Computational Mechanisms of Selective Attention during Reinforcement Learning

Angela Radulescu

a dissertation
presented to the faculty
of princeton university
in candidacy for the degree
of Doctor of Philosophy

recommended for acceptance
by the Department of
Psychology

Primary adviser: Yael Niv

Secondary adviser: Nathaniel Daw

June 2020

© Copyright by Angela Radulescu, 2020. All rights reserved.

Abstract

The multidimensional nature of our environment raises a fundamental question in the study of learning and decision-making: how do we know which dimensions are relevant for reward, and which can be ignored? For instance, an action as simple as crossing the street might benefit from selective attention to the speed of incoming cars, but not the color or make of each vehicle. This thesis proposes a role for selective attention in restricting representations of the environment to relevant dimensions. It further argues that such representations should be guided by inferred structure of the environment. Based on data from a paradigm designed to assess the dynamic interaction between learning and attention, the thesis introduces a novel sequential sampling mechanism for how such inference could be realized.

The first chapter discusses selective attention in the context of Partially-Observable Markov Decision Processes. Viewed through this lens, selective attention provides a mapping from perceptual observations to state representations that support behavior.

Chapter 2 provides evidence for the role of selective attention in learning such representations. In the 'Dimensions Task,' human participants must learn from trial and error which of several features is more predictive of reward. A model-based analysis of choice data reveals that humans selectively focus on a subset of task features. Age-related differences in the breadth of attention are shown to modulate the speed with which humans learn the correct representation.

Next, a method is introduced for directly measuring the dynamics of attention allocation during multidimensional reinforcement learning. fMRI decoding and eye-tracking are combined to compute a trial-by-trial index of attention. A model-based analysis reveals a bidirectional interaction between attention and learning: attention constrains learning; and learning, in turn guides attention to predictive dimensions.

Finally, Chapter 4 draws from statistical theory to explore a novel mechanism for selective attention based on particle filtering. The particle filter keeps track of a single hypothesis about task structure, and updates it in light of incoming evidence. To offset the sparsity of the representation suggested by gaze data, the particle filter is augmented with working memory for recent observations. Gaze dynamics are shown to be more consistent with the particle filter than with gradual trial-and-error learning. This chapter offers a novel account of the interaction between working memory and selective attention in service of representation learning, grounded in normative inference.

Contents

Listing of figures

To my parents, Eugenia and Silviu.

Acknowledgments

First and foremost, I thank my advisor, Yael Niv. I have had the privilege to be mentored by Yael for many years, first during my time as lab manager, and then as I continued on as a graduate student in her lab. Yael's influence on every part of my scientific journey cannot be overstated. She brought focus, clarity and rigor to often meandering research pursuits. Her gift for explaining abstract topics in the clearest possible way, and her superpower of unraveling the mysteries of data, have made me the researcher I am today. Yael is not only a brilliant scientist, but also a supportive mentor, and incredibly generous with her time and ideas. While at first I lacked any semblance of computational expertise, Yael quickly put my insecurities at ease, and conveyed to me the tools and concepts that became the backbone of years of research. She fosters a dynamic and collaborative lab environment, in which it is easy to grow and thrive as a scientist. And she generously offered me all the resources and intellectual freedom I needed to pursue a broad range of scientific questions, providing the right balance of autonomy and guidance. On a more personal note, I cannot imagine an advisor more attentive to her trainees' needs. When I needed to return home to reconnect with my father, she was supportive and gave me the time to do so. When I struggled with mental health issues in my second year, she was caring and pragmatic. As I journeyed towards self-acceptance regarding aspects of my identity, she made a point of being inclusive and affirming. Yael always puts her trainees first, often making me wonder how she can possibly have time and energy left for anything else. It is largely thanks to her that I gathered the knowledge, enthusiasm and courage that I take with me in future research.

I thank Nathaniel Daw, who has been a joy to work with insofar as our conversations always served as a very effective Occam's Razor. I more than once reflected for days on a passing remark that Nathaniel made in a meeting, only to have an "a-ha" moment weeks later. The model described in Chapter 4 is the result of direct collaboration with Nathaniel, during which he kindly guided me as I elaborated on his ideas on the link between approximate inference and reward learning. This inquiry not only led to a satisfying theory of attention learning, but also made me a stronger modeler and wielder of Bayes' rule. Nathaniel has taught me to seek simple, elegant explanations of otherwise opaque data.

During my first two years of graduate school, I benefited from the support and mentorship of Nicholas Turk-Browne. Nick's engaging cognitive psychology seminar still forms the basis of my knowledge of the field. His instinct for sharp and clear experiment design is a standard I always aspire to.

Even though I never formally was a member of Ken Norman's lab, I very much appreciated him welcoming me to the *CompMem* family. Our early scientific collaboration cemented my conviction that Princeton is the right academic environment for me to pursue graduate studies. And if I know anything about neural networks, it is thanks to Ken's demystifying and entertaining explanations.

Casey Lew-Williams has been an absolute wonder to have in my corner. I was privileged to assist in teaching his Child Development course, and learned a great deal about pedagogy from him. Outside teaching, knowing that Casey's door is always open made the graduate school experience far less daunting.

I would also like to thank Diana Tamir and Tom Griffiths, for sharing their comments and insights during the dissertation writing processes.

Outside Princeton, I thank Elke Weber, Jackie Gottlieb and Daphna Shohamy, who were early influences on my choice to pursue a doctorate degree.

Last but not least, I owe deep gratitude to James Hillis, my manager during an internship at *Facebook Reality Labs*. He gave me a shot at applying what I know in "the wild", and taught me the value of considering engineering outcomes and real-world behavior when developing theories of the mind. His guidance and respect for my ideas has greatly influenced and invigorated my thinking, and led to a fruitful research collaboration that I take with me in future work.

Graduate school would not have been the same without the presence of many colleagues and collaborators who have become friends.

The Niv Lab as a whole is a wonderful place to be a graduate student. Many members, past and present, have contributed to my work, either through direct collaboration, reading a paper draft, or making an insightful comment during lab meetings. I have so many to thank: Dan Bennett, Branson Byers, Mingbo Cai, Stephanie Chan, Reka Daniel, Vivian deWoskin, Guy Davidson, Carlos Diuk, Sarah DuBrow, Nicole Drummond, Eran Eldar, Val Felso, Andra Geana, Sam Gershman, Gecia Hermsdorff, Kat Holmes, Alana Jaskir, Angela Langdon, Yuan Chang Leong, Julie Newman, Nina Rouhani, Nico Schuck, Mel Sharpe, Yeon Soon Shin, Mingyu Song, Michael Todd, Bob Wilson, Sam Zorowitz. Doing science in your company has been the best part of graduate school.

I especially wish to thank Reka Daniel and Yuan Chang Leong. Much of the work presented in this thesis has benefited from their efforts and insights. Without them, Chapter 3 would not exist. Our study of the Faces-Landmarks-Tools task set a high standard for col-

laborations for years to come, and I am extremely grateful for our continued friendship and scientific exchanges.

I'm also fortunate enough to have collaborated on a synthesis of my thesis work with Ian Ballard. His extremely useful suggestions for analyses, incisive explanations of papers and concepts and commanding knowledge of cognitive neuroscience were invaluable in shaping a cohesive "big-picture" view of disparate research threads.

During the second half of my graduate training, I began exploring some of the implications of my work for computational psychiatry. This endeavor has been decisively shaped by discussion and collaboration with Daniel Bennett and Sam Zorowitz. Dan and Sam are sharp, deep thinkers who wield computational techniques with ease, and show great respect for how hard it is to conduct meaningful psychiatric research. I particularly thank Dan, for helping me mold vague intuitions into a precise empirical study of how mood influences value-based attention; and for sharing his encyclopedic knowledge of psychiatry and clinical psychology. And I thank Sam for being the catalyst of a long overdue transition to Python; developing *NivLink* together is one of my proudest achievements in graduate school.

Over the last year, Bas van Opheusden and Fred Callaway have become collaborators on a project on how some of the insights developed in this thesis might apply to naturalistic environments. Working with them under the auspices of *Facebook Reality Labs* has been fun and rewarding, giving me a fresh perspective on how attention might support decision-making. I thank them for that, and for converting me to a power Slack user.

I am grateful for the amazing and vibrant PSY/PNI+ community past and present, a unique mix of fascinating science and personality. In no particular order, I thank the following people for memorable academic and/or personal interactions: Mai Nguyen (cohort shout-out!), Ahmed El Hady, Mark Thornton, Debbie Yee, Abby Novick, Anne Mennen, Laura Bustamante, Clarice Robenalt, Lili Cai, Nathan Parker, Aaron Bornstein, Michael Arcaro, Cameron Ellis (cohort shout-out!), Ida Momennejad, Maijia Honig, Anna Schapiro, Megan de Bettencourt, Wouter Kool, Sam McDougle, Robin Gomila, Jon Berliner, Brandy Briones (cohort shout-out!), Kara Enz, Zidong Zhao, Judy Fan, Mariam Aly, Victoria Ritvo (cohort shout-out!), Sebastian Michelmann, Ilana Witten, Dakota Blackman, Yotam Sagiv, NaYeon Kim, Olga Lositsky, Joel Martinez, DongWon Oh, Xiaofang Yang, Aaron Kurosu (cohort shout-out!), Cristina Domnisoru, Sebastian Musslick, Qihong Lu, Judith Mildner, Evan Russek, Oliver Vikbladh, Jose Ribas-Fernandez, Francisco Pereira, Jonathan Pillow, Athena Akrami, Shiva Rouhani, Lindsay Hunter, Alex Libby, Gecia Hermsdorff, Alex Piet, Kim Stachenfeld, Alex Song, Diana Liao, Matt Panichello, Luis Piloto, Taylor Webb, Jessie Schwab (cohort shout-out!), Gary Kane, Hanna Hillman, Marius-Catalin Iordan, Elise Piazza, Carlos Correa, Rolando Masis, Talmo Pereira, Flora Bouchacourt, Hessam Akhlaghpour and Pavlos Kollias.

A huge thank you to the support staff across PSY/PNI: Keisha Craig, Tina McCoy, David Carter, RoseMarie Stevenson, Sami Mezger, Alex Lewis, Jeanne Heather, Paryn Wallace, just to name a few.

I was fortunate enough to be part of *The Prison Teaching Initiative*, an incredible organization that gave deeper meaning to my scholarship. I thank all the students who took *PSY101*, asking questions and challenging us to be better teachers. I am grateful for the friendships I formed through *PTI*, which helped me connect to the Princeton community at large. I will miss the long car rides and spontaneous food adventures that broke the monotony of graduate life. I thank Laura Bustamante, Stephanie Chan, Ian Davies, RL Goldberg, Robin Gomila, Tanja Kassuba, Matt King, Jill Knapp, Sam McDougle, Tara Ronda, Nina Rouhani, Matt Spellberg, Jill Stockwell, Taylor Webb, and Sarah Wilterson. Finally, I thank Annegret Dettwiler for providing yet another source of steady and kind mentorship throughout my time at Princeton.

Thank you also to other members of my team at Facebook, especially Serena Bochereau, Deb Boehm-Davis, Matt Boring, Ruta Desai, Tanya Jonker, Gabor Lengyel, Karl Ridgeway, Majed Samad and Mike Shvartsman (of Princeton fame). Each of them contributed in important ways to my research and made the time spent in Redmond a fun and formative one.

As I am nearing the end of this journey, I am reminded that getting this far would not have been possible without the encouragement of several people who at various times provided invaluable friendship, advice and support:

My office mates, Nina Rouhani and Yeon Soon Shin.

James Antony, Sarah DuBrow, Kara Enz, Andra Geana, Monika Schönauer, and Jamal Williams.

My housemates at Bank St throughout the years, Stephanie Chan, Andras Gyenis, Julia Langer, Ryan Ly, Kevin Miller, and Kelsey Ockert.

My friends away from Princeton, who are always ready to lend a patient ear, Alexandra Ioan, Adona Iosif, Lyuda Kovalchuke, Andra Mihali, Embry Owen, Maria Popa, Cristi Proistosescu, Sorina Vasile, and Daniella Zalcman.

I am lucky to have the most amazing sister, Maria, and parents, Eugenia and Silviu. Maria and I shared an apartment for a couple of years during graduate school, and I am so grateful for the time we got to spend together in New York as aspiring adults. Without my mother Eugenia's inspiring tenacity and commitment to science, I likely would never have left Romania, let alone pursue a childhood dream to train at Princeton. Her fierce love for her family is only matched by her passion for research. My father Silviu's steady encouragement has helped me maintain momentum in putting together this dissertation. It is he who gave me my first computer, and, together with my mother, nurtured my early scientific inclinations. Without them, none of this would be possible.

Finally, I am profoundly indebted to my wonderful partner, Juliet. It is her guidance and encouragement that carried me over the finish line during an especially challenging time. Juliet, it's been so much easier to tell my story since we embarked on our adventure together. Thank you for making loving fun!

The art of being wise is the art of knowing what to overlook.

William James

0

Introduction

Millennial decision-making can be complicated. We have a myriad of apps at our disposal providing us with ever more goals to strive towards. Our attention is in high demand by different facets of social media. Our working memory is taxed by constant multitasking. To go about our daily lives, we must somehow discount the distal sense of catastrophe surrounding climate change (and the not-so-distal one brought on by a global pandemic). Yet, even in these conditions, we strive to make sense of the world and make good decisions. At the core of this flexibility is the ability to select relevant sources of information, and use them for appropriate action selection. The central theme of this dissertation is the dynamic interaction of selective attention and learning. I share what I have learned in my attempt to glean an answer to a fundamental question about this interaction: **what is worth learning about?**

0.1 Defining attention

Every organism receives input from the world through a limited number of sensors. Human vision for example is possible because when light hits the retina, specialized photore-

ceptors transform photons into action potentials. From there, information travels along the visual pathway. By the time neural signals reach the primary visual cortex, they have been optimized for highly specialized functions, such as edge detection and depth perception (Kuffler, 1953; Blake & Logothetis, 2002). The physical peculiarities of the visual system determine many aspects of how we see. For instance, we only have three types of color-sensitive photoreceptors, which limit the color spaces we can perceive. The decreasing acuity in our peripheral vision means that we need to constantly move our eyes ("foveate") in order to increase the quality of our perception and maintain a coherent picture of the world. So at least partially, evolution has selected visual information for us.

Nevertheless, the 2D projection of the environment is rich with sources of information: a single visual scene may contain common features such as objects of different colors and shapes; but also less likely entities, like starfish. My work approaches the problem of how we sift through this information and learn to focus only on features that are relevant for the task at hand. I take an implicitly normative stance throughout: selective attention can be construed as a cognitive action (Dayan, 2012; Frank, 2011). Like any other action, it ought to serve some adaptive purpose. And that purpose depends on one's goals.

Attention can be concisely defined as the selective processing of a subset of features at any stage between sensation and action (Gottlieb, 2012). In my doctoral work, I have primarily studied the selective processing of visual features when choosing between options, and during explicit reward feedback. But the hope is to develop a more general theory that can be applied to a broad class of tasks and representational spaces.

0.2 Thesis overview

This thesis comprises of a background chapter, three main chapters that together offer an account of the role of attention during reinforcement learning, and a conclusions chapter.

In Chapter 1, I provide much of the theoretical background useful for understanding the rest of the thesis. I explicitly frame selective attention in the computational language that has been used to study goal-directed action. In particular, I propose a role for attention in carving state representations for reinforcement learning. Drawing on a rich literature on human category learning, I suggest that such **representation learning** is guided by inferring the structure of the environment.

In Chapter 2, I study how the interaction of reinforcement learning and selective attention critical in forming task representations changes with age. In particular, I present results from a multidimensional learning paradigm in which humans must identify which features are relevant for accruing reward. I show how age-related differences in representation learning can be captured using trial-by-trial fitting of computational models to behavioral choice data. In particular, older adults seem to employ a narrower focus of attention when learning what to attend to.

The main model used to study age differences in Chapter 2 is not expressive enough to fully distinguish selective attention from passive forgetting of past experiences. To overcome this limitation, in Chapter 3 I provide a set of novel empirical methods for studying the interaction between learning and attention. I use data from two different modalities, functional magnetic resonance imaging (fMRI) and eye-tracking, to directly measure the

dynamics of attention learning. I develop a method for fitting computational models not just to choices, but to trial-by-trial attention dynamics. Using this method, I show that attention and learning are engaged in a bidirectional interaction: human reinforcement learning is constrained by attention, and selective attention dynamics are sensitive to reward. Finally, I show that separable attention processes constrain human reinforcement learning during the choice and learning stages of each decision.

In Chapter 4, I present a new computational model of attention learning grounded in statistical inference theory. This model explicitly instantiates the hypothesis in Chapter 1, and is inspired by a rich tradition of treating Monte-Carlo methods as psychologically plausible algorithmic solutions to otherwise intractable Bayesian inference problems. I suggest attention learning can be viewed as particle filtering: humans sequentially entertain one or few hypotheses about which features of a task are relevant, and revise these hypotheses in light of feedback. Previous data from animal learning suggest a particle filter with one, or few particles. To allow the single-particle model to maintain enough evidence to switch hypotheses efficiently, I augment it with memory for recent observations. I show that this memory-augmented particle filter is consistent with trial-by-trial attention dynamics measured in the paradigm described in Chapter 3. This final chapter proposes a novel mechanism for attention learning, and highlights the role of memory in enabling inference of which features in the environment are relevant for reward.

Finally, in Chapter 5 I summarize the conclusions and contributions of the thesis as a whole, as well as directions for future research.

We demand rigidly defined areas of doubt and uncertainty!

Douglas Adams, The Hitchhiker's Guide to the Galaxy

1

Selective attention for state representation

Attention is both intuitively easy to grasp and famously elusive to define (James, Burkhardt, Bowers, & Skrupskelis, 1890). Empirical studies usually conceptualise attention as an (observed) consequence of limited computational resources. For example, in the Posner spatial cuing task, a cue signals the relevance of a particular spatial location. In response to this cue, participants may make an internal decision to shift attention, and devote more computational resources to that location. Attentional shifts can be overt, which involves physically moving one's eyes to collect sensory data; or covert, which involves redistributing computational resources between different sources of sensory data. Controlling attention in this way allows participants to respond faster when a second cue appears at the same location (Posner, 1980). Such facilitation is taken as evidence of attention. As in the Posner Task, across many paradigms, attention is defined in terms of the observed behavioral effects (e.g. faster responding, focus of eye-gaze). However, few attempts exist to explicitly model *why* attention should be devoted to some features (e.g. location) but not others. For such an analysis, we turn to reinforcement learning theory.

In the following sections, I lay out a framework for defining attention from the point

of view of a goal-directed agent that is trying to make optimal decisions in a multidimensional world. I introduce Markov Decision Processes (MDPs), and algorithms for solving MDPs that have been most widely used to model human and animal learning and decision-making. Finally, in a section based on a recently published paper (Radulescu, Niv, & Ballard, 2019), I briefly review their limitations, and data suggesting a role for selective attention in carving state representations in an MDP by inferring relevant structure in the environment.

1.1 Markov Decision Processes in learning and decision-making

An agent's internal model of the way the world unfolds as a consequence of its actions can be formalized as a Markov Decision Process (MDP). This framework spans cognitive science, artificial intelligence and robotics. Algorithmic solutions to MDPs provide a powerful common language for describing how intelligent agents should behave in a variety of environments. Markov Decision Processes (MDPs) consist of tuples $\{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}\}$.

The state space $\mathcal{S}$ denotes a set of states of the external world. We can define *state* as the subset of environmental features relevant to the agent's goal. For a rat trying to find food in the New York subway, the state may consist of its current position on the train tracks; for a deep neural network playing a game of Pong (Mnih et al., 2015), the state may consist of an array of color pixel values.

Given a set of possible actions $\mathcal{A}$, (e.g., turning left and turning right), the transition function $\mathcal{T}$ governs how the state of the world changes when the agent performs an action. For example, moving along the tracks would typically lead the rat to a subway station. Knowledge of $\mathcal{T}$ enables an agent to plan action sequences more efficiently. If the

rat found itself at 50th St, and knew that turning left leads to Times Square, while turning right leads to 59th St, it might generate a simple plan to turn left, and forage for pizza.

A key property of MDPs is that the current state s_t contains all the information needed to determine the probability of ending up in any state at time $t + 1$, given the agent's actions. This is known as the Markov property. For instance, we can say that Times Square is a "Markov state" if and only if where the rat has been before Times Square has no bearing on which stations it will reach in the future if he turns left and walks straight.

Finally, the reward function $\mathcal{R}$ takes as input the state of the environment and the current action a_t, and returns r_t, a scalar representing the immediate utility of performing that action in the current state. In human and animal learning paradigms, rewards are typically construed as primary extrinsic signals, such as juice, shocks, money, or, for our prototypical subway rat, the abundance of pizza.

For a given MDP, a policy π defines a mapping between states and actions. An agent is said to be behaving in a normative manner if the state-action mappings it has learned help it receive the most long-term reward. Finding such policies is difficult because the agent automatically faces a credit assignment problem: when a behavior is protracted, the agent needs to determine which actions led to the outcome. For example, the subway rat might have "turned left", but also "jumped over the tracks" before it got to the pizza. Which of these actions was critical, and which was less important for leading the rat into a better state than it found itself previously?

The quantity in reinforcement learning that summarizes how "good" it is to be in a particular state is known as the *value function*, $\mathcal{V}^{\pi}(s)$. The value of a state is the sum of future (discounted) rewards that an agent can expect to collect when following a specific policy

from that state on. Here, "discounted" means that rewards that are farther in the future are only worth a fraction of immediate ones. We return to New York one last time to illustrate this concept: given the rat's goal to reach the pizza, the value of being at 50th St is higher than being at 59th St Columbus Circle.

Until relatively recently (Daw, Niv, & Dayan, 2005), the dominant theory of human learning had been an algorithm for learning state values known as TD-Learning (Sutton, 1988). Given an MDP with a discrete and finite set of states, TD-Learning updates the value $\mathcal{V}(s)$ of the current state following each outcome, using a temporal-difference learning rule (Niv, 2009). Seminal work by Schultz and colleagues has sparked a flurry of progress in mapping the neural substrates of TD-Learning, leading to the influential "reward prediction error theory" of dopamine function (Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997): when a neutral stimulus such as a tone is followed by an unexpected reward, midbrain dopamine neurons fire; repeating this stimulus-reward pattern leads to the stimulus becoming "conditioned"; that is, it elicits a dopamine response, despite not being intrinsically rewarding. In MDP terms, the stimulus (or state) has now acquired predictive value. The framework elegantly explains why animals learn to associate arbitrary stimuli with reward, clarifying the neural basis of classical conditioning (Pavlov, 1927). This foundational theory has been extended to action selection, most notably in the form of two algorithms, Q-Learning and Actor-Critic (Daw et al., 2005; Daw, 2011; Joel, Niv, & Ruppin, 2002; Collins & Frank, 2014). These RL algorithms have proven powerful descriptors of how humans and animals learn to make good decisions from trial and error.

1.2 Revisiting state

One aspect of human and animal RL theory that might give us pause is that the same algorithms that explain behavior and neural activity on simple learning tasks learn much more slowly as the number of states in the environment grows (Bellman, 1957; Sutton, 1988). This is because the multidimensional nature of the environment quickly leads to a combinatorial explosion: representing all features of the environment (e.g. all possible shapes, or possible colors, and so on) yields too many possible states, each defined by a specific feature configuration.

Recall however our previous definition of state: the subset of environmental features relevant to the agent's goal. This definition rests on the assumption that some features are relevant for behavior, while others are not. So an agent can reduce the dimensionality of the state space by representing only some features and not others. But this gives rise to a second problem: which features ought to be included in the state?

Imagine for instance that you are learning to take photographs. In the beginning, you may not be able to distinguish relevant aspects of the task from those that matter less for taking good photos. But in time, with enough practice and feedback, you will learn to pay attention to the features that matter. Once you take enough photos, you will learn that the light source is more important than the brand of your camera. Experience will eventually lead you to discover that photographing near a window requires different settings, depending on whether you are photographing at midday, or minutes before sunset. All else equal, the nature of light is an aspect of photography that meaningfully distinguishes between different situations. To borrow an ancient metaphor, light carves nature at its joints (Plato, ca

360BCE): if photographers pay attention to it, they will eventually reach their goal.

In multidimensional settings typical of our every day experience, the dynamics of the world might thus be better described as a Partially Observable Markov Decision Process (POMDP) (Kaelbling, Littman, & Cassandra, 1998). In a POMDP, instead of receiving the exact state as input, agents have access to noisy observations $\mathcal{O}$. An agent's *model* of the world can then consist of all or a subset of these computations (Hamrick, 2019):

Representing a mapping between observations and states:

$$p(s_t|o_{1:t}) \tag{1.1}$$

Representing the transition function:

$$p(s_{t+1}|s_t, a_t) \tag{1.2}$$

Representing the reward function:

$$p(r_t|s_t, a_t) \tag{1.3}$$

The rest of this thesis seeks to model how humans accomplish the first aspect of model building, which is to learn a useful mapping between observations and states (Equation 1.1). The process of learning this mapping is known as **representation learning** (Bengio, Courville, & Vincent, 2013).

Equation 1.1 can be interpreted in two ways: in the early POMDP literature, it encoded a probability distribution over true states of the environment. The agent then had the addi-

tional task of inferring what state it is in, given the most recent observations. For example, a wildlife photographer might try to guess the position of an approaching animal from the rustling of leaves.

Alternatively, we can interpret s_t as a *state representation* internal to an agent. This view is convenient, because it does not require us to commit in advance to an exhaustive state space that may not ever be truly knowable. Instead, we can treat Equation 1.1 as a mapping between raw observations and state representations that changes with task demands. The usefulness of this mapping can then be solely determined by reward: if distinguishing between the absence or presence of a feature changes the outcome, then the agent should include it in the state representation (McCallum, 1997). Of course, one might still ask, what is the observation space the agent is mapping from? One conjecture I will make for the purposes of this thesis is that the observation space is only constrained by the "primitive" features that an agent's perceptual system has access to during a given task. For humans, this space might consists of objects with semantic labels which have certain attributes like shape, color, etc. I refrain from speculating how such a feature space might itself be learned, in large part because this endeavor necessarily requires taking a developmental view that is beyond the scope of this thesis (Sanborn, Chater, & Heller, 2009). Still, given an observation space invariant to reward over the course of a task*, I suggest that selective attention can be understood in light of Equation 1.1.

*To illustrate what I mean by this, consider a negative example: we will likely not be able to "condition" a participant to stop perceiving an object over the course of a short laboratory experiment.

1.3 Empirical studies of human representation learning

The flexibility with which humans adapt their state representation to different situations has been studied in multidimensional learning tasks, which I review in detail in Radulescu, Niv, and Ballard (2019). In such tasks, human participants make responses based on stimuli that vary along several dimensions. Observations consist of both the sensory properties of the stimuli, and the reward outcome that follows each action. Efficient learning depends on how participants use raw observations to construct an appropriate mapping between percepts and internal state representations (Equation 1.1). For instance, in Schuck et al. (2015) participants were instructed to respond manually to the location of a patch of colored squares within a square reference frame. A latent deterministic mapping was induced between the color of the patch and the correct response. Despite extended practice with the location-based policy, some participants spontaneously adapted their state representation to the structure of the task, using color to respond faster. This strategy shift was preceded by an increase in color information content in the medial prefrontal cortex. Humans can thus flexibly change strategies from one feature to another in the absence of explicit instructions.

The kinds of paradigms that have been used to study how humans form state representations resemble classic tasks in human category learning. In category learning tasks, participants are required to sort multidimensional stimuli one at a time into one of several categories. Category membership usually depends on the presence or absence of one or more features, as well as on the relationship between features. For example given two category labels "dax" and "bim", a red square would be classified as a "dax" if "all red objects are

daxes" or as a "bim" if "only red circles are daxes" (Ballard, Miller, Piantadosi, Goodman, & McClure, 2018). How humans infer category structure has successfully been modeled with Bayesian models of categorization (Sanborn, Griffiths, & Navarro, 2010; Anderson, 1991; Mansinghka et al., 2016; Goodman, Tenenbaum, Feldman, & Griffiths, 2008).

Tasks probing state representation learning may differ in framing (e.g. decision-making vs. categorization), the size of the observation space (i.e. how many dimensions stimuli can vary on), the nature of the feedback (scalar reward vs. a category label; stochastic vs. deterministic) and the instructions the participant receives about the structure. But they also are alike in that each trial consists of a perceptual observation, an action and a reward outcome. And they share the key property that the participant needs to disambiguate observations by learning and representing an appropriate mapping between perceptual observations and states. How fast participants learn depends on learning to carve the perceptual observation space into a compact state representation appropriate for the task.

In this dissertation, I propose that representations of task structure learned through Bayesian inference are the source of selective attention during learning (Figure 1.1). In POMDP terms, selective attention is a useful mapping function from observation to states (Equation 1.1). This mapping changes from task to task, as the agent infers new models for how rewards are generated, allowing her to make useful distinctions in perceptual space, and biasing attention towards features that are causally related to predictable fluctuations in reward (Mackintosh, 1975; McCallum, 1997). To mitigate the computational intractability of Bayesian inference, I further propose that changes in attention correspond to an approximate sequential sampling mechanism known as a particle filter (Doucet & Johansen, 2009; Sanborn & Chater, 2016; Speekenbrink, 2016). I show how this type of

sequential sampling mechanism can be realized in a multidimensional environment with a simple underlying structure, in which one of three dimensions is relevant, and within that dimension, one of three features predicts reward.

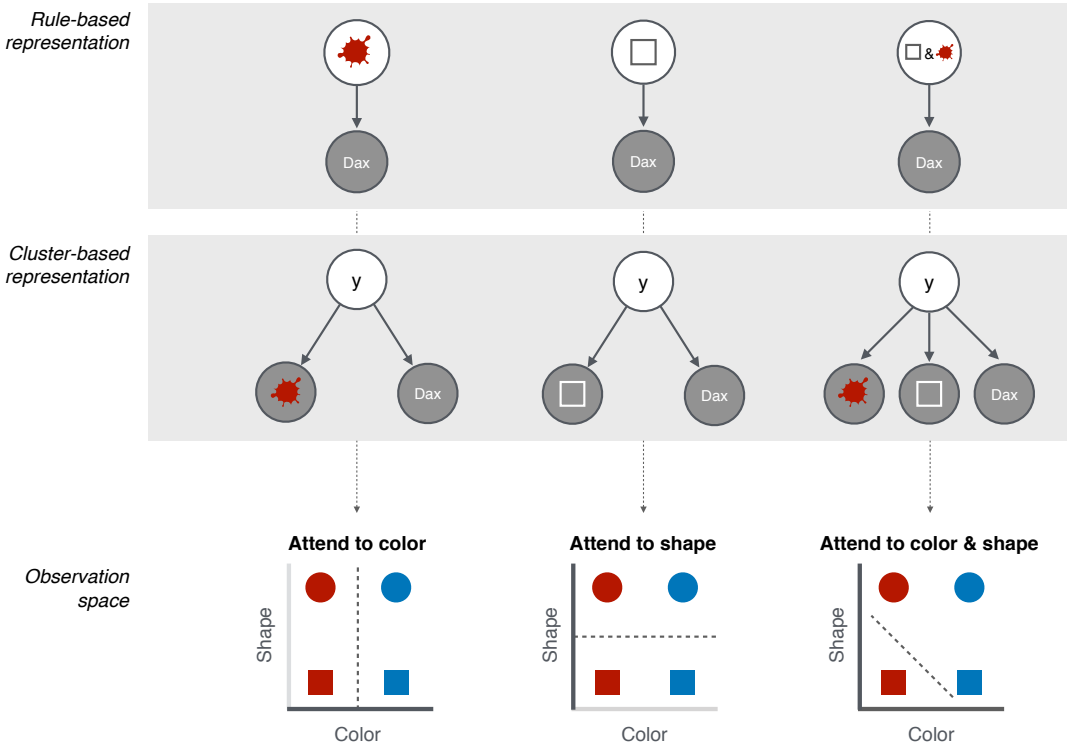


Figure 1.1: .

From structured representations to selective attention. Rules inferred by Bayesian probabilistic programming models (Goodman et al., 2008; Lake, Salakhutdinov, & Tenenbaum, 2015) (top row) or clusters inferred by Bayesian non-parametric models (Gershman & Blei, 2012; Franklin & Frank, 2018) (middle row) lead to different perceptual distinctions in observation space (bottom row). Left column: if the agent infers that red objects are "daxes", or that red and "dax" cluster together, then she can ignore shape and only attend to color when categorizing a stimulus as a "dax" or a "bim". Middle column: similarly if the agent infers that squares are "daxes" (middle left), or that square and "dax" cluster together, then she can ignore color and only attend to shape. Right column: finally if the agent infers that red squares are "daxes", or that red, square and "dax" cluster together, then she should attend to both color and shape.

Age is an issue of mind over matter. If you don't mind, it doesn't matter.

Mark Twain

2

Representation learning over the lifespan

2.1 Introduction

Representation learning refers to the process of discovering aspects of the environment relevant for the task at hand (Bengio et al., 2013). In the paper that this chapter is based on (Radulescu, Daniel, & Niv, 2016), I asked, how do humans learn representations in complex environments? To gain insight into what factors might be at play, I studied both younger and older adults. Aging is known to independently affect both simple trial-and-error learning and selective attention, which I argued in Chapter 1 should play a key role in learning state representations. A cross-sectional aging study therefore allows us to more precisely tease out the contribution of each process to representation learning.

Previous studies suggest that healthy aging affects the ability to associate stimuli with expected future rewards (Mata, Josef, Samanez-Larkin, & Hertwig, 2011; Mell et al., 2005; Samanez-Larkin & Knutson, 2014). When required to learn stimulus–reward associations from feedback, older adults consistently need more trials to reach the same level of performance as younger adults, and exhibit slower reaction times (RTs). Previous work has also emphasized that dopamine neurons—which have been implicated in reinforcement

learning (Niv, 2009)– are gradually lost over the course of the life span (Eppinger, Hämmerer, & Li, 2011; Li, Lindenberger, & Bäckman, 2010). Drawing on these findings, age-related behavioral differences in RL tasks have been linked to a reduced efficacy in reward prediction-error signaling in the human striatum (Eppinger, Schuck, Nystrom, & Cohen, 2013). Using a pharmacological manipulation, Chowdhury and colleagues further showed that dopaminergic drugs can restore this signal, and boost the performance of older adults to levels comparable to those observed in younger adults (Chowdhury et al., 2013).

But a deficit in simple stimulus–reward learning might not be at the heart of the difficulties that older adults show when learning in the real world. Instead, older adults may struggle exerting the attentional control required to update and maintain task representations. Behaviorally, older adults exhibit lower performance on tasks that require internally generating and maintaining task-relevant information (Braver & Barch, 2002; Hampshire, Gruszka, Fallon, & Owen, 2008), as well as suppressing task-irrelevant distractors (Campbell, Grady, Ng, & Hasher, 2012; Gazzaley, Cooney, Rissman, & D'esposito, 2005; Schmitz, Cheng, & De Rosa, 2010). A recent review summarized evidence that older adults compensate for these lapses in cognitive control by relying more on the external environment to provide task-appropriate representations (Lindenberger & Mayr, 2014). At the neural level, it has been suggested that changes in the interaction between DA and the prefrontal cortex (PFC) can account for observed differences in attentional modulation and inhibition of irrelevant stimuli (Braver & Barch, 2002; Dennis & Cabeza, 2012; Li et al., 2010). Taken together, these findings suggest that age may strongly affect the interaction between reward learning and attention.

In two experiments described in the chapter, younger and older adults performed a

learning task in which only one stimulus dimension was relevant to predicting reward, and within it, 1 "target" feature was the most rewarding. Participants had to learn the correct task representation through trial and error. In Experiment 1, stimuli varied on 1 or 3 dimensions and participants received hints that revealed the target feature, the relevant dimension, or gave no information. Group-related differences in accuracy and RTs differed systematically as a function of the number of dimensions and the type of hint available. In Experiment 2, I used trial-by-trial computational modeling of the learning process to test for age-related differences in learning strategies. Behavior of both young and older adults was explained well by a reinforcement-learning model that uses selective attention to constrain learning. However, the model suggested that older adults restricted their learning to fewer features, employing more focused attention than younger adults. Furthermore, this difference in strategy predicted age-related deficits in accuracy.

2.2 General methods: The Dimensions Task with Color-Shape-Texture

I studied how strategies for representation learning change over the lifespan using a paradigm previously developed in our lab known as the Dimensions Task (Niv et al., 2015). On each trial of the task, participants were presented with three visual stimuli. Stimuli differed along either one or three dimensions (color, shape, and texture, Figure 2.1). Within each dimension, a given stimulus could have one of three features (e.g., red, green, and yellow). On each trial, participants chose between stimuli that consisted of random combinations of features (e.g., red square with polka dots). Importantly, at any time point, only one dimension of the stimuli determined reward. Specifically, one "target" feature within this "relevant" dimension was more rewarding than the others: choosing the stim-

ulus that contained the target feature led to 75% chance of receiving 1 point (and 0 points otherwise), whereas choosing either of the other two stimuli was rewarded by 1 point with only 25% chance. Participants were fully informed of these reward probabilities, and the existence of a relevant dimension and target stimulus within it. To maximize the number of points earned, participants had to learn the identity of the target feature and use it to select the correct stimulus on each trial. Participants were asked to make their choice within 2 seconds, after which the trial timed out and the next trial began. To acquire repeated measurements of learning within each participant, we divided the task into several "games". The identity of the target feature stayed constant throughout a game. Once the game ended, participants were allowed a short, self-paced break and were notified that the relevant dimension and target feature would now be changing. This task is related to the Wisconsin Card Sorting Task that has previously been used to study cognitive flexibility in older adults (Fristoe, Salthouse, & Woodard, 1997; Rhodes, 2004), with the key difference being that rewards were probabilistic much like in the weather prediction task (Knowlton, Squire, & Gluck, 1994). The design prolonged the learning process such as to allow the use of computational modeling to analyze the dynamics of learning.

2.3 General methods: exclusion criteria

A game was considered "learned" if the participant chose the stimulus containing the target feature in each of the last six trials of the game. In both experiments, we excluded participants who learned fewer than 20% of all games, missed more than 10% of the trials, or performed at chance in any of the tasks. Chance was defined as less than 38% accuracy (two standard deviations above the mean of a binomial distribution with $p = 1/3$ and N of

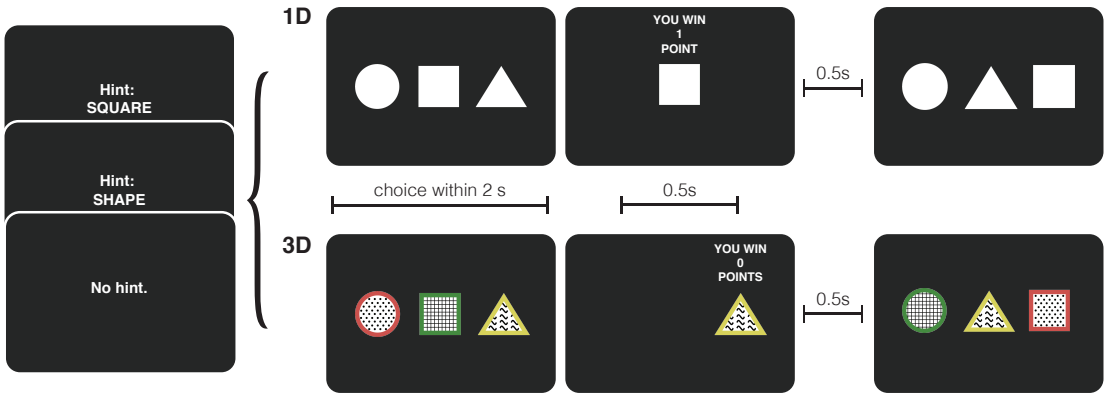


Figure 2.1: The Dimensions Task with Color-Shape-Texture. In Experiment 1, at the start of each game participants were given a 'hint' regarding the target feature, the dimension of the target feature, or else no hint was given. On each trial, participants chose between three stimuli that varied along a single dimension (e.g., shape) in the 1D case, or along three dimensions (shape, color, and texture) in the 3D case. Participants received binary reward feedback, winning either one or zero points on every trial, with reward probability depending on whether they chose the stimulus that contained the target feature. The game ended when the participant reached a performance criterion, or after 25 trials. A new game began with a signaled rule change followed by a new hint screen. Experiment 2 had the same structure, except that all games involved three-dimensional stimuli, no hints, and lasted 30 trials regardless of performance.

1000 trials, matching the average number of trials performed by participants).

2.3.1 General methods: apparatus

Participants sat approximately 50 cm from an LCD monitor and responded on a standard Macintosh keyboard using three adjacent keys corresponding to the left, middle, and right stimulus respectively. Stimuli were presented and responses were registered using MATLAB (The MathWorks, Natick, MA) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

2.4 General methods: statistical analyses

To quantify effect sizes, unless otherwise noted, we report the following: (a) Hedges' *g* for independent-sample t-tests, a measure more robust to small samples (Hedges, 1981;

Hentschke & Stüttgen, 2011). (b) partial η^2 for analyses of variance (ANOVAs), (c) Pearson correlation coefficients, (d) standardized regression coefficients.

2.5 Experiment 1 methods

The aim of the first experiment was to separately assess the contributions of reinforcement learning and attention to age differences in the performance of trial and error learning in a multidimensional environment. I tested each participant on five versions of the task, manipulating the number of dimensions along which stimuli varied (one or three, henceforth abbreviated as 1D and 3D) and the availability of hints that could be used to reduce the computational demands of probabilistic learning (Figure2.1). Throughout, reward contingencies and motor requirements were kept constant.

2.5.1 Participants

33 younger adults (23 female, 10 male; mean age 23 years; age range 19–37) and 33 older adults (16 female, 17 male; mean age 69.4 years; age range 62–80) participated in the experiment for either monetary compensation or course credit (younger adults). Older participants were recruited from among members of the Community Auditing Program at Princeton University.

All participants reported normal or corrected-to-normal color vision, were enrolled in an undergraduate program or held at least a university degree, had no history of psychiatric disorders, and provided informed consent. The experiment was approved by the Princeton University Institutional Review Board. The older adult cohort was screened for early onset dementia using a shortened version of Raven's Progressive Matrices (Raven et al., 1998).

One older adult who scored less than a 5 (out of 18) on this test was excluded from further analysis. Additionally, 2 younger adults and 7 older adults were excluded from further analysis as per the task performance criteria above, yielding a final sample of 31 younger adults and 25 older adults.

2.5.2 Stimuli and procedure

Games in the task were divided into five randomly intermingled conditions with 10 repetitions each.

In the first condition ("feature 1D"), stimuli varied on one dimension (e.g., the 3 distinct shapes; Figure 2.1 top). Before the game, a "hint" screen revealed which of the features within this dimension was the target (e.g., "square"; Figure 2.1). This condition was thus equivalent to performing a visual search for a predefined feature, and had no learning component: to maximize reward, participants simply had to select the target feature.

In the second condition ("feature 3D"), I again cued participants regarding the target feature, but presented them with three-dimensional stimuli each varying along color, shape, and texture (Figure 2.1 bottom). Comparing group behavior between the 1D scenario above and the 3D case allowed me to ask whether older adults show a disadvantage when distractor dimensions are present even when no learning is required.

In the third condition ("dimension 1D"), stimuli varied on a single dimension, but instead of being told the identity of the target feature, participants had to learn it from trial and error. This condition is equivalent to a 3-armed bandit task akin to the kinds of tasks that have previously been studied in older adults to characterize deficits in reinforcement learning (Chowdhury et al., 2013; Mell et al., 2005).

The fourth condition ("dimension 3D") involved three dimensional stimuli and a hint disclosing which dimension is relevant for predicting reward. Participants were thus required to learn the identity of the target feature within that dimension, as in the dimension 1D task. However, because distractor dimensions were present (Figure 2.1, bottom), this task required sustained attention to one dimension. To do well, participants had to restrict learning to the cued dimension and ignore the other distracting dimensions.

Finally, in the fifth condition ("full 3D"), participants were presented with three-dimensional cues, and received no information as to the relevant dimension or target feature. This condition is identical to the Color-Shape-Texture "dimensions task" that has been developed by Niv and colleagues to investigate the interaction between selective attention and reinforcement learning in younger adults (Niv et al., 2015; Wilson & Niv, 2012).

Prior to the experimental session, participants were given a tutorial that described the reward structure and informed them about the different conditions. Following the tutorial, participants completed several sample games. They were then tested on 50 games of the task with condition randomized such that within each block of five games, each condition appeared once for a total of 10 games per condition. Each game consisted of a minimum of 8 and a maximum of 25 trials. A correct trial was defined as one in which the participant chose the stimulus containing the target feature. Once a criterion of 8 consecutive correct trials was reached, the game had a 50% chance of ending on any subsequent trial. Games lasted at most 25 trials. The target feature was chosen randomly, avoiding relevant dimension repeats from game to game. Rewards were drawn pseudorandomly such that within each block of eight trials the frequency of presented rewards matched the reward probabilities specified in the design. Once a valid response was registered, stimuli that were not

chosen were removed from the screen and after a brief delay the outcome was presented for 0.5 seconds. A new trial started after 0.5 seconds.

2.5.3 Reaction time analysis

The ex-Gaussian distribution has been proposed as an analysis tool for RT that disentangles variance arising from two separate cognitive processes: The *transduction* component, indexed by a positive shift in the Gaussian mean, can be thought of as the time it takes to process the sensory information plus the time required to physically make the response once a decision has been made. The *decision* component, reflected in the exponential skew, is a proxy for the time it takes to represent the task and decide which response to make (Lacouture & Cousineau, 2008; Luce et al., 1986).

Maximum likelihood estimation was used to fit individual RT distributions separately within each condition. By analyzing RT data in this way, I sought to dissociate age effects in how the task is represented from perceptual and motor differences. I hypothesized that as representational demands increase with the number of dimensions, older adults would be selectively impaired in the decision component of RT.

2.6 Experiment 1 results

I first examined the effect of condition on overall accuracy. A 2 (age-group) x 5 (condition) mixed-effects ANOVA (Figure 2.2A) revealed a main effect of age ($F(1, 54) = 6.80, p < .01, \eta^2 = .11$), a main effect of condition ($F(4, 216) = 705.66, p < .001, \eta^2 = .93$), and a significant interaction between age group and learning condition ($F(4, 216) = 5.50, p < .001, \eta^2 = .09$).

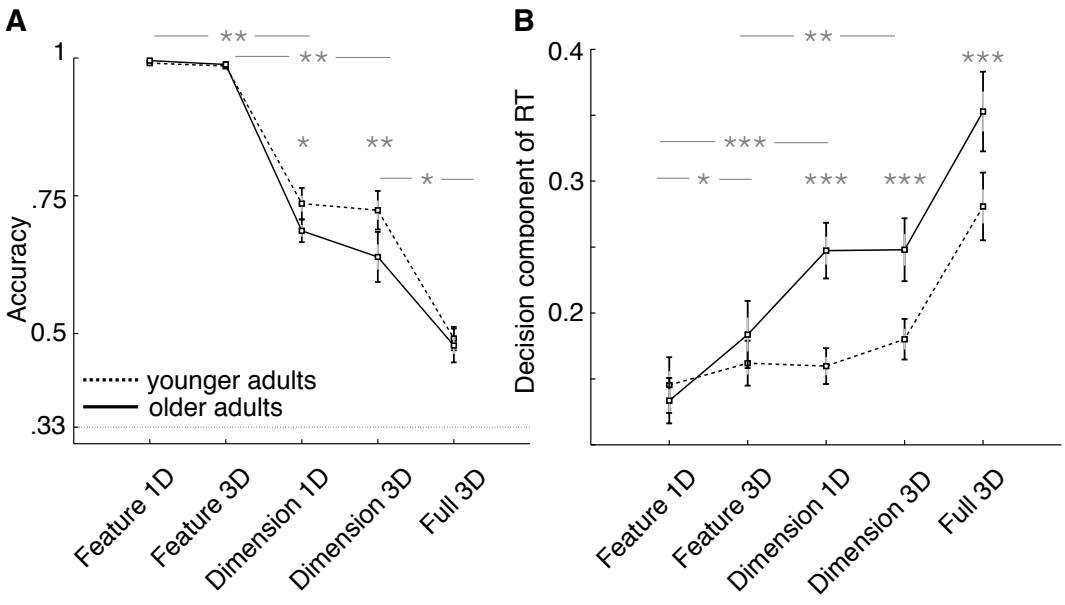


Figure 2.2: Performance of younger adults (dashed) and older adults (solid) in Experiment 1. A. Average accuracy in each of the five task conditions. Dotted line indicates chance performance. **B.** Decision component of RT by task condition. Error bars indicate one SEM (gray) and 95% confidence intervals (black). Asterisks indicate significant interactions (lines) and differences between groups within each condition $({}^{*}p < .05, {}^{**}p < .01, {}^{***}p < .001)$.

To better delineate the various determinants of accuracy differences between older and younger adults, I next performed 2 (age-group) x 2 (condition) ANOVAs examining the effect of each progression in task difficulty. I first compared the two "feature" conditions, in which trial-and-error learning was not necessary because the target feature was disclosed via the hint. In the "feature 1D" condition, both groups performed at ceiling. Although I did find a main effect of condition when adding the distractor dimensions in the "feature 3D" case ($M_{feature1D} = .993, M_{feature3D} = .987, F(1, 54) = 14.122, p < .001, \eta^2 = .21$), there was no main effect of group ($F(1, 54) = 2.38, p = .13, \eta^2 = .04$) and no interaction between age and condition ($F(1, 54) = .28, p = .60, \eta^2 = .01$). These results suggest that the presence of extradimensional distractors did not specifically impair accuracy in older adults, and also confirms that participants understood the instructions and used the hint correctly.

I next focused on performance differences when participants had to learn the target feature from trial and error, but were cued as to the relevant dimension. In the 1D case, this amounts to a simple 3-way choice task with binary rewards and fixed reward probabilities. As expected from previous work showing age-related impairments in probabilistic learning, I observed a significant group effect on overall accuracy when comparing the "feature 1D" to the "dimension 1D" condition (main effect of group : $F(1, 54) = 5.56, p = .02, \eta^2 = .09$; interaction : $F(1, 54) = 7.74, p = .01, \eta^2 = .13$).

In the 3D case, the dimension hint helps participants assign credit for a reward to only one of the 3 features of a stimulus: if the reward-relevant dimension is color, feedback for, say, choosing a green square with polka dots, can be correctly assigned to the color "green," while ignoring the "square" and "polka dot" features that act as distractors. Thus

the "dimension-3D" case is similar to a 3-way choice task, only with known distractors. As expected, here too I observed a significant group effect on accuracy when comparing the "feature 3D" to the "dimension 3D" condition (main effect of group : $F(1, 54) = 7.18, p = .01, \eta^2 = .12$; interaction : $F(1, 54) = 8.77, p = .01, \eta^2 = .14$). These age-related deficits in learning from trial and error have previously been attributed to a reduced efficacy of dopamine-dependent prediction error signals (Eppinger et al., 2013).

Finally, I compared the "dimension 3D" condition with the "full 3D" condition. In the latter, participants were also required to identify the relevant dimension from trial and error. As a result of this extra demand, I expected the performance of older adults to drop more precipitously than that of younger adults, as compared to the "dimension 3D" case. In line with this prediction, the analysis revealed a main effect of group ($F(1, 54) = 5.73, p = .02, \eta^2 = .10$) and of condition ($F(1, 54) = 183.93, p < .001, \eta^2 = .77$). I also observed a significant interaction between group and condition ($F(1, 54) = 5.89, p = .02, \eta^2 = .10$). Surprisingly however, this interaction was in the direction opposite from what was initially predicted. That is, when the dimension hint was removed, older adults incurred a smaller additional cost in accuracy than younger adults, performing as well as younger adults on the task ($M_{older} = .48, SD_{older} = .09; M_{younger} = .49, SD_{younger} = .06, t(54) = .65; p = .52, g = .17$). To rule out the possibility of either a floor or ceiling effect driving this interaction, I performed a within-group median split on accuracy . Both older and younger adults were distributed symmetrically around their respective group mean, giving no indication that the observed result was due to boundary effects.

I next analyzed response times using the ex-Gaussian distribution. Examining RTs revealed more subtle effects of condition than were apparent in overall accuracy. I submitted

the fitted skew parameters indexing the decision component of the RT to a 2 (age-group) x 5 (condition) mixed-effects ANOVA (Figure 2.2B). Paralleling the accuracy results, this initial test revealed a main effect of age ($F(1, 54) = 17.53, p < .001, \eta^2 = .25$), a main effect of condition ($F(4, 216) = 108.15, p < .001, \eta^2 = .67$), and a significant interaction between age group and learning condition ($F(4, 216) = 10.68, p < .001, \eta^2 = .17$). An independent-samples t-test indicated no group difference between older ($M = .13, SD = .05$) and younger ($M = .15, SD = .06$) adults for the decision component in the "feature 1D" condition, $t(54) = .81, p = .42, g = .22$, indicating that by analyzing RT using the ex-Gaussian distribution, the variance associated with the cost of visual search and response mapping was successfully removed. (Nevertheless, all interaction results reported here for the decision component also held when using raw RT as the dependent variable).

I then performed 2 (age-group) x 2 (condition) ANOVAs paralleling those we reported above for accuracy, to separately assess the effect of each manipulation on the decision component of the RT. Although in both the "feature 1D" and "feature 3D" conditions older and younger adults performed at ceiling as reflected by average accuracy, I did observe a group by condition interaction in the decision component of RT ($F(1, 54) = 5.12, p = .03, \eta^2 = .09$), suggesting that older adults required more time to decide on their choice when the stimulus consisted of multiple features.

As expected, I also found a group by condition interaction when comparing both the "feature 1D" and the "dimension 1D" conditions ($F(1, 54) = 53.89, p < .001, \eta^2 = .50$), and when comparing the "feature 3D" condition with the "dimension 3D" condition ($F(1, 54) = 7.2, p = .01, \eta^2 = .18$). These results mirror the accuracy effects, and suggest that having to learn the target feature from feedback affected older adults' decision time

significantly more than it did younger adults'.

Finally, I did not observe an interaction between group and condition when comparing the "dimension 3D" with the "full 3D" condition ($F(1, 54) = .05, p = .82, \eta^2 = .00$). This finding suggests that in the full dimensions task, older adults respond slower than younger adults, but not more so than in a simple probabilistic learning setting in which the relevant dimension is known. Together, the results of Experiment 1 suggest that although older adults are significantly more impaired than younger adults in simple trial-and-error learning, they do not show an additional impairment when required to learn which dimension is relevant to determining reward. In this latter case, their accuracy was not significantly different than that of younger adults, and although they did take significantly longer to respond than younger adults when the relevant dimension was unknown, this group difference was not greater than in the cued dimension case. These results are in line with previous reports of age-related deficits in reinforcement learning (Eppinger et al., 2013; Mell et al., 2005) and reveal that, contrary to expectations, attentional demands do not confer differential additional hardship on older adults. As previous work has suggested that attention processes do change during healthy aging, one possibility is that older adults adapt their strategies such as to allow them to perform the full 3D representation learning task better than would otherwise be expected.

2.7 Experiment 2 methods

2.7.1 Participants

Twenty-eight younger adults (17 female, 11 male; mean age 23.9 years; age range = 20–31) and 30 older adults (10 female, 20 male; mean age 70.1 years; age range = 65–80) partic-

ipated in the second experiment. All participants reported normal or corrected-to-normal color vision, were enrolled in a university or held at least a university degree, had no history of psychiatric disorders, and provided informed consent. The protocol was approved by the Princeton University Institutional Review Board.

In addition to completing the main task, participants in both groups also completed several psychometric tests and questionnaires: (a) a computerized version of the Digit-Symbol substitution test (Salthouse, 1992), (b) a 2-back task (Nystrom et al., 2000), (c) the Spot-the-Word test (Baddeley, Emslie, & Nimmo-Smith, 1993), (d) the BIS-BAS questionnaire (Carver & White, 1994), and (e) a shortened version Raven's Progressive Matrices (Raven et al., 1998). As in the first experiment, the older adult cohort was screened for early onset dementia using the Raven's Progressive Matrices. Exclusion criteria were identical to those of Experiment 1. One younger adult and 3 older adults were excluded from the analysis, yielding a final sample of 27 younger adults and 27 older adults.

2.7.2 Stimuli and procedure

Stimuli were identical to the "full 3D" condition in Experiment 1. On each trial, participants were presented with multidimensional stimuli varying in color, shape, and texture. One of the three dimensions was used to determine rewards. Within this reward-relevant dimension one target feature had a 75% probability of reward, whereas all other features had a 25% probability of reward. Participants received no information about the identity of the relevant dimension. Each participant played, on average, 35–50 games for a total of approximately 1,500 trials per participant. The length of a game was fixed at 30, excluding missed trials. The total duration of the experiment was capped at 40 min. Once a valid re-

sponse was registered, the stimuli that were not chosen were removed from the screen and the outcome was immediately presented for 0.5 seconds. A new trial started after 0.3 seconds.

2.7.3 Model-based analysis

Previous work has shown that in the Color-Shape-Texture of the Dimensions Task, RL models that allow for effects of selective attention explain subjects' behavior better than either a naïve RL model that learns values for each of the 27 possible stimuli or a Bayesian ideal-observer model that makes statistically optimal use of information (Wilson & Niv, 2012; Niv et al., 2015). Here I was interested in testing for an effect of age on the width of the "attentional filter." Unlike in the Faces-Landmarks-Tools version presented in Chapter 2, in this study neither fMRI decoding nor eye-tracking were available as a direct measure of participants' selective attention. To test for group differences in attentional strategies, I employed a model-based analysis.

I compared between two RL models, a "feature RL" (fRL) model that attends uniformly to all three dimensions, and a "feature RL with decay" (fRL+decay) model that emulates selective attention to dimensions that include consistently chosen features (see below). Both models track a weight W for each feature f and calculate the value $V(S)$ of stimulus S as the sum of the weights of its three features $W(f)$, where each stimulus has one feature per dimension. In the fRL model, on every trial, once the outcome for the chosen stimulus is displayed, the weights corresponding to the three features of the chosen stimu-

lus are updated according to:

$$W^{New}(f) = W^{Old}(f) + \eta(R(t) - V(S)) \quad \forall f \in S_{chosen} \tag{2.1}$$

where η is the step size or learning rate parameter and $R(t)$ is the reward (1 or 0 points) on the current trial.

The fRL+decay model is identical to the fRL model, except that it also decays to zero the weights of features that do not appear in the chosen stimulus:

$$W^{New}(f) = (1 - d)\, W^{Old}(f) \quad \forall f \notin S_{chosen} \tag{2.2}$$

where d is the decay rate. For both models, at the beginning of each game, the weights are initialized at zero. On each trial, the probability of selecting each of the three available stimuli is calculated using a softmax distribution:

$$p(S_{chosen} = S_i) = \frac{e^{\beta V(S_i)}}{\sum_j e^{\beta V(S_j)}} \tag{2.3}$$

where the inverse temperature parameter β captures the noise i the subjects' choices. Thus the fRL model has two free parameters, $\theta_{fRL} = \{\eta, \beta\}$, and the fRL+decay model has three free parameters, $\theta_{fRL+decay} = \{\eta, \beta, d\}$.

The total likelihood of the data is given by the product of the trial-by-trial choice probabilities (Daw, 2011):

$$p(C_{1:T}|\theta) = \prod_{t=1}^{T} p(S_{chosen} = S_i) \tag{2.4}$$

Importantly, the decay rate d dictates the width of an implicit attentional filter. To un-

derstand this mechanism, it is instructive to consider two hypothetical consecutive trials, t and $(t+1)$, in which a participant might choose stimuli such that only one feature — for example red — appears in the chosen stimulus on both trials. A decay rate of zero reduces the fRL+decay model to simple fRL and means that although the two features that co-occurred with red on trial t did not appear on trial $(t+1)$, their weights remain unchanged on that second trial. At the other extreme, a decay rate of 1 means that on trial $(t+1)$ all weights except those of features of the most recently chosen stimulus are set to 0, in effect erasing the values learned on trial t for features other than red. This is tantamount to a narrow attention filter that, across trials in which red is consistently chosen, only accumulates value for that feature, effectively attending only to color information. Intermediate values of d smoothly interpolate between these two extremes, with higher decay rates corresponding to more focused attention as reflected in high weights for fewer (recently chosen) features. Another way to view this model is that decay emulates attention with a one-trial delay; that is, the features the subject chooses on trial $t+1$ are used to infer what they attended to on trial t, and decay the learning that was done at t to nonattended dimensions. Again, such model-based inferences are necessary because when we do not have direct access to participants' attention (Niv et al., 2015).

To compare between models based on their predictive accuracy, I used participants' trial-by-trial choices to fit the parameters that maximize the likelihood of each subject's choices (Equation 2.4). In practice, this amounts to minimizing the log of the quantity in Equation 2.4. As the models had different numbers of parameters, I compared models using a leave-one-game-out cross-validation approach: for every participant and every game, I fit the model to all data excluding that game. The model and its maximum likelihood pa-

rameters were then used to assign likelihood to the trials of the left-out game. I repeated this procedure for each game and divided the resulting total likelihood by the number of trials N to yield the geometric average of the likelihood per trial. This is a quantity that varies between 0 and 1, and roughly corresponds to the average probability with which the model predicted the choices of the participant ($\frac{1}{3}$ is chance). This quantity was then used for model comparison, with the model that best predicts participants' behavior deemed the winning model. With this model in hand, I once again applied maximum likelihood parameter estimation, this time using all available data for each participant, to obtain individual parameters for every participant in each group. All optimizations were carried out using MATLAB's *fmincon* function.

Finally, to restrict the fitting as much as possible to trials in which the participants were still learning (rather than simply selecting the target feature), and to be able to compare the results here with those from Experiment 1, model fitting and model comparison analyses were done after imposing a post-hoc learning criterion of 8 correct trials in a row and capping the length of each game at 25 (as was the case in Experiment 1). All model-based results reported here also hold without this modification.

2.8 Experiment 2 results

Independent-sample t-tests showed that the average accuracy of older adults was significantly lower than that of younger adults ($M_{older} = .42, SD_{older} = .04; M_{younger} = .46, SD_{younger} = .04, t(52) = 3.49; p = .001, g = .94$). Older adults also learned fewer games than younger adults ($M_{older} = .39, SD_{older} = .09; M_{younger} = .49, SD_{younger} = .09; t(52) = 4.2, p = .001, g = 1.13$). These results were contrary to the findings in

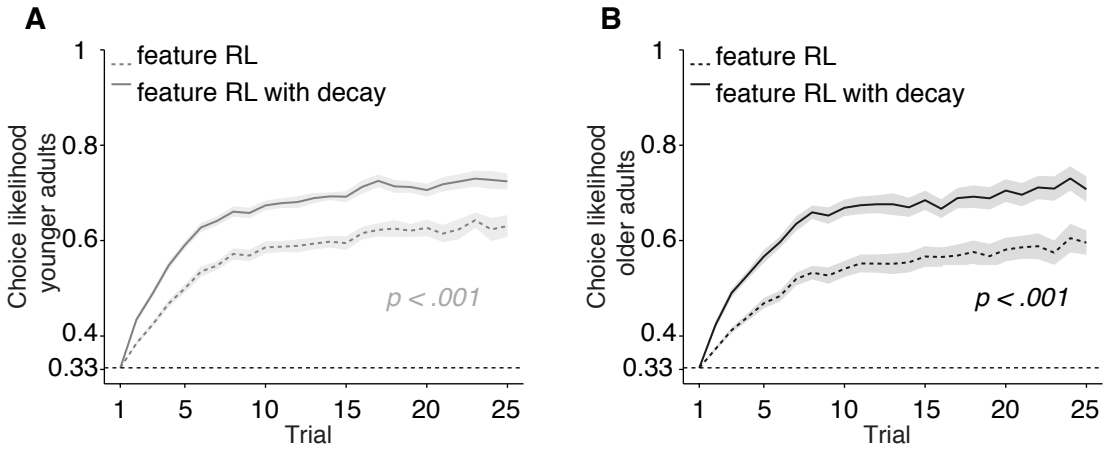


Figure 2.3: Model comparison for younger (left) and older (right) adults. Likelihood per trial as a function of trial within a game for **A.** younger adults and **B.** older adults. For both groups, the data heavily favor the fRL+decay model. Dashed line: chance (33%); shading: SEM; p-value corresponds to a repeated measures ANOVA (model x trial).

the first experiment, and suggested that failing to detect a difference in accuracy in Experiment 1 might have been due to the smaller number of games (10 per participant compared to 35–40 in the second experiment). Nevertheless, because Experiment 1 established that reward learning is significantly more impaired in older adults, I was interested in analyzing the differential contributions of learning and selective attention to task performance in older adults versus younger adults.

To determine which learning strategy best describes trial-by-trial behavior for each group, I performed a within-group model comparison by taking the average cross-validated likelihood per trial for each model and submitting it to a repeated measures ANOVA (Figure 2.4). In both groups the fRL+decay model predicted participants' choices better than the fRL model (younger adults: $F(1, 1248) = 42.31, p = .001, \eta^2 = .46$, Figure 2.4A; older adults: $F(1, 1248) = 23.32, p = .001, \eta^2 = .47$, Figure 2.4B). This finding is in line with previous work showing that strategies that incorporate attentional mechanisms

better explain behavior in the task (Wilson & Niv, 2012; Niv et al., 2015). Importantly, no significant difference emerged in the average per-trial likelihood of the fRL+decay model ($F(1, 1248) = .66, p = .42, \eta^2 = .018$). In other words, the fRL+decay captured both groups' strategy equally well and could thus be used to assess group differences in fit parameters.

Table 2.1: Best-fit values for fRL+decay model across groups (mean ± SEM). Bottom: Results of the corresponding Mann-Whitney test for group differences for each parameter.

Group	Learning rate (η)	Decay rate (d)	Inverse temperature (β)
Younger adults	$.11 \pm .01$	$.45 \pm .017$	$12.52 \pm .68$
Older adults	$.12 \pm .01$	$.56 \pm .03$	14.06 ± 1.65
	$p = .49, r = .005$	$p = .002, r = .40$	$p = .41, r = .03$

To test for group differences in the breadth of attention for learning, I compared the decay rate parameters for the two groups. A Mann–Whitney test indicated that older adults had significantly higher decay rates than younger adults (older adults median = .52, younger adults median = .42, U = 192.0, p = .002, r = .40, Figure 4A). This suggests that older adults utilize narrower stimulus representations in trial-and-error learning in multidimensional environments. The results of all group tests on fit parameters are summarized in Table 2.1.

If the model indeed captures aspects of subjects' strategy that are relevant to behavior, it should be able to reproduce the qualitative patterns observed in the data. To test this, we used individual fit parameters drawn from each group to simulate 54 agents (27 per group), and computed their average learning curves. We found that the model could perform the task at a level comparable to that of participants, slightly undershooting the performance of

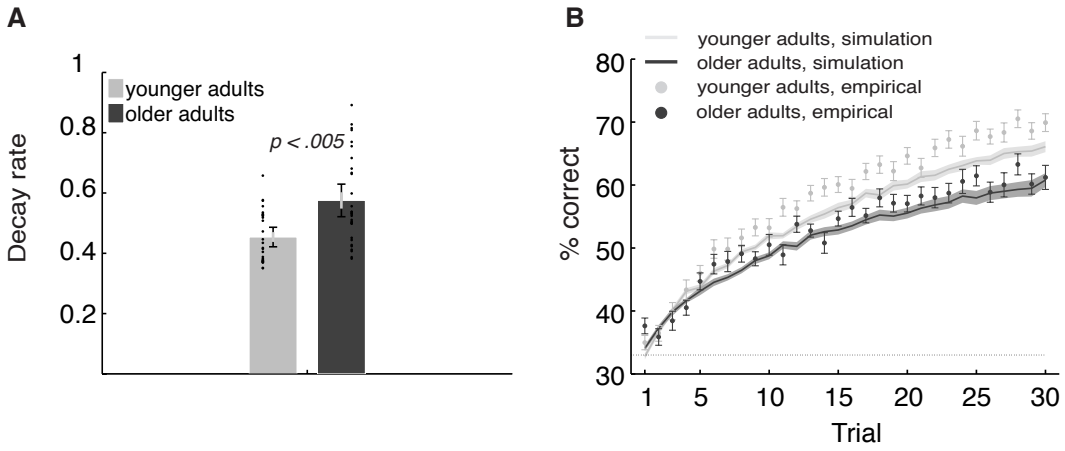


Figure 2.4: Age-related differences in decay rate and empirical vs. simulated learning curves. A. Group difference in decay rate. Error bars: SEM (gray) and 95% confidence interval (black). Black dots show the decay rate estimates for each participant. **B.** Average learning curves for participants performing the task (N = 27 for each group) and for simulated agents (27 per group) performing 40 games of the task with individual fit parameters within each group. Single data points indicate the empirical performance, and continuous lines indicate simulated performance. Dotted line: chance (33%), error bars and shading: SEM.

younger adults, but capturing that of older adults. Importantly, the model reproduced the general performance differences between older and younger adults (Figure 2.4).

To assess the specific effect of the decay rate parameter on task performance in our data, we took a multiple linear regression approach. In particular, each participant's accuracy was regressed on learning rate, decay rate and inverse temperature estimated from the fRL decay model. The results of the regression indicated that the three parameters explained 49% of the variance ($R^2 = .49, F(3, 50) = 16.0, p = .001$). We found that lower decay rates predicted higher accuracies ($\beta = -.52, p = .001$), suggesting that even after taking into account possible effects of learning rate and inverse temperature, the width of the attentional filter modulates performance in the task. We also found that learning rate did not significantly predict accuracy ($\beta = .077, p = .61$), whereas inverse temperature

did ($\beta = -.53, p < .001$). Including both the learning rate and the inverse temperature parameters in the regression was necessary to isolate the effect of decay on performance. However, because these parameters are not completely separable in the model (that is, they cannot be precisely estimated independently of each other (Daw, 2011)), we cannot make strong claims about any observed effect of learning rate or inverse temperature alone. Importantly, we repeated the above analysis within each group, and found that decay remained a significant predictor of performance within both younger ($\beta = -.59, p < .01$) and older adults ($\beta = -.35, p < .01$). Taken together, our results suggest that more focused attention during learning in older adults can, in part, explain the observed decrease in task performance.

Finally, we investigated whether the decay rate reflects a deficit in working memory rather than more focused attention. To test this, we regressed decay rate as estimated from the dimensions task on the working memory score measured using the 2-back task, and included age as a covariate in the regression. We found that age ($\beta = .09, p < .05$), but not working memory ($\beta = -.02, p = .19$), significantly predicted decay rate, suggesting that age differences in the decay rate parameter are not due to working memory impairments in older adults.

2.9 Discussion

In this chapter, I described a study of how age affects reinforcement learning in multidimensional environments that characterize real-world learning and decision-making scenarios. In Experiment 1, I tested young and older adults on a set of probabilistic learning tasks in which I manipulated the number of stimulus dimensions and the availability of hints

about the identity of the target feature or relevant dimension. The aim was to dissociate the relative contribution of reinforcement learning and representation learning processes to age-related differences in task performance. I found that age differences in both accuracy and RT depended on the extent to which reward learning was required to solve the task. Surprisingly, adding representation learning to the demands of the task did not affect the performance of older adults more than it did the performance of younger adults. The results of the first experiment therefore suggested that older adults might adapt to deficits in reinforcement learning such as to reduce the burden on this mechanism in multidimensional environments.

To test this hypothesis, in Experiment 2, I modeled choice data of a new group of participants who performed the full three-dimensional representation learning and reinforcement learning task without any hints regarding the identity of the target feature or relevant dimension. I found that the behavior of both groups was well described by a reinforcement learning model that emulates an attentional filter by decaying the value of unchosen options to zero. Group differences in the decay rate suggested that older adults employ more focused attention—they are more likely to maintain high values for single features rather than combinations of features. But, this difference in strategy came at a cost: more focused attention at least partially explained the lower performance of older adults in our task.

Two mechanistic explanations are consistent with higher decay rates in older adults: participants could employ narrower selective attention at the time of learning, attributing the reward to fewer stimulus features; or they could be more likely to forget recently learned feature–reward associations. Although further work is necessary to precisely distinguish between the two, both lead to a strategy closer to serial hypothesis testing (Wilson & Niv,

2012) in which older adults attend to single features when learning from reinforcement, whereas younger adults may be learning about multiple chosen features at once. Ignoring the relationship between reward feedback and incidental features of the chosen stimulus (i.e., unattended features that were not responsible for it being chosen) may be detrimental if participants have learned the wrong task representation. For instance, a participant could be focusing on the red color when trying to maximize reward, and not notice that, in fact, rewards are obtained more often when the red stimulus happens to be a square. Learning about incidental features would enable more efficient switching to other potentially rewarding features. In this sense, narrowly focused attention could pose difficulties for older adults. However, this is not always the case, and in some situations a narrower focus of attention may be normative. Although in the specific task used in this study such focused attention is not statistically optimal (Niv et al., 2015), the findings are consistent with a recent proposal that in older adults, general models of the world that have been learned over the lifespan reduce the need to rely on sensory updating (Moran, Symmonds, Dolan, & Friston, 2014). Focusing on fewer aspects of the environment during learning can, in fact, be seen as an adaptation to the structure of real-world tasks, where correct performance might often depend on only few attributes. This is especially true if the representations older adults have learned are useful for generalizing across different situations. Mata and colleagues term this idea "ecological rationality." They make a compelling case for the argument that age-related deficits in strategy use may not necessarily be due to impaired decision making, and that decision strategies can only be evaluated relative to the environment in which they are used (Mata et al., 2012). A striking example is work by Worthy and Maddox, who show that older adults perform better than younger adults in a task with com-

plex structure that favors a Win-Stay-Lose-Shift strategy, which older adults are more likely to employ, over a reinforcement learning strategy (Worthy, Gorlick, Pacheco, Schnyer, & Maddox, 2011; Worthy & Maddox, 2012). An interesting avenue of future research would thus be to characterize the statistics of natural tasks that older adults have learned to engage in. Ecological rationality view suggests that older adults should be just as good as younger adults at learning new tasks in which previously learned structure could be "recycled."

The idea that older adults may display more focused attention in certain situations has been suggested before, in work examining age differences in category learning. For instance, Glass and colleagues argue that when older adults are trained to categorize exemplars from two prototypes, they take into account fewer stimulus dimensions (Glass, Chotibut, Pacheco, Schnyer, & Maddox, 2012). My findings suggest that this strategy also manifests during sequential learning and decision making, therefore laying the groundwork for a number of future questions: is more focused attention in older adults accompanied by less or more frequent attention switching, as compared to younger adults? And if so, is there an age difference in how much feedback is needed to redirect attention, as suggested by studies that report a tendency to perseverate in older adults (Rhodes, 2004; Ridderinkhof, Span, & Van Der Molen, 2002).

In light of the results in this chapter, perseveration may be attributed to a more rigid focus of attention that prevents the formation of alternative representations, because it filters out incidental learning. Although less efficient, this could reflect an adaptation to the reduced efficacy of dopaminergic signaling (Li et al., 2010) in which selective attention is deployed during learning so as to tax mechanisms subserved by dopamine as little as possible.

Finally, the finding that older adults are more likely to filter out information may seem at odds with a broad literature documenting age-related deficits in suppressing task-irrelevant distractors (Campbell et al., 2012; Gazzaley et al., 2005; Schmitz et al., 2010). Lindenberger and Mayr have suggested that the inability of older adults to ignore visual distractors is linked to a broader developmental trend in which older adults shift from internal cognitive control to relying more on the environment to provide appropriate task representations (Lindenberger & Mayr, 2014). However, increased focus and increased distractibility could both result from an attentional system that cannot allocate attention to multiple items, but rather can only maintain narrow, rigid hypotheses. In such conditions, a distractor that captures attention would do so more strongly, and to the exclusion of the task otherwise being performed, leading to apparent distractibility. When attention can be maintained more broadly, to the task-relevant stimuli and also to incidental features, the effect of attention-grabbing distractors is mitigated.

Another way to reconcile the idea of increased environmental reliance with narrower attention when learning task sets concerns the broader question of how tasks are represented in the brain (Wilson, Takahashi, Schoenbaum, & Niv, 2014). In a recent paper, Mayr and colleagues have suggested an intriguing explanation for age-related increases in task-switching RT costs: older adults may not fully represent the relevant task states, opting for a simpler structure at the expense of flexibility (Mayr, Spieler, & Hutcheon, 2015). The Dimensions Task was explicitly designed to study how participants learn to represent a task. Narrow attention in our case has the effect of preventing complex stimulus–reward associations from forming (e.g., the participant is more likely to learn that red predicts reward, instead of red and polka dots predict reward). This narrowness limits internal representa-

tions, and simplifies the task as much as possible. The trade-off is that such simple representations may not allow for flexibility in learning new tasks that require de-aliasing similar percepts. That is, older adults might have difficulty when choosing the correct action required to learn about a second, disambiguating feature. Finally, the process of learning to attend, is different from maintaining (instructed) attention in the face of distraction—the results in this chapter suggest that older adults may filter out important relationships between reward feedback and incidental cues, while at the same time they may erroneously focus attention on incidental distractors.

A growing interest in applying RL methods to the study of cognitive aging has bridged knowledge about dopaminergic loss in older adults and deficits in trial and error learning (Shohamy & Wimmer, 2013). This chapter focused on how age-related impairments in RL might play out in multidimensional environments where, in addition to trial-and-error learning, one must learn the relevant task representations. In such cases, attentional mechanisms have been hypothesized to interact with RL so as to allow more efficient learning (Niv et al., 2015; Wilson & Niv, 2012; Geana & Niv, 2014). I found that aging is accompanied by a narrowing of attention during reinforcement learning, perhaps in order to adapt to impairments in neural trial-and-error learning mechanisms. These results suggest a role for selective attention in representation learning. However, the evidence is indirect, relying on a computational model that does not distinguish between selective attention and passive forgetting of values. I address this limitation in the next chapter.

Tears are lovely prisms. In every tear, there's a rainbow.

Anon

3

Learning to attend, attending to learn

3.1 Introduction

In Chapter 2, I showed how humans faced with high-dimensional learning problems seem to solve them with ease, despite cognitive changes that come with age. But the conclusions were limited by a computational model that did not fully disentangle the role of selective attention and working memory in learning task representations. Moreover, the task did not allow a direct measurement of attention. In the paper that this chapter is based on (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017), Yuan Chang Leong and I set out to more directly study the question of how humans learn to efficiently deploy attention when learning in multidimensional environments. Given evidence that selective attention is critical for narrowing down the dimensionality of the task (Jones & Canas, 2010; Wilson & Niv, 2012; Niv et al., 2015; Radulescu et al., 2016), we designed a study to directly measure the dynamics of selective attention during learning.

Selective attention prioritizes a subset of environmental dimensions for learning while generalizing over others, thereby reducing the number of different states or stimulus configurations that the agent must consider. However, attention must be directed toward dimensions of the environment that are important for the task at hand (i.e., dimensions that

predict reward) to provide learning processes with a suitable state representation (Gershman & Niv, 2010; Wilson et al., 2014). What dimensions are relevant to any particular task is not always known and might itself be learned through experience. In other words, for attention to facilitate learning, we might first have to learn what to attend to. We therefore hypothesized that a bidirectional interaction exists between attention and learning in high-dimensional environments.

To test this hypothesis, we had human participants perform a modified Dimensions Task with compound stimuli—each comprised of a face, a landmark, and a tool—while scanning their brain using fMRI. As before, at any one time, reward depended on only one of the three stimulus dimensions, mimicking real-world learning problems where only a subset of dimensions in the environment is relevant for the task at hand. Using eye tracking and multivariate pattern analysis (MVPA) of fMRI data, we obtained a quantitative measure of participants' attention to different stimulus dimensions on each trial. We then used trial-by-trial choice data to test whether attention biased participants' valuation of stimuli, their learning from prediction errors, or both processes. Estimates of participants' choice value and outcome-related prediction errors using the best-fitting model were generated and regressed against brain data to further determine the influence of attention on neural value and prediction error signals. Finally, we analyzed trial-by-trial changes in the focus of attention to study how attention was modulated by ongoing experience and to search for neural areas involved in the control of attention.

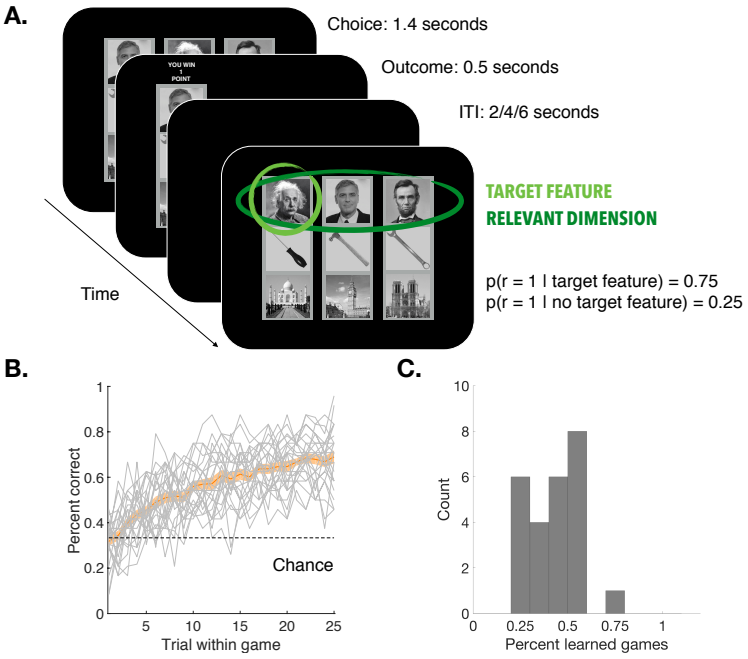


Figure 3.1: The Dimensions Task with Faces-Landmarks-Tools. A. Schematic illustration of the task. On each trial, the participant was presented with three stimuli, each defined along face, landmark, and tool dimensions. The participant chose one of the stimuli, received feedback, and continued to the next trial. The relevant dimension and target feature changed every 25 trials (announced: "new game"), requiring new "representation learning" **B.** Learning curves. The fraction of trials on which participants chose the stimulus containing the target feature (i.e., the most rewarding feature) increased throughout games. Dashed line: random choice, shading: SEM, grey lines: individual participants. **C.** Percent learned games. The fraction of games on which participants selected the stimulus containing the target feature on 6 of the last 6 trials of each game.

3.2 Methods: The Dimensions Task with Faces-Landmarks-Tools

On each trial, participants were presented with three compound stimuli, each defined by a feature on each of the three dimensions (faces, landmarks, and tools), vertically arranged into a column (Figure 3.1). Row positions for the dimensions were fixed for each participant and counterbalanced across participants. Stimuli were generated by randomly assigning a feature (without replacement) on each dimension to the corresponding row of each stimulus. Participants had 1.5 s to choose one of the stimuli, after which the outcome was presented for 0.5 s. If participants did not respond within 1.5 s, the trial timed out. The inter-trial interval (ITI) was 2 s, 4 s, or 6 s (truncated geometric distribution, mean = 3.5 s), during which a fixation cross was presented. Stimulus presentations were timelocked to the beginning of a repetition time (TR). In any one game, only one dimension was relevant for predicting reward. Within that dimension, one target feature predicted reward with high probability. If participants chose the stimulus containing the target feature, they had a 0.75 probability of receiving reward. If they chose otherwise, they had a 0.25 probability of receiving reward. The relevant dimension and target feature were randomly determined for each game. Participants were instructed of the structure of the task in advance: they knew that they were looking for one target feature within one relevant dimension, as well as the reward probability associated with choosing the target feature. They were told when a new game started but were not told which dimension was relevant or which feature was the target feature. Participants performed four functional runs of the task, each consisting of six games of 25 trials each.

3.3 Methods: measuring attention

To obtain a trial-by-trial index of attention, we combined fMRI decoding and eye-tracking to directly measure participants' focus of attention as they played the task. Taking advantage of differential activation in the human ventral visual stream in response to faces, landmarks and tools, a classifier was trained on fMRI data from a localizer task in which participants were told which dimension was relevant (details on the the localizer task follow below). This classifier was then used decode which dimension participants were attending to on every trial of the uninstructed task. The trial-by-trial output of the classifier was combined with a second measure of attention obtained by averaging and normalizing looking time to each dimension within a trial. This method yielded a trial-by-trial composite measure of a participant's focus of attention as they learned the structure of the task (Figure 3.2).

3.3.1 Methods: fMRI Data Acquisition and Preprocessing

MRI data were collected using a 3T MRI scanner (Siemens Skyra). Anatomical images were acquired at the beginning of the session (T1-weighted MPRAGE, TR = 2.3 s, echo time [TE] = 3.1 s, flip angle = 9, voxel size 1 mm3). Functional images were acquired in interleaved order using a T2*-weighted echo planar imaging(EPI) pulse sequence (34 transverse slices, TR = 2 s, TE = 30ms, flip angle = 71, voxel size 3 mm3). Image volumes were preprocessed using FSL/FEATv.5.98 (FMRIB software library, FMRIB). Preprocessing included motion correction, slice-timing correction, and removal of low-frequency drifts using a temporal high-pass filter (100ms cutoff). For MVPA analyses, the classifier was trained and tested in each participant's native space. For all other analyses, functional volumes were

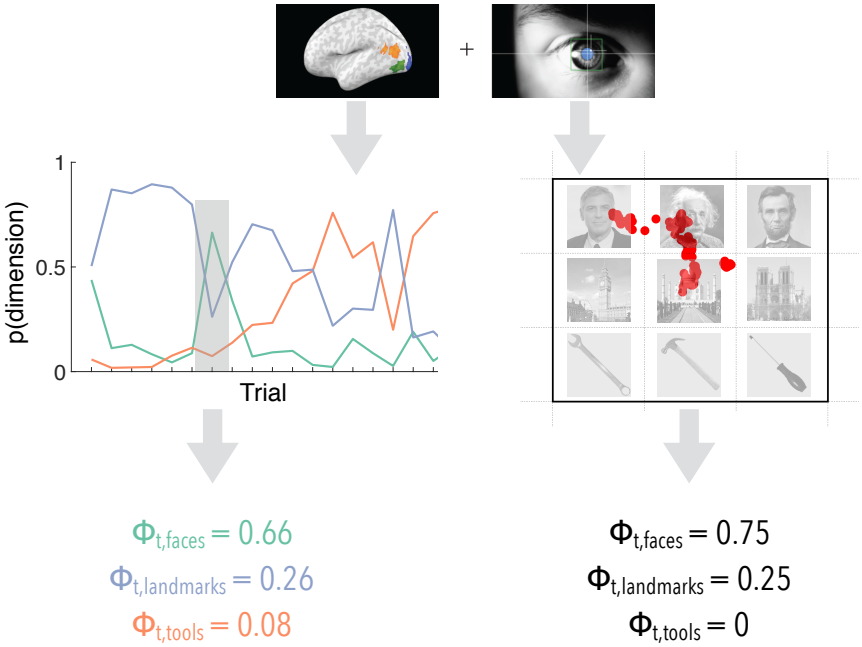


Figure 3.2: Using fMRI decoding and eye-tracking to measure attention. Left: Sample timecourse obtained by applying an SVM classifier trained on the localizer task to trials in the uninstructed task and returning the class probabilities for each dimension. In the trial marked in grey for instance, the classifier returned a probability of 0.66 for faces, 0.26 for landmarks, and 0.08 for tools. **Right:** Sample raw gaze data collected during the trial marked in grey on the left. In this example, the participant looked mostly at faces, but directed some attention to landmarks. This distribution of attention across the three dimensions is reflected in the relative looking time to faces (0.75), landmarks (0.25) and tools (0) respectively, computed by summing the raw number of samples and normalizing the counts to sum to 1. While the two measures are correlated ($r = 0.33, p < .001$), they are not redundant, likely because fMRI and eye-tracking have different sources of noise. Indeed in a model-based analysis, we show that a model which uses the composite measure outperforms one which uses the eye-tracking or MVPA measure alone (see Choice Models section).

first registered to participants' anatomical image (rigid-body transformation with 6 of freedom) and then to a template brain in Montreal Neurological Institute (MNI) space (affine transformation with 12 of freedom).

3.3.2 Methods: fMRI Localizer task

A modified one-back task was used to identify patterns of fMRI activation in the ventral visual stream that were associated with attention to faces, landmarks, or tools. Participants observed a display of the nine images similar to that used for the main task. On each trial, they had to attend to one particular dimension. Participants were instructed to respond with a button press if the horizontal order of the three images in the attended dimension repeated between consecutive trials. The order of the images on each trial was pseudo-randomly assigned so that participants would respond on average every three trials. Participants were told which dimension to attend to at the start of each run, and the attended dimension changed every one to five trials, signaled by a red horizontal box around the new attended dimension. The sequence of attended dimensions was counterbalanced (Latin square design) to minimize order effects. On each trial, the stimulus display was presented for 1.4 s, during which participants could make their response. Participants received feedback (500 ms) for hits, misses, and false alarms, but not for correct rejections (where a fixation cross was presented for 500ms instead). Each trial was followed by a variable ITI (2 s, 4 s, or 6 s, truncated geometric distribution, mean = 3.51 s). Participants performed two runs of the localizer task (135 trials each) after completing the main task.

3.3.3 Methods: fMRI MVPA Analysis

A linear support vector machine (SVM) was trained on data from the localizer task to classify the dimension that participants were attending to on each trial based on patterns of BOLD activity. Analysis was restricted to voxels in a ventral visual stream mask consisting of the bilateral occipital lobe and ventral temporal cortex. The mask was created in MNI space using anatomical masks defined by the Harvard-Oxford Cortical Structural Atlas as implemented in FSL. The mask was then transformed into each participant's native space using FSL's FLIRT implementation, and classification was performed in participants' native space. Cross-validation classification accuracy on the localizer task was 87.4% (SE = 0.9%; chance level: 33%). The SVM was then applied to data from the Dimensions Task to classify participants' trial-by-trial attention to the three dimensions. Classification was performed using the SVM routine LinearNuSVMC (Nu = 0.5) implemented in the PyMVPA package(Hanke et al., 2009). On each trial, the classifier returned the probability that the participant was attending to each of the dimensions (three numbers summing to 1). To model subject-specific noise inherent in the BOLD data, a weighted sum was computed between these probabilities and uniformly distribution attention. The weight ω_{MVPA} was a free parameter fit to each participant's behavioral data (see Choice Models section below). As ω_{MVPA} decreased, the MVPA measure of attention contributed less to the final attention vector. Fitting a subject-specific ω_{MVPA} parameter thus provided us with a data-driven method to weigh the MVPA measure based on how much it contributed to explaining choices.

3.3.4 Methods: eye-tracking

Eye-tracking data were acquired using an iView X MRI-LR system (SMI SensoMotoric Instruments) with a sampling rate of 60 Hz. System output files were analyzed using in-house MATLAB code. A horizontal rectangular area of interest (AOI) was defined around each horizontal dimension in the visual display. Data were preprocessed by low-pass filtering (10 Hz cutoff) to reduce high-frequency noise, discarding data from the first 200 ms after the onset of each trial to account for saccade latency and taking the proportion of time participants' point of gaze resided within each AOI as a measure of attention to the corresponding dimension. The level of noise in the eye-tracking measure can vary systematically between participants. As with the fMRI data, to account for subject-specific noise, we computed a weighted sum between the raw measure and uniform attention (one-third to each dimension). The weight ω_{ET}, which served to smoothly interpolate between uniform attention and the eye-tracking measure of attention, was a free parameter fit to each subject's behavioral data (see Choice Models section below). Lower values of ω_{ET} meant the eye-tracking attention measure contributed less to the final attention vector.

3.3.5 Methods: composite measure of attention

To combine the two measures of attention, a composite measure was computed as the product of eye-tracking and MVPA measures of attention, renormalized to sum to 1. Taking a product means that each of the two measures contributes to the composite according to how strongly the measure is biased toward one dimension and not others. For example, a uniform $(1/3, 1/3, 1/3)$ measure contributes nothing to the composite measure for that trial. In contrast, if one measure is extremely biased to one dimension (e.g., $1, 0, 0$), it over-

rides the other measure completely. In the Choice Models section below, we provide the exact equations used to compute the composite measure, and show how it can be used to separately modulate computing and updating the value of each stimulus.

3.4 Methods: choice models

To elucidate the role attention might play in the computations that underlie learning and decision-making, four RL models were tested by fitting them trial-by-trial to participants' choices (Sutton & Barto, 2018; Daw, 2011). All four models assumed that participants learned to associate each feature with a value and linearly combined the values of features to obtain the value of a compound Face-Landmark-Tool stimulus:

$$V_{(t)}(S_i) = \sum_d \Phi_t(d) \cdot v_t(d, S_i) \tag{3.1}$$

$V_t(S_i)$ is the value of the stimulus i on trial t, $\Phi_t(d)$ is the attention weight of dimension d and $v_t(d, S_i)$ denotes the value of the feature in dimension d of stimulus S_i. Following feedback, a prediction error, δ_t, was calculated as the difference between observed reward, r_t, and the expected value of the chosen stimulus $V_t(S_c)$:

$$\delta_t = r_t - V_t(S_c) \tag{3.2}$$

δ_t was then used to update the feature values of the chosen stimulus:

$$v_{(t+1)}(d, S_c) = v_{(t)}(d, S_c) + \eta \cdot \Phi_t(d) \cdot \delta_t \tag{3.3}$$

The update is weighted by attention to the respective dimensions and scaled by a learning rate η, which was fit to each participants' behavioral data. Because one prediction error and one update were computed per trial, this model can be viewed as an instance of the Rescorla-Wagner learning rule (Rescorla, Wagner, et al., 1972), which in turn is a special case of TD-learning with a discount factor set to 0 (Ludvig, Sutton, & Kehoe, 2012).

In the ACL (Attention at Choice and Learning) model, both value computation and value update were biased by attention weights. In the AC (Attention at Choice) model, the attention measure was used for value computation, but all $\Phi_t(d)$ were set to one-third during value update such that the three dimensions were updated equally during learning. In the AL (Attention at Learning) model, the attention measure was used for value update, but all $\Phi_t(d)$ were set to one-third for value computation, weighting all dimensions equally at choice. In the UA (Uniform Attention) model, $\Phi_t(d)$ were set to one-third for both value computation and value update. For all models, choice probabilities were computed according to a softmax action selection rule:

$$\pi_t(i) = \frac{e^{\beta V_t(S_i)}}{\sum_a e^{\beta V_t(S_a)}} \tag{3.4}$$

In the equation above, $\pi_t(i)$ is the probability of choosing stimulus i, a enumerates over the three available stimuli, and β is a free inverse-temperature parameter that determines how strongly choice is biased toward the maximal-valued stimulus.

The trial-by-trial composite measure of attention to each dimension $\Phi_t(d)$ was computed as follows. First, the raw MVPA measure was smoothed:

$$\Phi_t(d)_{MVPA,smoothed} = \Phi_t(d)_{MVPA,raw} \cdot \omega_{MVPA} + (1 - \omega_{MVPA}) \cdot [1/3, 1/3, 1/3] \tag{3.5}$$

The same transformation was applied to the eye-tracking measure with a different smoothing parameter:

$$\Phi_t(d)_{ET,smoothed} = \Phi_t(d)_{ET,raw} \cdot \omega_{ET} + (1 - \omega_{ET}) \cdot [1/3, 1/3, 1/3] \tag{3.6}$$

Finally, the two smoothed measures were combined by element-wise multiplication and re-normalizing:

$$\Phi_t(d) = \frac{\Phi_t(d)_{MVPA,smoothed} \odot \Phi_t(d)_{ET,smoothed}}{\sum_d \Phi_t(d)_{MVPA,smoothed} \odot \Phi_t(d)_{ET,smoothed}} \tag{3.7}$$

3.5 Methods: fitting and comparing the choice models

The three choice models incorporating the composite attention measure (ACL, AL and AC) had four free parameters – ω_{MVPA}, ω_{ET}, β and η – while the uniform attention model (UA) had two free parameters, β and η. As in Chapter 2, model comparison used a leave-one-game-out cross-validation procedure: for each participant and for each game, the model was fit to participants' choices from all other games by minimizing the negative log-likelihood of the choices (Daw, 2011). For the minimization procedure, we used MATLAB's *fmincon* function.

Given the best fit parameters, the likelihood of each choice in the held-out game was computed. The total likelihood of the data of each participant, computed for each game as it was held out, was then divided by the number of trials that the participant played to obtain the geometric average of the likelihood per trial. Using cross-validation allowed us to compare between models based on their likelihood per trial without over-fitting and thus we did not need to correct for model complexity.

3.6 Methods: attention models

The main benefit of the approach outlined in this chapter is that it allows us to directly measure the focus of attention as participants learn task representations. The section above treats the MVPA and eye-tracking measures as direct readouts of attention. Combining these attention measures with RL models of choice allows us to ask what role attention plays in the computations that underlie trial and error learning and decision making.

But with the attention measure in hand, we can also study the other side of this interaction: given past choices and rewards, how is attention determined in the first place? To begin chipping away at this question, in models presented in this section, I treated the attention measure as an outcome variable. That is, instead of making predictions about what the participants choose trial by trial, I tried to predict how attention to each of the face, landmark and tool dimensions changes as a function of recent experience. Specifically, I built a series of RL models that tested the hypothesis that attention fluctuates with recent rewards, while controlling for another factor known to modulate attention: choice (Krajbich, Armel, & Rangel, 2010).

The first two models tested the hypothesis that attention only depends on prior choices. The "Full Choice History" model allocated attention based on a leaky choice count. This model instantiated the hypothesis that participants are more likely to attend to features that have been chosen most often. On each trial, counts for each of the three chosen features were incremented by 1, and counts for the remaining six unchosen features were decayed toward 0 at a subject-specific decay rate. Attention to each dimension was then determined by the softmax of the maximal count on each dimension. That is, on each trial, I took the

highest counts among the three features of each dimension and passed them through a softmax function (see Equation 3.4) to obtain three attention weights that sum up to 1. This model had two free parameters: a decay rate for the choices, and the softmax inverse temperature for distributing attention in proportion to highest choice counts within each dimension.

The "Recent Choice History" model added an additional assumption: only recent choices determine attention. The model used a delta-rule update to adjust the weights of the chosen features toward 1. For each chosen feature, the weight $w_t(d, S_{chosen})$ was updated as:

$$w_{t+1}(d, S_{chosen}) = w_t(d, S_{chosen}) + \eta_a[1 - w_t(d, S_{chosen})] \tag{3.8}$$

where η_a is a free update rate parameter. Here, too, the weights of the unchosen features were decayed toward 0 at a subject-specific decay rate, and the predicted attention weights were determined using softmax on the maximum weights in each dimension. In addition to the decay rate for choices and softmax temperature free parameters, this model also had a learning rate for updating choice counts.

The next two models tested the hypothesis that attention is only directed to chosen features, but only they have been rewarded.

In contrast to the "Full Choice History" model, the "Full Reward History" model allocated attention based on a leaky reward count: on rewarded trials only, counts of chosen features were incremented by 1 and counts of unchosen features were decayed toward 0 at a subject-specific decay rate. No learning or decay occurred on unrewarded trials. Again, softmax was applied to the maximum counts in each dimension to determine attention. This model also had two free parameters: a decay rate, and the softmax inverse temperature

for distributing attention in proportion to highest reward counts within each dimension.

In the "Recent Reward History" model, analogous to the "Recent Choice History" model, on each rewarded trial a delta-rule update was used to adjust the weights of the chosen features toward 1. As before, weights of the unchosen features were decayed toward 0. This model builds in the additional assumption that participants learn to attend to features that were chosen recently, but only if the choice was rewarded. In addition to the decay rate and softmax temperature, this model also had a learning rate for updating reward counts.

Finally, in the "Value" model, attention was hypothesized to follow feature values that are learned from trial and error. Specifically, feature values were initialized to 0 and updated via feature reinforcement learning with decay, the same model presented in Chapter 2 (Niv et al., 2015; Radulescu et al., 2016): on each trial, a prediction error was calculated as the difference between the obtained reward and the value of the chosen stimulus. The value of each stimulus was assumed to be the sum of the values of all its features (Equation 3.1). The value of chosen features was updated based on the prediction error scaled by a subject-specific update rate, while the value of unchosen features was decayed toward 0 at a subject-specific decay rate. While in Chapter 2 stimulus values in the feature RL + decay model only determined choice, here I assumed that feature values also guided attention. As in the models above, the maximum feature value in each dimension was passed through a softmax function to obtain the predicted attention vector. This model is different from the "Recent Reward History" model in that it learns not only from positive, but also from negative prediction errors, basing attention learning on a stimulus-level prediction error. The "Value" model had three free parameters: a learning rate for updating feature values, a decay rate and a softmax temperature for determining how attention is distributed between

dimensions in proportion to the highest-valued features.

3.7 Methods: fitting and comparing the attention models

I compared these five attention models to a baseline zero-parameter model in which attention is always uniform $(1/3, 1/3, 1/3)$. As with our models of choice behavior, the attention models were evaluated using leave-one-game-out cross-validation: for each participant and for each game, I fit the free parameters of the model to all but that game by minimizing the Root-Mean-Square Deviation (RMSD) of the predicted attention weights from the measured attention weights. I then used the model to predict attention weights for the left-out game to determine the mean RMSD per trial for each model (Figure 3.3. I fit the models separately to the raw (preprocessed but not smoothed) eye-tracking attention measure, the raw (unsmoothed) MVPA attention measure, and the composite measure (with smoothing parameters ω_{ET} and ω_{MVPA} determined according to the best fit to choice behavior)*. Other distance methods are possible, and are theoretically more appropriate for compositional data (Aitchison, 1982). All model comparison results presented here are robust to the choice of distance metric.

3.8 Methods: fMRI analyses

Three general linear models (GLMs) were implemented as design matrices for analysis of the fMRI data:

GLM1 served to investigate whether the computation and update of the expected value

*Aside from the smoothing parameters, the learning rate and decay rate were optimized based on gaze data. In the Conclusions chapter, we discuss the issue of jointly fitting gaze and choice data

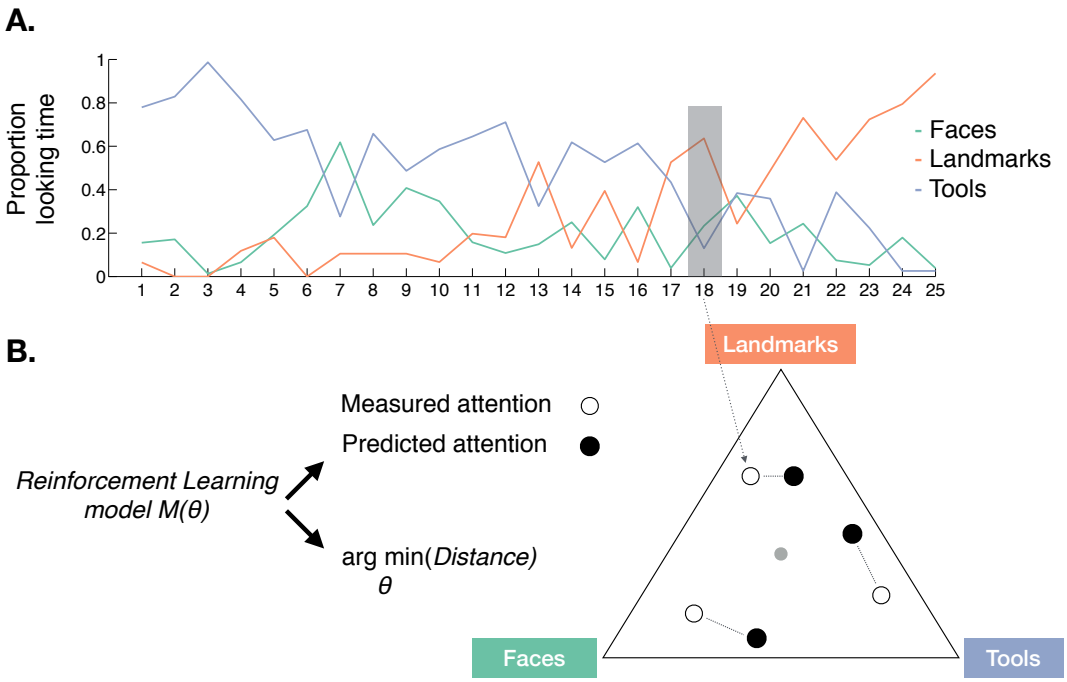


Figure 3.3: Modeling 'attention learning'. A. sample time course obtained from eye-tracking during one game, as the participant was trying to learn the target feature from trial and error. Even though the target feature stayed constant throughout this game, the participant switched attention at least once, focusing on tools early in the game, but then gradually switching to landmarks. **B.** Schematic of model comparison procedure for attention data. The attention index on each trial is a 3D vector of weights that sum to 1 (also known as a compositional vector (Aitchison, 2003)), which can be visualized on a 2-simplex (right). The projection of the attention weights on the trial highlighted in (A) is marked by an arrow. In this particular trial, attention was directed mostly to landmarks. In general, the closer a point is to the corners of the triangle, the more focused attention is on one of the three dimensions. The grey dot in the middle denotes uniform attention $(1/3, 1/3, 1/3)$. As a distance measure for model comparison, I used the root mean square distance (RMSD) between the measured and predicted attention distribution on each trial.

signal in the brain was biased by attention. For each participant, estimates for the expected value of the chosen stimulus on each trial were generated using the UA, AC, AL, and ACL models. These value estimates were entered into the GLM as parametric modulators of the stimulus onset regressor. Because, in linear regression, variance shared by different regressors is automatically not attributed to any of the regressors, the regressors were not orthogonalized. GLM1 could therefore identify regions that are associated with the value estimates of each model, while simultaneously controlling for the value estimates of the other models. Through contrasting the different model regressors, this GLM allows pinpointing areas that correlate with values from one model significantly better than they correlate with values from another model.

Reaction time, trial outcome, outcome onset, and head movement parameters were also added as nuisance regressors. With the exception of head movement parameters, all regressors were convolved with the hemodynamic response function. Missed-response trials were not modeled as there was no chosen stimulus in those trials. The GLM was estimated throughout the whole brain using FSL/FEAT v.5.98 available as part of the FMRIB software library (FMRIB). Results were corrected for multiple comparisons using a family-wise error cluster-corrected threshold of $p < 0.05$ (FSL FLAME 1), with a cluster-forming threshold of $p < 0.001$. Unless otherwise stated, all GLM analyses included the same nuisance regressors and were corrected for multiple comparisons using the same procedure.

GLM2 served to investigate whether prediction error signals were also biased by attention. This GLM was identical to GLM1 except that instead of including estimates of expected values, estimates of trial-by-trial prediction errors were generated using the four models and entered into the GLM as parametric modulators of the outcome onset regres-

sor.

GLM3 modeled switch and stay trials as stick functions at the onset of the respective trials. Switch trials were defined as trials in which the maximally attended dimension was different from the previous trial. All other trials were modeled as stay trials. A contrast identified clusters that were more active during switch versus stay trials.

3.9 Results: attending to learn

Through trial and error, participants learned to choose the stimulus containing the target feature over the course of a game (Figures 3.1B and C). A learned game was defined as one in which participants chose the target feature on every one of the last five trials. By this metric, participants learned on average 11.3 (SE = 0.7) out of 25 games. The number of learned games did not depend on the relevant dimension ($F(2, 24) = 0.886, p = 0.42$).

In this section, I present behavioral and neural evidence suggesting that decision-making and learning in multidimensional settings are both constrained by attention.

3.9.1 Both choice and learning are biased by attention

The ACL model, in which the composite measure of attention modulated both choice and learning, outperformed the other three models. Average likelihood per trial for the ACL model was highest for 21 of 25 subjects, on average significantly higher than that for the AC ($t_{24} = 4.72, p < 0.001$), AL ($t_{24} = 6.70, p < 0.001$), and UA ($t_{24} = 8.61, p < 0.001$) models. Both the AC and AL models also yielded significantly higher average likelihood per trial than the UA model (AC: $t_{24} = 8.03, p < 0.001$; AL: $t_{24} = 7.20, p < 0.001$) (Figure 3.4A). Moreover, the average likelihood per trial of the ACL model diverged sig-

nificantly from that of the other models early in the game, when performance was still well below asymptote (as early as trial 2 for AL and UA, and from trial 7 for AC). These results were not driven by the learned portion of games (in which participants may have focused solely on the relevant dimension), as they held when tested on unlearned games only.

The ACL model used the same set of attention weights for choice and learning; however, previous theoretical and empirical work suggest that attention at choice might focus on stimuli or features that are most predictive of reward (Mackintosh, 1975), whereas at learning, one might focus on features for which there is highest uncertainty (Pearce & Hall, 1980). To test whether attention at learning and attention at choice were separable, I took advantage of the higher temporal resolution of the eye-tracking measure. I considered eye positions from 200 ms after stimulus onset to choice as indicating "attention at choice" and eye positions during the 500 ms of outcome presentation as a measurement of "attention at learning." Attention at choice and attention at learning on the same trial were moderately correlated (average $r = 0.56$), becoming increasingly correlated over the course of a game ($F(24, 24) = 4.95, p < 0.001$). This suggests that as participants figured out the relevant dimension, they attended to the same dimension in both phases of the trial. When the ACL model was fit using attention at choice to bias value computation and attention at learning to bias value update, the model performed slightly, but significantly, better than the ACL model that used whole-trial attention weights for both choice and learning. This suggests that attentional processes at choice and at learning may reflect dissociable contributions to decision making.

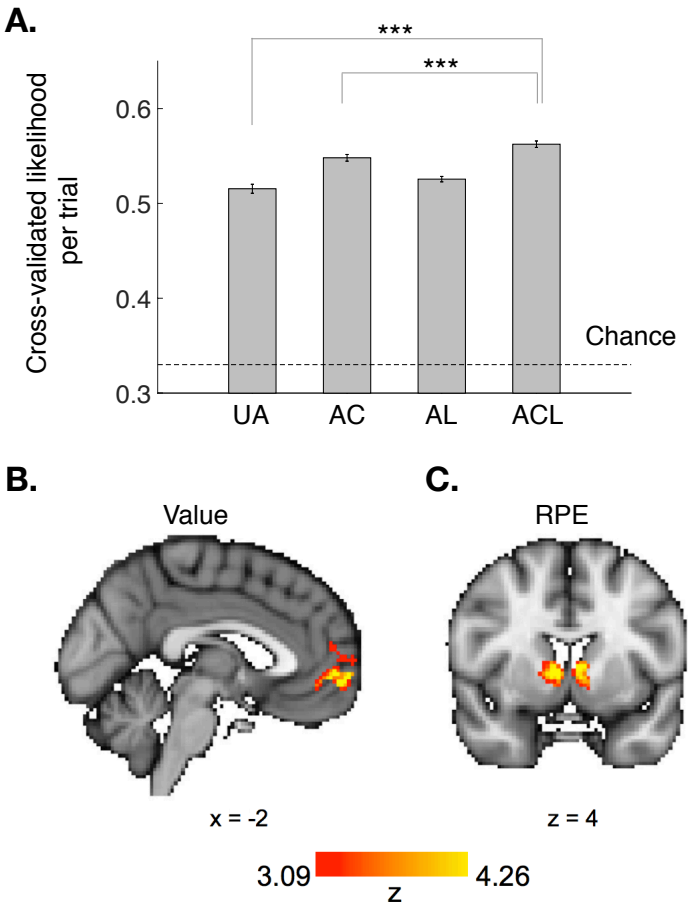


Figure 3.4: Behavioral and neural evidence for attentional selection during learning. A. Average choice likelihood per trial for each model shows that the Attention at Choice and Learning (ACL) model predicted the data significantly better than other models (paired t tests, $p < 0.001$). **B.** BOLD activity in the vmPFC was significantly correlated with the value estimates of the ACL model, controlling for the value estimates of the AC, AL, and UA models. **C.** BOLD activity in the striatum correlated with reward prediction errors generated by the ACL model, controlling for reward prediction error estimates from the AC, AL, and UA models. In sum, the ACL model's predictions for value and prediction errors best corresponded to their respective neural correlates.

3.9.2 Attention biases neural value and reward prediction errors signals

To test which model was most consistent with the neural value representation, the trial-by-trial value estimates of the chosen stimulus generated by all four models were entered into a single GLM (GLM1 in the fMRI Analyses section). This allowed searching the whole brain for clusters of brain activity whose variance was uniquely explained by one of the models while simultaneously controlling for the value estimates of the other models. Results showed that activity in the vmPFC was significantly correlated with the value estimates of the ACL model (Figure 3.4B), suggesting that the computation and update of the value representation in vmPFC was biased by attention. No clusters were significantly correlated with the value estimates of the AC and AL model; one cluster in the visual cortex was significantly correlated with the value estimates of the UA model. Next, we investigated whether neural prediction error signals were also biased by attention. For this, trial-by-trial prediction error regressors generated by each of the four models were entered into a single whole-brain GLM (GLM2 in the fMRI Analyses section). Prediction errors generated by the ACL model were significantly correlated with activity in the striatum (Figure 3.4C), the area most commonly associated with prediction error signals in fMRI studies (O'Doherty et al., 2004; Niv, Edlund, Dayan, & O'Doherty, 2012; Pagnoni, Zink, Montague, & Berns, 2002). Prediction error estimates of the other models were not significantly correlated with any cluster in the brain. Together, these results provide neural evidence that attention biases both the computation of subjective value as well as the prediction errors that drive the updating of those values.

3.10 Results: learning to attend

In previous analyses, I demonstrated that attention biased both choice and learning. In the subsequent analyses, I focus on the other side of the bidirectional relationship, examining how learning modulates attention.

3.10.1 Attention dynamics are sensitive to reward

To understand how subjects decided what features to attend too, I compared different models of the trial-by-trial dynamic allocation of attention. In particular, I tested whether attention allocation could be better explained by choice history (i.e., attention was enhanced for features that have been previously chosen), reward history (i.e., attention was enhanced for features that had been previously rewarded), or learned value (i.e., attention was enhanced for features associated with higher value over the course of a game). Cross-validated model comparison revealed that both the eye-tracking and attention data were best explained by a model that tracked feature values. In particular, the "Value" model outperformed the next-best "Recent Reward History" model for both the eye-tracking (lowest root-mean-square deviation [RMSD] in 17/25 subjects, paired-sample t-test, $t(24) = 2.77, p < 0.05$, Figure 3.5A) and composite attention measured (lowest RMSD in 18/25 subjects, paired-sample t-test, $t(24) = 2.41, p < 0.05$). For the MVPA data, the "Value" model did not significantly improve upon the predictions of the Recent Reward History model (lowest RMSD in 16/25 subjects, paired-sample t test, t(24) = 1.02, p = 0.31, Figure 3.5B); however, it still performed significantly better than the "Recent Choice History" model (paired-sample t-test, $t(24) = 3.83, p < 0.001$).

These results suggest that attention is dynamically modulated by ongoing learning. As

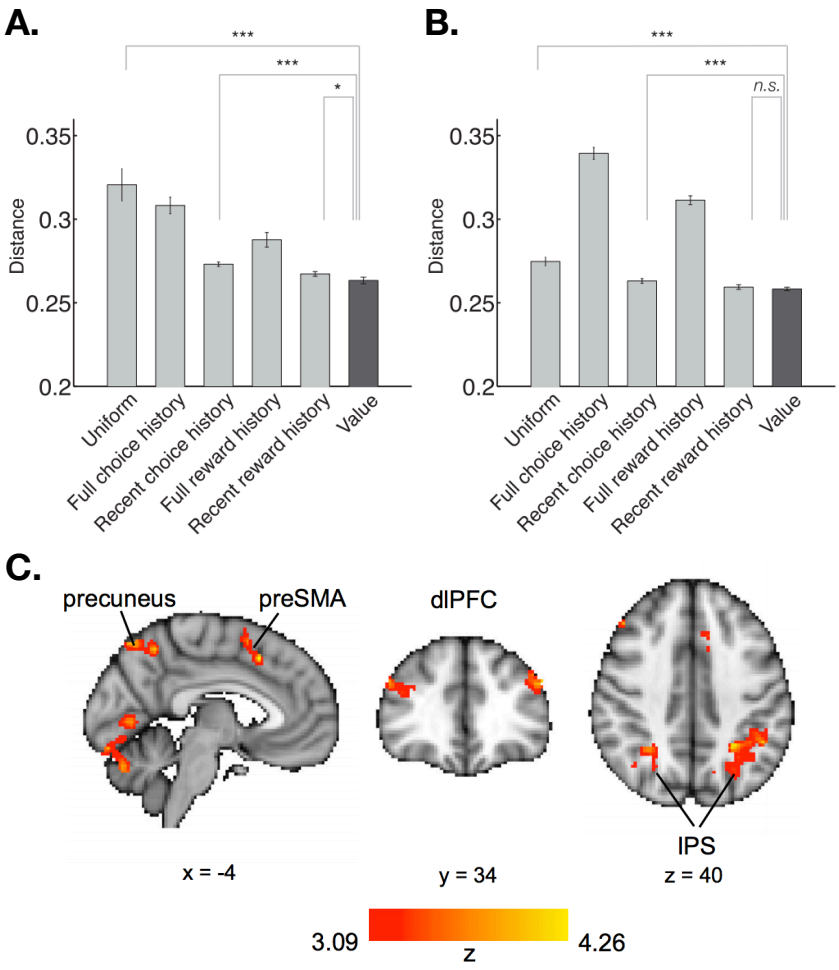


Figure 3.5: Behavioral and neural evidence for reward-sensitive attention dynamics. A. and B. Comparison of models of attention fitted separately to the eye-tracking (A) and the MVPA (B) measures, according to the root-mean-square deviation (RMSD) of the model's predictions from the empirical data (lower values indicate a better model). Plotted is the subject-wise average per-trial RMSD calculated from holdout games in leave-one-game-out cross-validation (see Fitting and Comparing the Attention Models section). The Value model (dark grey) has the lowest RMSD. Error bars, 1 SEM. ***p < 0.001; **p < 0.01; *p < 0.05. **C.** BOLD activity in the frontoparietal network was higher on switch trials than on stay trials.

participants learned to associate value with features over the course of a game, attention was directed toward dimensions with features that acquired high value (which, in this task, are also the features that are most predictive of reward (Mackintosh, 1975)). The greater the feature values in a dimension, the stronger the attention bias was toward that dimension. Finally, attention was better explained as a function of learned value rather than simpler models of reward or choice history.

3.10.2 Neural correlates of attention switches

Modeling results suggested that ongoing learning and feedback dynamically modulated participants' deployment of attention. How might the brain be realizing these attention dynamics? To answer this, I searched for brain areas that were more active during switches in attention. Trials on which the maximally attended dimension was different from that of the preceding trial were labeled as switch trials and the rest as stay trials. A contrast searching for more activity on switch rather than stay trials (GLM3 in the fMRI Analyses section) showed clusters in the dorsolateral prefrontal cortex (dlPFC), intraparietal sulcus (IPS), frontal eye fields (FEF), presupplementary motor area (preSMA), precuneus, and fusiform gyrus (Figure 3.5C). These brain regions are part of a frontoparietal network that has been implicated in the executive control of attention (Corbetta & Shulman, 2002; Petersen & Posner, 2012). These results suggest that this attentional control system also supports top-down allocation of attention during learning and decision making in multidimensional environments.

3.11 Discussion

As discussed and shown in Chapter 2, learning and attention play complementary roles in facilitating adaptive decision making (Niv et al., 2015; Wilson & Niv, 2012). Here, to test directly for the interaction between attention and reward learning, we combined computational modeling, eye tracking, and fMRI to study the interaction between trial-and-error learning and attention in a decision-making task. We used eye tracking and pattern classification of multivariate fMRI data to measure participants' focus of attention as they learned which of three dimensions of task stimuli was instrumental to predicting and obtaining reward. Model-based analysis of both choice and neural data indicated that attention biased how participants computed the values of stimuli and how they updated these values when obtained reward deviated from expectations. The strength and focus of the attention bias was, in turn, dynamically modulated by ongoing learning, with trial-by-trial allocation of attention best explained as following learned value rather than the history of reward or choices. Blood-oxygen-level-dependent (BOLD) activity in a frontoparietal executive control network correlated with switches in attention, suggesting that this network is involved in the control of attention during reinforcement learning.

This study builds on a growing body of literature in which RL models are applied to behavioral and neural data. Converging evidence suggests that the firing of midbrain dopamine neurons during reward-driven learning corresponds to a prediction error signal that is key to learning (D. Lee, Seo, & Jung, 2012; Schultz et al., 1997; Steinberg et al., 2013). The dopamine prediction error hypothesis has generated much excitement, as it suggests that RL algorithms provide a formal description of the mechanisms underlying learning. How-

ever, it is becoming increasingly apparent that this story is far from complete (Dayan & Niv, 2008; Langdon, Sharpe, Schoenbaum, & Niv, 2018). In particular, RL algorithms suffer from the "curse of dimensionality": they are notoriously inefficient in realistic, high-dimensional environments (Bellman, 1957; Sutton & Barto, 2018).

How can the RL framework be extended to provide a more complete account of real-world learning? A key insight is that learning can be facilitated by taking advantage of regularities in tasks. For example, humans can aggregate temporally extended actions into subroutines that reduce the number of decision points for which policies have to be learned (Botvinick, 2012). The results here point to a parallel strategy whereby participants employ selective attention to simplify the state representation of the task. While real-world decisions often involve multidimensional options, not all dimensions are relevant to the task at hand. By attending to only the task-relevant dimensions, one can effectively reduce the number of environmental states to learn about. In the Dimensions Task, for example, attending to only the face dimension simplifies the learning problem to one with three states (each of the faces) rather than 27 states corresponding to all possible stimulus configurations. Selective attention thus performs a similar function as dimensionality-reduction algorithms that are often applied to solve computationally complex problems in the fields of machine learning and artificial intelligence (Ponsen, Taylor, & Tuyls, 2009).

Drawing on theories of visuospatial attention (Desimone & Duncan, 1995), we conceptualized attention as weights that determine how processing resources are allocated to different aspects of the environment. In the computational models of choice behavior, these weights influenced value computation in choice and value update in learning. Several previous studies have taken a similar approach to investigate the relationship between attention

and learning (Jones & Canas, 2010; Marković, Gläscher, Bossaerts, O'Doherty, & Kiebel, 2015; Wilson & Niv, 2012; Wunderlich, Beierholm, Bossaerts, & O'Doherty, 2011), and recently, we demonstrated that neural regions involved in control of attention are also engaged during learning in multidimensional environments, providing neural evidence for the role of attention in learning (Niv et al., 2015). These prior studies, however, have relied on inferring attention weights indirectly from choice behavior or from self-report.

Here, we obtained a direct measure of attention, independent of choice behavior, using eye-tracking and MVPA analysis of fMRI data. Attention and eye movements are functionally related (Kowler, Anderson, Dosher, & Blaser, 1995; D. Smith, Rorden, & Jackson, 2004) and share underlying neural mechanisms (Corbetta et al., 1998; Moore & Fallah, 2001). Attention is also known to enhance the neural representation of the attended object category (O'Craven, Downing, & Kanwisher, 1999), which can be decoded from fMRI data using pattern classification (Norman, Polyn, Detre, & Haxby, 2006). Therefore, as a second proxy for attention, we quantified the level of category-selective neural patterns of activity on each trial. By incorporating attention weights derived from the two measures into computational models fitted to participants' choices, we provide evidence for the influence of attention processes on both value computation and value updating during RL.

Previous work has shown that value computation is guided by attention (Krajbich et al., 2010) and that value signals in the vmPFC are biased by attention at the time of choice (Hare, Malmaud, & Rangel, 2011; Lim, O'Doherty, & Rangel, 2011). For example, Hare et al. (2011) found that when attention was called to the health aspects of food choices, value signals in the vmPFC were more responsive to the healthiness of food options, and participants were more likely to make healthy choices. Here, we extended those findings and

demonstrated that attention biases not only value computation during choice, but also the update of those values following feedback. Another neural signal guiding decision making is the reward prediction error signal, which is reflected in BOLD activity in the striatum, a major site of efferent dopaminergic connections (O'Doherty et al., 2004; Seymour et al., 2004). This prediction-error signal was also biased by attention, providing additional evidence that RL signals in the brain are attentionally filtered.

But how does the brain know what to attend to? To facilitate choice and learning, attention has to be directed toward stimulus dimensions that are relevant to obtaining reward, such that learning processes operate on the correct state representation of the task. However, at the beginning of each game in the Dimensions Task, participants did not know which dimension is relevant. The results presented here, as well as those presented in Chapter 2, suggest that without explicit cues, participants can learn to attend to the dimension that best predicts reward and dynamically modulate both what they attend to and how strongly they attend based on ongoing feedback. These findings are consistent with the view of attention as an information seeking mechanism that selects information that best informs behavior (Gottlieb, 2012). In particular, a model in which attention was allocated based on learned value provided the best fit to the empirical attention measures. Notably, this model is closely related to a model of associative learning that assumes attention is directed to features that are most predictive of reward (Mackintosh, 1975).

An alternative view is that attention should be directed to the most uncertain features in the environment — that is, the features that participants know the least about and that have been associated with more prediction errors (Pearce & Hall, 1980). In support of this theory, errors in prediction have been shown to enhance attention to a stimulus and in-

crease the learning rate for that stimulus (Esber et al., 2012; Holland & Gallagher, 2006). The seemingly contradictory Mackintosh and Pearce-Hall theories of attention have both received extensive empirical support (Dayan, Kakade, & Montague, 2000). Dayan et al. (2000) offered a resolution by suggesting that when making choices, one should attend to the most reward-predictive features, whereas when learning from prediction errors, one should attend to the most uncertain features. When attention at choice and attention at learning were assessed separately, the two measures were correlated. Nevertheless, a model with separate attention weights at choice and learning fit participants' data better than the same model that used the same whole-trial attention weights at both phases. This result suggests a dissociation between attention at choice and learning, although further work is clearly warranted to determine how attention in each phase is determined. In particular, the current task was not optimally designed for separately measuring attention at choice and attention at learning as the outcome was only presented for 500 ms, during which participants also had to saccade to the outcome. The task was also not well suited to test the Pearce-Hall framework for attention at learning because, in the Dimensions Task, the features associated with more prediction errors are features in the irrelevant dimensions that participants were explicitly instructed to try to ignore.

Neural results suggested that the flexible deployment of attention during learning and decision making is controlled by a frontoparietal network that includes the IPS, FEF, precuneus, dlPFC, and preSMA. This network has been implicated in the executive control of attention in a variety of cognitive tasks (Corbetta & Shulman, 2002; Petersen & Posner, 2012), and the dlPFC in particular is thought to be involved in switching among "task sets" by inhibiting irrelevant task representations when task demands change (Dias, Robbins, &

Roberts, 1996; Hyafil, Summerfield, & Koechlin, 2009). Our findings demonstrate that the same neural mechanisms involved in making cued attention switches can also be triggered in response to internal signals that result from learning from feedback over time. A possible interpretation is that the frontoparietal executive control network flexibly adjusts the focus of attention in response to ongoing feedback, such that learning can operate on the correct task representation in multidimensional environments.

In summary, this chapter provided behavioral and neural evidence for a dynamic relationship between attention and learning: attention biases what we learn about, but we also learn what to attend to. By incorporating attention into the reinforcement learning framework, we provided a solution for the seemingly computationally intractable task of learning and decision making in high-dimensional environments. This chapter also demonstrated the potential of using eye-tracking and MVPA to measure trial-by-trial attention in cognitive tasks. Combining such measures of attention with computational modeling of behavior and neural data will be useful in future studies of how attention interacts with other cognitive processes to facilitate adaptive behavior.

Without a filter, one is just chaos walking.

adapted from Patrick Ness

4

Selective attention as particle filtering

As shown in Chapters 2 and 3, SELECTIVE ATTENTION plays a role in representation learning of task-relevant features. Yet a formal theory is still missing for how humans learn what to attend to. In this chapter, parts of which have been published as a short conference paper (Radulescu, Niv, & Daw, 2019), I lay out a formal account of 'attention learning' grounded in principles of statistical inference. In particular, I model the dynamics of selective attention as a memory-augmented particle filter, a flexible sequential sampling algorithm that can provide approximate solutions to complex inference problems (Sanborn & Chater, 2016; Speekenbrink, 2016; A. Smith, 2013). I show that trial-by-trial attention to features measured using eye-tracking is better fit by a one-particle particle filter, compared to the reinforcement learning mechanism introduced in Chapters 2 and 3. This is because inference based on a single particle captures the sparse allocation of attention, predominantly to one dimension at each point in time. Unlike gradual reinforcement learning, the particle filter can also accommodate the rapid switching of attention between dimensions. However, because a single particle maintains insufficient information about past events to switch hypotheses as efficiently as do participants, the data are best fit by the filter augmented with a memory buffer for recent observations. This proposal suggests a new role

for working memory in enabling tractable, resource-efficient approximations to normative inference, and proposes a tight link between memory and attention functions in realistic tasks.

4.1 Introduction

Results in previous chapters highlight a role for selective attention in shaping reinforcement learning in multidimensional environments. In the work that follows, I suggest that selective attention during human reinforcement learning arises from sequential sampling of hypotheses about which features of a task are relevant. I formalize the attentional selection process as a memory-augmented particle filter (Doucet & Johansen, 2009; Bonawitz, Denison, Gopnik, & Griffiths, 2014; Speekenbrink, 2016). Particle filters offer a tractable approximation to rational inference and in the case of only a few particles, resemble sequential hypothesis testing (Wilson & Niv, 2012). The key idea of a particle filter is to represent the target probability distribution using a finite number of point estimates, or particles. The ensemble of particles is dynamic: estimates that are inconsistent with recent evidence are filtered out. With additional experience, the ensemble comes to better approximate the target distribution.

In general, the quality of the approximation increases both with time, and with the number of particles. Here, I show that a single-particle model does well in capturing the dynamics of human attention allocation, due to the sparsity of the representation and the model's ability to rapidly switch hypotheses about the identity of the reward-predictive feature. But such sparsity is in tension with the main normative appeal of particle filters, which is to treat the ensemble of particles as approximating the exact posterior at each step.

One way to compensate for using fewer particles is through the choice of 'proposal distribution' for re-sampling particles. In general, the closer the match between the proposal distribution and the target posterior distribution that is being approximated, the better the approximation to the posterior will be (Speekenbrink, 2016). In the Dimensions Task (Chapters 2 and 3), the proposal distribution defines an implicit switching rule for staying with the current hypothesis about which feature is more predictive of reward, or switching to a different one. One suggestion in a similar setting was to use the exact posterior as the proposal distribution, effectively endowing the model with the ability to switch hypotheses in proportion to the true posterior probability (Bonawitz et al., 2014). But this is unrealistic as a process-level model, since it relies on access to the very distribution the particle filter is attempting to approximate.

Thus, I replace this assumption with a novel memory mechanism that modifies the proposal distribution to incorporate a set of the most recent observations. This modification both solves the efficiency problem associated with single-particle models and highlights a new role for memory in enabling approximate inference. I develop a method for fitting memory-augmented particle filters to trial-by-trial eye-tracking data, and compare the particle filter to the reinforcement learning account of selective attention in the Dimensions Task (Chapters 2 and 3). I find that the memory-augmented particle filter more closely matches the trial-by-trial dynamics of attention allocation, suggesting a role for memory in guiding attention to task relevant features.

4.2 Particle filter model

Consider the task environment described in Chapters 2 and 3: multidimensional stimuli vary along several discrete dimensions d (e.g. Faces, Landmarks and Tools). Each dimension can take on f features per dimension (e.g. Einstein, ...). The structure of the task is defined such that only one dimension is relevant for reward, and one target feature f^* within that dimension is most rewarding. At each time point t:

$$p(r_t|f_t^*) \neq p(r_t|\neg f_t^*). \tag{4.1}$$

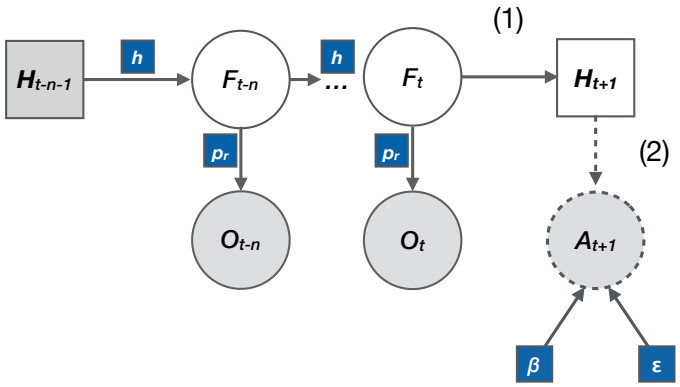
When $p(r_t|f_t^*) > p(r_t|\neg f_t^*)$, the presence of f^* signals a higher probability of future reward.

While in the Dimensions Task changes in the target feature are signaled, here we consider a more general setting in which the identity of the target feature changes according to:

$$\begin{aligned} p(f_{t+1}^*|f_t^*) &\sim U\{1, K\} \text{ with probability } h \\ p(f_{t+1}^*|f_t^*) &= f_t^* \text{ with probability } 1-h \end{aligned} \tag{4.2}$$

That is, there is a fixed h probability (hazard rate) that f^* will change in the future, and a $(1-h)$ probability that f^* will stay the same.

Under the generative model defined by Equations 4.1 and 4.2, we can model participants as sequentially approximating the belief state $p(f^*)$ using particle filtering. Instead of maintaining and updating the full posterior distribution over f^* (Niv et al., 2015), we assume that they keep track of a single particle H, which represents their hypothesis about the identity of the target feature. Participants have access to observations of the form $O_t = \{C_t, R_t\}$



(1) **Particle update**
$H_{t+1} \sim p(F_t \mid O_{t-n:t}, H_{t-n-1} = k)$

(2) **Attention model**
$A_{t+1} \sim Dirichlet(Softmax(\mathbb{1}[H_{t+1} = k], \beta), \varepsilon)$

Figure 4.1: Graphical model for the particle filter. Circular nodes in white represent the latent target feature. Circular nodes in gray represent observations. The square node in gray represents the previous hypothesis. The square node in white represents the updated hypothesis. Nodes in blue denote fixed parameters of the model. Solid and dotted lines distinguish between observations experienced by the participant during the task (i.e. choices and rewards), and data that the experimenter gets to observe when fitting the model (e.g. looking times).

– the choice and reward experienced on each trial, which they can use to update their hypothesis about the identity of f^*.

Particle filters learn by sampling new particles from a proposal distribution, and discarding particles that are inconsistent with new evidence. Over time, particles settle on hypotheses that, in ensemble, approximate the true posterior. Previous work has shown that individual behavior in classic associative learning experiments supports single-particle models (Daw & Courville, 2008). But a single particle is impoverished in that it does not maintain enough history to learn efficiently in our task. We thus consider a memory-augmented particle filter in which the proposal distribution at each time point is given by the probabil-

ity of the current target feature being k, conditional on the n most recent observations, and the hypothesis $n-1$ trials back (Figure 4.1):

$$H_{t+1} \sim p(F_t | O_{t-n:t}, H_{t-n-1} = k) \tag{4.3}$$

The intuition behind using this proposal distribution is that a participant who has been testing the hypothesis that k is the target feature will switch to a different hypothesis based on current observation O_t, as well as observations $O_{t-n:t-1}$ stored in working memory. The proposal distribution can also be viewed as the statistically optimal answer to the question, "if k was the target feature $t-n-1$ trials ago, what is the most likely target feature now, given the outcomes I just observed?"

We now turn to how the proposal distribution in Equation 4.3 can be computed. Since the dynamics of the target feature correspond to the latent state in a Hidden Markov Model, we can use the Forward algorithm (Ghahramani, 2001). For every state k, the forward probability $\alpha_t(k)$ is given by:

$$\alpha_t(k) = p(O_t | F_t = k) \sum_j p(F_t = k | F_{t-1} = j) \alpha_{t-1}(j) \tag{4.4}$$

We can understand this recursion as follows: the sum expresses the prior probability of the target feature being k (the total number of possible k's is the product of the number of dimensions d and the number of features per dimension f). To calculate the prior probability, we enumerate over all the possible target features in the previous trial, j, from which we could have transitioned to k on this current trial with probability $p(F_t = k | F_{t-1} = j)$. This is weighted by $\alpha_{t-1}(j)$, the posterior probability that the target feature was indeed j in the

previous trial. The transition probability depends on h, a free parameter that governs the participant's fixed belief about the rate of change in the target feature (Equation 4.2).

The prior is multiplied by $p(O_t|F_t = k)$, the likelihood of the observation given that k is the target. This depends on p_r, a free parameter that governs the participant's fixed belief about how likely rewards are given the presence or absence or the target feature (Equation 4.1). Equation 4.4 therefore defines a recursion over the posterior probability of each target feature, $\alpha_t(k)$. On each trial, the recursion runs over n time steps back (or fewer if there were fewer timesteps so far) and ends with a normalization step to yield the desired conditional probability:

$$p(F_t|O_{t-n:t}, H_{t-n-1} = k) = \frac{\alpha_t(k)}{\Sigma_j \alpha_t(j)} \tag{4.5}$$

Finally, the free parameter n can be interpreted as the working memory capacity of the model. At each time point, the model computes the switching rule given the current observation, and the $n - 1$ most recent observations from memory.

4.3 Fitting the particle filter to gaze data

So far, I have described a model of approximate inference over task-relevant features from the point of view of a participant learning in multidimensional environments. But how might this inference process modulate the allocation of attention to different features? In Chapter 1, I suggested that selective attention carves out state representations guided by the structure of a particular task: that is, attention prioritizes features that are predictive of reward. The particle filter specified in the previous section provides a mechanism for inferring structure in a multidimensional environment in which one dimension is relevant. I next show how we can use trial-by-trial model fitting of the particle filter to ask, does such

inference indeed guide selective attention? Specifically, I test whether the (latent) dynamics of the hypothesis participants were considering at each time point are reflected in trial-by-trial looking times.

Trial-by-trial maximum likelihood estimation (MLE) requires computing the likelihood of a data sequence $\mathcal{D}_{1:T}$ under a set of parameters θ (Daw, 2011). While MLE is standard practice in the human reinforcement learning literature (Wilson & Collins, 2019), evaluating the likelihood of data under models with stochastic latent states is typically intractable because the state space of possible trajectories grows exponentially with the number of trials (c.f. Findling, Skvortsova, Dromnelle, Palminteri, and Wyart (2019); van Opheusden, Acerbi, and Ma (2020)). Since we cannot directly observe what hypothesis the participant was considering, we need to marginalize out our uncertainty about $H_{1:T}$. This computation is quite costly, because it requires summing over all possible hidden state values at all times, yielding K^T terms. However, if we factorize the join likelihood as follows,

$$p(D_{1:T}|\theta) = \sum_k p(D_{1:T}, H_T = k|\theta) = \sum_k \tilde{\alpha}_T(k) \tag{4.6}$$

we notice that the term inside the summation is the joint probability of being in state k at time T, and all observations up to that time point. Because the state space is discrete, we can again compute this joint probability efficiently using the recursion given by the Forward algorithm for inference in hidden Markov models:

$$\tilde{\alpha}_T(k) = p(D_T|H_T = k) \sum_j p(H_T = k|H_{T-1} = j)\tilde{\alpha}_{T-1}(j) \tag{4.7}$$

Here, we keep the α notation standard for forward probabilities that we used in Equa-

tion 4.4. But note that in the previous section, α denoted the forward probability from the point of view of the participant, necessary for computing the proposal distribution (Equation 4.3). The recursion ran for n trials, where n is the fixed working memory capacity (Equation 4.4). Here, $\tilde{\alpha}$ denotes the forward probability from the point of view of the experimenter, and runs over all T trials of the task.

Let us unpack the terms of this second recursion (compare to 4.4). The sum expresses the prior probability of the hypothesis being k (again, the total number of possible k's is the product of the number of dimensions d and the number of features per dimension f). To calculate the prior probability, we enumerate over all the possible hypotheses in the previous trial, j, from which the participant could have switched to k on this current trial with probability $p(F_t = k|F_{t-1} = j)$. This is weighted by $\tilde{\alpha}_{t-1}(j)$, the posterior probability that the hypothesis was indeed j in the previous trial. This transition term can be computed for each feature using the formula for the proposal distribution in Equation 4.3. Because this computation relies on the Forward recursion in Equation 4.4, obtaining the likelihood requires us to run two nested Forward algorithms: an "inner" one for calculating proposal distributions, and an "outer" for calculating the likelihood of observed data. Inferring the hypothesis here can be viewed as inference in non-homogenous-HMMs (nHMMs), in which the transition probabilities depend on recent rewards (Kour & Morris, 2019).

The prior is multiplied by $p(D_T|H_T = k)$, the likelihood of the data given that k is the hypothesis. Recall from Chapter 3 that the data we observe are summarized as relative looking times (Figure 4.3)*. So we need to additionally specify a linking function between the latent dynamics of the hypothesis and these trial-by-trial looking times. We use the Dirich-

*In the Conclusions chapter, we turn to the issue of jointly fitting gaze and choice data

let distribution, a generalization of the Beta distribution over N-dimensional compositional vectors (i.e. vectors of proportions that sum to 1) (Aitchison, 1982; Stojić, Orquin, Dayan, Dolan, & Speekenbrink, 2020):

$$p(D_t|H_t = k) = Dirichlet(Softmax(1^{[H_{t+1}=k]}, \beta), \varepsilon) \tag{4.8}$$

The linking function above uses a softmax to determine how much of probability mass to place on the current hypothesis. It can be thought of as capturing directed exploration: the lower β is, the more attention is directed to features other than the current hypothesis (Guo & Brunskill, 2019). In addition, the fixed precision ε models random probability of exploring other hypotheses.

In sum, Equation 4.8 formalizes the idea that hypotheses about task structure guide attention: attention is mostly focused on the current hypothesis, and sometimes explores alternative hypotheses. The free parameters of the particle filter model are $\theta \in \{n, h, p_{high}, p_{low}, \beta_{hypothesis}, \varepsilon\}$: the working memory capacity, the prior probability of the change in the target feature, a probability of reward given the target is present, a probability of reward given the target is absent, a softmax temperature and a precision.

4.4 Alternative model: feature reinforcement learning with decay

In Chapter 2, I introduced Feature Reinforcement Learning with decay (fRL+decay) as a candidate mechanism for learning what to attend to in a multidimensional setting. As a brief reminder, fRL+decay, assumes the participant learns a feature weight W_f for each of the nine features. The predicted value for the chosen stimulus is the sum of its feature

weights. After each observation, the weights of the chosen features are updated based on a prediction error – the difference between the obtained reward and the predicted value (multiplied by learning rate η). Weights of unchosen features decay toward zero in proportion to decay rate d. This results in a set of updated weights W_{t+1}.

In Chapter 3, I modeled *dimensional attention* as a softmax over maximum feature weights in each dimension. Here I model *feature attention* as a softmax over the vector of updated feature weights W_{t+1}, where the inverse temperature β dictates the degree to which attention is focused attention on features with higher values:

$$\varphi_t(k) = \frac{e^{\beta W_t(k)}}{\sum_i e^{\beta W_t(i)}}. \tag{4.9}$$

As with the particle filter, the output from the softmax determines the parameters of a Dirichlet distribution that translates the feature weights to a prediction of eye gaze for each feature:

$$p(D_t|W_t) = Dirichlet(\varphi_t, \varepsilon) \tag{4.10}$$

The softmax temperature β dictates how likely it is that attention is directed to features that accrue a high value, and the precision ε models random noise. So the free parameters of this model are $\theta \in \{\eta, d, \beta, \varepsilon\}$: a learning rate, a decay rate, a softmax temperature and a precision. The total likelihood of the data is simply the product of the trial-by-trial probabilities (Daw, 2011) (see also Chapters 2 and 3):

$$p(D_{1:T}|\theta) = \prod_{t=1}^{T} p(D_t|W_t, \theta) \tag{4.11}$$

The fRL+decay model is an exemplar in a wide class of "connectionist" (artificial neural-network) architectures that have proposed 'attention learning' happens via trial and error (Cohen, Dunbar, & McClelland, 1990; Kruschke, 1992; Roelfsema & Ooyen, 2005). These include modern approaches based on deep reinforcement learning, which, notably, are more efficient when basic mechanisms of attentional selection are introduced (Vaswani et al., 2017). But while the fRL+decay model has mechanisms for "sparsifying" the state representation, it still assumes a gradual learning process very different from stepping dynamics exhibited by the particle filter (Zoltowski, Latimer, Yates, Huk, & Pillow, 2019). Thus, the comparison between the two models is informative with respect to a fundamental question about attentional dynamics: does attention arise from fast-switching serial hypothesis testing? Or are attentional biases acquired through gradual, error-correcting learning that starts from a wide attention distribution and narrows over time?

4.5 Fitting procedure

Both models were fit to gaze data by minimizing the negative log of the likelihood function (particle filter: Equation 4.6; fLR+decay: Equation 4.11). I used the minimize function (L-BFGS-B algorithm) in the SciPy Python package. As in Chapter 3, I take a leave-one-game-out cross validation approach: starting from different initial conditions, the model is fit to all games but one, and tested on the left-out game. The log likelihoods of all left-out games are then summed to get the cross-validated log likelihood. To investigate the robustness of fit results, this procedure was repeated 5 times using different initial conditions. Because it provides a full generative model of how gaze data are generated, the fitting procedure presented here improves upon the regression-based approach I presented

in Chapter 3.

4.6 Dataset: The Dimensions Task with Faces-Landmarks-Tools, v2.0

I tested the particle filter on gaze data from the a multidimensional learning study similar to the one presented in Chapter 3. Human participants were tasked with learning from trial and error which of nine features was most predictive of reward (Figure 4.2). The design of the task was similar, with a few modifications: participants could respond within 2 seconds, and their choice was indicated by a grey rectangle that stayed on until the 2-second mark; the outcome was presented for 2 seconds and consisted of a colored rectangle around the chosen column (green for a rewarded trial, and red for a non-rewarded trial); the unchosen features remained on the screen for the duration of the outcome presentation, allowing participants to attend to both chosen and unchosen features; and finally, in this dataset we only collected eye-tracking data as a measure of attention.

Aside from these modifications, as before, on each trial, participants selected one of 3 columns, each including a face, a landmark, and a tool. Choosing the column containing the target feature yielded a reward with 0.75 probability. Choosing any of the other two columns was rewarded with only 0.25 probability. All features were visible on every trial, with feature combinations within columns determined randomly on each trial. Each block of 20 trials was defined as a 'game' during which the target feature stayed constant. The target feature randomly changed between games, and this was announced to participants. Participants were instructed about the reward contingencies before the experiment, but were not instructed for each game regarding the relevant dimension or the target feature.

In Chapter 3, I demonstrated the viability of using eye-tracking to measure trial-by-trial

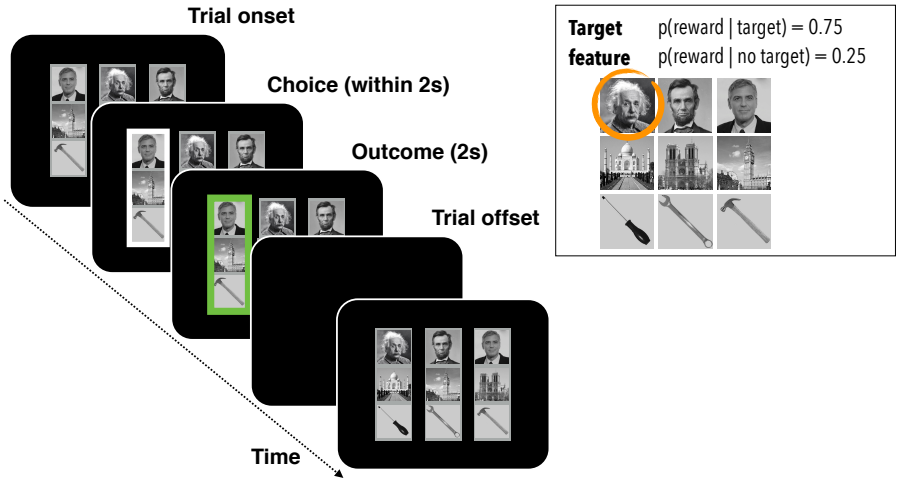


Figure 4.2: The Dimensions Task with Faces-Landmarks-Tools, v2.0. Left: As in Chapter 3, participants had to learn from trial and error which of 9 features was predictive of most reward. In this version, the outcome consisted of a rectangle around the chosen column stimulus, and was presented for 2 seconds. Unchosen columns remained on screen during outcome presentation.

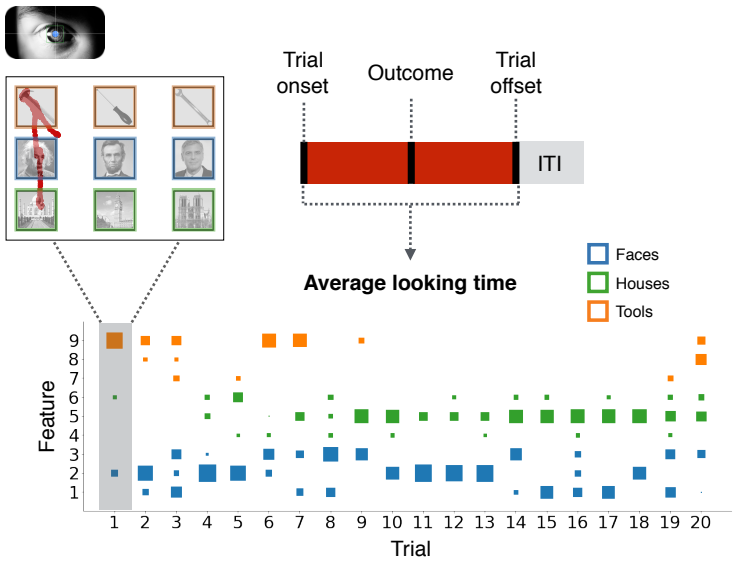


Figure 4.3: Sample attention dynamics for one game of the task Inset shows the looking patterns corresponding to the first trial of the game. Within each trial, looking time was binned and averaged across the spatially-resolved features. The size of the squares indicates the relative looking time spent looking at each feature, averaged over the whole trial.

changes in attention to different dimensions of the task (Faces, Landmarks or Tools). Here I extend this approach by employing high-frequency eye-tracking to derive a trial-by-trial measure of feature-level attention (Figure 4.3).

4.7 Results

I first investigated whether the particle filter can reliably learn the task as quickly as humans do (Figure 4.4). I simulated the choice behavior of particle filter agents with different memory capacities on the same stimulus sequence as human participants were exposed to. For illustration, I fixed the hazard rate h at 0.001, to compensate for the assumption that h does not exactly match the generative dynamics of the task (i.e there are no unsignaled changes in the target feature). I used a probability of reward $p_r = 0.99$, which means agents treat the likelihood function as more deterministic than in the true model of the task (where $p_r = 0.75$). This was done to mitigate effects of stochasticity that might arise after learning, when the agent, unlike humans, does not know to stop testing hypotheses. In practice, p_r can be fit to individual participants' data. Particle filter agents were randomly initialized to one of the 9 possible features, and made choices using a softmax choice rule with an inverse temperature of 100 (i.e. the model almost always chooses the stimulus containing the current hypothesis).

I found that the performance of the model improved with memory capacity, and approached human performance for the 5-back condition both in terms of speed of learning (Figure 4.4B) and accuracy on the last 6 trials of a game (Figure 4.4D).

Next, I compared the performance of the particle filter and fRL+decay models in predicting trial-by-trial fluctuations in relative looking times (Figure 4.5). I found that the par-

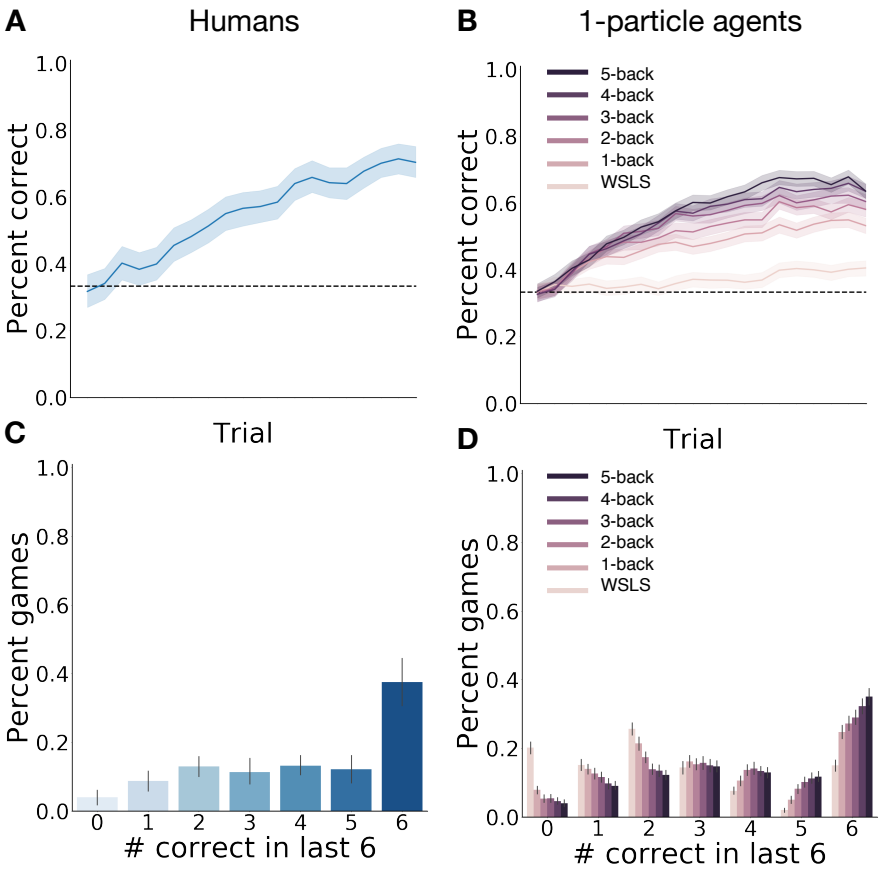


Figure 4.4: Particle filter task performance. **A**: Learning curves for 21 human participants. Performance was assessed as the proportion of trials in which the participant chose the stimulus containing the target feature. Shading: SEM. **B**: Learning curves for particle filter agents with different memory capacities experiencing the same stimulus sequence as human participants (100 agents per value of n). 0-back agents are equivalent to a win-stay-lose-shift (WSLS) strategy. Shading: SEM. **C**: Histogram of the average number of correct choices participants made in the last 6 trials of each game. Games in which they made the correct choice in 6 of the last 6 trials can be considered "learned". **D**: Average number of correct choices in the last 6 trials of a game by the particle filter model, as a function of memory capacity.

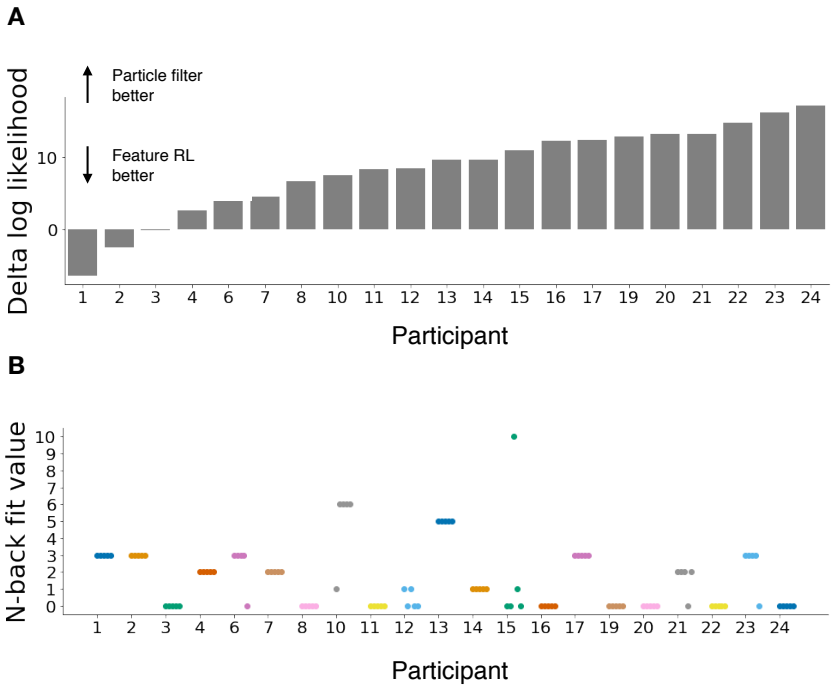


Figure 4.5: Model comparison A: Per participant difference in log-likelihood between particle filter model and feature reinforcement learning with decay model. **B**: Fit value of the N-back working memory capacity for each participant, after each iteration of the model fit. Each color represents one participant. Each dot corresponds to one of 5 iterations.

ticle filter outperformed fRL+decay for all but three participants, suggesting that shifts in attention were more consistent with hypothesis testing based on evidence from recent trials than with gradual error-driven learning (Figure 4.5A). I also found significant variability in the estimated memory capacity of each participant, with 11 out of 21 participants being better fit by particle filters with $n > 0$ (Figure 4.5B). One reason for which many participants seem to be fit by particle filters with no memory (i.e. $n = 0$) is that n determines the amount of recent evidence that drives hypotheses switches. While so far I have treated it as fixed, in principle n could vary depending on the stage of the learning process. For instance, n might be lower early on, when the participant is testing various hypotheses, than late in the game when learning is complete; alternatively, n could change depending on whether a participant was rewarded or not in the recent past (Bonawitz et al., 2014). While further work is necessary to fully investigate these variants, as well as potential interactions between n and other parameters of the particle filter, these results are suggestive of a role for working memory in storing recent experiences relevant for guiding selective attention.

Finally, I tested a key empirical prediction of the particle filter model. Recall that the relative looking times to each of the 9 features yield a trial-by-trial index of a participant's focus of attention. As such, descriptive statistics of these time series can offer clues as to the underlying dynamics of attention learning. In particular, we would expect that if participants switch attention in accordance to the mechanics of a one-particle particle filter, the relative looking time for the feature with the highest weight should be 1 or close to 1, indicating (almost-)full attention to one feature. A secondary prediction can be made about the difference between the maximum feature weight and that of the second highest feature weight: if the participant is switching hypotheses from one feature to another, then we

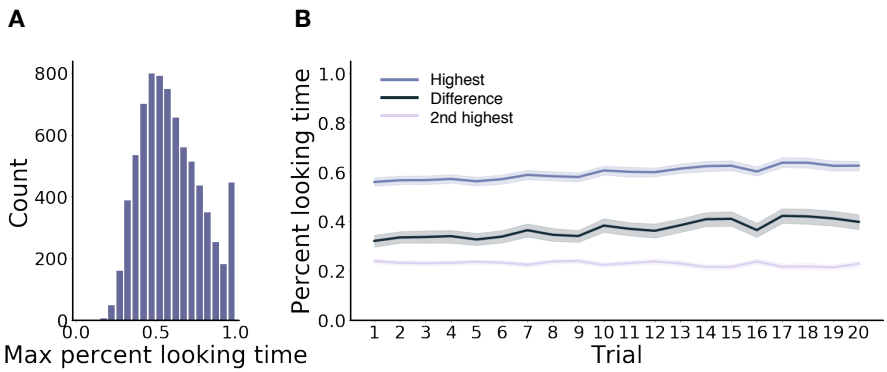


Figure 4.6: Attention is focused, its dynamics consistent with abrupt switching A: Histogram of the maximum relative looking time to each of the 9 features across all participants. **B**: Difference between the highest and second highest relative looking time to each of the 9 features as a function of trial within a game. The difference is relatively constant across a game, suggesting hypothesis testing rather than gradual learning.

would expect the difference in relative looking times to stay relatively constant throughout a game. Both of these predictions were borne out in the empirical data (Figure 4.6).

In sum, the results in this chapter provide evidence that approximate inference over task-relevant features guides selective attention during trial and error learning, and point to a new mechanism by which memory of recent experiences informs this inference.

Everything is going to be fine in the end. If it's not fine it's not the end.

Oscar Wilde

5

Contributions and future directions

I started this thesis by defining selective attention as a function that maps from perceptual observations to state representations in reinforcement learning. I suggested that such a mapping might be learned by inferring the structure of the environment: features of the observation space should be represented to the extent that they are useful for predicting reward. This first contribution is thus a conceptual one: framing selective attention as carving state representations in service of decision-making joins in a tradition of research that asks not *how* selective attention is deployed, but *why* it might be directed to some features and not others (McCallum, 1997; Dayan et al., 2000; Najemnik & Geisler, 2005; Gottlieb & Oudeyer, 2018; Callaway & Griffiths, 2019).

In Chapter 2, I provided empirical evidence for the role of attention in state representation learning. I studied younger and older adults' behavior in the 'Dimensions Task', a multidimensional learning environment in which one dimension was relevant for predicting reward. I showed that a reinforcement learning model which includes an indirect mechanism for selective attention is more consistent with choice data than a model which assumes uniform attention. Moreover, older participants' selective attention strategies differed from younger adults in a manner that was predictive of performance.

In Chapter 3, I described a series of methodological contributions that enabled us to more directly study how attention carves state representations. Using fMRI and eye-tracking, we obtained a trial-by-trial index of selective attention in a modified version of the 'Dimensions Task.' We used this index in conjunction with reinforcement learning models of choice to precisely pinpoint a role for attention in biasing both the valuation of multidimensional stimuli, and in biasing learning about different features. This technical contribution made it possible to begin asking questions about the other side of the interaction between learning and attention: given past experience, how is attention determined in the first place? I developed a preliminary model comparison for predicting gaze and BOLD attention data, which revealed that selective attention is sensitive to recent outcomes, controlling for choice (Krajbich et al., 2010; Krajbich, Lu, Camerer, & Rangel, 2012). Taken together, these results provide evidence for a bidirectional interaction between selective attention and reinforcement learning: attention constrains how we choose and what we learn about, and learning in turn directs our attention to task-relevant features. On the technical side, this study adds to a growing literature on using additional data modalities in conjunction with choice to disentangle different decision-making strategies.

Finally, Chapter 4 describes the main theoretical contribution of this thesis. I suggested in Chapter 1 that attention might be guided by inference over task-relevant features. Chapter 4 makes this hypothesis explicit and provides an algorithmic solution, in the form of a computational model grounded in principles of statistical inference. I propose that latent attention dynamics in the 'Dimensions Task' are consistent with particle filtering. This algorithm keeps track of one particle (or hypothesis) about which feature within the relevant dimension is most predictive of reward. On each trial, the particle filter computes a

switching rule (i.e. a proposal distribution for new hypotheses) based on recent evidence. I fit the particle filter model to a trial-by-trial index of attention measured from gaze (as shown in Chapter 3). The results of this analysis showed that attention dynamics reflected in trial-by-trial looking times are more consistent with particle filtering than with the feature reinforcement learning model tested in Chapter 3. Specifically, the one-particle particle filter captures rapid switching and focused attention, two prominent qualitative features of the attention data. Critically, augmenting the particle filter with memory enables it to compensate for only keeping track of one hypothesis, and to learn as quickly as human participants do. These results suggest a tight link between selective attention and working memory (van Ede, Chekroud, & Nobre, 2019; Panichello & Buschman, 2020), and show how this link might support state representation learning (McCallum, 1997). The findings in this chapter also open up the intriguing possibility that previously proposed optimal strategies for attention allocation could be realized, but over approximate representations of learned beliefs (Najemnik & Geisler, 2005; Nelson & Cottrell, 2007; Braunlich & Love, 2018; Callaway & Griffiths, 2019). On a final technical note, while stochastic decision-making models with continuous latent variables are notoriously difficult to fit to trial-by-trial data (Findling et al., 2019), this work shows that, in discrete settings at least, particle filters are tractable enough to fit using established likelihood-maximization approaches. This insight opens the door for testing and refining a larger class of sampling-based models which have been proposed as computational models of memory (Gershman & Daw, 2017; Bornstein, Khaw, Shohamy, & Daw, 2017).

5.1 Joint fitting of gaze and choice data

One theme throughout this dissertation has been the use of additional data modalities (e.g. eye-tracking, fMRI) to directly measure the dynamics of attention allocation. Candidate models of attention described in Chapters 3 and 4 sought to predict trial-by-trial fluctuations in gaze data, summarized as relative looking times either to features or dimensions. Parameters of both the fRL+decay model and particle filter model were optimized based on how well they could predict gaze. But this approach ignored the other source of data from the Dimensions Task: participants' choices. While not a systematic focus of investigation so far, this does mean best fit parameter values could differ depending on which data streams models are fit to.

In ongoing work, I am addressing this limitation by jointly fitting models to choice and gaze data. This amounts to maximizing a likelihood function that takes into account both trial-by-trial choices and relative looking times. The best-fit parameters obtained by this procedure are optimized simultaneously for both choice and gaze prediction. A similar approach has been applied before to integrate behavioral and fMRI data in the same computational modeling framework (Turner et al., 2013).

Recall Equation 4.7, which provided a recursive algorithm for exactly computing the likelihood of the data under the particle filter. Joint fitting means considering both the likelihood of gaze φ_T and choice c_T:

$$\tilde{\alpha}_T(k) = p(\varphi_T, c_T | H_t = k) \sum_j \tilde{\alpha}_{T-1}(j) p(H_T = k | H_{T-1} = j) \tag{5.1}$$

We can further factor the likelihood term in the product, since φ_T and c_T are conditionally

independent given a particular hypothesis:

$$\tilde{\alpha}_T(k) = p(\varphi_T|H_T = k)p(c_T|H_T = k)\sum_j \tilde{\alpha}_{T-1}(j)p(H_T = k|H_{T-1} = j) \tag{5.2}$$

In other words, on each timestep of the algorithm, evidence in favor of a given hypothesis is a product of the likelihood of the choice data and looking time data.

The joint probability of relative looking times φ and choices c under a fRL+decay model can be written as:

$$p(\varphi_{1:T}, c_{1:T}|\theta) = \prod_{i=1}^{T} p(c_t, \varphi_t|\theta) \tag{5.3}$$

By conditional independence given θ, we have:

$$\prod_{i=1}^{T} p(c_t|\theta)p(\varphi_t|\theta) = \prod_{i=1}^{T} p(c_t|\theta)\prod_{i=1}^{T} p(\varphi_t|\theta) \tag{5.4}$$

Taking the log gives:

$$log(\prod_{i=1}^{T} p(c_t|\theta)) + log(\prod_{i=1}^{T} p(\varphi_t|\theta)) = log\sum_{i=1}^{T} p(c_t|\theta) + log\sum_{i=1}^{T} p(\varphi_t|\theta) \tag{5.5}$$

The expression above is just a sum of the log-likelihoods of each data series, computed independently.

Using the analytical expressions for the joint likelihood derived in this chapter, future work will seek to integrate looking time and choice data into a single computational framework.

5.2 A neural circuit model of representation learning

5.2.1 A cognitive neuroscience view of particle filtering

In this thesis, I have presented an algorithmic-level model of representation learning based on sequential approximate inference. In the following section, I discuss how such a model may be implemented in a neural architecture grounded in what we know about the human brain (Figure 5.1). I am indebted to Ian Ballard for the scientific exchange that shaped the ideas in this section (Radulescu, Niv, & Ballard, 2019).

Consider a particle that represents a compositional hypothesis about which features are relevant (e.g "red square"). In the cognitive neuroscience literature, this kind of representation is known as a "rule" (Goodman et al., 2008; Badre, Kayser, & D'Esposito, 2010). Connectionist models of the basal ganglia–prefrontal cortex circuitry (Collins & Frank, 2013; Cohen, Braver, & Brown, 2002; O'Reilly & Frank, 2006) describe how rules stored in working memory can be selected via known corticostriatal circuitry (see O'Reilly and Frank (2006) for anatomical detail). Anterolateral prefrontal cortical pools can represent different rules. These different pools compete via mutual lateral inhibition. The outcome of this competition is biased by the relative strength of each pools' connectivity with the striatum. Pools with stronger cortical–striatal connectivity will generate a stronger striatal response, which in turn increases the strength of thalamic feedback onto these pools. This recurrent circuit allows a pool representing a task rule to inhibit competing pools and control behavior (Alexander, DeLong, & Strick, 1986). If an unexpected reward occurs, dopamine release in the striatum strengthens the synapses from the most active cortical pool to the striatum (Wickens & Kötter, 1995; Wickens, Reynolds, & Hyland, 2003). In

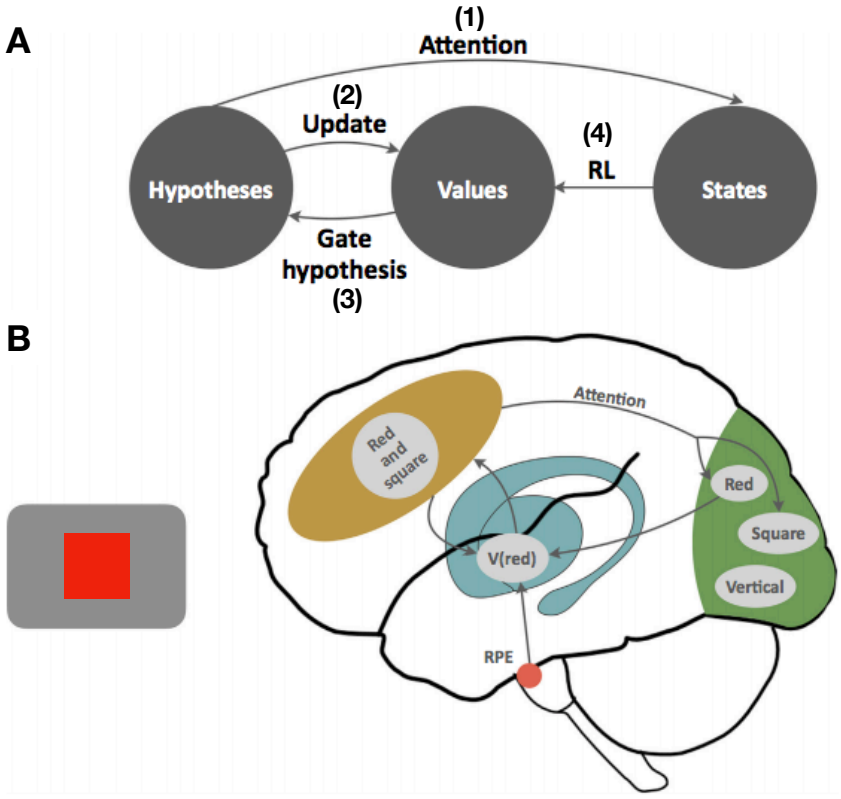


Figure 5.1: A neural circuit model of representation learning. A: I propose that: (i) hypotheses about task structure are the source of top-down attention (Arrow 1); and (ii) attention sculpts the state representation by prioritizing some features of the environment for reinforcement learning (Arrows 2 and 4). In turn, in accordance with previous models of prefrontal-striatal interactions, learned values of states participate in the gating of which hypotheses are considered (Arrow 3). **B**: A simple model showing how the interacting systems architecture in (A) could be realized in interacting neural circuits. Yellow area corresponds to the lateral prefrontal cortex, blue to the basal ganglia, green to the sensory cortex, and red to the dopaminergic midbrain. A prefrontal hypothesis that 'red and square' leads to reward biases top-down attention to the 'red' and 'square' features in the sensory cortex, which in turn biases the state representation in the striatum towards these features both during valuation (posterior corticostriatal synapses) and updating (prefrontal corticostriatal synapses). In turn, values stored in the striatum influence prefrontal rule selection.

this way, rule representations that lead to reward are more likely to win over alternative representations in the future (Collins & Frank, 2013; Alexander et al., 1986). This model describes how a reinforcement learning system could gate the representation of a hypothesis about task structure into the cortex. This hypothesis can then guide "top-down" selective attention during learning (Radulescu, Niv, & Ballard, 2019) (Figure 1.1).

Hypotheses about task structure can constrain feature-based reinforcement learning by directing attention to specific component features and not others. For example, a hypothesis that 'red stimuli predict reward' would increase the strength and fidelity of the representation of color in the sensory cortex. If an unexpected outcome follows a red square, the heightened representation of 'red' will cause a larger update to posterior corticostriatal projections from 'red' neurons than from 'square' neurons. As a result, reinforcement learning will operate over a feature-based representation with biased attention to the color 'red'. If later in the task the hypothesis is updated to 'red squares predict reward', both red and square features will be attended to more strongly than other features. In this way, hypotheses can sculpt the state representation underlying reinforcement learning.

Reinforcement learning could, in turn, contribute to the selection of hypotheses via two mechanisms. First, learning can adjust the frontal corticostriatal weights of "top-down" projections from cortical pools representing alternative rules, as has been proposed in a recent neural network model (Collins & Frank, 2013). Second, reinforcement learning over features represented in the sensory cortex can contribute "bottom-up" to rule selection. For instance, the rule 'red squares predict reward' will be influenced by simple reinforcement learning linking 'red' with reward and 'square' with 'reward. This way, when a rule is discarded, reinforcement learning would support the selection of an alternative rule based on

a common feature that has predicted past observations. This mechanism alleviates the need to exactly remember previous trials and evaluate alternative rules against these memories, thereby endowing the hypothesis testing system with implicit memory. In this chapter, I instantiated and tested a model that only has explicit memory of a fixed number of different episodes, but hybrid implicit-explicit mechanisms are also possible (Shohamy & Daw, 2015), for instance by computing the proposal distribution based on a mixture of memory-based inference and reinforcement learning (Song, Cai, & Niv, 2019). By designating hypotheses as the source of top-down attention, this model provides a mechanistic account of how reinforcement learning is influenced by both structured knowledge and attention. This idea is closely related to recent work suggesting that working memory contents in lateral prefrontal circuits act as the source of top-down attention to the constituent sensory circuits (Kiyonaga & Egner, 2013). Working memory plays an important role in constraining reinforcement, and the model I present here predicts that learning is influenced both by the number of hypotheses that can be simultaneously considered (Lloyd, Sanborn, Leslie, & Lewandowsky, 2017), and the number of episodes that can be remembered at any given time (Todd, Niv, & Cohen, 2009).

5.2.2 Neurally plausibility of particle filtering

A strength of particle filters is that they can approximate any given Bayesian model by using a finite number of particles, each of which expresses a particular hypothesis about the state of the world. For example, in the Dimensions Task, each particle represents a single hypothesis about which feature is relevant. After an observation, the particle either samples a new hypothesis or stays with the current one. This decision depends on how likely the

observation is under the current hypothesis. A particle encoding the belief that 'red leads to reward' would be more likely to stay with its current hypothesis after observing a red square followed by a reward and more likely to switch to a new hypothesis after observing a red square that followed by no reward. This update algorithm is computationally simple because it incorporates only each particle's belief about the world.

In addition to their computational simplicity, particle filters are an appealing model for representation learning because they can include a prior that preferentially samples simpler rules. Moreover, they capture the phenomenological report that people consider alternative hypotheses (Armstrong, Gleitman, & Gleitman, 1983). Particle filters also closely resemble a serial hypothesis testing model that has previously been shown to describe human behavior in the Dimensions Task well (Wilson & Niv, 2012). In addition, they provide a single framework for implementing representation learning over different types of models, including both probabilistic programming models and Bayesian nonparametric models (Lloyd et al., 2017; Sanborn et al., 2010).

Given an infinite number of particles, particle filters converge to the true posterior probability. Remarkably, recent work has demonstrated that the use of a single or very few particles can describe human behavior well. This is because humans face a practical problem: rather than learning the true probability distribution over all possible rules, people need only find a rule that explains enough observations to make good decisions (Lieder, Griffiths, & Hsu, 2018). This could explain why behavior across an entire group may be Bayes optimal but individual choices are often not (Courville & Daw, 2008). If each individual tracks just one or a few hypotheses, only the group behavior will aggregate over enough 'particles' to appear Bayes optimal (Courville & Daw, 2008).

Our neural model proposes that hypotheses are gated by corticostriatal circuitry that is, in turn, influenced by reinforcement learning. This architecture could form the basis of a particle filter algorithm. Specifically, particle filters sample hypotheses based on how well each hypothesis accounts for previous observations. Feature weights learned via reinforcement learning could enable the sampling of hypotheses that have already explained some observations. Unlike particle filter accounts of sensory integration, which propose that individual spikes of feature-selective neurons represent particles (Huang & Rao, 2016; Kutschireiter, Surace, Sprekeler, & Pfister, 2017; Legenstein & Maass, 2014; T. S. Lee & Mumford, 2003), in our model particles correspond to distributed prefrontal representations of rules. The particle filter algorithm is a flexible mechanism for inference that could apply to different timescales (from milliseconds to trials) and different types of problems (e.g., perception, categorization).

Although corticostriatal connectionist models can exhibit properties similar to a Bayesian structure-learning model (Collins & Frank, 2013), the corticostriatal gating mechanism need not perfectly implement a particle filter, and the differences may be informative. For example, in a task where the motor response mapping varies (e.g., reward sometimes follows the left-hand and sometimes the right-hand response), a corticostriatal gating model would correctly predict that if recent right hand selections were rewarded, the subject is more likely to respond 'right' regardless of the category of the current stimulus (Lau & Glimcher, 2005). A particle filter implementing a probabilistic programming model of representation learning would not predict this effect. A fruitful area for future research will be to examine other ways in which constraints imposed by the corticostriatal architecture can predict deviations from Bayesian inference.

5.3 Naturalistic feature spaces for state inference

In closing, a note on a conjecture I made early on in the thesis: the observation space is only constrained by the "primitive" features that an agent's perceptual system has access to during a reward learning task (Chapter 1). In the Dimensions Task, that space is conveniently defined by discrete stimuli and clear category structure. This design choice enables precise model-building and testing. Going forward, assessing the validity of candidate models in real-life scenarios challenges us to engage with decades-old questions in the psychology of perception: how are real-world objects represented (Biederman, 1987; Spelke, 1990)? What dimensions define them, and to what extent are they separable (Shepard, 1991)? How are object features integrated into a coherent percept (Treisman & Gelade, 1980)? What are the semantic associations that underlie real-world scene understanding (Greene & Oliva, 2009)? What symbolic representations rely on such associations (Goodman et al., 2008; Ballard et al., 2018)? In future work, I hope to consider these questions in naturalistic scenarios, in which humans can use their full prior knowledge of the world to inform their decisions (Radulescu, van Opheusden, Callaway, Griffiths, & Hillis, 2020).

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139–160.

Aitchison, J. (2003). A concise guide to compositional data analysis. In *Cda workshop, girona*.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual review of neuroscience*, *9*(1), 357–381.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, *98*(3), 409.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308.

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The spot-the-word test: A robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology*, *32*(1), 55–65.

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2), 315–326.

Ballard, I., Miller, E. M., Piantadosi, S. T., Goodman, N. D., & McClure, S. M. (2018). Beyond reward prediction errors: Human striatum updates rule values during learning. *Cerebral Cortex*, *28*(11), 3965–3975.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. *Models of information processing in the basal ganglia*, *215*.

Bellman, R. (1957). A markovian decision process. *Journal of mathematics and mechanics*, 679–684.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, *94*(2), 115.

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience*, *3*(1), 13–21.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, *74*, 35–65.

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*(1), 1–9.

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current opinion in neurobiology*, *22*(6), 956–962.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433–436.

Braunlich, K., & Love, B. C. (2018). Bidirectional influences of information-sampling and concept learning.

Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Biobehavioral Reviews*, *26*(7), 809–817.

Callaway, F., & Griffiths, T. (2019). Attention in value-based choice as optimal sequential sampling.

Campbell, K. L., Grady, C. L., Ng, C., & Hasher, L. (2012). Age differences in the frontoparietal cognitive control network: implications for distractibility. *Neuropsychologia*, *50*(9), 2212–2223.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the bis/bas scales. *Journal of personality and social psychology*, *67*(2), 319.

Chowdhury, R., Guitart-Masip, M., Lambert, C., Dayan, P., Huys, Q., Düzel, E., & Dolan, R. J. (2013). Dopamine restores reward prediction errors in old age. *Nature neuroscience*, *16*(5), 648.

Cohen, J. D., Braver, T. S., & Brown, J. W. (2002). Computational perspectives on dopamine function in prefrontal cortex. *Current opinion in neurobiology*, *12*(2), 223–229.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, *97*(3), 332.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological review*, *120*(1), 190.

Collins, A. G., & Frank, M. J. (2014). Opponent actor learning (opal): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological review*, *121*(3), 337.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., ... others (1998). A common network of functional areas for attention and eye movements. *Neuron*, *21*(4), 761–773.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, *3*(3), 201–215.

Courville, A. C., & Daw, N. (2008). The rat as particle filter. In *Advances in neural information processing systems* (pp. 369–376).

Daw, N. (2011). Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII*, *23*(1).

Daw, N., & Courville, A. (2008). The pigeon as particle filter. *Advances in neural information processing systems*, *20*, 369–376.

Daw, N., 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. (2012). How to set the switches on this thing. *Current opinion in neurobiology*, *22*(6), 1068–1074.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature neuroscience*, *3*(11), 1218–1223.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, *18*(2), 185–196.

Dennis, N., & Cabeza, R. (2012). Frontal lobes and aging: deterioration and compensation. *Principles of frontal lobe function*, *2*, 628–652.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*(1), 193–222.

Dias, R., Robbins, T., & Roberts, A. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. *Nature*, *380*(6569), 69–72.

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, *12*(656-704), 3.

Eppinger, B., Hämmerer, D., & Li, S.-C. (2011). Neuromodulation of reward-based learning and decision making in human aging. *Annals of the New York Academy of Sciences*, *1235*, 1.

Eppinger, B., Schuck, N. W., Nystrom, L. E., & Cohen, J. D. (2013). Reduced striatal responses to reward prediction errors in older compared with younger adults. *Journal of Neuroscience*, *33*(24), 9905–9912.

Esber, G. R., Roesch, M. R., Bali, S., Trageser, J., Bissonette, G. B., Puche, A. C., ... Schoenbaum, G. (2012). Attention-related pearce-kaye-hall signals in basolateral amygdala require the midbrain dopaminergic system. *Biological psychiatry*, *72*(12), 1012–1019.

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature neuroscience*, *22*(12), 2066–2077.

Frank, M. J. (2011). Computational models of motivated action selection in corticostriatal circuits. *Current opinion in neurobiology*, *21*(3), 381–386.

Franklin, N. T., & Frank, M. J. (2018). Compositional clustering in task structure learning. *PLoS computational biology*, *14*(4), e1006116.

Fristoe, N. M., Salthouse, T. A., & Woodard, J. L. (1997). Examination of age-related deficits on the wisconsin card sorting test. *Neuropsychology*, *11*(3), 428.

Gazzaley, A., Cooney, J. W., Rissman, J., & D'esposito, M. (2005). Top-down suppression deficit underlies working memory impairment in normal aging. *Nature neuroscience*, *8*(10), 1298–1300.

Geana, A., & Niv, Y. (2014). Causal model comparison shows that human representation learning is not bayesian. In *Cold spring harbor symposia on quantitative biology* (Vol. 79, pp. 161–168).

Gershman, S. J., & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology*, *68*, 101–128.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, *20*(2), 251–256.

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. In *Hidden markov models: applications in computer vision* (pp. 9–41). World Scientific.

Glass, B. D., Chotibut, T., Pacheco, J., Schnyer, D. M., & Maddox, W. T. (2012). Normal aging and the dissociable prototype learning systems. *Psychology and aging*, *27*(1), 120.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Gottlieb, J. (2012). Attention, learning, and the value of information. *Neuron*, *76*(2), 281–295.

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758–770.

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472.

Guo, Z. D., & Brunskill, E. (2019). Directed exploration for reinforcement learning. *arXiv preprint arXiv:1906.07805*.

Hampshire, A., Gruszka, A., Fallon, S. J., & Owen, A. M. (2008). Inefficiency in self-organized attentional switching in the normal aging population is associated with decreased activity in the ventrolateral prefrontal cortex. *Journal of Cognitive Neuroscience*, *20*(9), 1670–1686.

Hamrick, J. B. (2019). Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, *29*, 8–16.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). Pymvpa: a python toolbox for multivariate pattern analysis of fmri data. *Neuroinformatics*, *7*(1), 37–53.

Hare, T. A., Malmaud, J., & Rangel, A. (2011). Focusing attention on the health aspects of foods changes value signals in vmpfc and improves dietary choice. *Journal of Neuroscience*, *31*(30), 11077–11087.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *journal of Educational Statistics*, *6*(2), 107–128.

Hentschke, H., & Stüttgen, M. C. (2011). Computation of measures of effect size for neuroscience data sets. *European Journal of Neuroscience*, *34*(12), 1887–1894.

Holland, P. C., & Gallagher, M. (2006). Different roles for amygdala central nucleus and substantia innominata in the surprise-induced enhancement of learning. *Journal of Neuroscience*, *26*(14), 3791–3797.

Huang, Y., & Rao, R. P. (2016). Bayesian inference and online learning in poisson neuronal networks. *Neural computation*, *28*(8), 1503–1526.

Hyafil, A., Summerfield, C., & Koechlin, E. (2009). Two mechanisms for task switching in the prefrontal cortex. *Journal of Neuroscience*, *29*(16), 5135–5142.

James, W., Burkhardt, F., Bowers, F., & Skrupskelis, I. K. (1890). *The principles of psychology* (Vol. 1) (No. 2). Macmillan London.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural networks*, *15*(4-6), 535–547.

Jones, M., & Canas, F. (2010). Integrating reinforcement learning with models of representation learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 32).

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*(1-2), 99–134.

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic bulletin & review*, *20*(2), 228–242.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & memory*, *1*(2), 106–120.

Kour, G., & Morris, G. (2019). Estimating attentional set-shifting dynamics in varying contextual bandits. *BioRxiv*, 621300.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision research*, *35*(13), 1897–1916.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, *13*(10), 1292.

Krajbich, I., Lu, D., Camerer, C., & Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in psychology*, *3*, 193.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, *16*(1), 37–68.

Kutschireiter, A., Surace, S. C., Sprekeler, H., & Pfister, J.-P. (2017). Nonlinear bayesian filtering and learning: a neuronal dynamics for perception. *Scientific reports*, *7*(1), 1–13.

Lacouture, Y., & Cousineau, D. (2008). How to use matlab to fit the ex-gaussian and other probability functions to a distribution of response times. *Tutorials in quantitative methods for psychology*, *4*(1), 35–45.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Langdon, A. J., Sharpe, M. J., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions for dopamine. *Current Opinion in Neurobiology*, *49*, 1–7.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior*, *84*(3), 555–579.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual review of neuroscience*, *35*, 287–308.

Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448.

Legenstein, R., & Maass, W. (2014). Ensembles of spiking neurons with noise support optimal probabilistic inference in a dynamically changing environment. *PLoS computational biology*, *10*(10).

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*(2), 451–463.

Li, S.-C., Lindenberger, U., & Bäckman, L. (2010). Dopaminergic modulation of cognition across the life span. *Neuroscience & Biobehavioral Reviews*, *34*(5), 625–630.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, *125*(1), 1.

Lim, S.-L., O'Doherty, J. P., & Rangel, A. (2011). The decision value computations in the vmpfc and striatum use a relative value code that is guided by visual attention. *Journal of Neuroscience*, *31*(37), 13214–13223.

Lindenberger, U., & Mayr, U. (2014). Cognitive aging: is there a dark side to environmental support? *Trends in cognitive sciences*, *18*(1), 7–15.

Lloyd, K., Sanborn, A., Leslie, D., & Lewandowsky, S. (2017). Why does higher working memory capacity help you learn? In *Cogsci*.

Luce, R. D., et al. (1986). *Response times: Their role in inferring elementary mental organization* (No. 8). Oxford University Press on Demand.

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2012). Evaluating the td model of classical conditioning. *Learning & behavior*, *40*(3), 305–319.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological review*, *82*(4), 276.

Mansinghka, V., Shafto, P., Jonas, E., Petschulat, C., Gasner, M., & Tenenbaum, J. B. (2016). Crosscat: A fully bayesian nonparametric method for analyzing heterogeneous, high dimensional data. *The Journal of Machine Learning Research*, *17*(1), 4760–4808.

Marković, D., Gläscher, J., Bossaerts, P., O'Doherty, J., & Kiebel, S. J. (2015). Modeling the evolution of beliefs using an attentional focus mechanism. *PLoS computational biology*, *11*(10).

Mata, R., Josef, A. K., Samanez-Larkin, G. R., & Hertwig, R. (2011). Age differences in risky choice: a meta-analysis. *Annals of the New York Academy of Sciences*, *1235*, 18.

Mata, R., Pachur, T., Von Helversen, B., Hertwig, R., Rieskamp, J., & Schooler, L. (2012). Ecological rationality: a framework for understanding and aiding the aging decision maker. *Frontiers in Neuroscience*, *6*, 19.

Mayr, U., Spieler, D. H., & Hutcheon, T. G. (2015). When and why do old adults outsource control to the environment? *Psychology and aging*, *30*(3), 624.

McCallum, R. (1997). Reinforcement learning with selective perception and hidden state.

Mell, T., Heekeren, H. R., Marschner, A., Wartenburger, I., Villringer, A., & Reischies, F. M. (2005). Effect of aging on stimulus-reward association learning. *Neuropsychologia*, *43*(4), 554–563.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, *16*(5), 1936–1947.

Moore, T., & Fallah, M. (2001). Control of eye movements and spatial attention. *Proceedings of the National Academy of Sciences*, *98*(3), 1273–1276.

Moran, R. J., Symmonds, M., Dolan, R. J., & Friston, K. J. (2014). The brain ages optimally to model its environment: evidence from sensory learning over the adult lifespan. *PLoS computational biology*, *10*(1).

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Nelson, J. D., & Cottrell, G. W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, *70*(13-15), 2256–2272.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, *10*(9), 424–430.

Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., & Cohen, J. D. (2000). Working memory for letters, shapes, and locations: fmri evidence against stimulus-based regional organization in human prefrontal cortex. *Neuroimage*, *11*(5), 424–446.

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fmri evidence for objects as the units of attentional selection. *Nature*, *401*(6753), 584–587.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, *18*(2), 283–328.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97–98.

Panichello, M. F., & Buschman, T. J. (2020). Selective control of working memory in prefrontal, parietal, and visual cortex. *bioRxiv*.

Pavlov, I. P. (1927). Conditioned reflex. *Feldsher Akush*, 6–12.

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, *87*(6), 532.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, *10*(4), 437–442.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual review of neuroscience*, *35*, 73–89.

Plato. (ca 360BCE). *Plato's phaedrus*. Cambridge University Press.

Ponsen, M., Taylor, M. E., & Tuyls, K. (2009). Abstraction and generalization in reinforcement learning: A summary and framework. In *International workshop on adaptive and learning agents* (pp. 1–32).

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, *32*(1), 3–25.

Radulescu, A., Daniel, R., & Niv, Y. (2016). The effects of aging on the interaction between reinforcement learning and attention. *Psychology and aging*, *31*(7), 747.

Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic reinforcement learning: the role of structure and attention. *Trends in cognitive sciences*.

Radulescu, A., Niv, Y., & Daw, N. (2019). A particle filtering account of selective attention during learning. In *Computational cognitive neuroscience*.

Radulescu, A., van Opheusden, B., Callaway, F., Griffiths, T., & Hillis, J. (2020). From heuristic to optimal models in naturalistic visual search. In *Bridging ai and cognitive science @ iclr*.

Raven, J. C., et al. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford pyschologists Press.

Rescorla, R. A., Wagner, A. R., et al. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.

Rhodes, M. G. (2004). Age-related differences in performance on the wisconsin card sorting test: a meta-analytic review. *Psychology and aging*, *19*(3), 482.

Ridderinkhof, K. R., Span, M. M., & Van Der Molen, M. W. (2002). Perseverative behavior and adaptive control in older adults: Performance monitoring, rule induction, and set shifting. *Brain and cognition*, *49*(3), 382–401.

Roelfsema, P. R., & Ooyen, A. v. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural computation*, *17*(10), 2176–2214.

Salthouse, T. A. (1992). What do adult age differences in the digit symbol substitution test reflect? *Journal of Gerontology*, *47*(3), P121–P128.

Samanez-Larkin, G. R., & Knutson, B. (2014). Reward processing and risky decision making in the aging brain.

Sanborn, A., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences*, *20*(12), 883–893.

Sanborn, A., Chater, N., & Heller, K. A. (2009). Hierarchical learning of dimensional biases in human categorization. In *Advances in neural information processing systems* (pp. 727–735).

Sanborn, A., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144.

Schmitz, T. W., Cheng, F. H., & De Rosa, E. (2010). Failing to ignore: paradoxical neural effects of perceptual load on early attentional selection in normal aging. *Journal of Neuroscience*, *30*(44), 14750–14758.

Schuck, N. W., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P. A., Haynes, J.-D., & Reverberi, C. (2015). Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron*, *86*(1), 331–340.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

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.

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis.

Shohamy, D., & Daw, N. (2015). Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences*, *5*, 85–90.

Shohamy, D., & Wimmer, G. E. (2013). Dopamine and the cost of aging. *Nature neuroscience*, *16*(5), 519–521.

Smith, A. (2013). *Sequential monte carlo methods in practice*. Springer Science & Business Media.

Smith, D., Rorden, C., & Jackson, S. R. (2004). Exogenous orienting of attention depends upon the ability to execute eye movements. *Current Biology*, *14*(9), 792–795.

Song, M., Cai, M., & Niv, Y. (2019). Learning what is relevant for rewards via serial hypothesis testing. In *Computational cognitive neuroscience*.

Speekenbrink, M. (2016). A tutorial on particle filters. *Journal of Mathematical Psychology*, *73*, 140–152.

Spelke, E. S. (1990). Principles of object perception. *Cognitive science*, *14*(1), 29–56.

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature neuroscience*, *16*(7), 966.

Stojić, H., Orquin, J. L., Dayan, P., Dolan, R. J., & Speekenbrink, M. (2020). Uncertainty in learning, choice, and visual fixation. *Proceedings of the National Academy of Sciences*, *117*(6), 3291–3300.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, *3*(1), 9–44.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Todd, M. T., Niv, Y., & Cohen, J. D. (2009). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In *Advances in neural information processing systems* (pp. 1689–1696).

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206.

van Ede, F., Chekroud, S. R., & Nobre, A. C. (2019). Human gaze tracks attentional focusing in memorized visual space. *Nature human behaviour*, *3*(5), 462–470.

van Opheusden, B., Acerbi, L., & Ma, W. J. (2020). Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *arXiv preprint arXiv:2001.03985*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wickens, J., & Kötter, R. (1995). Cellular models of reinforcement.

Wickens, J., Reynolds, J. N., & Hyland, B. I. (2003). Neural mechanisms of reward-related motor learning. *Current opinion in neurobiology*, *13*(6), 685–690.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547.

Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in human neuroscience*, *5*, 189.

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–279.

Worthy, D. A., Gorlick, M. A., Pacheco, J. L., Schnyer, D. M., & Maddox, W. T. (2011). With age comes wisdom: Decision making in younger and older adults. *Psychological science*, *22*(11), 1375–1380.

Worthy, D. A., & Maddox, W. T. (2012). Age-based differences in strategy use in choice tasks. *Frontiers in neuroscience*, *5*, 145.

Wunderlich, K., Beierholm, U. R., Bossaerts, P., & O'Doherty, J. P. (2011). The human prefrontal cortex mediates integration of potential causes behind observed outcomes. *Journal of neurophysiology*, *106*(3), 1558–1569.

Zoltowski, D. M., Latimer, K. W., Yates, J. L., Huk, A. C., & Pillow, J. W. (2019). Discrete stepping and nonlinear ramping dynamics underlie spiking responses of lip neurons during decision-making. *Neuron*, *102*(6), 1249–1258.