conditions. The mean RT within the resulting samples (1077 ms for type D, 1075 ms for type E) did not differ significantly across the two conditions (paired two-tailed t-test, $p = 0.98$), nor did the proportion of inaccurate responses, as defined earlier (49.98% for type D vs. 55.10% for type E, paired two-tailed t-test $p = 0.08$). An ERP analysis

focusing on this matched data subset of slow responses yielded a robust PPE effect ($p = 0.02$, Figure 4.7D).

EEG correlates of shifts of attention. Figure 4.8 shows the electrode potential at Cz for conditions involving a shift of attention (average of conditions D and E) and the no jump condition.

4.5.2 fMRI experiment 1

As reported in the main text, Experiment 1 yielded a significant parametric PPE effect in several regions. One additional aspect of the results that deserves comment is the fact that these same regions did not display a statistically significant categorical effect. That is, while their activation scaled with the magnitude of the subgoal-distance increase induced by type D jumps, the mean activation induced by type D jumps was not significantly greater than that induced by type E jumps. Two possible explanations can be offered for this aspect of the results. First, it should be noted that the average increase in subgoal distance across all trials in the experiment was well above zero. Taking this into account, on a precise HRL account, type E jumps should in fact have induced a small positive PPE. For simplicity, in deriving our experimental predictions, we assumed that the PPE was calculated against a reference or expected subgoal-distance change of zero. This difference between the assumptions of our model and a strict HRL account may at least partially account for the details of our GLM results. On a more prosaic level, it should be noted that, across trials, the increase in subgoal distance was heavily skewed to the right. This may have undermined power for detecting a mean effect of jump type, making it easier to detect the parametric effect that we in fact obtained in the main GLM analyses. Further experimentation is called for to evaluate the merit of these two interpretations.

As noted in the main text, the design of our neuroimaging experiments reflected a presumption that participants would represent and perform the delivery task in a hierarchical manner. However, as also intimated the main text (see Note 1), we also view our experimental results as providing evidence supporting that assumption. Specifically, our behavioral study provided evidence against a non-hierarchical or 'flat' RL account involving primary reward at subgoal attainment, and the EEG and fMRI results could not be easily explained by a flat RL account with no reward at subgoal.

One concern that may arise from this line of reasoning centers on our interpretation of the neuroimaging data. Obviously, our neuroimaging experiments were designed to test for a neural correlate of the PPE. However, given that the PPE is assumed to arise from hierarchical processing, it may appear necessary for us to have established *independent of the imaging experiments* that subjects represent the delivery task hierarchically. We have claimed that the imaging data provide evidence both for the PPE and for the logically prior proposition that the delivery task is performed hierarchically. Isn't there necessarily some circularity in this analysis?

Despite the appeal of this intuition, there is in fact nothing circular in our interpretation of the data. To show this formally, let us define the following terms:

A: The event that the task is represented hierarchically

$\overline{A}$: The event that the task is not represented hierarchically

B: The event that the task gives rise to a PPE

$\overline{B}$: The event that the task does not give rise to a PPE

D: Our neuroimaging findings

On purely logical grounds, it is clear that:

$B \Rightarrow A$: If B were true, then A would necessarily also be true

$\overline{A} \Rightarrow \overline{B}$: If A were false, then B would necessarily also be false

Given these two premises, basic probability yields the following two conclusions:

$$p(B|D) = \frac{p(D|B)p(B|A)p(A)}{p(D|B)p(B|A)p(A) + p(D|\overline{B})p(\overline{B})} \tag{4.3}$$

$$p(A|D) = \frac{p(D|B)p(B|A)p(A) + p(D|A \cap \overline{B})p(A \cap \overline{B})}{p(D|B)p(B|A)p(A) + p(D|\overline{B})p(\overline{B})} \tag{4.4}$$

Equation 4.1 gives the posterior probability of the PPE hypothesis, given the neuroimaging data. Equation 4.2 gives the probability of hierarchical processing, given those same data. Two points are worth noting. First, there is no circular or reciprocal dependency between the two equations.[1] Given the appropriate likelihoods and prior probabilities, the equations can be evaluated in parallel. It is thus logically sound to draw parallel conclusions from the imaging data concerning both hierarchical processing and the PPE.

Second, both probabilities depend inversely on $p(D|\overline{B})$, the probability that the data might have been obtained in the absence of a PPE. This indicates the importance of ruling out alternative explanations for theimaging results. It is here that the behavioral study comes in, since it rules out an interpretation of the imaging data based on primary reward at subgoal.

Naturally, both probabilities (1) and (2) also depend on $p(A)$, the a priori probability that the delivery task is performed hierarchically. Previous research makes it reasonable to consider this probability to be fairly high: As we have recently reviewed elsewhere (Botvinick, 2008; Botvinick, Niv, & Barto, 2009), decades of research in cognitive psychology (e.g., G. A. Miller, Galanter, & Pribram, 1960; R. P. Cooper & Shallice, 2000; Zacks, Speer, Swallow, Braver, & Reynolds, 2007),

[1]The two expressions do of course share terms, and will thus be correlated, but this is no indication of circularity or tautology. As an aside, also note that $p(B|D) = p(AB|D)$; our experiment may be seen as evaluating the *joint* hypothesis AB.

developmental psychology (e.g. Saffran & Wilson, 2003), neuropsychology (e.g., Schwartz et al., 1995; Badre et al., 2009), functional neuroimaging (e.g., Badre & D'Esposito, 2007; Koechlin, Ody, & Kouneiher, 2003), and neurophysiology (e.g., Fujii & Graybiel, 2003; Fuster, 2004) indicate that hierarchical representation is ubiquitous, and perhaps even obligatory in human behavior. The possibility that our experimental task, with its very salient goal-subgoal structure, might constitute an exception to this general rule seems improbable.

Nevertheless, the importance of the hierarchy assumption prompted us to consider whether our data might provide some additional, independent and convergent evidence for hierarchical processing. One opportunity, in this regard, is suggested by recent neurophysiological research, which has discovered phasic activity within the dorsolateral prefrontal cortex and dorsolateral striatum at sequence boundaries (Barnes et al., 2011; Fujii & Graybiel, 2003; Jin & Costa, 2010). We reasoned that, if participants in our experiment represented the delivery task hierarchically, such activity should occur at the point of subgoal attainment, since this marks the completion of one subsequence and he onset of another. Importantly, the moment of subgoal attainment in our task also requires a shift in visual attention; to control for this factor, we used package-jump events (pooling across jump-types E and D) as a baseline, since these events also require a shift in visual attention but do not lie at a subtask boundary. The resulting contrast revealed relative activation at subgoal attainment ($p < 0.01$, corrected as in previous analyses) at three points within dorsolateral prefrontal cortex (Talairach coordinates 63, 7, 25; −61, 4, 30; and −51, 40, 19) and bilaterally within dorsolateral striatum (15, −14, 25; −12, 11, 19). Relative activation was also observed in left anterior parietal cortex spanning the intraparietal sulcus, in the right precuneus, in bilateral middle occipital gyri, and in the cerebellum.

Interestingly, the prefrontal areas identified in this contrast lie near to areas identified in recent neuroimaging studies aimed at isolating regions responsible for instantiating hierarchical representations of action (Badre & D'Esposito, 2007; Koechlin et al., 2003). We refrain from drawing strong conclusions from this apparent correspondence, given the many differences between the task and analysis employed here and ones involved in those previous studies. However, the finding of phasic activation in these frontal regions at the subtask boundary within our task does appear to offer some convergent support for our assumption that participants represented the delivery task in a hierarchical fashion.

4.5.3 Supplemental figures

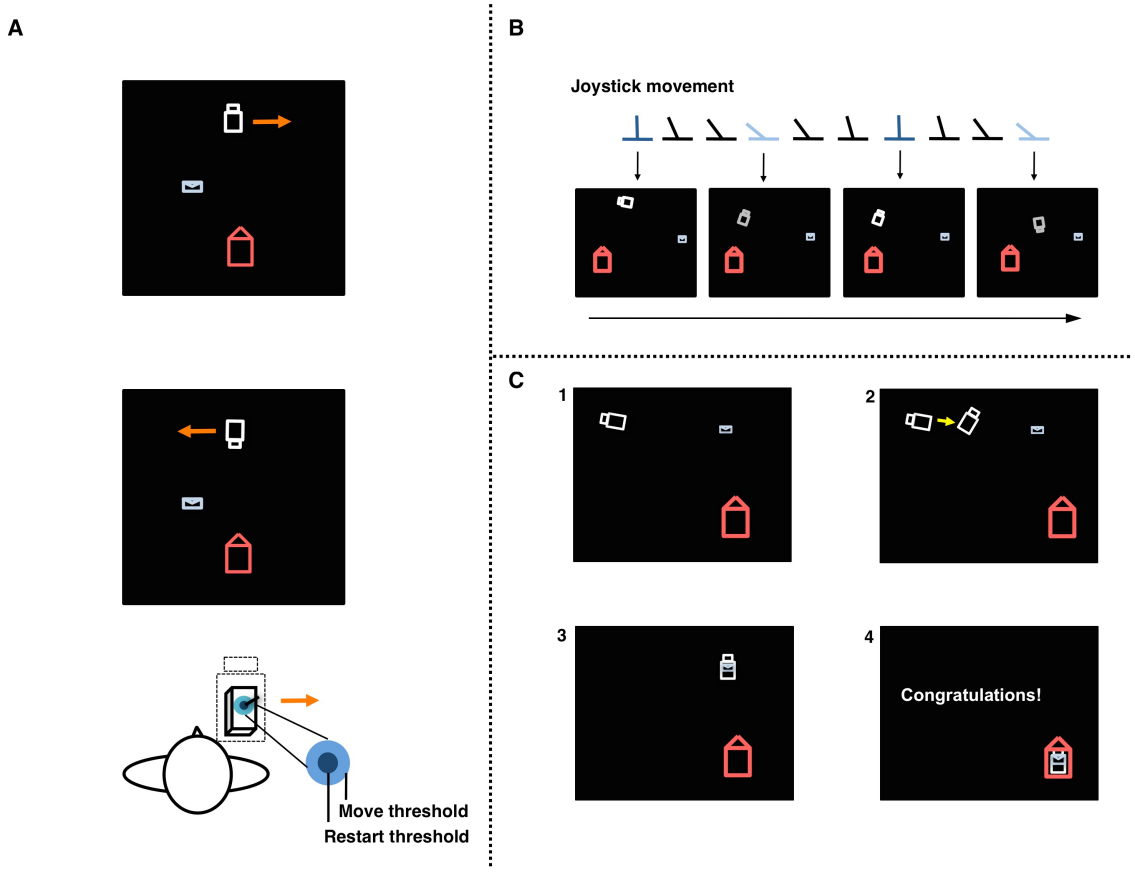

Figure 4.6: Task and relationship between joystick and truck movement. **A**. Illustration of the effect of a movement command to the right of the joystick – as shown in the figure, actual displacement on the screen depends on orientation of the truck; after every movement command the joystick had to be reset from the outer "Move threshold" (light blue) to the "Restart threshold" (dark blue). **B**. An example of two movement commands. **C**. Illustration of the initial arrangement of truck, package and house – 1; of the first movement – 2 –and of subgoal and goal attainment –3 and 4 respectively.

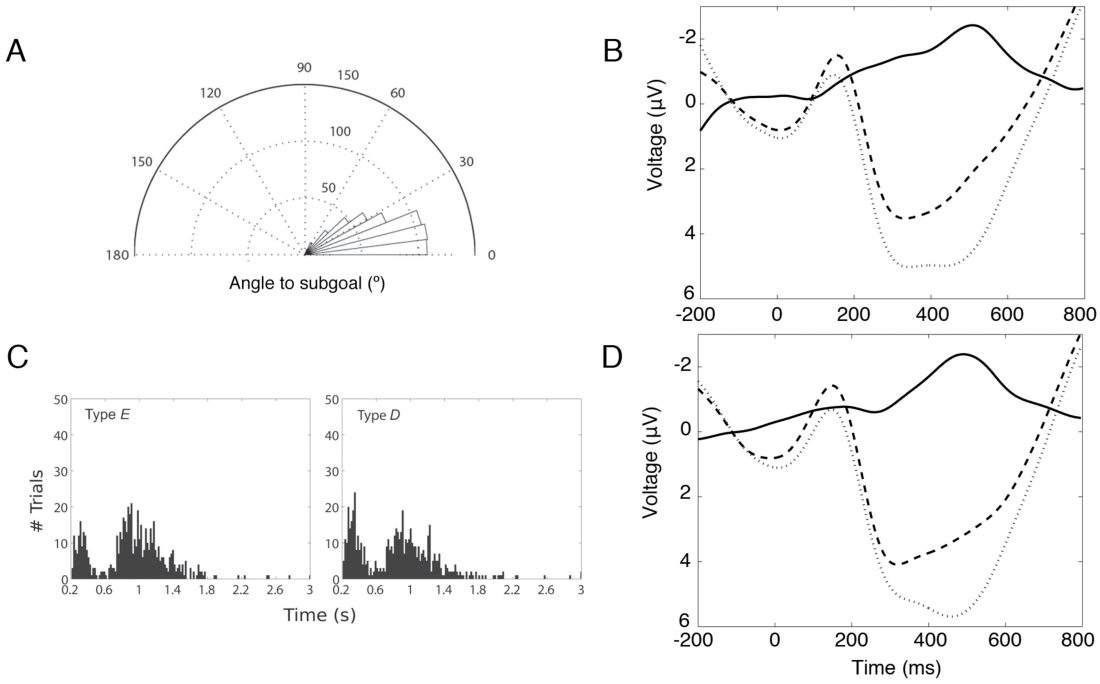

Figure 4.7: Accuracy, reaction times, and evoked potentials conditioned on these variables. **A**. Polar accuracy plot for the movement command before the subgoal jump. 0° is a perfect movement in the direction of the subgoal. Left and right commands are shown collapsed. **B**. Evoked potentials at electrode Cz, aligned to jump events and difference wave, conditioned on highly accurate responses. Dashed line corresponds to class *D* events, grey solid line to E events and the black solid to the difference D-E. **C**. Reaction time distributions for type *E* and *D* jumps. **D**. Evoked potentials at electrode Cz, aligned to jump events and difference wave, conditioned on slow responses. Dashed line corresponds to class *D* events, grey solid line to *E* events and the black solid to the difference *D*-*E*.

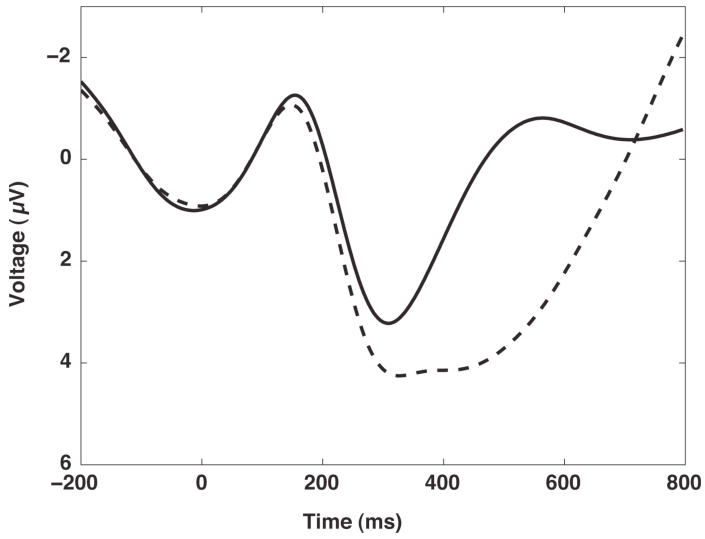

Figure 4.8: ERP for conditions involving a shift of attention (*E* and *D*; dashed line) and the condition with no jump (solid line) in electrode Cz. 0 ms is the moment when the package flashes yellow and a tone is played.

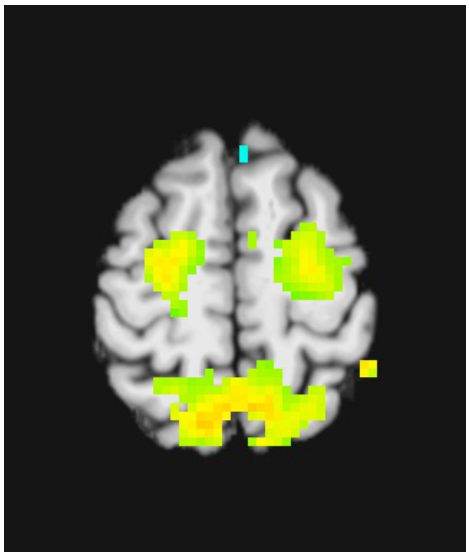

Figure 4.9: Axial view ($z = 53$) of the BOLD activity for events *D* and *E* ($p < 0.01$ corrected) contrasted with no jump condition. Talairach coordinates for the peak voxel for the clusters shown are (18, −66, 51) intraparietal sulcus, (−27, −12, 54) and (24, −15, 51) for frontal eye fields (see main text for discussion).

Chapter 5

Optimal behavioral hierarchy

Solway, A*., Diuk, C*., Cordova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M.

5.1 Abstract

Human behavior has long been recognized to display hierarchical structure: actions fit together into subtasks, which cohere into extended goal-directed activities. Arranging actions hierarchically has well established benefits, allowing behaviors to be represented efficiently by the brain, and allowing solutions to new tasks to be discovered easily. However, these payoffs depend on the particular way in which actions are organized into a hierarchy, the specific way in which tasks are carved up into subtasks. We provide a mathematical account for what makes some hierarchies better than others, an account that allows an *optimal* hierarchy to be identified for any set of tasks. We then present results from four behavioral experiments, suggesting that human learners spontaneously discover optimal action hierarchies.

*Alec Solway and Carlos Diuk contributed equally to this work.

5.2 Author summary

In order to accomplish everyday tasks, we often divide them up into subtasks: To make spaghetti, we (1) get out a pot, (2) fill it with water, (3) bring the water to a boil, and so forth. But how do we learn to subdivide our goals in this way? Work from computer science suggests that the way a task is subdivided or decomposed can have a dramatic impact on how easy the task is to accomplish: Certain subgoals will work better than others. Despite this general insight, little work has been done to pinpoint exactly what makes good subgoals good. We address that question here, by asking what might make a certain task decomposition *optimal*, i.e., better than all others? We provide a mathematical answer to this question, based on methods for comparing between statistical models. We then present four behavioral experiments, showing that human learners spontaneously discover optimal task decompositions.

5.3 Introduction

Since the earliest days of psychology and neuroscience, a core objective within both fields has been to understand the formal structure of behavior (Hebb, 1949; G. A. Miller et al., 1960; Tinbergen, 1951; Tolman, 1932). In pursuing this question, both in humans and in other animals, a crucial and recurring observation has been that behavior displays a hierarchical organization. Simple actions fit together into coherent subtasks, which themselves combine to accomplish higher-level goals (Botvinick, 2008; Shallice & Cooper, 2011). This kind of tiered or nested structure is readily apparent in our everyday activities: Turning on the stove forms part of boiling water, which in turn forms part of cooking pasta. It has also been quantified in detailed formal analyses of behavior, both in the laboratory and in the

field (Whiten, Flynn, Brown, & Lee, 2006; Zacks, Kurby, Eisenberg, & Haroutunian, 2011).

The ubiquity of hierarchical structure in behavior presumably reflects an adaptive benefit. Consistent with this, computational analyses have revealed at least two important advantages that can be gained by organizing behavior hierarchically. First, hierarchical representations of behavior can be more compact or efficient than non-hierarchical (flat) representations, allowing complex behaviors to be encoded more economically at the neural level (Graybiel, 1998). Second, hierarchical representations of action can facilitate the discovery of new adaptive behaviors, either through learning or through on-line planning (Barto, Konidaris, & Vigorito, 2013; Botvinick, Niv, & Barto, 2009; Foster & Dayan, 2002) or problem-solving (Anderson et al., 2004; Boutilier, Dean, & Hanks, 1999; Laird, 2012; Newell & Simon, 1972).

An illustration of this latter point is provided in Figure 5.1. The example centers on a artificial reinforcement learning agent (Sutton & Barto, 1998) that navigates from vertex to vertex in the grid shown in panel A. The agent must learn, through trial and error, to move from the start location highlighted in green to a rewarded goal location, highlighted in red. The black data-series in panel B charts the agent's improvement over successive trials. In contrast, the blue data-series tracks learning in a hierarchical reinforcement learning agent (Botvinick, Niv, & Barto, 2009; Sutton et al., 1999). This agent is furnished with subtask representations or subroutines for navigating to each of the "doorway" locations marked in blue in panel A (simulation code available online at `www.princeton.edu/~matthewb`). It can thus behave hierarchically, choosing among subroutines that in turn specify concrete, low-level actions. As is clear from the learning curve, the hierarchical agent converges on shortest-path behavior much more quickly than the flat agent.

As previously mentioned, a proper hierarchical representation can significantly aid not only learning, but also planning. In reinforcement learning (Sutton &

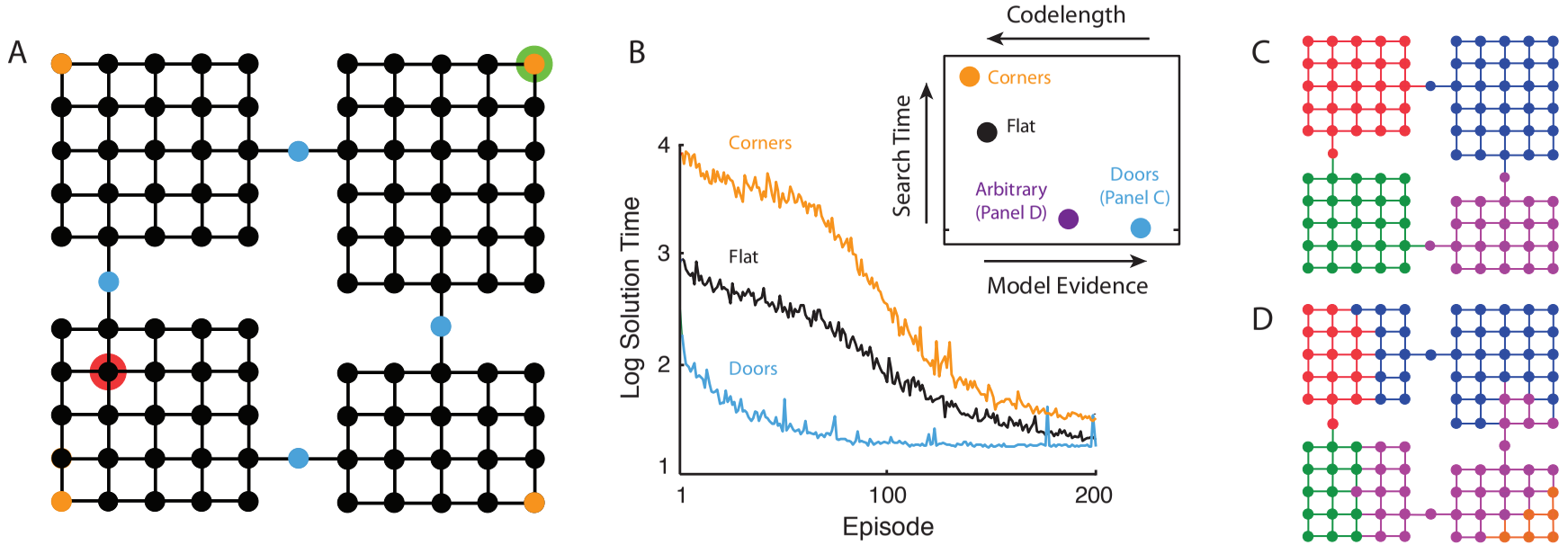

Figure 5.1: **A**. Rooms domain. Vertices represent states (green = start, red = goal), and edges feasible transitions. **B**. Mean performance of three hierarchical reinforcement learning agents in the rooms task. *Inset*: Results based on four graph decompositions. Blue: decomposition from panel C. Purple: decomposition from panel D. Black: entire graph treated as one region. Orange: decomposition with orange vertices in panel A segregated out as singleton regions. *Model evidence* is on a log scale (data range -7.00×10^4 to -1.19×10^5). *Search time* denotes the expected number of trial-and-error attempts to discover the solution to a randomly drawn task or subtask (geometric mean; range 685 to 65947; tick mark indicates the origin). *Codelength* signifies the number of bits required to encode the entire data-set under a Shannon code (range 1.01×10^5 to 1.72×10^5). **C**. Optimal decomposition. **D**. An alternative decomposition.

Barto, 1998), there is a strong connection between these two facets of behavior, and the relation is perhaps best exemplified by the Dyna framework (Sutton, 1990). More generally, planning proceeds by unrolling a model of the environment that consists of a transition function that dictates movements between states and a reward function that dictates how good or bad those movements are. In Dyna, this is accomplished by using the model to run *simulations*, and then learning from those simulations as if they were real. The "learning" curve in Figure 5.1B may thus also be seen as the number of simulation episodes that must be run in the service of planning for a given level of performance.

While this example illustrates the point that hierarchy can facilitate the discovery of new adaptive behaviors, there is an important caveat: Not all hierarchies are created equal. The wrong hierarchical representation can actually undermine adaptive behavior. This point is again illustrated in Figure 5.1B. The orange dataseries in the figure tracks the course of learning for a second hierarchical agent. This agent, like the one just considered, is furnished with a set of subroutines. However, here each subroutine involves navigating not to a doorway but into a corner (one of the locations highlighted in orange in panel A). In contrast to the doorway agent, this corner agent learns much more slowly than the flat agent. Obviously, it is not hierarchy *per se* that facilitates adaptive behavior. It matters very much which specific set of hierarchical representations an agent carries.

These observations bring to the surface a fundamental point concerning behavioral hierarchy: While hierarchy can facilitate learning, it also introduces a new learning problem, the problem of discovering beneficial rather than disruptive subtask representations.

Computational work in the area of hierarchical reinforcement learning has given rise to a number of approaches aimed at discovering useful behavioral hierarchies, leveraging ideas from information theory, graph theory, and developmental psy-

chology (Botvinick, Niv, & Barto, 2009; Barto et al., 2013; Simsek & Barto, 2008; van Dijk & Polani, 2011; Vigorito & Barto, 2010). For example, Simsek and Barto (2008) describe a method based on betweenness, a graph centrality metric which measures the fraction of shortest paths that go through each vertex of a graph. They construct what they call an interaction graph, representing possible state transitions, and compute a weighted betweenness metric that depends on the costs associated with each path. Local maxima, which often appear in "bottleneck" states (described further below), represent subgoal locations that can be utilized in hierarchical representations. Van Dijk and Polani (2011) take an information theoretic approach and define subgoals as states in which there is a significant change in the amount of relevant goal information, a measure of the amount of information that needs to be maintained about the goal at each step in order to perform well. Still other work has suggested that useful task decompositions might be learned through analyses of the causal structure of the environment, or via curiosity-driven learning mechanisms (Vigorito & Barto, 2010).

However, such work has never directly confronted a key underlying question: What exactly *should* the agent learn? Given that some hierarchies are better than others, can one specify for any given behavioral domain the best hierarchy overall? In other words, what would it mean for a behavioral hierarchy to be optimal? It is this question that we confront in the present work.

5.4 Results and discussion

5.4.1 Formal approach

In order to set the stage, we briefly introduce some additional terminology from the reinforcement learning literature. The goal of a reinforcement learning agent is to find a reward maximizing *policy*, a mapping from states to actions, in an

environment obeying certain Markovian dynamics. In particular, it is assumed the environment consists of a set of states, $\mathcal{S}$, a set of actions, $\mathcal{A}$, a transition function $\mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S} : Pr(S_{t+1} = s' | S_t = s, A_t = a)$, and a reward function $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathcal{R} : E[r_{t+1} | S_{t+1} = s', S_t = s, A_t = a]$, where E is expectation and r is scalar reward. There are several ways of incorporating hierarchy into reinforcement learning; we adopt the options framework approach (Sutton et al., 1999) in this paper. An option may be thought of as a temporally extended action and consists of: an initiation set containing the states from which it may be invoked, a termination function $\mathcal{S} \rightarrow [0, 1]$ specifying the probability of terminating the option in each state, and a policy. Once invoked, the agent's behavior is controlled by the option specific policy until it terminates, at which point the higher level policy again takes over. Options may also be nested, resulting in arbitrarily deep hierarchies. In this paper, we will use the terms option and subtask interchangeably. *Root-level* policy will refer to the policy at the top level (outside of all options), in contrast to option-level or subtask policies.

In any optimization problem, the crucial first step is to identify the objective. In the present case, this means asking: What exactly should an optimal hierarchy optimize? The rooms example in Figure 5.1 suggests a sensible answer to this question: An optimal hierarchy should maximize the efficiency with which an agent can discover new reward-maximizing behaviors. To make good on this idea, a method is needed for scoring or ranking candidate hierarchies on this property.

In order to solve this problem, we reframe it in terms of Bayesian model selection, where a set of candidate models are compared in their ability to account for a set of target data (MacKay, 2003). In the present case, the set of candidate models comprises all possible combinations of options the agent can be furnished with. The data, in turn, are a target set of optimal behaviors representing the solutions to an *ensemble* of tasks faced by the candidate agent. That is, the agent knows it

may face one of a variety of tasks, but it doesn't know which one, and it seeks a hierarchical representation that will be best *on average*. More concretely, a task is a sequence of states, and the data are composed of state–action pairs representing the optimal behaviors (policies) in those states. In what follows, we first describe how Bayesian model selection can be applied in this context. We then explain how model selection achieves the desired optimum, maximizing the ease with which adaptive behaviors can be discovered.

In Bayesian model selection, each candidate model is assumed to be associated with a set of parameters, and the fit between the model and the target data is quantified by the marginal likelihood or *model evidence*:

$$Pr(data|model) = \sum_{\theta \in \Theta} Pr(data|model, \theta) Pr(\theta|model), \tag{5.1}$$

where Θ is the set of feasible model parameterizations. In the present setting, where the models in question are different possible hierarchies, and the data are a set of target behaviors, the model evidence becomes:

$$Pr(behavior|hierarchy) = \sum_{\pi \in \Pi} Pr(behavior|hierarchy, \pi) Pr(\pi|hierarchy), \tag{5.2}$$

where Π spans the set of behavioral policies available to the candidate agent, given its inventory of subtask representations (this includes the root policy for each task in the target ensemble, as well as the policy for each subtask itself). Within the framework we propose, the optimal hierarchy is the one that maximizes the model evidence, as formulated in Equation 5.2.

In order to illustrate this approach, we consider an agent like the one in the rooms example from Figure 5.1: an agent whose actions equate to deterministic, reversible transitions between discrete states, visualizable as vertices in a graph. We assume, for concreteness, that the ensemble of tasks that the agent faces comprises

the set of all shortest-path problems within the graph. In order to build an inventory of subtask representations, the agent is permitted to decompose the graph into a set of connected components (see Figure 5.1), defining regions within the state-space of its environment. The agent is then furnished with a subtask representation for each available method of transitioning between regions (Hauskrecht, Meuleau, Kaelbling, Dean, & Boutilier, 1998, see Methods for further detail). For example, given the partitioning shown in Figure 5.1C, the rooms agent would obtain two subtask representations for each room, each with one doorway as its goal.

Applying Bayesian model selection under this problem formulation, the data to be modeled take the form of state–action pairs, where the states represent all of the shortest paths within the state-transition graph. In order to mark task boundaries, this concatenation is supplemented by a set of task-unique symbols, associated with indices specifying where each new task begins. The set of models (behavioral hierarchies) corresponds to the set of all possible decompositions of the graph. In this context, the model evidence assumes a surprisingly compact form:

$$Pr(behavior|hierarchy) = \prod_i (\bar{k}_i^T k_i^S)^{-1} \tag{5.3}$$

where i indexes vertex identifiers within the data; k_i is the degree of the vertex appearing as data element i; $\bar{k}_i$ is k_i plus the number of subtasks initiable at i; and T and S are indicator functions of i, assuming a value (1 or 0) that indicates whether each element constrains the agent's task-level action policy (T) or a subtask-level policy (S). As detailed under Methods and in the online supplement, each of the terms in Equation 5.3 can be quantified based strictly on the target data and the graph itself.

Figure 5.1B (inset) applies Equation 5.3 to the rooms domain, plotting the model evidence for four agent hierarchies. The hierarchy with the greatest evidence

corresponds to the partition shown in Figure 5.1C. This partition, with subgoals corresponding to the doors, in fact represents the optimal behavioral hierarchy in this particular domain. Another example is shown in Figure 5.2F. This shows the task graph for the Tower of Hanoi, a puzzle in which disks must be moved from a start arrangement to a goal arrangement, without ever placing any disk upon a smaller one. The optimal hierarchy for this task divides the state space into three regions, each corresponding to one position of the largest disk.

Crucially, by maximizing the model evidence, these hierarchies also turn out to satisfy our original desideratum, maximizing the agent's ability to efficiently discover target behaviors. Specifically, the optimal hierarchy minimizes the geometric mean number of trial-and-error attempts necessary for the agent to discover the optimal policy for any selected task or subtask (see Figure 5.1B, inset, for illustrative data). An explicit proof of this point is provided in the online supplement. However, the conclusion follows from the fact that every candidate hierarchy induces a probability distribution over behaviors (see Eq. 5.2), and that the optimal hierarchy, by definition, places the greatest probability mass on the agent's target behavior. This further implies that the optimal hierarchy will minimize the number of trials needed, on average, to discover the target behavior.

It also happens that the optimal hierarchy, by maximizing the model evidence, is guaranteed to minimize the expected number of information-theoretic bits needed to specify the target behavior. That is, if we treat the target behavior as a stream of data, we can encode this stream using a set of symbols representing the top level and option policies (see e.g., Rosvall & Bergstrom, 2008, for a related example outside reinforcement learning). Depending on the set of options available, some encodings are more compact than others. The hierarchy that maximizes the model evidence induces an encoding that is the *most* compact. This once again follows directly from the fact that every candidate hierarchy induces a probability distribution

over behaviors, and that the optimal hierarchy places the greatest probability mass on the target behavior. The optimal hierarchy will thus accord this behavior the shortest code length under a Shannon code assignment (Cover & Thomas, 2012), also implying the shortest expected description for any task-specific behavior (i.e., shortest path). Figure 5.1B (inset) shows the expected description length for several agent hierarchies in the rooms domain. As is clear from the figure, the hierarchy that maximizes the efficiency of representation also maximizes the efficiency of learning. This is no coincidence: It is a well established result from learning theory, echoed in empirical observations of human behavior, that ease of learning is directly related to descriptive complexity (Feldman, 2000; Kemp, 2012). Indeed, this connection has inspired previous efforts to identify useful subtask representations through data compression (Schmidhuber, 1992; Thrun & Schwartz, 1995; van Dijk & Polani, 2011; van Dijk, Polani, & Nehaniv, 2011).

A salient aspect of the specific hierarchies we have considered so far (Figures 5.1C, 5.2F), is that they carve the state-space at topological bottlenecks, narrow segments bridging between densely interconnected clusters of vertices. Further examples are shown in Figure 5.2, panels A, C, and D. The decompositions discovered here by Bayesian model selection strikingly resemble those arising from graph-theoretic algorithms for community detection, which explicitly aim to isolate tightly connected clusters within complex networks. Indeed, compression of walks on graphs has been employed as one method of community detection (Rosvall & Bergstrom, 2008). In the present case, where graph decompositions correspond to behavioral hierarchies, the prominence of bottlenecks is intuitive, in the sense that subtask representations are useful precisely to the extent that they carve tasks "at their joints." Recognizing this parallel, some work in hierarchical reinforcement learning has used community structure in order to identify useful subtasks (Kazemitabar & Beigy, 2009; Moradi, Shiri, Rad, Khadivi, & Hasler, 2012;

Simsek & Barto, 2008). The present results place this past research on a normative basis, by showing that the behavioral hierarchies resulting from community or bottleneck detection approximate hierarchies that provably maximize the agent's ability to discover reward-maximizing behaviors. In fact, the two approaches are complementary: While the present work provides a normative basis for understanding which partitions are best, previous work on bottleneck detection offers heuristic algorithms that may find such partitions more efficiently than searching through the entire space of possible hierarchies. Of course, the approaches will not always coincide, and understanding how and when they differ is an interesting challenge for future work.

5.4.2 Behavioral experiments

Having introduced a framework for identifying optimal behavioral hierarchies, we turn to the question of whether human learners decompose novel tasks in an optimal fashion. Some encouragement for this possibility comes from previous work in which related formal principles have been proposed to underlie learning in other domains, including vision (Feldman, 2009; Orbán, Fiser, Aslin, & Lengyel, 2008), working memory (Brady, Konkle, & Alvarez, 2009; Mathy & Feldman, 2012), language (Finley & Newport, 2010), and others (Chater & Vitányi, 2003; Robinet, Lemaire, & Gordon, 2011). Still more germane is a recent study in which participants were asked to parse sequences of visual stimuli whose ordering, unbeknownst to them, was determined by a random walk in the graph shown in Figure 5.2A (Schapiro et al., 2013). Participants marked the transitions between the five-vertex clusters as natural breaking points, consistent with the idea that human sequence perception spontaneously detects temporal community structure.

In order to examine hierarchy learning in the context of goal-directed action, we conducted four new behavioral experiments. In each of these, undergraduate

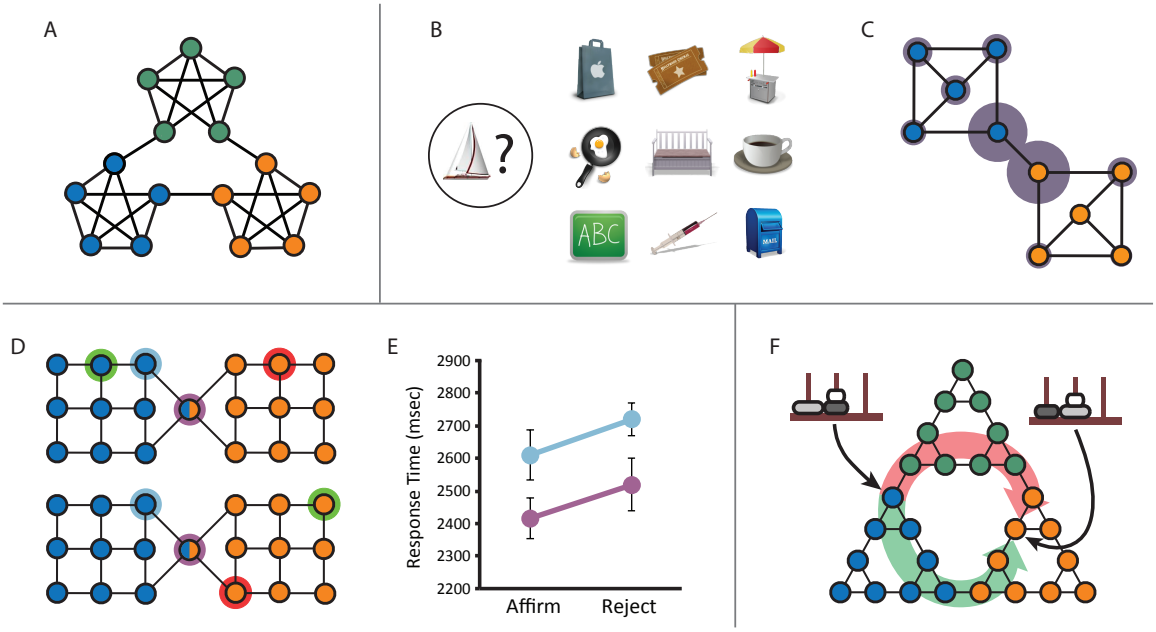

Figure 5.2: **A**. Graph studied by Schapiro et al. (2013), showing the optimal decomposition. **B**. Task display from Experiment 1. Participants used the computer mouse to select three locations adjacent to the probe location. **C**. Graph employed in Experiment 1, showing the optimal decomposition. Width of each gray ring indicates mean proportion of cases in which the relevant location was chosen. **D**. Graph studied in Experiments 2 and 3, showing the optimal decomposition (two regions, with central vertex grouped either to left or right). *Top:* Illustration of a "delivery" assignment from Experiment 3 (green = start, red = goal), where bottleneck (purple) and non-bottleneck (blue) probes called for a positive response. *Bottom:* An assignment where bottleneck and non-bottleneck probes called for a negative response. **E**. Mean correct response times from Experiment 3. *Affirm:* trials where the probe fell on the shortest path between the specified start and goal locations. *Reject:* trials where it did not. Purple: bottleneck probes. Blue: non-bottleneck probes. **F**. State-transition graph for the Tower of Hanoi puzzle, showing the optimal decomposition and indicating the start and goal configurations of the kind studied in Experiment 4. A different set of colors was used for the beads in the actual experiment. Furthermore, as explained under Methods, the beads were the same size. The changes were made here for display purposes.

participants learned about and chose actions within graph-like domains. Our general prediction, probed in different ways in each experiment, was that participants would develop a hierarchical representation of each domain aligning with the one predicted by our theoretical framework. As in the rooms domain, the set up in all four experiments is that the agent is able to make deterministic reversible transitions between (discrete) states, and that the task ensemble consists of shortest path problems between all pairs of states. Although this is our present focus, it is not a general limitation of the framework. The optimality guarantees outlined above and detailed in the online supplement apply to arbitrary tasks.

In our first experiment, a group of forty participants prepared to make a set of "deliveries" by learning the layout of a small town. The town comprised a set of ten locations, each associated with a distinctive visual icon (Figure 5.2B). Participants were never given a bird's eye view of the town. Instead, during an initial training period, participants were drilled on the adjacency relations among individual locations. On each trial a randomly selected location was highlighted, and the participant's task was to select the three locations immediately adjacent to this probe (see Figure 5.2B). Following this training period, participants were informed that they would next be asked to navigate through the town in order to make a series of deliveries between randomly selected locations, receiving a payment for each delivery that rewarded use of the fewest possible steps. Before making any deliveries, however, participants were asked to choose the position for a "bus stop" within the town. Instructions indicated that, during the subsequent deliveries, participants would be able to "take a ride" to the bus stop's location from anywhere in the town, potentially saving steps and thereby increasing payments. Participants were asked to identify three locations as their first-, second- and third-choice bus-stop sites.

Crucially, the pattern of adjacencies to which participants were exposed was based on the graph shown in Figure 5.2C. As is obvious upon inspection, the graph has a single bottleneck at its center, and an optimal partition reflecting this fact (indicated by color in the figure). Bayesian model selection identifies two graph vertices, lying at this bottleneck, as optimal subgoal locations. Given the structure of the task and the goal of navigating rapidly to an *a priori* unknown location, the optimal strategy is to place the bus stop at one of these locations. The objective of the experiment was to evaluate whether participants could detect the bottleneck and exploit it in this way. It is important to stress that participants were never given a bird's-eye view of the town, or even direct information about relative Cartesian positions. The topology of the town graph had to be inferred solely from local adjacency information. Furthermore, all of the locations had exactly three neighbors and received on average equal exposure during training. There was thus nothing salient about any of them. Despite this challenge, participants showed a marked tendency to place the bus-stop at the locations predicted (see Figure 5.2C). After adjusting for chance, the two bottleneck locations were identified as first-choice locations 4.4 times as often as the remaining locations ($\chi^2(1, N = 40) = 35.16$, $p < 0.001$). Among participants who were able at the end of the experiment to draw the underlying graph perfectly, 94% chose a bottleneck location first ($\chi^2(1, N = 17) = 58.37$, $p < 0.001$).

The results of this initial experiment are consistent with the notion that human learners identify and exploit optimal task decompositions or behavioral hierarchies. However, it might be argued that the bus stop manipulation prompted a special, task-specific orientation. Two further experiments investigated whether human learners identify and exploit optimal hierarchies spontaneously, without such a prompt. In Experiment 2, ten participants completed a set of deliveries, with no mention of bus stops, within a town whose layout was based on the bot-

tleneck graph in Figure 5.2D. Some deliveries were completed step by step, using a graphical interface that showed participants their current location and allowed them to select among adjacent locations. However, on another subset of trials participants were shown all town location icons concurrently and asked either to (1) indicate all locations lying on the shortest path between a specified start and goal *in any order*, or (2) identify any *single* location lying on this path. In the former condition, participants showed a strong tendency to select the bottleneck location first (84% of correct responses on relevant trials; Monte Carlo test, $p < 0.007$). And in the single-location condition, participants again showed a strong tendency to select the bottleneck (74% of correct responses on relevant trials; Monte Carlo test, $p < 0.01$). These findings suggest that participants planned their routes hierarchically, "thinking first" of transition-points between subregions, and then planning the specific steps needed to reach those transition points (Wiener & Mallot, 2003). More importantly, the observed behavior confirms that participants decomposed the task space in an optimal fashion, consistent with the Bayesian model selection account.

These conclusions were reinforced by the results of a third experiment. Here, 21 participants made deliveries within a town based again on the graph from Experiment 2. Interleaved with step-by-step delivery trials like those in Experiments 1 and 2 were trials in which participants were presented with a start location and a goal location, and asked whether a third location would lie on the shortest path from one to the other (see Figure 5.2D). Correct response times were faster when the probe location lay at the boundary between subregions in the optimal parse than when it lay elsewhere in the graph (Figure 5.2D-E; $F(1, 1474) = 6.838, p < 0.01$), again consistent with the idea that route planning occurred initially at the level of the regions arising from the optimal decomposition, followed later by finer-grained

selection. Further statistical analysis, detailed in the supplement, showed that this main effect was not explained by differences in probe frequency.

In a final experiment, we tested whether the predictions of the optimal hierarchy framework extend beyond the domain of spatial navigation. Here, we leveraged the Tower of Hanoi task. As shown earlier, the optimal decomposition of this task separates it into three regions (Figure 5.2F). Consider the problem defined by the start and goal states shown in Figure 5.2F. As also shown in the figure, there are two shortest-path solutions to this problem, each involving the same number of steps. The two paths differ, however, in terms of the number of boundaries they traverse between regions: One traverses one such boundary, the other two. Based on the idea that planning occurs first at the level of the regions defined by the optimal hierarchy, and that maintaining subgoals in memory is costly (Anderson & Douglass, 2001), we predicted that participants faced with this particular problem would prefer the path crossing only a single region boundary. This prediction was confirmed in an experiment involving thirty-five participants, who solved a series of Tower of Hanoi problems. When the problems of interest occurred, participants pursued the single-boundary solution in 72% of cases (right-tail sign test, $p < 0.05$). Seventeen subjects traversed the single-boundary route most often, while only seven showed the opposite asymmetry (one-tailed t-test, $p < 0.05$).

The results of these four experiments support the conclusion that human learners discover optimal task decompositions and leverage these decompositions in planning action sequences. The data suggest that novel behavioral domains are spontaneously decomposed into subdomains or regions, and that planning initially focuses on transitions between these, typically via topological bottlenecks. More specifically, the decompositions selected are optimal in the sense specified in the Bayesian model selection account.

Although our focus has been on a reinforcement learning (Sutton & Barto, 1998) characterization of learning and planning, this view includes more classic notions of planning both in artificial intelligence (Boutilier et al., 1999) and in psychology (Newell & Simon, 1972). Such problems may be cast in the reinforcement learning framework (i.e. as Markov decision processes) by encoding the goal state in the reward function (e.g. by setting reward to be 0 in the goal state and -1 everywhere else). In psychology, a number of theorists have attempted to understand planning in the context of broader unified frameworks for cognition, such as ACT-R (Anderson et al., 2004) and Soar (Laird, 2012). In ACT-R, both goals and subgoals are specified by the task model. In Soar, subgoals specifically related to solving impasses in decision making are automatically acquired. The present paper outlines a normative framework for understanding task decompositions, and this information could in theory be applied to either ACT-R or Soar, specifying the type of decompositions each should *strive* to achieve.

This raises a final issue of note: It is not our proposal that human learners discover optimal hierarchies by literally computing the Bayesian model evidence given foreknowledge of target behaviors, as in Equation 5.3. The present experimental results thus raise the important question of what discovery procedure human learners actually employ in order to approximate the same result. One possible answer comes from recent work on statistical learning, which shows that simply learning to predict future events can support discovery of community structure and topological bottlenecks in novel behavioral domains (Schapiro et al., 2013). An inviting direction for further work is to test whether this learning procedure might underlie the kind of hierarchy induction observed in the present experiments.

5.5 Methods

5.5.1 Computational analysis

Problem formulation

The Bayesian model selection approach compares agents equipped with different action hierarchies, but faced with the same ensemble of tasks. Our application of the approach focused on tasks taking the form of episodic Markov decision problems or MDPs (Sutton & Barto, 1998). A MDP comprises a set of states, a set of actions, a transition function that specifies the results proceeding from selection of specific actions in specific states, and a reward function attaching a scalar reward to each state transition. The challenge posed is to discover an action policy – a mapping from states to actions – that maximizes expected cumulative reward. For simplicity, we focused on MDPs with discrete, tabular representations of state, deterministic transition functions, and fully reversible actions (where reversibility means that for every action causing a transition from state i to j, there is another action taking j to i). This focus allowed us to represent any particular problem domain in the form of an undirected graph, with a vertex for each state and edges marking feasible actions (i.e., transitions between states). Within this setting, we considered a task ensemble ($\mathcal{T}$) comprising the set of all shortest-path problems within the graph, each task being specified in terms of a start vertex (state), a goal vertex, and a negative reward associated with traversal of each edge (randomly sampled between -1.01 and -0.99, and held constant across tasks).

The agents considered under our model comparison approach were assumed to take the form of hierarchical reinforcement learning (HRL) agents, implementing the options framework introduced in Sutton et al. (1999, see also Barto & Mahadevan, 2003; Botvinick, Niv, & Barto, 2009). Each agent was assumed to carry a set of policies, including (1) a deterministic root-level policy for each task in the target

ensemble $\pi_t, t \in \mathcal{T}$, and (2) a fixed, agent-unique set of option policies $\pi_o, o \in O$, the same across all tasks. Note that, while the HRL framework includes learning procedures for tuning action policies at root and subtask levels, the "agents" we considered did not deploy such learning algorithms; they merely carried already optimal policies structured as in HRL.

As described earlier, the option set for each agent was fully determined by an agent-specific decomposition of the state-transition graph into connected components or regions. Following Hauskrecht et al. (1998), we define a region's *entrance vertices/states* as the set of all vertices within the region that have at least one neighbor outside of it. Conversely, the *exit vertices/states* for a region comprise those vertices outside the region that receive at least one edge from a vertex inside it. Armed with these definitions, we viewed any region within a given graph decomposition as inducing a set of option representations, each having the region's entrance states plus the start state for the current task as its initiation set, the exit states for the region as its termination set, and one specific exit as its subgoal (see Botvinick, Niv, & Barto, 2009; Sutton et al., 1999, for a full explanation of these terms). Each option policy was assumed to cover only the states/vertices lying within the relevant region, with an action set including only atomic actions, i.e., transitions between neighboring states/vertices. The action set for root-level policies, in contrast, included options as well as atomic actions.

As introduced earlier, Bayesian model selection takes into account a dataset and a set of candidate models, each associated with a set of parameters (see Equation 5.1). In the present application, the target data comprise the set of behaviors representing solutions to the target task ensemble, here the set of all shortest paths within the state-transition graph. The set of models to be compared includes one HRL agent for each possible decomposition of the state-transition graph.[1] Each

[1] As noted above, this formulation assumes that the termination function for each option is zero everywhere except a single state (a subgoal state in a nearby region), and that the space of hierarchies

candidate agent is associated with a policy set consisting of each task ($\pi_t, t \in \mathcal{T}$) and option ($\pi_o, o \in O$) policy. The parameters, Π, dictate the values of these policies, with a separate parameter specifying the optimal action for a single state in either the task or option policy. In the case of π_o, the number of possible values for each parameter equals the number of atomic actions available in the relevant state, which in turn is equal to the number of edges projecting from the corresponding vertex in the graph representation (i.e. the vertex degree). In the case of π_t, the number of possible values equals the number of atomic actions available plus the number of options available, a sum we designate as $\bar{k}_i$. The number of options available corresponds to the number of exit states for the region: there is a separate option for navigating to each neighboring state in each adjacent region.

Calculation of model evidence

Given the above formulation, the model evidence can be written as in Equation 5.2. We assume a uniform distribution over model parameterizations. Thus, we can rewrite the model evidence as follows:

$$\frac{\sum_{\pi \in \Pi} Pr(behavior|hierarchy, \pi)}{|\Pi|}. \tag{5.4}$$

We assume the parameter space to span only deterministic policies. This means that any specific model parameterization will be either perfectly compatible with the data (i.e., $Pr(behavior|hierarchy, \pi) = 1$) or categorically incompatible ($Pr(behavior|hierarchy, \pi) = 0$). If we denote the set of all compatible parameteriza-

is only one level deep, among other things. The actual space of possible hierarchies is far more vast. We restrict the search space in this way in the present experiments for simplicity, as searching even over this restricted space is difficult. We leave it to future work to explore more fine-grained predictions having to do with multiple levels of hierarchy and other elaborations. This is not a limitation of the overall framework that we outline, but of the implementation of the particular examples and experiments that are of present focus. In particular, all of the optimality guarantees apply to the general case.

tions as Π^+, then

$$\frac{\sum_{\pi \in \Pi} Pr(behavior|hierarchy, \pi)}{|\Pi|} = \frac{|\Pi^+|}{|\Pi|}. \tag{5.5}$$

Thus, in order to calculate the model evidence, it suffices to determine the proportion of all parameterizations that are compatible with the data.

Imagine starting with all of the parameters unset. The number of possible parameterizations, $|\Pi|$, is equal to the product of the number of values that each parameter can take. The target data then arrive sequentially, one vertex–action pair at a time. Because the policy parameters for the relevant task must specify the actions that actually occur in the data, each newly arriving element has the potential to reduce the set of compatible parameterizations. That is, each element of the data has the potential to fix the value of one or more parameters, ruling out the remaining inconsistent values. Mathematically speaking, the consequence is to divide the number of compatible parameterizations by the number of values the parameter can take. If we designate this number as ϕ_i, with i indexing the data element, then the number of compatible parameterizations is

$$\frac{|\Pi|}{\prod_i \phi_i}, \tag{5.6}$$

and the model evidence is thus

$$Pr(behavior|hierarchy) = \prod_i \phi_i^{-1}. \tag{5.7}$$

All that remains is to specify how ϕ_i should be chosen for any element i. To see this, consider that any state transition in the data can constrain (1) the root policy π_t for the current task, (2) an option-specific policy π_o, (3) both the root policy and an option policy, or (4) neither. Because each state-specific parameter in π_t can take on

$\bar{k}_i$ values, case (1) requires $\phi_i = \bar{k}_i$. Because each state-specific parameter in π_o can take on k_i values, case (2) requires $\phi_i = k_i$. It naturally follows that case (3) requires $\phi_i = \bar{k}_i k_i$. In case (4), where no constraint is added, $|\Pi|$ should remain unchanged, thus requiring $\phi_i = 1$. Taking these points on board,

$$\phi_i = \bar{k}_i^T k_i^S, \tag{5.8}$$

where T and S are indicator functions of i: T assumes value 1 if data-element i imposes a constraint at the root (task) level; S does so if the element imposes a constraint at the option (subtask) level; and each is otherwise zero. One further requirement we impose is that if a subsequence can be represented by an option, then the option policy is invoked (constrained) rather than the root-level policy (except on the first step, where both policies are constrained). This assumption is mild, but necessary for the optimality arguments to logically follow: If an agent is equipped with an option, it is assumed the option is used and not ignored.

The online supplement illustrates calculation of the model evidence through a concrete example. Also presented in the supplement are formal proofs of the optimality assertions advanced in the main body of the paper.

5.5.2 Ancillary procedures

As noted earlier, shortest paths were generated by adding a small amount of frozen noise to the edge weights. This approach was taken to avoid ties and to assure that the same path would always be followed between any two nodes, a condition necessary in order for deterministic option policies to transfer between tasks. We found that for some but not all graphs tested, the optimal partition varied slightly depending on the choice of shortest paths. For example, in the rooms domain (Figure 5.1), although the optimal hierarchy always separated out the four rooms

at the doorways, it sometimes also included an isolated singleton near the center of one of the rooms. This coincides with previous work grounded in information theory, which suggests that in addition to bottlenecks, such locations are salient because they correspond to local maxima in goal information transitions (van Dijk & Polani, 2011, 2013). However, such results were idiosyncratic and appeared to arise from shortest-path choices that happened to channel behavior across small sets of edges, effectively creating behavioral bottlenecks not reflected in the topology of the graph itself. Because this is not a general property, it is unlikely to be reflected in the mean effects across participants that we focus on in the present paper. We thus do not pursue this point further. Instead, in order to avoid this nuisance effect, we searched for sets of shortest paths whose edge-traversal statistics closely approximated the betweenness-centrality statistics of the set of all edges in the underlying graph, quantifying the goodness of fit using Euclidean distance. This optimization was performed based on a sample of 100,000 shortest path sets. That is, for each graph, we first computed the edge betweenness statistics based on all possible shortest paths, and then separately based on each candidate set of consistent shortest paths (as described above). The set of edge betweenness statistics based on each set of paths was treated as a point in a high dimensional space, and the set of consistent shortest paths closest to the set of all paths in this space was selected.

In order to search the space of partitions for the partition yielding the highest model evidence, we followed Brandes et al. (2008) and reformulated the problem as binary integer programming. In this formulation, the optimization is over a set of binary variables, with one variable for each edge in the graph. Setting a variable to '1' turns the corresponding edge 'on', and the connected components correspond to separate communities. We used a genetic algorithm to search the resulting space, as implemented by the GA library in R (Scrucca, 2013). Each generation

consisted of 2000 individuals. The search was halted when the best score remained unchanged for 20 generations. Because genetic algorithms are not guaranteed to find the global optimum, we ran the search for each graph multiple times, each time with a different starting population. The space of partitions for each of the graphs from the navigation experiments and for the graph from Schapiro et al. (2013, see Figure 2A) was searched 1000 times. The space of partitions for the Tower of Hanoi graph was searched 500 times, and that for the rooms graph was searched 20 times.

In Experiments 2-3 the optimal parse involved two regions, with the bottleneck vertex assimilated to one. However note that, given the graph's symmetry, this implies the existence of two parses with equal model evidence: One incorporating the bottleneck vertex into one region, the other parse incorporating it into the other region. Figure 5.2D was designed to reflect this tie outcome.

Behavioral experiments

Ethics statement: All experimental procedures, including procedures for informed consent, were approved by the Princeton University Institutional Review Board.

Experiment 1

Subjects. Forty adults from the Princeton community (21 female; ages 18-21) participated. All gave written consent and were either given course credit or a nominal payment for their participation.

Materials and procedure. Participants were told that they were going to navigate through a virtual town to make deliveries. Each in a set of locations (icons) was mapped to a node in a graph (see Figure 5.2C). The present experiment began with an extended training phase, which took place before any deliveries were actually assigned. During each trial within the training period, participants saw all

ten location icons in a randomly sorted array (see Figure 5.2B). One of the locations was highlighted, and the participant was asked to identify that location's three immediate neighbors in the town, using mouse clicks, in a self-paced fashion. Three selections were permitted. Each icon selected was immediately circled in green if it was a neighbor, and in red if not. If three correct selections were made, a new location was highlighted (sampled randomly without replacement), initiating the next trial. If, after each three selections, any selection was incorrect, the original index item remained highlighted while the rest of the display was reinitialized, and the participant made another attempt at identifying the relevant neighbors. When participants managed to identify all three neighbors correctly on the first round of any trial, they received a "point." The training phase continued for a total of 55 minutes (for the first twenty participants) or 40 minutes (for remaining participants). The time was reduced partway through the experiment because we noticed that performance was on average at ceiling after 40 minutes.

Following the training phase, the experimenter introduced the delivery task, informing participants that deliveries would involve randomly selected initial and target locations, and that a "point" would be awarded for deliveries completed in the fewest possible steps. Participants were shown an example display, which showed icons indicating the current location, goal location and all locations adjacent to the current location, and 20 subjects (but not the remainder[2]) completed a *single* practice delivery trial, using key presses to transition from the current location to an adjacent location until the goal was reached.

The notion of a bus-stop was also introduced at this point. Participants were told that before embarking on the delivery task, they would be asked to position a bus stop within the town. They were told that a well-chosen location could

[2]We realized that practice trials with deliveries made through the bottleneck could result in a recency bias, and thus opted to run another 20 subjects without this potential confound. However, the results for both sets of subjects were similar, and we opted to combine them.

help them navigate efficiently. During subsequent delivery trials, the participant was informed, their chosen location would appear in the display, and they could transition to it in one step, at any time, by pressing the 0 key. If they used the bus-stop to complete the delivery in fewer steps than the shortest path attainable without a bus-stop jump, they would receive an additional bonus point for the delivery.

Following provision of this information, participants were asked to provide their first, second and third choices for the bus-stop location. Participants then completed three delivery trials, to confirm that they had understood the task description. At the close of the experiment, participants were asked to draw a map of the town in the form of a graph, with nodes representing locations and edges indicating adjacency relations.

Supplementary results from this and the subsequent experiments are reported in the online supplement.

Experiment 2

Subjects. Ten adults from the Princeton community (5 female, ages 18-21) participated. All gave written consent and were either given course credit or a nominal payment for their participation.

Materials and procedure. As in Experiment 1, participants were told that they were going to navigate through a virtual town to make deliveries. And once again, each in a set of locations (icons) was mapped to a node in a graph, in this case a graph of size nineteen (see Figure 5.2D). The experiment consisted of 19 blocks. It began with 6 blocks of 6 deliveries each. Each delivery trial was exactly as in Experiment 1, but without the bus-stop destination. In order to ensure proper learning of the town structure, this time participants were aided by periodic

displays of a birds-eye view of the town's underlying map. Note that this visual aid makes bottleneck discovery straightforward. However, in this experiment we were not interested in bottleneck discovery per se, but rather in the question of whether participants used this knowledge to plan action sequences hierarchically. On the first 6 blocks, participants could look at the map as long as they wished. From the end of block 7 through the end of the task, they could look at the map for 3 seconds. This latter design decision was imposed with the assumption that participants should already be familiar with the town layout in later trials, reducing the time spent looking at the map and allowing for more trials to be collected.

Starting at the end of block 6 and through block 19 (the last block), some of the trials were normal delivery trials, and some were "path identification" tasks (the type of trial was determined with probability 0.5). In "path identification" trials participants were shown a grid with all 19 locations in random order with a start location identified with a green box and a target location with a red box. In 40% of these trials, participants were asked to identify, using mouse-clicks, all the locations that would lie on a shortest path between the start and target, in any order. Participants could choose locations by clicking on them and un-choose them by clicking again. Chosen locations were marked by a gray square around them. The trial ended either when the chosen locations formed a shortest path, or after a maximum of 15 clicks. In the other 60% of "path identification" trials, participants were shown the grid with the start and target locations and asked to click on just one location that lay in some shortest path between them. The trial ended after one click, regardless of whether the choice was correct or not. In all path identification trials participants received feedback indicating whether their choices were correct. A 40/60 split was chosen because the trials on which participants were asked for the full path were significantly longer in duration, limiting the number of data

points that we would be able to collect overall.

Experiment 3

Subjects. Twenty-one adults from the Princeton community (11 female, ages 18-21) participated. All gave written consent and were either given course credit or a nominal payment for their participation.

Materials and procedure. As in Experiments 1 and 2, participants were told that they were going to navigate through a virtual town to make deliveries. And once again, each in a set of locations (icons) was mapped to a node in a graph, in this case the graph of size nineteen used in Experiment 2 (see Figure 5.2D). Following the structure of Experiment 2, this experiment began with 6 blocks of 6 deliveries each, followed by a birds-eye view of the town's map, to aid in learning the town's distribution. At the end of these first 6 blocks, participants could look at the map as long as they wished. From the end of block 7 through the end of the task, they could look at the map for 3 seconds.

Each delivery trial was exactly as in Experiment 1 and 2. From block 6 through block 19 (the last block), at the end of each set of deliveries participants were asked ten Yes/No questions of the form "If you had to navigate from A to B, would you go through C?" Locations A, B and C were depicted graphically using their corresponding icon. The questions of interest were chosen randomly from a pool of four types, plus some extra filler questions: queries could be about local (A and B on the same side of town) or non-local (A and B on opposite sides) locations, and they could be about the bottleneck (C corresponding to bottleneck location) or about another non-bottleneck node. Type 1 queries were about non-local deliveries and the probe node C was the bottleneck (therefore the correct answer was always Yes). Type 2 queries were non-local deliveries and the through node was either of

the nodes adjacent to the bottleneck on the target side of town (correct response was always Yes). Type 3 and Type 4 queries were local ones (A and B on same side), with or without the bottleneck as the through node, respectively (correct response was always No). A set of extra filler queries involved local deliveries, sometimes with the bottleneck as either start or target, with the through node selected from the same side of the city (correct answer could be Yes or No, depending on the participant's choice of shortest path).

All response times faster than 250 msec or slower than 7000 msec were excluded, and the remaining response times were log-transformed. Participants answered the queries correctly 98% of the time, and we excluded from our analyses the few incorrect responses. Our central predictions were that in queries where the bottleneck was the queried through node (Types 1 and 3), participants would be faster to correctly respond Yes or No than in queries involving the adjacent nodes (Types 2 and 4).

Experiment 4

Subjects. Thirty-five adults from the Princeton University community (15 female, ages 18-46) participated in this study. All gave written consent and received a nominal payment for their participation.

Materials and procedure. Participants were trained to perform a computer-based version of the three-disk Tower of Hanoi (ToH) puzzle. The display showed a rectangular base supporting three posts, with three beads (isoluminant in red, green and blue) threaded onto the posts. Participants solved a series of puzzles, moving beads from post to post, one at a time, to transform initial configurations into target configurations. To move a bead, keys corresponding to its current and desired new positions were pressed in series (the J, K, and L keys and right index,

middle and ring fingers were used for this purpose). In addition to the current bead configuration, the display also included an image of the goal configuration in the upper right portion of the screen.

Participants were required to follow a set of rules restricting the range of legal moves. Specifically, if the three colors are designated C1, C2 and C3, the rules specified that C2 could never be placed on top of C1 and that C3 could never be placed on top of either C1 or C2. The specific colors assigned to these three roles was counterbalanced across subjects. (The standard ToH task involves disks of different diameters rather than different colors; we used colors in preparation for a follow-up fMRI study, where considerations of visual similarity will be important). If an illegal move was attempted, a brief tone was sounded and no change would occur in the display.

After an initial orientation, participants performed a series of randomly selected ToH problems consisting of random start and goal configurations. This phase of the session lasted twenty minutes and was entirely self-paced. No limit was imposed on the number of moves allowed. However, participants received a monetary bonus of 2 cents for each puzzle solved, and were rewarded with a performance bonus of 3 cents for reaching the goal state in the minimum numbers of moves. At the end of each game, subjects were informed of their earnings (e.g., "You have earned 2¢!"). Participants then pressed the space bar to begin a new trial.

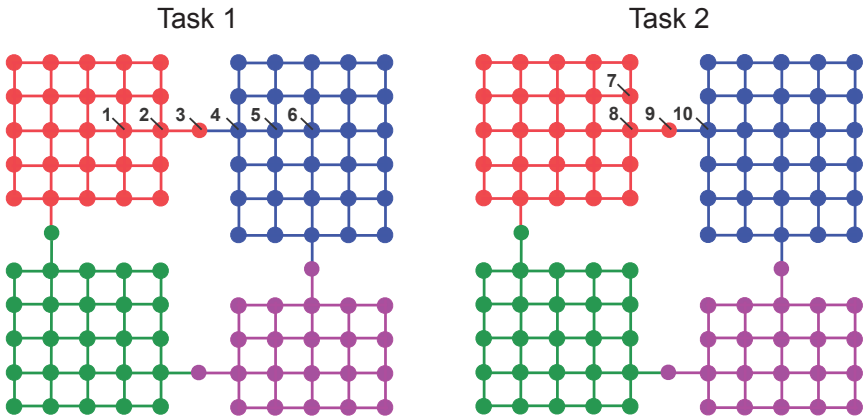

Figure 5.3: Calculating model evidence example

5.6 Supplementary methods and results

5.6.1 Computational analysis

Calculation of model evidence

The procedure employed to calculate the model evidence is described in detail in the main text. Here we provide a concrete illustration. To this end, we revisit the rooms domain and consider scoring the particular partition shown in Figure 5.3. Imagine that the data begin with the two task-specific paths shown in that figure.[3]

Consider Task 1 first. The node appearing first on the path has four primitive actions available (north, south, east, west) and two options (go to blue region, go to green region). We can begin coding this path by considering taking the action *east* from element 1. However, as noted in the main text, our coding scheme assumes that if an option is available, it will be used. Because the option *go to blue region* also applies, it is selected. The transition from element 1 to 2 thus constrains both the task-specific policy π_{t1} (setting it to the option *go to blue region*) and the option-specific policy π_o for the *go to blue region* option (setting it to the action

[3]Recall that in general, the data consist of optimal *policies*. When transitions are deterministic, the data may equivalently consist of a sequence of state transitions; the optimal policies are then implied. In what follows, we take the latter view to simplify the exposition.

east). We compute ϕ_1 using Equation 5.8 from the main text. Because there are six choices at the top level (four actions and two options), $\bar{k}_1$ is 6, and because there are four choices at the option level (just the four actions), k_1 is 4. We thus have $\phi_1 = \bar{k}_1 k_1 = 6 \cdot 4 = 24$.

The transitions from element 2 to 3 and from 3 to 4 are guided by the option selected at element 1 (recall that behavior continues to be controlled by an option until it terminates, which in this case happens only when the subgoal in the blue region is reached). Consequently, these steps further constrain the option policy, but not the root level policy ($\phi_2 = k_2 = 4$, $\phi_3 = k_3 = 2$). The root level policy must again be invoked in the transition from element 4 to 5, setting $\bar{k}_4$ to 6 (the standard four actions are available, as are the options *go to red region* and *go to purple region*). However, because the ensuing sequence does not exit the blue region, no option can be selected, and the policy is instead set to the primitive action *east*. The transition from 5 to 6 is also guided by the root level policy, and $\phi_5 = \bar{k}_5 = 4$ (recall that the initiation set for options includes only the entrance states; there are thus no options available).

Element 7 begins a new task, and once again it is evident that an option is selected. Indeed, it is the same option that was selected at the outset of Task 1 (the sequence leads to the same exit node 10). The transition from 7 to 8 thus constrains both the root (setting it to the option *go to blue region*) and option (setting it to the action *south*) policies, and $\phi_7 = \bar{k}_7 k_7 = 5 \cdot 3 = 15$. On the transitions from 8 to 9 and 9 to 10, the same option remains in control. It is important to note, however, that these steps do *not* constrain this option's policy. The reason is that these same steps were taken, under the same option, in Task 1 (transitions 2-3 and 3-4). Because the option, like all options, can be reused across tasks, the constraints imposed on the option policy by those earlier steps already assure consistency with the repeated transitions in Task 2. Since these steps do not impact Π^+, $\phi_8 = \phi_9 = 1$.

Calculation of the factors ϕ_i would continue from this point, spanning all subsequent tasks contained in the dataset. The model evidence would then be directly computable, using Eq. 5.7.

Optimality

We show here that by choosing an action hierarchy that maximizes the model evidence, one maximizes the agent's ability to discover adaptive behaviors.[4] Note that the adaptive behaviors in question span two levels. First there is the challenge of discovering an optimal root-level policy for each task in the relevant target ensemble. Second, because the resulting set of root level policies will in general call options, there is the additional challenge of learning optimal policies for those options themselves. The problem of discovering adaptive behaviors can thus be decomposed into a set of sub-problems, one for each task (the challenge at the task level is to discover the optimal root-level policy, *given* options already furnished with optimal policies) plus one for each option. We refer to the union of these sub-problems as the *target set*. We aim to show that the hierarchy that maximizes the model evidence maximizes the expected log probability that the agent, proceeding by trial and error, will discover the optimal policy for a problem sampled randomly from the target set, prior to the arrival of any chosen deadline.

We begin with the model evidence itself: $Pr(behavior|hierarchy)$. Note that this can be regarded as the probability that the agent will generate the target behavior (the behavior described in the target data) based on simple trial-and-error. Each action in the process is generated by randomly selecting one value for the relevant policy parameter. Taking on board the hierarchical structure of the available

[4]Our focus here is on discovering optimal behavior *for the first time*. After the optimal policies are known, they may be stored and used each time the same task is encountered. The argument is thus independent of task *frequency*.

policies, the model evidence can be thought of as:

$$Pr(behavior|hierarchy) = \prod_{t \in \mathcal{T}} p_t \prod_{o \in \mathcal{O}} p_o, \tag{5.9}$$

where p_t is the probability of randomly guessing the correct root-level policy for task t, and p_o is the probability of randomly guessing the correct option-level policy for option o. Taking the log on both sides gives,

$$\log Pr(behavior|hierarchy) = \sum_{t \in \mathcal{T}} \log p_t + \sum_{o \in \mathcal{O}} \log p_o = \sum_{j \in \mathcal{T} \cup \mathcal{O}} \log p_j. \tag{5.10}$$

Dividing by the size of the target set,

$$\frac{\log Pr(behavior|hierarchy)}{|\mathcal{T} \cup \mathcal{O}|} = \frac{\sum_{j \in \mathcal{T} \cup \mathcal{O}} \log p_j}{|\mathcal{T} \cup \mathcal{O}|}, \tag{5.11}$$

which, because we assume random sampling from the target set, implies

$$\log Pr(behavior|hierarchy) \propto E[\log p_{j \in \mathcal{T} \cup \mathcal{O}}]. \tag{5.12}$$

Thus, maximizing the model evidence also maximizes the expected log probability of correctly guessing the contents of a root or option level policy drawn at random from the target set. This will obviously remain true if multiple independent guesses are permitted. It is also easy to show, by extension, that maximizing the model evidence minimizes the geometric mean number of trial-and-error attempts necessary for the agent to discover the optimal policy for a task or option randomly drawn from the target set. Relating the present theoretical approach to more sophisticated procedures, such as temporal-difference learning or Monte Carlo tree search, is an objective for ongoing work.

As asserted in the main text, maximizing the model evidence is also optimal in a second sense: It minimizes description length. Specifically, it minimizes the number of information-theoretic bits necessary to describe optimal behavior, given an ensemble of target tasks. "Description" here means specifying the policy for each task and each option, indicating which particular action should be taken in each state. From the definition of ϕ_i in the preceding section, it follows that the number of bits required by element i of the data is $\log_2 \phi_i$. The number of bits needed to describe the entirety of the data is therefore

$$-\log_2 \left(\prod_i \phi_i^{-1} \right), \tag{5.13}$$

where i ranges over data elements. The expression in parentheses here is the model evidence (see Eq. 5.7). Thus, maximizing the model evidence minimizes the number of bits necessary to specify a policy for the agent that is consistent with the data.

Although our focus has been on problems with deterministic reversible transitions, a similar set of arguments apply to the stochastic case. As described previously, the data consist of optimal task policies. With deterministic transitions, the policies have to be defined only for states along the shortest paths, as the agent cannot be knocked off-course. In the general case, the optimal policy has to be defined for all states that have a greater than zero percent probability of being visited. The arguments above then naturally follow. Maximizing the model evidence aids the discovery of optimal policies on average across tasks, and minimizes the number of bits necessary to store these policies.

5.6.2 Supplementary experimental results

Experiment 1

During the training phase, participants cycled through the full set of locations an average of 18.3 times. Of the forty participants, 23 (58%) identified one of the two bottleneck locations as their first bus-stop choice, far above the number that would be predicted to occur by chance, $\chi^2(1, N = 40) = 35.16$, $p < 0.001$. Additional analyses indicated that bus-stop choices did not differ significantly with training duration (40 vs. 55 minutes). Of the 15 participants who selected the two bottleneck locations as their first two bus-stop choices, 11 selected an adjacent node – that is, one of the nodes with the highest graph centrality among those available – as their third choice. Two of the remaining participants violated the task instructions in order to once again select a bottleneck location. Participants who made no errors on the final two cycles were classified as highly successful learners. Twenty-three participants met this criterion. Among highly successful learners, 18 (78%) selected a bottleneck location as their first bus-stop choice, again well above chance, $\chi^2(1, N = 23) = 48.79$, $p < 0.001$. Within the group of highly successful learners, eighteen participants rendered the underlying graph perfectly at the end of the experiment, and of these, 17 (94%) chose a bottleneck bus-stop first $\chi^2(1, N = 17) = 58.37$, $p < 0.001$.

Experiment 2

Among single-location trials, we used a Monte Carlo procedure to test for a tendency to select the bottleneck location. Vertex-to-location mappings were repeatedly ($N = 100,000$) randomly permuted within each trial across the entire sample, and for each permutation the number of bottleneck choices was recorded. This resulted in a null distribution of choice frequencies. The result reported in the

main text reflects a fixed-effects analysis. In order to test for consistency across the subject population, the same Monte Carlo procedure was used to construct a null distribution for each participant, and a right-sided p-value was derived from this distribution, using the actual number of trials on which the participant selected the bottleneck location. The resulting set of p-values was compared against 0.5 using a Wilcoxon signed rank test. This procedure confirmed a strong tendency, across participants, to select the bottleneck location ($p < 0.003$).

The same analysis strategy was applied for any-order trials, in order to test for a tendency to select the bottleneck location first. Once again, the result reported in the main text reflects a fixed-effects analysis. However, a random-effects analysis based one the same Monte Carlo approach as described for single-location trials indicated a consistent effect across participants ($p < 0.02$).

Experiment 3

A GLM analysis was conducted on log RT data from correct responses, testing for main effects of probe type (bottleneck vs. non-bottleneck) and response (affirm vs. reject). Four further factors were included, in order to assure that any main effect of probe type did not reflect a confound between probe type and stimulus familiarity. These factors each coded for the cumulative number of occurrences, up to the current trial, of a particular stimulus type. Using the term *multiplicity* for this number of occurrences, the factors were, (1) multiplicity of the current probe, (2) multiplicity of the current start location, (3) multiplicity of the current goal location, (4) multiplicity of the current combination of start, goal and probe. A factor was also included for trial number, and subject was included as a random effect. As reported in the main text, this GLM analysis revealed a main effect of probe type ($F(1, 1474) = 6.838$, $p < 0.01$). A main effect of trial was also observed ($F(1, 1474) = 21.723$, $p << 0.001$). No other main effect reached statistical significance ($p > 0.05$). To check

whether the main effect of probe type might reflect a speed-accuracy tradeoff, we compared response accuracy for bottleneck probes against non-bottleneck probes. Mean accuracy was virtually identical for the two conditions (0.967 versus 0.970), a t-test indicated no significant difference ($p = 0.58$).

Experiment 4

Subjects completed an average of 79.24 trials (range 35-148). The trials of interest were those where two shortest-path solutions existed, one of which traversed a single region boundary and the other of which traversed two boundaries (see Figure 5.2F). A total of 22 different problems fitting this description occurred during the experiment, all involving shortest-path solutions of either six or seven steps. Twenty-five participants received at least one relevant task assignment. Within this group, the range of such trials per subject ranged from one to seven (median 1.96), with a total sample of 47 trials.

Each trial was classified based on whether the path chosen traversed one or two region boundaries. Thirty-four trials (72%) involved a two-boundary solution. A fixed-effects, one-tailed sign test rejected the null hypothesis that one- and two-boundary solutions were equally frequent ($p = 0.0015$). In order to test for a consistent effect across participants, each participant was classified according to whether he or she more frequently selected a two-boundary solution on trials of interest. Seventeen participants fell into this category, while seven showed the opposite asymmetry (one participant chose one- and two-boundary solutions equally frequently). A one-tailed sign test was consistent with a bias toward single-boundary solutions ($p = 0.032$).

Chapter 6

Future directions

A significant amount of work over the last two decades has focused on understanding animal and human learning from the perspective of simple model-free reinforcement learning algorithms (Glimcher, 2011; Maia, 2009; Niv, 2009; Schultz, 2010). Although important, they do not explain much of observed behavior. More recently, other aspects of the reinforcement learning framework have been brought to bear on behavioral and neural data. One recent addition is the use of model-based reinforcement learning to cast what has previously been termed *goal-directed behavior* in computational terms (Dolan & Dayan, 2013). This work, so far in its infancy, has concentrated on understanding how the brain learns the world model prescribed by the framework (Bornstein & Daw, 2012, 2013; Glascher et al., 2010), and on enumerating the BOLD correlates of model-based value functions (Daw et al., 2011; D. A. Simon & Daw, 2011b; Wunderlich et al., 2012). This thesis contributes to this ongoing effort by beginning to ask more algorithmic questions, about how the brain actually converts the world model into a decision.

Chapter 2 presents the problem from a Bayesian perspective, offering to explain a large collection of existing data in terms of probabilistic inference. Although promising, much work remains to be done to fully realize this approach. Perhaps

the strongest issue of contention involves the details by which the brain implements the sort of inference that has been postulated. We built our work on top of Rao (2005), who outlined how a network of integrate-and-fire neurons can implement hierarchical Bayesian inference. Rao's work is based on the assumption that neurons can multiply their inputs with reasonable precision. A number of mechanisms for how this might be accomplished have been proposed (Gabbiani et al., 2004; Koch & Poggio, 1992; Mel, 1992; Srinivasan & Bernard, 1976). However, the issue is far from settled, and furthermore, it is not clear what the maximum number of inputs is that can be multiplied. A number of proposals are based on having only a few inputs, and although Rao claims that "the model described above can be extended to perform on-line belief propagation and inference for arbitrary graphical models" (Rao, 2005, p. 1116), networks more complex than the ones he presents require multiplying more than the two inputs used in his example. In our model, the maximum number of multiplications is problem dependent, and in particular relies on the number of states under consideration. Despite this, the general predictions outlined in Chapter 2 still stand, because they require only that the brain perform belief propagation *in some way*, not necessarily the particular way that is outlined. Nevertheless, a better understanding of the details by which the brain performs this type of inference would be of significant value.

A second and somewhat related point concerns the movement from the Bayesian network and normative computation to the noisy neural network model.[1] The first part of the chapter, and the work of Botvinick and An (2009) on which it is based, describes how to optimally integrate the information available (i.e. the world model). The neural network implementation somewhat artificially breaks this notion of optimally by introducing noise ad hoc, under the assumption that neurons are inherently noisy. A more interesting approach would attempt to

[1] I would like to thank an anonymous reviewer for bringing up this point.

bridge the specific type of inference being performed with Bayesian accounts of neural noise (Ma et al., 2006).

Third, the work we present assumes the decision maker already has access to a world model. We suggested in broad strokes a few ways in which the model could be learned, and detailed work directly addressing this question has been performed by other groups (Bornstein & Daw, 2012, 2013; Glascher et al., 2010). Of particular interest to the present framework, mentioned only in passing previously, is whether learning the state space takes place at the level of the *parameters* of the Bayesian network, or at the level of its *structure*. The distinction between the two can be rather thin, as each parameter can simply be represented by a single node. However, it could also be the case that each node s is a highly structured probability distribution, and is represented by a multitude of other nodes in interesting configurations. At an algorithmic level such configurations can be learned using general methods for Bayesian structure learning (Koller & Friedman, 2009), although this raises yet additional questions about implementation (Pouget, Beck, Ma, & Latham, 2013). The simpler possibility of having point estimates corresponds to the empirical Bayes approach. If more complex structure is indeed learned, it would certainly have ramifications for the algorithmic details outlined in the chapter. Exploring such details is left to future work.

Chapter 3 begins from the evidence accumulation perspective and from the work of Rangel and colleagues on simple choice (Milosavljevic et al., 2010; Krajbich et al., 2010; Krajbich & Rangel, 2011). It presents two new experiments that extend the simple choice paradigm to multiple steps, and develops a new accumulator model to explain the observed relationship between choice and reaction time data. The Discussion section of the chapter reveals just how nascent our understanding of algorithmic questions is, and describes many avenues for improvement. I do not detail them again here, but briefly, they include understanding how the state space

is reduced through pruning and sampling, the role that attention plays in biasing decision making, *why* choice data is variable and where this noise comes from (similar to the question about neural noise above), and the manner in which the model-based system interacts with the episodic memory system. These questions are rather general overall, and can be pursued using the framework of either Chapter 2 or 3.

Two other questions which apply equally well to both frameworks also deserve mention. The first involves how the model-based system interacts with the model-free system. Previous work has begun to address this question (Daw et al., 2005; S. W. Lee, Shimojo, & O'Doherty, 2014; D. A. Simon & Daw, 2011a), and incorporating it with the algorithmic ideas outlined here would be of interest.[2] A second question concerns the statistical optimality of the two frameworks. The drift diffusion model has been shown to implement the sequential probability ratio test: For a set error level, it makes the fastest possible decisions, and for a given response time, it minimizes the amount of error (Bogacz et al., 2006). Multi-alternative versions that are approximately optimal also exist (Ditterich, 2010). It is not clear the extent to which either the framework in Chapter 2 (beyond the binary choice case detailed there) or Chapter 3 is optimal in this way. However, as noted in Chapter 2, we do not know of *any* current algorithm that is provably optimal in the multi-step setting. Understanding what such an algorithm looks like and the extent to which the proposals presented here are able to approximate it (or not) would be of interest.

In the longer term, it may be beneficial to understand model-based reinforcement learning from within a single overarching framework, or to at least be able to translate ideas from one framework into another. The ideas in Chapters 2 and 3 are tied through the evidence accumulation perspective, but it is not immediately

[2]Chapter 2 outlines a few ideas on how inputs from the model-free system can be incorporated into the Bayesian framework. However, these suggestions are rather speculative, and we did not perform simulations testing their predictions.

clear how they relate formally. It would be of interest to explore their similarities and differences in more detail, and to understand exactly where they diverge. The predictions of each could then be tested empirically.

Chapter 4 tests the idea that the brain codes pseudo-reward prediction errors as predicted by the options framework of hierarchical reinforcement learning (Sutton et al., 1999). Both the EEG and fMRI data suggest that dopamine may be responsible for coding pseudo-reward much in the same way it codes standard reward, but neither methodology can be used to tell for certain. More invasive methods that can track dopamine directly, either in human patients (Kishida et al., 2011; Zaghloul et al., 2009) or in animal models, will have to be used in the future. Botvinick, Niv, and Barto (2009) outline the many others aspects of the options framework that are ripe for testing, including how the brain maintains: the option (or hierarchy of options) currently being executed, the total reward accumulated while an option is executed, and option specific policies and value functions. Also mentioned is positive and negative transfer, a detailed study of which would be of interest within the context of the options framework. This refers to the idea that training in some tasks may be either a benefit or a hindrance to performance in other tasks, depending on the structure and goals that are shared. A related question concerns how shared structure between subtasks (i.e. between different options, rather than between different overall tasks) can be exploited when multiple subtasks have to be learned simultaneously.

A related study was performed by Diuk et al. (2013), who describe an experiment in which reward prediction errors simultaneously resulted at two levels of a hierarchy. The same system (putatively dopamine) is suggested to simultaneously code both errors. Another avenue for future research, suggested by these authors, is to understand how this information is segregated.

Chapter 5 steps back to answer the question of which options are useful to have. It is suggested that a useful set of options maximally aids learning *on average* across a range of possible tasks the organism may encounter. The Bayesian model selection framework is leveraged to find this set, and it is shown that people are able to approximate the optimum in four behavioral experiments. However, the story we present is purely at the computational level (Marr, 1982), outlining the upper bound that people *should strive* to achieve. Future work will need to address the algorithmic question, how the brain actually realizes the predictions of the framework.

Finally, it would be of great interest to synthesize the directions presented here and develop an algorithmic account of how the brain performs hierarchical model-based reinforcement learning. I hope this thesis presents the preliminary steps towards this goal.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology*, *34*(2), 77–98.

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology*, *33*(2), 109–121.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Albert, M. V., & Fincham, J. M. (2005). Tracing problem solving in real time: fmri analysis of the subject-paced Tower of Hanoi. *Journal of Cognitive Neuroscience*, *17*(8), 1261–1274.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Anderson, J. R., & Douglass, S. (2001). Tower of hanoi: Evidence for the cost of goal retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1331–1346.

Arana, F. S., Parkinson, J. A., Hinton, E., Holland, A. J., Owen, A. M., & Roberts, A. C. (2003). Dissociable contributions of the human amygdala and orbitofrontal cortex to incentive motivation and goal selection. *The Journal of Neuroscience*, *23*(29), 9632–9638.

Asaad, W. F., Rainer, G., & Miller, E. K. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, *84*, 451–459.

Atance, C. M., & O'Neill, D. K. (2001). Episodic future thinking. *Trends in Cognitive Sciences*, *5*(12), 533–539.

Attias, H. (2003). Planning by probabilistic inference. In *Proceedings of the 9th international workshop on artificial intelligence and statistics.* Key West, FL.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200.

Badre, D., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*, *19*(12), 2082–2099.

Badre, D., & Frank, M. J. (2012). Mechanisms of hierarchical reinforcement learning in cortico–striatal circuits 2: Evidence from fMRI. *Cerebral Cortex*, *22*(3), 527–536.

Badre, D., Hoffman, J., Cooney, J. W., & D'Esposito, M. (2009). Hierarchical cognitive control deficits following damage to the human frontal lobe. *Nature Neuroscience*, *12*(4), 515–522.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Baker, T. E., & Holroyd, C. B. (2011). Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by the feedback error-related negativity and N200. *Biological Psychology*, *87*(1), 25–34.

Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel visual computation. *Nature*, *306*, 21–26.

Balleine, B. W. (1992). Instrumental performance following a shift in primary motivation depends on incentive learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *18*(3), 236–250.

Balleine, B. W. (2005). Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. *Physiology & Behavior*, *86*(5), 717–730.

Balleine, B. W., & Dickinson, A. (1994). Role of cholecystokinin in the motivational control of instrumental action in rats. *Behavioral Neuroscience*, *108*(3), 590–605.

Balleine, B. W., & Dickinson, A. (1998a). Consciousness: The interface between affect and cognition. In J. Cornwell (Ed.), *Consciousness and human identity* (pp. 57–85). Oxford, England: Oxford University Press.

Balleine, B. W., & Dickinson, A. (1998b). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4), 407–419.

Balleine, B. W., & Dickinson, A. (1998c). The role of incentive learning in instrumental outcome revaluation by sensory-specific satiety. *Learning & Behavior*, *26*(1), 46–59.

Balleine, B. W., Killcross, A. S., & Dickinson, A. (2003). The effect of lesions of the basolateral amygdala on instrumental conditioning. *The Journal of Neuroscience*, *23*(2), 666–675.

Bargh, J. A., Green, M., & Fitzsimons, G. (2008). The selfish goal: Unintended consequences of intended goal pursuits. *Social Cognition*, *26*(5), 534–554.

Barlow, H. B. (1969). Pattern recognition and the responses of sensory neurons. In M. Clynes (Ed.), *Annals of the new york academy of sciences* (Vol. 156, pp. 872–881). New York, NY: New York Academy of Sciences.

Barnes, T. D., Mao, J.-B., Hu, D., Kubota, Y., Dreyer, A. A., Stamoulis, C., ... Graybiel, A. M. (2011). Advance cueing produces enhanced action-boundary patterns of spike activity in the sensorimotor striatum. *Journal of Neurophysiology*, *105*(4), 1861–1878.

Barone, P., & Joseph, J. P. (1989). Prefrontal cortex and spatial sequencing in macaque monkey. *Experimental Brain Research*, *78*(3), 447–464.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). The MIT Press.

Barto, A. G., Konidaris, G. D., & Vigorito, C. M. (2013). Behavioral hierarchy: Exploration and representation. In G. Baldassarre & M. Mirolli (Eds.), *Computational and robotic models of hierarchical organization of behavior* (pp. 13–46). Springer.

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, *13*(4), 341–379.

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*, *107*(50), 21767–21772.

Baxter, M. G., & Murray, E. A. (2002). The amygdala and reward. *Nature Reviews Neuroscience*, *3*(7), 563–573.

Baxter, M. G., Parker, A., Lindner, C. C., Izquierdo, A. D., & Murray, E. A. (2000). Control of response selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex. *The Journal of Neuroscience*, *20*(11), 4311–4319.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141.

Beck, J. M., & Pouget, A. (2007). Exact inferences in a neural implementation of a hidden Markov model. *Neural Computation*, *19*(5), 1344–1361.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Bekkering, H., Wohlschlager, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *The Quarterly Journal of Experimental Psychology*

Section A: Human Experimental Psychology, *53*(1), 153–164.

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *The Journal of Neuroscience*, *21*(8), 2793–2798.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, *191*(3), 391–431.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). Neuro-dynamic programming. *Athena Scientific*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science*, *311*(5763), 1020–1022.

Blodgett, H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology*, *4*, 113–134.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1655–1670.

Bornstein, A. M., & Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience*, *35*(7), 1011–1023.

Bornstein, A. M., & Daw, N. D. (2013). Cortical and hippocampal correlates of deliberation during model-based decisions for rewards in humans. *PLOS Computational Biology*, *9*(12), e1003387.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12*(5), 201–208.

Botvinick, M. M., & An, J. (2009). Goal-directed decision making in prefrontal cortex: A computational framework. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21, pp. 169–176).

Botvinick, M. M., Huffstetler, S., & McGuire, J. T. (2009). Effort discounting in human nucleus accumbens. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 16–27.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280.

Botvinick, M. M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, *402*, 179–181.

Botvinick, M. M., & Plaut, D. C. (2006). Such stuff as habits are made on: A reply to Cooper and Shallice (2006). *Psychological Review*, *113*(4), 917–927.

Botvinick, M. M., & Plaut, D. C. (2009). Empirical and computational support for context-dependent representations of serial order: Reply to Bowers, Damian, and Davis (2009). *Psychological Review*, *116*(4), 998–1001.

Boutilier, C., Dean, T., & Hanks, S. (1999). Decision-theoretic planning: structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, *11*(1), 1–94.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual work-

ing memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487–502.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, *20*(2), 172–188.

Bray, S., Shimojo, S., & O'Doherty, J. P. (2010). Human medial orbitofrontal cortex is recruited during experience of imagined and real rewards. *Journal of Neurophysiology*, *103*(5), 2506–2512.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, *30*(2), 619–639.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49–57.

Bugmann, G., Taylor, J. G., & Denham, M. (1995). Route finding by neural nets. In J. G. Taylor (Ed.), *Neural networks* (pp. 217–230). Henley-on-Thames: Waller.

Bullock, D., & Rhodes, B. J. (1993). Actions and responses: The dual psychology of behavior. In N. Eilan, R. McCarthy, & B. Brewer (Eds.), *Spatial representation* (pp. 277–293).

Bunge, S. A. (2004). How we use rules to select actions: a review of evidence from cognitive neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 564–579.

Bunge, S. A., & Wallis, J. D. (2007). *Neuroscience of rule-guided behavior*. Oxford, England: Oxford University Press.

Bunge, S. A., Wallis, J. D., Parker, A., Brass, M., Crone, E. A., Hoshi, E., & Sakai, K. (2005). Neural circuitry underlying rule use in humans and nonhuman primates. *The Journal of Neuroscience*, *25*(45), 10347–10350.

Busemeyer, J. R. (1985). Decision making under uncertainty: a comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(3), 538–564.

Busemeyer, J. R., & Diederich, A. (2002). Survey of decision field theory. *Mathematical Social Sciences*, *43*(3), 345–370.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432–459.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*(5), 313–323.

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences*, *7*(5), 225–231.

Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, *377*, 59–62.

Castro, R. M., Kalish, C., Nowak, R., Qian, R., Rogers, T. T., & Zhu, X. (2008). Human active learning. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21, pp. 241–248). Cambridge, MA: MIT Press.

Cavada, C., Compañy, T., Tejedor, J., Cruz-Rizzolo, R. J., & Reinoso-Suárez, F. (2000). The anatomical connections of the macaque monkey orbitofrontal cortex: A review. *Cerebral Cortex*, *10*(3), 220–242.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344.

Chater, N., & Oaksford, M. (Eds.). (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford, England: Oxford University Press.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Christoff, K., Keramatian, K., Gordon, A. M., Smith, R., & Mädler, B. (2009). Prefrontal organization of cognitive control according to levels of abstraction. *Brain Research*, *1286*, 94–105.

Christoph, G. R., Leonzio, R. J., & Wilcox, K. S. (1986). Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *The Journal of Neuroscience*, *6*(3), 613–619.

Colwill, R. M., & Rescorla, R. A. (1985a). Instrumental responding remains sensitive to reinforcer devaluation after extensive training. *Journal of Experimental Psychology: Animal Behavior Processes*, *11*(4), 520–536.

Colwill, R. M., & Rescorla, R. A. (1985b). Postconditioning devaluation of a reinforcer affects instrumental responding. *Journal of Experimental Psychology: Animal Behavior Processes*, *11*(1), 120–132.

Colwill, R. M., & Rescorla, R. A. (1986). Associative structures in instrumental learning. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 20, pp. 55–104). San Diego, CA: Academic Press.

Colwill, R. M., & Rescorla, R. A. (1988). The role of response-reinforcer associations increases throughout extended instrumental training. *Animal Learning & Behavior*, *16*(1), 105–111.

Cooper, G. F. (1988). A method for using belief networks as influence diagrams. In *Proceedings of the fourth conference on uncertainty in artificial intelligence* (pp. 55–63).

Cooper, J. C., & Knutson, B. (2008). Valence and salience contribute to nucleus accumbens activation. *NeuroImage*, *39*(1), 538–547.

Cooper, R. P., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*(4), 297–338.

Cooper, R. P., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*(4), 887–916.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, *58*(3), 306–324.

Corbit, L. H., & Balleine, B. W. (2003). The role of prelimbic cortex in instrumental conditioning. *Behavioural Brain Research*, *146*(1-2), 145–157.

Corbit, L. H., & Balleine, B. W. (2005). Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer. *The Journal of Neuroscience*, *25*(4), 962–970.

Corbit, L. H., Ostlund, S. B., & Balleine, B. W. (2002). Sensitivity to instrumental contingency degradation is mediated by the entorhinal cortex and its efferents via the dorsal hippocampus. *The Journal of Neuroscience*, *22*(24), 10976–10984.

Coutureau, E., & Killcross, S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioural Brain Research*, *146*(1-2), 167–174.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. Hoboken, NJ: John Wiley & Sons.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173.

Csibra, G., & Gergely, G. (2007). "obsessed with goals": functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, *124*(1), 60–78.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental

area. *Science*, *319*(5867), 1264–1267.

Daw, N. D., & Frank, M. J. (2009). Reinforcement learning and higher level cognition: Introduction to special issue. *Cognition*, *113*(3), 259–261.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Dayan, P. (1992). The convergence of TD (λ) for general λ. *Machine Learning*, *8*(3-4), 341–362.

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks*, *22*(3), 213–219.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Cambridge, MA: The MIT Press.

Dayan, P., & Hinton, G. E. (1997). Using expectation-maximization for reinforcement learning. *Neural Computation*, *9*(2), 271–278.

Dayan, P., Hinton, G. E., Neal, M., Radford, & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196.

De Araujo, I. E. T., Kringelbach, M. L., Rolls, E. T., & McGlone, F. (2003). Human cortical responses to water in the mouth, and the effects of thirst. *Journal of neurophysiology*, *90*(3), 1865–1876.

Deneve, S. (2008). Bayesian spiking neurons I: Inference. *Neural Computation*, *20*(1), 91–117.

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell

sequences during ripples. *Nature Neuroscience*, *10*, 1241–1242.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *308*(1135), 67–78.

Dickinson, A., & Dawson, G. R. (1989). Incentive learning and the motivational control of instrumental performance. *The Quarterly Journal of Experimental Psychology*, *41*(1), 99–112.

Dickinson, A., & Mulatero, C. W. (1989). Reinforcer specificity of the suppression of instrumental performance on a non-contingent schedule. *Behavioural Processes*, *19*(1-3), 167–180.

Dietterich, T. G. (1998). The MAXQ method for hierarchical reinforcement learning. In *Proceedings of the 15th international conference on machine learning* (pp. 118–126). Morgan Kaufmann.

Ditterich, J. (2010). A comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. *Frontiers in Neuroscience*, *4*(184).

Diuk, C., Botvinick, M. M., Barto, A., & Niv, Y. (2010). Hierarchical reinforcement learning: An fMRI study of learning in a two-level gambling task. *Society for Neuroscience Meeting*.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M. M., & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *The Journal of Neuroscience*, *33*(13), 5797–5805.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Dorris, M. C., & Glimcher, P. W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron*, *44*(2),

365–378.

Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., . . . Petersen, S. E. (2006). A core system for the implementation of task sets. *Neuron*, *50*(5), 799–812.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7), 961–974.

Doya, K., Ishii, S., Pouget, A., & Rao, R. (2006). *Bayesian brain: probabilistic approaches to neural coding*. The MIT Press.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology*, *30*(3), 257–303.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, J. (2009). Bayes and the simplicity principle in perception. *Psychological Review*, *116*(4), 875–887.

Finley, S., & Newport, E. L. (2010). Morpheme segmentation from distributional information. In *Boston University conference on language development (BUCLD) online proceedings supplement*.

Foster, D., & Dayan, P. (2002). Structure in the space of value functions. *Machine Learning*, *49*(2-3), 325–346.

Fox, C. R., & Poldrack, R. A. (2008). Prospect theory and the brain. In P. W. Glimcher, C. F. Camerer, R. A. Poldrack, & E. Fehr (Eds.), *Neuroeconomics: Decision making and the Brain* (pp. 145–174). Academic Press.

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*(2), 300–326.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the*

Royal Society B: Biological Sciences, *360*(1456), 815–836.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLOS ONE*, *4*(7), e6421.

Fujii, N., & Graybiel, A. M. (2003). Representation of action sequence boundaries by macaque prefrontal cortical neurons. *Science*, *301*(5637), 1246–1249.

Furmston, T., & Barber, D. (2009). Solving deterministic policy (PO)MPDs using expectation-maximisation and antifreeze. In *Workshop on learning and data mining for robotics.* Bled, Slovenia.

Fuster, J. M. (1997). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe*. Philadelphia, PA: Lippincott-Raven.

Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in Cognitive Sciences*, *8*(4), 143–145.

Gabbiani, F., Krapp, H. G., Hatsopoulos, N., Mo, C.-H., Koch, C., & Laurent, G. (2004). Multiplication and stimulus invariance in a looming-sensitive neuron. *Journal of Physiology-Paris*, *98*(1–3), 19–34.

Gaussier, P., Revel, A., Banquet, J.-P., & Babeau, V. (2002). From view cells and place cells to cognitive map learning: processing stages of the hippocampal system. *Biological Cybernetics*, *86*(1), 15–28.

Geisler, W. S. (2003). Ideal observer analysis. In L. M. Chalupa (Ed.), *The visual neurosciences* (pp. 825–838). Cambridge, MA: The MIT Press.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256.

Girard, B., Filliat, D., Meyer, J.-A., Berthoz, A., & Guillot, A. (2005). Integration of navigation and action selection functionalities in a computational model of cortico-basal-ganglia–thalamo-cortical loops. *Adaptive Behavior*, *13*(2), 115–

130.

Glascher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Glimcher, P. W. (2008). Choice: Toward a standard back-pocket model. In P. W. Glimcher, C. F. Camerer, R. A. Poldrack, & E. Fehr (Eds.), *Neuroeconomics: Decision making and the Brain* (pp. 503–521). Academic Press.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(Supplement 3), 15647–15654.

Gluth, S., Rieskamp, J., & Büchel, C. (2012). Deciding when to decide: Time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *The Journal of Neuroscience*, *32*(31), 10686–10698.

Glymour, C. N. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: The MIT Press.

Goel, V., & Grafman, J. (1995). Are the frontal lobes implicated in "planning" functions? interpreting data from the Tower of Hanoi. *Neuropsychologia*, *33*(5), 623–642.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*(1), 10–16.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–32.

Gopnik, A., & Schulz, L. (Eds.). (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford, England: Oxford University Press.

Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, *301*(5636), 1104–1107.

Grafman, J. (2002). The human prefrontal cortex has evolved to represent components of structured event complexes. In J. Grafman (Ed.), *Handbook of neuropsychology.* Amsterdam: Elsevier.

Graybiel, A. M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory*, *70*(1-2), 119–136.

Green, C. S., Benson, C., Kersten, D., & Schrater, P. (2010). Alterations in choice behavior by manipulations of world model. *Proceedings of the National Academy of Sciences*, *107*(37), 16401–16406.

Grent, T., & Woldorff, M. G. (2007). Timing and sequence of brain activity in top-down control of visual-spatial attention. *PLOS Biology*, *5*(1), e12.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Gul, F., & Pesendorfer, W. (2006). Random expected utility. *Econometrica*, *74*(1), 121–146.

Haber, S. N., Kim, K.-S., Mailly, P., & Calzavara, R. (2006). Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. *The Journal of Neuroscience*, *26*(32), 8368–8376.

Hamilton, A. F. d. C., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. *The Journal of Neuroscience*, *26*(4), 1133–1137.

Hamilton, A. F. d. C., & Grafton, S. T. (2008). Action outcomes are represented in human inferior frontoparietal cortex. *Cerebral Cortex*, *18*(5), 1160–1168.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation

of goal values and prediction errors. *The Journal of Neuroscience*, *28*(22), 5623–5630.

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., & Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences*, *108*(44), 18120–18125.

Haruno, M., & Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, *19*(8), 1242–1254.

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, *104*(5), 1726–1731.

Hasselmo, M. E. (2005). A model of prefrontal cortical mechanisms for goal-directed behavior. *Journal of Cognitive Neuroscience*, *17*(7), 1115–1129.

Hauskrecht, M., Meuleau, N., Kaelbling, L. P., Dean, T., & Boutilier, C. (1998). Hierarchical solution of markov decision processes using macro-actions. In *Proceedings of the fourteenth conference on uncertainty in artificial intelligence* (pp. 220–229). San Francisco, CA.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York, NY: John Wiley & Sons.

Helmholtz, H. (1860/1962). *Handbuch der physiologischen optik [Handbook of physiological optics] (J. P. C. Southall, Trans.; Vol. 3)*. New York, NY: Dover.

Hemmer, P., & Steyvers, M. (2009). A bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202.

Hestvik, A., Maxfield, N., Schwartz, R., & Shafer, V. (2007). Brain responses to filled gaps. *Brain and Language*, *100*(3), 301–316.

Hoffman, M. D., Freitas, N. D., Doucet, A., & Peters, J. R. (2009). An expectation maximization algorithm for continuous markov decision processes with

arbitrary reward. In *Proceedings of the 12th international workshop on artificial intelligence and statistics.* Clearwater Beach, FL.

Holland, P. C., & Gallagher, M. (2004). Amygdala–frontal interactions and reward expectancy. *Current Opinion in Neurobiology*, *14*(2), 148–155.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–708.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *NeuroReport*, *14*(18), 2481–2384.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*(5), 497–498.

Hori, Y., Minamimoto, T., & Kimura, M. (2009). Neuronal encoding of reward value and direction of actions in the primate putamen. *Journal of Neurophysiology*, *102*(6), 3530–3543.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). The MIT Press.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, *SSC-2*(1), 22–26.

Hsu, M., Krajbich, I., Zhao, C., & Camerer, C. F. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *The Journal of Neuroscience*, *29*(7), 2231–2237.

Hull, C. L. (1943). *Principles of behavior*. New York, NY: Appleton-Century.

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*(3), e1002410.

Inoue, M., & Mikami, A. (2006). Prefrontal activity during serial probe reproduction task: Encoding, mnemonic and retrieval processes. *Journal of Neurophysiology*, *95*, 1008–1041.

Izquierdo, A., Suda, R. K., & Murray, E. A. (2004). Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *The Journal of Neuroscience*, *24*(34), 7540–7548.

Jensen, F. V. (2001). *Bayesian networks and decision graphs*. New York, NY: Springer-Verlag.

Jin, X., & Costa, R. M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, *466*(7305), 457–462.

Jocham, G., & Ullsperger, M. (2009). Neuropharmacology of performance monitoring. *Neuroscience & Biobehavioral Reviews*, *33*(1), 48–60.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*(4), 535–547.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *The Journal of Neuroscience*, *27*(45), 12176–12189.

Johnson, A., van der Meer, M. A. A., & Redish, D. A. (2007). Integrating hippocampus and striatum in decision-making. *Current Opinion in Neurobiology*, *17*(6), 692–697.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 215–231.

Jordan, M. I. (1998). *Learning in graphical models*. Cambridge, MA: The MIT Press.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive science*, *16*(3), 307–354.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kalenscher, T., Tobler, P. N., Huijbers, W., Daselaar, S. M., & Pennartz, C. M. (2010). Neural signatures of intransitive preferences. *Frontiers in Human Neuroscience*, *4*(49).

Kazemitabar, S. J., & Beigy, H. (2009). Automatic discovery of subgoals in reinforcement learning using strongly connected components. In M. Kppen, N. Kasabov, & G. Coghill (Eds.), *Advances in neuro-information processing* (Vol. 5506, pp. 829–834). Springer.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*(4), 685–722.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Kesner, R. P. (2000). Subregional analysis of mnemonic functions of the prefrontal cortex in the rat. *Psychobiology*, *28*(2), 219–228.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, *13*(4), 400–408.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). The mirror-neuron system: a Bayesian perspective. *NeuroReport*, *18*(6), 619–623.

Kim, H., Sul, J. H., Huh, N., Lee, D., & Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *The Journal of Neuroscience*, *29*(47), 14701–14712.

Kishida, K. T., Sandberg, S. G., Lohrenz, T., Comair, Y. G., Sáez, I., Phillips, P. E. M., & Montague, P. R. (2011). Sub-second dopamine detection in human striatum. *PLOS ONE*, *6*(8), e23291.

Klossek, U. M. H., Russell, J., & Dickinson, A. (2008). The control of instrumental action following outcome devaluation in young children aged between 1 and 4 years. *Journal of Experimental Psychology: General*, *137*(1), 39–51.

Knill, D. C., & Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Knill, D. C., & Richards, W. (1996). *Perception as bayesian inference*. Cambridge, England: Cambridge University Press.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *The Journal of Neuroscience*, *25*(19), 4806–4812.

Kobayashi, S., de Carvalho, O. P., & Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *The Journal of Neuroscience*, *30*(2), 534–544.

Koch, C., & Poggio, T. (1992). Multiplying with synapses and neurons. In T. McKenna, J. Davis, & S. F. Zornetzer (Eds.), *Single neuron computation* (pp. 315–345). San Diego, CA: Academic Press.

Koechlin, E., & Hyafil, A. (2007). Anterior prefrontal function and the limits of human decision-making. *Science*, *318*(5850), 594–598.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, *302*(5648), 1181–1185.

Koene, R. A., & Hasselmo, M. E. (2005). An integrate-and-fire model of prefrontal cortex neuronal activity during performance of goal-directed decision making. *Cerebral Cortex*, *15*(12), 1964–1981.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. Cambridge, MA: The MIT Press.

Kool, W., Getz, S. J., & Botvinick, M. M. (2013). Neural representation of probability: Evidence from the illusion of control. *Journal of Cognitive Neuroscience*, *25*, 852–

861.

Kool, W., McGuire, J. T., Wang, G. J., & Botvinick, M. M. (2013). Neural and behavioral evidence for an intrinsic cost of self-control. *PLOS ONE*, *8*(8), e72626.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Kornienko, T. (2010). A cognitive basis for context-dependent utility: An adaptive magnitude evaluation approach. *Unpublished manuscript, Department of Economics, University of Edinburgh*.

Kouneiher, F., Charron, S., & Koechlin, E. (2009). Motivation and cognitive control in the human prefrontal cortex. *Nature Neuroscience*, *12*(7), 939–945.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857.

Krigolson, O. E., & Holroyd, C. B. (2006). Evidence for hierarchical error processing in the human brain. *Neuroscience*, *137*(1), 13–17.

Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. *Nature Reviews Neuroscience*, *6*(9), 691–702.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*(3), 210–226.

LaBar, K. S., Gitelman, D. R., Parrish, T. B., Kim, Y.-H., Nobre, A. C., & Mesulam, M. M. (2001). Hunger selectively modulates corticolimbic activation to food stimuli in humans. *Behavioral Neuroscience*, *115*(2), 493–500.

Laird, J. E. (2012). *The soar cognitive architecture*. Cambridge, MA: The MIT Press.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, *33*(1), 1–64.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The hixon symposium* (pp. 112–135). New York, NY: Wiley.

Lau, B., & Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, *58*(3), 451–463.

Lauwereyns, J., Watanabe, K., Coe, B., & Hikosaka, O. (2002). A neural correlate of response bias in monkey caudate nucleus. *Nature*, *418*, 413–417.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*, 687–699.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434–1448.

Lengfelder, A., & Gollwitzer, P. M. (2001). Reflective and reflexive action control in patients with frontal brain lesions. *Neuropsychology*, *15*(1), 80–100.

Levy, R. P., Reali, F., & Griffiths, T. L. (2008). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21, pp. 937–944). Cambridge, MA: MIT Press.

Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *The Journal of Neuroscience*, *31*(14), 5504–5511.

Litvak, S., & Ullman, S. (2009). Cortical circuitry implementing graphical models. *Neural computation*, *21*(11), 3010–3056.

Liu, Y. S., Holmes, P., & Cohen, J. D. (2008). A neural network model of the eriksen task: Reduction, analysis, and data fitting. *Neural Computation*, *20*(2), 345–373.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(4), 343–364.

Manski, C. F. (1977). The structure of random utility models. *Theory and Decision*, *8*(3), 229–254.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Freeman & Co.

Martinet, L.-E., Passot, J.-B., Fouque, B., Meyer, J.-A., & Arleo, A. (2008). Map-based spatial navigation: A cortical column model for action planning. In *Spatial cognition vi: Learning, reasoning, and talking about space* (Vol. 5248, pp. 39–55). Springer.

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, *122*(3), 346–362.

Matsumoto, K., Suzuki, W., & Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, *301*(5630), 229–232.

Matsumoto, K., & Tanaka, K. (2004). The role of the medial prefrontal cortex in achieving goals. *Current Opinion in Neurobiology*, *14*(2), 178–185.

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, *447*, 1111–1115.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346.

McDougall, W. (1923). *Outline of psychology*. New York, NY: Scribner.

Mel, B. W. (1992). NMDA-based pattern discrimination in a modeled cortical neuron. *Neural Computation*, *4*(4), 502–517.

Mel, B. W. (1993). Synaptic integration in an excitable dendritic tree. *Journal of Neurophysiology*, *70*(3), 1086-1101.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York, NY: Holt, Rinehart & Winston.

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, *5*(6), 437–449.

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a generic neural system for error detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798.

Monchi, O., Petrides, M., Strafella, A. P., Worsley, K. J., & Doyon, J. (2006). Functional role of the basal ganglia in the planning and execution of actions. *Annals of Neurology*, *59*(2), 257–264.

Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron*, *36*(2), 265–284.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *The Journal of Neuroscience*, *16*(5), 1936–1947.

Moradi, P., Shiri, M. E., Rad, A. A., Khadivi, A., & Hasler, M. (2012). Automatic skill acquisition in reinforcement learning using graph centrality measures.

Intelligent Data Analysis, *16*(1), 113–135.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, *9*(8), 1057–1063.

Muller, R. U., Stead, M., & Pach, J. (1996). The hippocampus as a cognitive graph. *The Journal of General Physiology*, *107*(6), 663–694.

Mumford, D. (1992). On the computational architecture of the neocortex II: The role of cortico-cortical loops. *Biological Cybernetics*, *66*(3), 241–251.

Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In C. Koch & J. Davis (Eds.), *Large-scale theories of the cortex* (pp. 125–152). Cambridge, MA: The MIT Press.

Murphy, K. (2001). *The Bayes Net Toolbox for Matlab.* `http://code.google.com/p/bnt/`.

Mushiake, H., Saito, N., Sakamoto, K., Itoyama, Y., & Tanji, J. (2006). Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron*, *50*(4), 631–641.

Nason, S., & Laird, J. E. (2005). Soar-RL: Integrating reinforcement learning with Soar. *Cognitive Systems Research*, *6*(1), 51–59.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, *341*, 52–54.

Nicola, S. M., Surmeier, D. J., & Malenka, R. C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annual Review of Neuroscience*, *23*, 185–215.

Nieuwenhuis, S., Slagter, H. A., Geusau, V., Alting, N. J., Heslenfeld, D. J., & Holroyd, C. B. (2005). Knowing good from bad: differential activation of

human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*, *21*(11), 3161–3168.

Ninokura, Y., Mushiake, H., & Tanji, J. (2004). Integration of temporal order and object information in the monkey lateral prefrontal cortex. *Journal of Neurophysiology*, *91*(1), 555–560.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends in Cognitive Sciences*, *10*(8), 375–381.

Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 321–349). Cambridge, England: Cambridge University Press.

O'Dohert, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53.

O'Doherty, J. P., Buchanan, T. W., Seymour, B., & Dolan, R. J. (2006). Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron*, *49*(1), 157–166.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, England: Oxford University Press.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750.

Ostlund, S. B., & Balleine, B. W. (2005). Lesions of medial prefrontal cortex disrupt

the acquisition but not the expression of goal-directed learning. *The Journal of Neuroscience*, *25*(34), 7763–7770.

Ostlund, S. B., & Balleine, B. W. (2007). Orbitofrontal cortex mediates outcome encoding in pavlovian but not instrumental conditioning. *The Journal of Neuroscience*, *27*(18), 4819–4825.

Ostlund, S. B., Winterbauer, N. E., & Balleine, B. W. (2009). Evidence of action sequence chunking in goal-directed instrumental conditioning and its dependence on the dorsomedial prefrontal cortex. *The Journal of Neuroscience*, *29*(25), 8280–8287.

Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *The Journal of Neuroscience*, *29*(44), 14004–14014.

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model. *Annual Review of Neuroscience*, *34*, 333–359.

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223–226.

Padoa-Schioppa, C., Jandolo, L., & Visalberghi, E. (2006). Multi-stage mental process for economic choice in capuchins. *Cognition*, *99*(1), B1–B13.

Parr, R., & Russell, S. (1998). Reinforcement learning with hierarchies of machines. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems* (Vol. 10, pp. 1043–1049). The MIT Press.

Pasquereau, B., Nadjar, A., Arkadir, D., Bezard, E., Goillandeau, M., Bioulac, B., ... Boraud, T. (2007). Shaping of motor responses by incentive values through the basal ganglia. *The Journal of Neuroscience*, *27*(5), 1176–1183.

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Phan, K. L., Wager, T. D., Taylor, S. F., & Liberzon, I. (2004). Functional neuroimag-

ing studies of human emotions. *CNS Spectrums*, *9*, 258–266.

Pickens, C. L., Saddoris, M. P., Gallagher, M., & Holland, P. C. (2005). Orbitofrontal lesions impair use of cue-outcome associations in a devaluation task. *Behavioral Neuroscience*, *119*(1), 317.

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *The Journal of Neuroscience*, *27*(37), 9984–9988.

Platt, M. L., Dayan, P., Dehaene, S., McCabe, K., Menzel, R., Phelps, E., ... Singer, W. (2008). Neural correlates of decision making. In C. Engel & W. Singer (Eds.), *Better than conscious? Decision making, the human mind, and implications for insitutions* (pp. 125–154). Cambridge, MA: The MIT Press.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63.

Polsky, A., Mel, B. W., & Schiller, J. (2004). Computational subunits in thin dendrites of pyramidal cells. *Nature Neuroscience*, *7*(6), 621–627.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, *16*(9), 1170–1178.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, *26*, 381–410.

Presson, P. K., & Benassi, V. A. (1996). Illusion of control: A meta-analytic review. *Journal of Social Behavior & Personality*, *11*(3), 493–510.

Puterman, M. L. (2005). *Markov decision processes: discrete stochastic dynamic programming*. Hoboken, NJ: Wiley.

Rangel, A. (2008). The computation and comparison of value in goal-directed choice. In P. W. Glimcher, C. F. Camerer, R. A. Poldrack, & E. Fehr (Eds.),

Neuroeconomics: Decision making and the brain (pp. 425–440). Academic Press.

Rangel, A. (2009). The neuroeconomics of simple goal-directed choice. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1075–1084). The MIT Press.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556.

Rangel, A., & Clithero, J. A. (2013). The computation of stimulus values in simple choice. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: Decision making and the brain* (Vol. 2, pp. 125–147). Academic Press.

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, *20*(2), 262–270.

Rao, R. P. N. (2005). Hierarchical Bayesian inference in networks of spiking neurons. *Advances in Neural Information Processing Systems*, *17*, 1113–1120.

Rao, R. P. N. (2007). Neural models of Bayesian belief propagation. *Bayesian brain: Probabilistic approaches to neural coding*, 239.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

Rao, R. P. N., Shon, A. P., & Meltzoff, A. N. (2007). A Bayesian model of imitation in infants and robots. *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*, 217–248.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Reid, A. K., & Staddon, J. E. R. (1998). A dynamic route finder for the cognitive

map. *Psychological Review*, *105*(3), 585–601.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current Research and Theory*, 64–99.

Reynolds, J. R., & O'Reilly, R. C. (2009). Developing PFC representations using reinforcement learning. *Cognition*, *113*(3), 281–292.

Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370–379.

Roberts, A. C. (2006). Primate orbitofrontal cortex and adaptive behaviour. *Trends in Cognitive Sciences*, *10*(2), 83–90.

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: A MDL-based cognitive model of inductive learning. *Cognitive Science*, *35*(7), 1352–1389.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: a dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370–392.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, *10*(12), 1615–1624.

Rolls, E. T. (2004). The functions of the orbitofrontal cortex. *Brain and Cognition*, *55*(1), 11–29.

Rolls, E. T. (2006). Brain mechanisms underlying flavour and appetite. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1471), 1123–1136.

Rolls, E. T., Everitt, B., & Roberts, A. (1996). The orbitofrontal cortex [and discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *351*(1346), 1433–1444.

Rolls, E. T., Grabenhorst, F., & Deco, G. (2010). Choice, difficulty, and confidence

in the brain. *Neuroimage*, *53*(2), 694–706.

Ross, S., Pineau, J., Paquet, S., & Chaib-Draa, B. (2008). Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research*, *32*, 663–704.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123.

Rudebeck, P. H., Walton, M. E., Smyth, A. N., Bannerman, D. M., & Rushworth, M. F. S. (2006). Separate neural pathways process different decision costs. *Nature Neuroscience*(9), 1161–1168.

Rummery, G. A., & Niranjan, M. (1994). On-line Q-learning using connectionist systems. *Cambridge University Engineering Department: Technical Report CUED/F-INFENG/TR 166*.

Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*(9), 410–417.

Russell, S., & Norvig, P. (2002). *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice Hall.

Rustichini, A. (2008). Neuroeconomics: Formal models of decision making and cognitive neuroscience. In P. W. Glimcher, C. F. Camerer, R. A. Poldrack, & E. Fehr (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 33–46). Academic Press.

Rustichini, A., Dickhaut, J., Ghirardato, P., Smith, K., & Pardo, J. V. (2005). A brain imaging study of the choice procedure. *Games and Economic Behavior*, *52*(2), 257–282.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*(2), 273–284.

Saito, N., Mushiake, H., Sakamoto, K., Itoyama, Y., & Tanji, J. (2005). Representation

of immediate and final behavioral goals in the monkey prefrontal cortex during an instructed delay period. *Cerebral Cortex*, *15*(10), 1535–1546.

Sakai, K. (2008). Task set and prefrontal cortex. *Annual Review of Neuroscience*, *31*, 219–245.

Sakai, K., & Passingham, R. E. (2003). Prefrontal interactions reflect future task operations. *Nature Neuroscience*, *6*, 75–81.

Salamone, J. D., Correa, M., Farrar, A., & Mingote, S. M. (2007). Effort-related functions of nucleus accumbens dopamine and associated forebrain circuits. *Psychopharmacology*, *191*(3), 461–482.

Salas, R., Baldwin, P., De Biasi, M., & Montague, P. R. (2010). BOLD responses to negative reward prediction errors in human habenula. *Frontiers in Human Neuroscience*, *4*(36).

Samejima, K., & Doya, K. (2007). Multiple representations of belief states and action values in corticobasal ganglia loops. In B. W. Balleine, K. Doya, J. O'Doherty, & M. Sakagami (Eds.), (Vol. 1104, pp. 213–228). New York, NY: New York Academy of Sciences.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337–1340.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 211–229.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, *8*, 657–661.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492.

Schmajuk, N. A., & Thieme, A. D. (1992). Purposive behavior and cognitive

mapping: a neural network model. *Biological Cybernetics*, *67*(2), 165–174.

Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Computation*, *4*(2), 234–242.

Schoenbaum, G., & Setlow, B. (2001). Integrating orbitofrontal cortex into prefrontal theory: common processing themes across species and subdivisions. *Learning & Memory*, *8*(3), 134–147.

Schoenbaum, G., Setlow, B., Saddoris, M. P., & Gallagher, M. (2003). Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. *Neuron*, *39*(5), 855–867.

Schultz, W. (2010). Review dopamine signals for reward value and risk: basic and recent data. *Behavioral and Brain Functions*, *6*(24).

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, *10*(3), 272–283.

Schütz-Bosbach, S., & Prinz, W. (2007). Prospective coding in event representation. *Cognitive Processing*, *8*(2), 93–102.

Schwartz, M. F., Montgomery, M. W., Fitzpatrick-DeSalme, E. J., Ochipa, C., Coslett, H. B., & Mayer, N. H. (1995). Analysis of a disorder of everyday action. *Cognitive Neuropsychology*, *12*(8), 863–892.

Scrucca, L. (2013). GA: a package for genetic algorithms in R. *Journal of Statistical Software*, *53*(4), 1–37.

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *The Journal of Neuroscience*, *27*(18), 4826–4831.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J.,

. . . Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(6992), 664–667.

Shachter, R. D., & Peot, M. A. (1992). Decision making using probabilistic inference methods. In *Proceedings of the eighth international conference on uncertainty in artificial intelligence* (pp. 276–283).

Shadlen, M. N. (2008). Neurobiology of decision making: An intentional framework. In C. Engel & W. Singer (Eds.), *Better than conscious? Decision making, the human mind, and implications for insitutions* (pp. 71–102). Cambridge, MA: The MIT Press.

Shadlen, M. N., & Newsome, W. T. (1998). The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *The Journal of Neuroscience*, *18*(10), 3870–3896.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *298*(1089), 199–209.

Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain*, *114*(2), 727–741.

Shallice, T., & Cooper, R. P. (2011). *The organization of mind*. Oxford, England: Oxford University Press.

Simon, D. A., & Daw, N. D. (2011a). Environmental statistics and the trade-off between model-based and TD learning in humans. In J. Shawe-Taylor, R. S. Zemel, O. Bartlett, F. C. N. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 24, pp. 127–135).

Simon, D. A., & Daw, N. D. (2011b). Neural correlates of forward planning in a spatial decision task in humans. *The Journal of Neuroscience*, *31*(14), 5526–5539.

Simon, H. A. (1987). Bounded rationality. In J. Eatwell, M. Milgate, & P. Newman

(Eds.), *The new palgrave dictionary of economics* (pp. 266–268). London, England: Macmillan.

Simsek, Ö., & Barto, A. G. (2008). Skill characterization based on betweenness. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21, pp. 1497–1504).

Singh, S. P., Barto, A. G., & Chentanez, N. (2004). Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 17, pp. 1281–1288).

Skinner, B. F. (1935). Two types of conditioned reflex and a pseudo type. *The Journal of General Psychology*, *12*(1), 66–77.

Sloman, S. (2005). *Causal models*. Oxford, England: Oxford University Press.

Smith, A., Li, M., Becker, S., & Kapur, S. (2004). A model of antipsychotic action in conditioned avoidance: A computational approach. *Neuropsychopharmacology*, *29*(6), 1040–1049.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*, *119*(1), 120–154.

Solway, A., Diuk, C., Cordova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (n.d.). Optimal behavioral hierarchy.

Spence, K. W. (1956). *Behavior theory and conditioning*. New Haven, CT: Yale University Press.

Srinivasan, M. V., & Bernard, G. D. (1976). A proposed mechanism for multiplication of neural signals. *Biological cybernetics*, *21*(4), 227–236.

Stalnaker, T. A., Calhoon, G. G., Ogawa, M., Roesch, M. R., & Schoenbaum, G. (2010). Neural correlates of stimulus–response and response–outcome associations in dorsolateral versus dorsomedial striatum. *Frontiers in Integrative Neuroscience*, *4*(12).

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *The Quarterly Journal of Experimental Psychology*, *62*(6), 1041–1062.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Stewart, T. C., West, R., & Lebiere, C. (2009). Applying cognitive architectures to decision making: How cognitive theory and the equivalence measure triumphed in the technion prediction tournament. In *The ninth international conference on cognitive modeling*.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, *304*(5678), 1782–1787.

Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning* (Unpublished doctoral dissertation). University of Massachusetts.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the 7th international conference on machine learning* (pp. 216–224).

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*(2), 135–170.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA: The MIT Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. The MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs:

A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*, 181–211.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain*. Stuttgart, Germany: Thieme.

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating consequences: brain systems that encode the causal effects of actions. *The Journal of Neuroscience*, *28*(26), 6750–6755.

Tanji, J., & Hoshi, E. (2008). Role of the lateral prefrontal cortex in executive behavioral control. *Physiological Reviews*, *88*, 37–57.

Tanji, J., Shima, K., & Mushiake, H. (2007). Concept-based behavioral planning and the lateral prefrontal cortex. *Trends in Cognitive Sciences*, *11*(12), 528–534.

Tatman, J. A., & Shachter, R. D. (1990). Dynamic programming and influence diagrams. In (Vol. 20, pp. 365–379).

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 301–322). Oxford, England: Oxford University Press.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York: Macmillan.

Thrun, S., & Schwartz, A. (1995). Finding structure in reinforcement learning. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 385–392).

Tinbergen, N. (1951). *The study of instinct*. Oxford, England: Clarendon Press.

Tolman, E. C. (1932). *Purposive behavior in animals and men*. New York, NY: Century.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Tolman, E. C. (1949). The nature and functioning of wants. *Psychological Review*,

56(6), 357–369.

Tolman, E. C., & Honzik, C. H. (1930). "Insight" in rats. *University of California Publications in Psychology*, *4*, 215-232.

Toussaint, M., Charlin, L., & Poupart, P. (2008). Hierarchical POMDP controller optimization by likelihood maximization. In *Proceedings of the 24th international conference on uncertainty in artificial intelligence* (Vol. 24, pp. 562–570).

Toussaint, M., & Storkey, A. (2006). Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proceedings of the 23rd international conference on machine learning* (pp. 945–952).

Towal, R. B., Mormann, M., & Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *Proceedings of the National Academy of Sciences*, *110*(40), E3858–E3867.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, *398*, 704–708.

Tsetsos, K., Usher, M., & Chater, N. (2010). Preference reversal in multiattribute choice. *Psychological Review*, *117*(4), 1275–1291.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281–299.

Ullsperger, M., & von Cramon, D. Y. (2003). Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. *The Journal of Neuroscience*, *23*(10), 4308–4314.

Unterrainer, J., & Owen, A. (2006). Planning and problem solving: from neuropsychology to functional neuroimaging. *Journal of Physiology-Paris*, *99*(4-6), 308–317.

Usher, M., Elhalal, A., & McClelland, J. L. (2008). The neurodynamics of choice,

value-based decisions, and preference reversal. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: prospects for Bayesian cognitive science* (pp. 278–300). Oxford, England: Oxford University Press.

Uylings, H., Groenewegen, H. J., & Kolb, B. (2003). Do rats have a prefrontal cortex? *Behavioural Brain Research*, *146*(1), 3–17.

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *The Journal of Neuroscience*, *27*(15), 4019–4026.

van Dijk, S. G., & Polani, D. (2011). Grounding subgoals in information transitions. In *IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)* (pp. 105–111).

van Dijk, S. G., & Polani, D. (2013). Informational constraints-driven organization in goal-directed behavior. *Advances in Complex Systems*, *16*(02n03), 1350016.

van Dijk, S. G., Polani, D., & Nehaniv, C. L. (2011). Hierarchical behaviours: getting the most bang for your bit. In G. Kampis, I. Karsai, & E. Szathmáry (Eds.), *Advances in artificial life: Darwin meets von neumann* (pp. 342–349). Springer.

van Praag, B. M. S. (1968). *Individual welfare functions and consumer behavior: A theory of rational irrationality*. Amsterdam, The Netherlands: North-Holland.

van Veen, V., Holroyd, C. B., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2004). Errors without conflict: implications for performance monitoring theories of anterior cingulate cortex. *Brain and Cognition*, *56*(2), 267–276.

Verma, D., & Rao, R. P. N. (2006a). Goal-based imitation as probabilistic inference over graphical models. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 1393–1400). Cambridge, MA: MIT Press.

Verma, D., & Rao, R. P. N. (2006b). Planning and acting in uncertain environments using probabilistic inference. In *IEEE/RSJ international conference on intelligent*

robots and systems (pp. 2382–2387). Beijing, China.

Vigorito, C. M., & Barto, A. G. (2010). Intrinsically motivated hierarchical skill learning in structured environments. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 132–143.

Voicu, H., & Schmajuk, N. (2002). Latent learning, shortcuts and detours: a computational model. *Behavioural Processes*, *59*(2), 67–86.

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

Wagenmakers, E.-J., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*(3), 332–367.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, *19*(3), 326–339.

Wallis, J. D. (2007). Orbitofrontal cortex and its contribution to decisionmaking. *Annual Review of Neuroscience*, *30*, 31–56.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*, 953–956.

Wallis, J. D., & Miller, E. K. (2003). From rule to response: neuronal processes in the premotor and prefrontal cortex. *Journal of Neurophysiology*, *90*(3), 1790–1806.

Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. E. M., & Rushworth, M. F. S. (2006). Weighing up the benefits of work: behavioral and neural analyses of effort-related decision making. *Neural Networks*, *19*(8), 1302–1314.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Unpublished doctoral dissertation). Cambridge University, Cambridge, England.

Weiss, Y., & Pearl, J. (2010). Belief propagation. *Communications of the ACM*, *53*(10), 94.

White, I. M., & Wise, S. P. (1999). Rule-dependent neuronal activity in the prefrontal

cortex. *Experimental Brain Research*, *126*(3), 315–335.

Whiten, A., Flynn, E., Brown, K., & Lee, T. (2006). Imitation of hierarchical action structure by young children. *Developmental Science*, *9*(6), 574–582.

Wickens, J., Kotter, R., & Houk, J. C. (1995). Cellular models of reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 187–214). The MIT Press.

Wiener, J. M., & Mallot, H. A. (2003). 'Fine-to-coarse' route planning and navigation in regionalized environments. *Spatial Cognition & Computation*, *3*(4), 331–358.

Williams, B. A. (1989). The effects of response contingency and reinforcement identity on response suppression by alternative reinforcement. *Learning and Motivation*, *20*(2), 204–224.

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science*, *338*(6104), 270–273.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1431), 593–602.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit–goal interface. *Psychological Review*, *114*(4), 843–863.

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, *15*(5), 786–791.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245–272.

Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D. F., & Büchel, C. (2006). Dissociable systems for gain-and loss-related value predictions and errors of

prediction in the human brain. *The Journal of Neuroscience*, *26*(37), 9530–9537.

Yamaguchi, S., Tsuchiya, H., & Kobayashi, S. (1995). Electrophysiologic correlates of visuo-spatial attention shift. *Electroencephalography and Clinical Neurophysiology*, *94*(6), 450–461.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959.

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex*, *15*(5), 535–544.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464–476.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action–outcome learning in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 505–512.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513–523.

Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Dynamics of attentional selection under conflict: toward a rational bayesian account. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 700–717.

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience*, *23*(12), 4057–4066.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007).

Event perception: a mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293.

Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2009). Human substantia nigra neurons encode unexpected financial rewards. *Science*, *323*(5920), 1496–1499.