Adaptive decision making in dynamic environments by artificial and biological agents

by

Nathan J. Wispinski

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychology
University of Alberta

Abstract

The ability to adaptively respond to changing environments is a fundamental aspect of intelligent behaviour. From catching a ball in motion to changing one's mind in the face of new information, adaptation requires several key cognitive mechanisms, such as the flexible integration of sensorimotor information and the ability to make predictions about the future. In this dissertation, I explore decision mechanisms underlying adaptive decision making in both artificial and biological agents, and the environmental pressures that may give rise to these mechanisms.

In Chapter 2, we assessed how well human participants can plan an upcoming movement based on a dynamic, but predictable stimulus. Our results showed that how people moved during their decisions reflected information in the moment, despite known neural and movement delays. These results suggest that humans rapidly and accurately integrate visuospatial predictions and estimates of their own temporal limitations to adapt their behaviour to a constantly changing environment.

In Chapter 3, we developed a deep reinforcement learning agent that learns via rewards to make adaptive decisions. In two tasks with different movement requirements, these artificial agents exhibit "changes of mind"—a behaviour thought to be a hallmark of flexible behaviour. Despite being trained solely with rewards in the absence of biological data, behaviour and neural mechanisms in these agents emerge during reward learning that closely resemble those in primates making similar decisions. These results suggest that the ability to make adaptive decisions similar to many biological agents emerges in artificial agents trained to maximize reward in the face of noisy, temporally evolving information.

In Chapter 4, we investigated deep reinforcement learning agents in an ecological patch foraging task. Our results showed that these artificial agents learn to patch forage adaptively in patterns similar to biological foragers, and approach optimal patch foraging behaviour. When investigating the mechanisms underlying this behaviour, we find dynamics that closely align with those from foraging theory, and neural recordings from foraging primates thought to give rise to biological adaptation during foraging.

Overall, I argue that the need to effectively act in dynamic environments contributes to the emergence of computational mechanisms in both artificial and biological learning systems that allow for adaptive behaviour. Further, as the ability to adaptively respond to changing environments is a fundamental aspect of intelligent biological behaviour, I discuss the implications of this work with respect to the emergence of human-like artificial intelligence.

Preface

This thesis is an original work by Nathan J. Wispinski. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name "Action and Attention Using EEG, Eye and Motion Tracking", No. Pro00059044, January 14, 2016, and Project Name "ACE 2", No. Pro00087329, February 25, 2019.

Chapter 2 of this thesis has been published as: Wispinski, N. J., Stone, S. A., Bertrand, J. K., Ouellette Zuk, A. A., Lavoie, E. B., Gallivan, J. P., & Chapman, C. S. (2021). Reaching for the known unknowns: Rapid reach decisions accurately reflect the future state of dynamic probabilistic information. *Cortex*, *138*, 253-265. doi:10.1016/j.cortex.2021.02.010. Data were collected by Craig Chapman and Jason Gallivan in the labs of Mel Goodale and Jody Culham at Western University in 2010. Experimental procedures were approved by Western University's Research Ethics Board. I was responsible for the majority of analysis, visualization, data curation, and writing. Scott Stone, Jenn Bertrand, Alex Ouellette Zuk, and Craig Chapman were involved with data analysis. All authors were involved in writing and manuscript composition.

Chapter 3 of this thesis has not been previously published. A version of this chapter has been presented as a talk at the 2022 Society for Neuroscience Conference as Wispinski, N. J., Stone, S. A., Singhal, A., Pilarski, P. M., & Chapman, C. S. (2021). Primate-like perceptual decision making through deep recurrent reinforcement learning. Simulations were performed at the University of Alberta, and human data were collected by myself and Garett Motley during 2019. Computational resources were

generously provided in part by WestGrid and Compute Canada, and more recently the Digital Research Alliance of Canada (`https://alliancecan.ca/`). Experimental procedures were approved by the University of Alberta's Research Ethics Office. I was responsible for the majority of experiment design, programming, analysis, visualization, data curation, and writing. Scott Stone and myself were responsible for model training. All authors were involved with experiment design, interpretation of results, writing, and manuscript composition.

Chapter 4 of this thesis has been published as: Wispinski, N. J., Butcher, A., Mathewson, K. W., Chapman, C. S., Botvinick, M. M., & Pilarski, P. M. (2023). Adaptive patch foraging in deep reinforcement learning agents. *Transactions on Machine Learning Research*. Simulations were performed at DeepMind during and after my time as a Research Scientist Intern in 2021. I was responsible for the majority of analysis, visualization, and writing. Agents and environment assets were adapted from Cultural General Intelligence Team et al. (2022) in collaboration with Andrew Butcher and many others. All authors were involved in experiment conceptualization, writing, and manuscript composition.

The "Bridging the gap between decision making and action" section of Chapter 1 has been previously published as: Wispinski, N. J., Gallivan, J. P., & Chapman, C. S. (2018). Models, movements, and minds: Bridging the gap between decision making and action. *Annals of the New York Academy of Sciences* [The Year in Cognitive Neuroscience series]. doi:10.1111/nyas.13973. The remainder of Chapters 1 and 5 are original works composed by Nathan J. Wispinski, and are previously unpublished.

Acknowledgements

I am grateful to so many people who helped in so many ways along this journey in my life.

First and foremost, thank you to Craig Chapman. We have worked together for over a decade now, and throughout all this time you have continued to be a great mentor, advocate, colleague, and friend. I am deeply appreciative of our time together, and grateful for every minute.

Thank you to my co-supervisor Anthony Singhal for your invaluable guidance and knowledge. And thank you to Patrick Pilarski—I truly appreciate your time, endless insights, support, and enthusiasm.

I would like to thank all the amazing people at UBC who went well out of their way to help me kindle my passion for research, including Jim Enns, Catherine Rawn, Todd Handy, Grace Truong, Elizabeth Dunn, and Ashley Whillans. Thank you to Joel Zylberberg, James Wright, Jeremy Caplan, Kelvin Jones, Kyle Mathewson, and Jason Gallivan. I would like to thank the entire ACELab—Jeff Sawalha, Ewen Lavoie, Alex Ouellette Zuk, Brea Chouinard, Quinn Boser, Riley Dawson, Gabi Oancea, and Beth Jantz. And thank you to my colleagues Alice Atkin, Nadia Ady, Becky Long, Abhishek Naik, and Alex Kearney. A huge thank you to Jenn Bertrand for all your insight, expertise, and encouragement throughout our many years of research together. And Scott Stone, thank you for being a fantastic colleague, and a truly great friend.

I would like to thank all the incredible people I met during my time at DeepMind, who helped me navigate a complex system, refine ideas, or changed the way I think in

some way. These include Ola Kalinowska, Mike Johanson, Kory Mathewson, Leslie Acker, Richard Everett, Arne Olav Hallingstad, Andrew Bolt, Mike Bowling, Dylan Brenneis, Adrian Collister, Elnaz Davoodi, Nik Hemmings, the entire DeepMind CGI Team, Alex Zacherl, Ed Hughes, Marlos Machado, Drew Purves, Kim Stachenfeld, Francis Song, Rich Sutton, Kevin Miller, and Jane Wang. I would especially like to thank Matt Botvinick for the many, many insights, and Andrew Butcher for your time, generosity, and wisdom.

I have been extremely fortunate to have had graduate funding throughout my studies. For their support, I would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC), the Killam Trusts, the Alberta Gambling Research Institute (AGRI), the Canadian Institute for Advanced Research (CIFAR), the Alberta Machine Intelligence Institute (Amii), the University of Alberta, and the University of Alberta's Department of Psychology. I am grateful for the computational resources provided in part by WestGrid and Compute Canada, and more recently the Digital Research Alliance of Canada. I would also like to thank the organizers, supporters, and other attendees at all the stellar summer schools I attended—CoSMO, CIFAR, Amii, York University's Centre for Vision Research, and Campus Alberta Neuroscience. I would also like to thank the faculty and staff of the Psychology, Kinesiology, Computing Science, and Neuroscience departments at the University of Alberta.

Thank you to all the amazing people in my life outside of research. I am so grateful to have such supportive and hilarious friends. Thank you Mom, Dad, Nic, Derek, Sheila, Brian, Baba, Oma, and the rest of my endlessly supportive family. Finally, thank you to my partner Miranda—you brightened every single day of this long journey, and I am endlessly grateful for your support and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Our world is highly dynamic. Seasons change and time creeps forward. While I move around in the world, so do others. Internal states, like feelings of alertness or hunger, change throughout the course of each day. In addition, even in very similar situations, no sensory experience is exactly the same. How are we to possibly act in such a complex and changing world? The problem feels immense, yet we as humans are able to act successfully all the time with minimal effort.

For example, if I am drinking a coffee and someone sets something down on the table, I am able to adapt my movements to avoid setting my coffee down in a now-occupied location. If I am driving and I see a traffic jam coming up, I am sometimes able to turn and take a less congested route. These actions may seem easy and intuitive, but are highly flexible. In contrast, failure to adapt to a changing world can cause a variety of headaches. Failing to adapt your movements to a new environment might mean that your coffee gets spilled or you languish in traffic when a perfectly good alternate route is available.

Further examples of non-adaptive behaviour come from other branches of the animal kingdom. For instance, greylag geese exhibit a "fixed action pattern" when they see an egg rolling out of its nest (Lorenz & Tinbergen, 1938). These geese extend their necks to gently pull the displaced egg back into the nest when it has fallen out.

However, geese continue this sequence of actions even when the egg is removed by a human experimenter midway through the motion and the goose no longer feels the egg under its neck.

A failure to adapt can even have dire consequences. Army ants, for all their complex collective behaviour, sometimes catastrophically fail to adapt to external and internal changes. If a group of ants loses a pheromone track while foraging, they can form an ant mill or "death spiral", where ants will follow each other in a circle until they eventually die of exhaustion (Delsuc, 2003). These ants fail to adapt appropriately to the loss of a pheromone track, and then continue failing to adapt to seeing repeating landmarks, to the lack of environmental change despite significant locomotion, or to changes in their internal state to that of exhaustion.

The difficulty of adaptive behaviour is also well-illustrated when looking at artificial systems. For example, some robots cannot deal with small environmental perturbations (Krotkov et al., 2018). When a box to pick up is not in the correct place, the robot can drop or crush the box. When starting a sequence of movements from a slightly different place than programmed, some robots cannot adapt to the changes in relative position between themselves and other objects. These failures of adaptive behaviour illustrate a critical lesson from artificial intelligence research—that adaptive and robust sensory and motor skills are some of the most computationally difficult to specify (Minsky, 1988; Moravec, 1988; Zador et al., 2023).

In general, biological systems excel at adaptation, while the adaptive abilities of artificial systems are far behind. How then do biological systems adapt so well—what are the underlying neural mechanisms? And if we were able to gain insight into these adaptive mechanisms in biological agents, could we leverage this knowledge to build better adaptive artificial systems? Answering these questions may have real-world applications to aid humanity. Research into biological adaptive behaviour not only helps us better understand ourselves, but may also provide ways to improve our decisions and shed insight into clinical treatments for neurological disorders. In

addition, research into artificial adaptive behaviour may help us build better virtual assistants that adapt to our needs, autonomously drive vehicles, or support humans in health care settings.

To contribute to these large-scale goals, researchers typically work on a small piece of the larger puzzle. Here I present my piece. I aim to uncover the neural mechanisms that allow for rapid sensorimotor adaptation during decision making, and find ways for artificial systems to learn these mechanisms on their own. I specifically focus on rapid sensorimotor adaptation, rather than other, slower forms of adaptation like learning or evolution. This is in part because of my personal interest in the topic—these rapid, flexible behaviours can seem almost automatic, and yet hinge on many complex computations involving visual perception, proprioception, social context, internal states, and high-level goals. This focus is also partly serendipitous—I had the good fortune to stumble upon stellar research supervisors who had expertise in high temporal resolution techniques like motion tracking, neuroimaging, and robotic control that provide great tools for investigating such fast-paced adaptive behaviour.

In addition, I focus on how artificial agents might learn rapid sensorimotor adaptation mechanisms on their own, instead of ways to explicitly program these agents with computational mechanisms inspired by biology. Again, this focus is in part because of personal interest and theory—biological agents learned these mechanisms themselves, so why can't agents? This focus is also serendipitous—throughout my graduate career, I was fortunate enough to learn from experts on reinforcement learning, which provides a unique framework for artificial agents to discover behaviours on their own through environmental interaction.

To recap, this work has two interwoven goals—to understand the computational mechanisms of rapid sensorimotor adaptation in biological systems, and to develop ways in which artificial systems can learn these mechanisms themselves. Throughout this thesis, my work starts with a focus on biological systems, and gradually ends with

a focus on artificial systems. However, I strongly believe the two are complementary and should not lose sight of each other (see The mutualistic relationship of artificial and biological intelligence research).

In the following sections within Chapter 1, I provide some background to scope this thesis. I'll discuss more on why I'm specifically focusing on rapid sensorimotor adaptation, rather than other forms of adaptation like learning or evolution (see Adaptive behaviour). Next, I'll argue that biological agents need to be able to rapidly adapt *because* the world is dynamic. This idea means that we need to focus on dynamic environments within this thesis to better understand adaptive behaviour (see A dynamic decision process for a dynamic world). Afterward, I include a review on what is known about rapid sensorimotor adaptation during decision making in the biological literature to describe what we know so far about how biological agents adapt so well (see Bridging the gap between decision making and action). To conclude Chapter 1, I turn to artificial agents, and discuss some ways in which adaptive behaviour has been achieved in artificial systems so far, and how there is still much work to be done (see Adaptation in artificial systems). In the chapters that follow, I present three studies in scientific paper form that each progress the goals of this thesis, then conclude this thesis with a General Discussion (Chapter 5) on the origins of adaptive behaviour and the relationship between biological and artificial intelligence research.

In Chapter 2, I push forward our understanding of rapid sensorimotor adaptation in biological systems. In this work, a beautiful dataset came across my desk which asked the question, "how quickly can humans adapt their movements based on a dynamic, but predictable stimulus?". The results surprised me—human performance was much better than I anticipated. The incredible degree of adaptive behaviour displayed in this experiment points to two biological mechanisms that I believe will be critical for adaptive artificial systems in many contexts. Specifically, humans have intimate knowledge of their own temporal limitations, and also have the ability to rapidly integrate visuospatial predictions into their movements.

In Chapter 3, I describe work that arose from my yearslong interest in a phenomenon emblematic of rapid sensorimotor adaptation—changes of mind. Changes of mind in decision making are when agents change their decision in the face of new information (Resulaj et al., 2009). For example, you might change your mind if you were leaning toward ordering a beer at a restaurant, and then your friend suggests sharing a bottle of wine. In this chapter we replicate results showing that humans change their minds in the face of new information to improve their decisions (Resulaj et al., 2009; van den Berg et al., 2016). Like in Chapter 2, this improves and reinforces our understanding of rapid sensorimotor adaptation in biological systems. At this point in the thesis, I begin to focus on artificial agents. Specifically, my original goal in this project was to develop artificial agents that could change their minds like humans. I had previously spent a substantial part of my time explicitly programming decision making rules for artificial agents to display changes of mind (Wispinski, 2017). I learned a lot through this work, but ultimately found it slightly unsatisfying on a personal level. Despite being programmed to exhibit adaptive behaviour, these agents were surprisingly rigid. We preprocessed sensory information for them, and came up with explicit ways for them to overcome ambiguous situations. This led me to investigate if artificial agents could learn to change their minds the same way that humans are thought to have learned to change their minds—by acting in dynamic environments. I was additionally inspired by three recent findings in deep learning that I regarded as critical individual pieces of this puzzle. Specifically, I was inspired by agents that learned for themselves primate-like visual processing (Rideaux & Welchman, 2020), decision-making (H. F. Song et al., 2017), and motor control (Lillicrap et al., 2015). Extending this work, we trained deep reinforcement learning agents to make decisions in a noisy, dynamic environment, and found that they learned themselves to change their mind when appropriate. In addition, these agents learned several other behaviours and decision making mechanisms found in primates. Finally, we showed that agents trained in ways that deviate from biology

did *not* learn to make adaptive decisions like primates. Overall, I argue these results provide a promising way to develop adaptive artificial systems—by training them to act in environments inspired by biology.

In Chapter 4, I further investigate this idea that artificial agents might learn adaptive mechanisms on their own via interacting within biologically inspired environments. In this work, I turn to one of the most fundamental decision problems that face biological agents—patch foraging. Almost all animals patch forage, and must do so successfully to survive. Theorists even speculate that biological foraging behaviour is not only adaptive, but optimally adaptive because of strong selective pressures (Charnov & Orians, 2006; Pearson et al., 2014; Stephens & Krebs, 2019). In this way, optimizing agents to forage in biologically inspired environments might encourage them to discover similar adaptive mechanisms. This is largely what we found—agents learned to adaptively patch forage and approached optimal foraging behaviour. Additionally, agent neural dynamics mirrored mechanisms thought to underlie adaptive patch foraging in theory (Davidson & El Hady, 2019), and in biological experiments (Hayden et al., 2011). These results further support the idea that environmental pressures inspired by biology provide a compelling path toward adaptive artificial agents.

Together, these results reinforce findings that biological agents possess the incredible ability to rapidly adapt to dynamic environments—a feature we desire in artificial agents. Theory suggests that these mechanisms in biological agents emerged through the pressures to act in a changing world. We roughly simulate the same process and find that artificial agents also begin to learn complex adaptive behaviours similar to biological agents. Overall, this work supports the idea that artificial and biological research still have much to give one another in our journey to understand and create (Hassabis et al., 2017)—an argument I return to make at the end of this thesis (see The mutualistic relationship of artificial and biological intelligence research).

1.2 Adaptive behaviour

Here I elaborate on some of the scoping decisions in this thesis. I specifically focus on rapid sensorimotor adaptation, rather than other forms of adaptation. Researchers in experimental psychology or neuroscience might call this within-trial adaptation. One example is in target jump tasks, where participants are asked to rapidly reach and touch a target on a screen. On some proportion of trials, the target teleports to a new location just before or while the participant is moving, requiring the participant to rapidly update their movement (Megaw, 1974; Sarlegna & Mutha, 2015). Another example comes from the random dot motion discrimination task in perceptual decision making (Newsome & Paré, 1988). Participants are shown dots that move to the left or to the right with some level of random noise, and asked to report in which direction the dots are moving. Because dot motion is noisy, participants need to consider multiple time samples of motion to make a good decision. Sometimes participants initially begin to indicate the dots are moving leftward before switching to ultimately report that the dots are moving to the right. Such behaviour has been termed "changes of mind", where participants can revise their decision in the face of new information (Resulaj et al., 2009). In other words, changes of mind indicate an ongoing decision process that continuously adapts an agent's preference to incoming sensory information. Finally, dynamic pursuit or avoidance tasks require the rapid online adaptation of movement as an animal needs to continuously avoid or pursue other objects in a dynamic environment (Fooken et al., 2016; Yoo et al., 2021). The key property of these forms of adaptive behaviour are that they rapidly take place online within a single decision. This adaptation likely utilizes previously-learned neural mechanisms and operates quickly within the dynamics of spiking cells (J. X. Wang et al., 2018).

In contrast, another, slower form of adaptive behaviour involves learning throughout the course of a task. One example is in motor adaptation tasks, where participants

need to reach toward a target by moving a controller such as a joystick or a robotic manipulandum. After acclimatization, a perturbation is applied to the controller such as a slight rotation so that movements do not have the same effect as before. In these tasks, participants have to learn to adapt to the perturbation to reach the target, and do so gradually over the course of several subsequent trials (Krakauer et al., 2019). Another example is in bandit tasks, where participants are presented with several discrete options (called arms) each with their own win probability unknown to the participant (Sutton & Barto, 2018). In order to maximize reward, participants must learn which arm has the highest win probability through trial-and-error. The kind of adaptive behaviour on display in these tasks is a slower learning process that takes place across several decisions, and involves short or long term learning mechanisms (J. X. Wang et al., 2018; Wolpert et al., 2001).

Adaptation can also take place over longer timescales such as across several tasks, or across a lifetime (Wolpert et al., 2001). For example, as an individual grows throughout development their limb sizes change, which requires adaptation to changing effectors. Humans are able to adapt to drift in action output, but also changes in sensory input such as age-related changes in visual function (Scialfa, 2002). Finally, adaptation can of course also take place between lifetimes. Here I am referring to evolutionary adaptation, where natural selection enhances the evolutionary fitness of organisms with respect to their environment (Darwin, 1872; Mayr, 1982). In this thesis, I mainly focus on the cognitive ability of rapid sensorimotor adaptation within a single trial within an individual agent, rather than other important aspects of adaptation outlined above, such as learning or evolutionary processes.

1.3 A dynamic decision process for a dynamic world

This thesis focuses on rapid sensorimotor adaptation within a single decision. However, this behaviour necessitates that there is a change in the environment within a single decision that requires adaptation. As such, this thesis largely ignores *static* en-

vironments, which are environments where goal-relevant information does not change within a single decision. In general, static environments can be simple or complex. Examples of static environments include image recognition tasks (e.g., as in many machine learning studies; CIFAR-10, etc.; Krizhevsky and Hinton, 2009), navigation to a stationary target location (as in artificial agents in video games: Devlin et al., 2021; Milani et al., 2023; or single-target reaching studies in humans: Scott, 2004), or many value-based decision making studies in humans (e.g., choosing between two images of snack foods; Krajbich and Rangel, 2011). In these environments, agents generally do not need to reconsider previous information and change goals on the fly accordingly.

In contrast, this thesis focuses on *dynamic* environments, which are environments where goal-relevant information *can* change within a single decision. Dynamic environments are those where changing information changes optimal actions or goals. For example, in the target jump tasks described above, the physical location of a target teleports to another location, requiring a rapid change in movement to the new target location. Other examples come from multi-agent tasks, where agents need to adapt to other agents that adapt to their own behaviour online. Examples include pursuit and avoidance tasks (Yoo et al., 2021), or vehicle racing (Wurman et al., 2022).

A dynamic environment can also appear similar to a static environment if it is noisy and partially observable. For example, in the random dot motion discrimination task on each trial, dots move in a single direction with some level of random noise. In this way the latent variable of this environment (i.e., the direction of dot motion) is static, but information conveyed to the agent is dynamic—in flux from time step to time step. As such, the agent needs to consider noisy information across time to evaluate which target is the correct goal of action. Perceptual discrimination tasks for biological agents, like the random dot motion task, have many variants that are even more dynamic (Cisek et al., 2009; Huk & Shadlen, 2005; Thura & Cisek, 2014; Yates et al., 2020). For example, some tasks change the direction of dot motion

midway through the trial unbeknownst to participants. In biological experiments, these tasks are often used as a proxy for the real, dynamic world in which animals typically act. For instance, imagine an animal hunting for prey in a forest. The animal's vision at each particular moment does not give the whole picture—the prey might be partially or fully obscured because of trees and bushes. In addition, the animal's neural processing of visual information is subject to attention and sensory noise. Even though the true goal-relevant state of the prey behind the foliage is not changing, the information conveyed to the animal is dynamic.

In the context of machine learning, classification or prediction tasks in tutorial environments are often static. For example, regression problems or image recognition tasks often use training and testing data from a single, unchanging distribution. However, real world problems are often dynamic—economic trends impact consumer behaviour, and image sensors drift over time. Popular large language models (LLMs) are sometimes faced with user questions about ever-evolving current events which are constantly beyond their training data. This often requires a dynamic process of continuously retraining or re-tuning models with more up-to-date data.

Reinforcement learning research provides an excellent framework for thinking about dynamic environments. As stated before, classic reinforcement learning tasks include bandit tasks, where agents are presented with discrete actions (called arms) that each give rewards with some win probability (Sutton & Barto, 2018). These tasks are often illustrated as a choice between some number of simple slot machines. Bandit tasks can be static, where each choice has a single, unchanging win probability. In these tasks, reinforcement learning agents are tasked with maximizing reward by finding the action with the highest win probability through trial-and-error. Bandit tasks can also be dynamic—the win probability of each unique choice can change over time according to some dynamic process like a random walk (Behrens et al., 2007; Daw et al., 2006; J. X. Wang et al., 2018). In these dynamic bandit tasks, the optimal action changes over time, and so agents need to adapt their choices over time through experience.

Finally, artificial agent research also includes motor control tasks, where agents need to control effectors to reach a goal (e.g., Tunyasuvunakool et al., 2020). Sometimes perturbations are applied, like a sideways force on the body or a new obstacle, which requires adapting the current control policy online to these environmental changes (Heess et al., 2017).

This is not to say that static environments are not important. Some tasks biological systems complete can be thought of as static. Additionally, some tasks that we would like artificial agents to perform are static, yet still important and difficult. However, dynamic tasks pose a set of challenges that mirror many properties we desire agents to complete in the real world. Moreover, responding to one challenge in a dynamic environment also prepares an agent to be better equipped to respond to other challenges. That is, if an agent is able to robustly adapt behaviour to dynamic changes within one task, the agent may be able to leverage these mechanisms to generalize to other tasks as well.

1.4 Adaptation in biological systems

A core goal of this thesis is to understand the mechanisms underlying rapid sensorimotor adaptation in biological agents. Despite their limitations, biological agents are incredibly flexible compared to artificial agents. In some cases, humans and other animals have even been shown to be optimally adaptive. For example, many animals have been shown to adapt their foraging patterns optimally to local or global environmental changes (Cowie, 1977; Krebs et al., 1974; Lottem et al., 2018; Pacheco-Cobos et al., 2019; Stephens & Krebs, 2019; Vertechi et al., 2020). In these studies, animals are tasked with patch foraging, where they need to trade off exploiting diminishing resources within a patch, and exploring the environment for more plentiful alternative resource patches (see Chapter 4). Animals have been shown to optimally adapt their explore and exploit behaviour to the resource-richness of different environments, consistent with mathematical work on optimal foraging (Charnov, 1976). Other work

has shown that humans are able to optimally adapt their choices in a dynamic bandit task, where the win probabilities of an arm change over time (Behrens et al., 2007). Finally, research has shown that human reaching behaviour aligns well with optimal control models, explaining how movements vary in space and how they adapt to perturbation of their effectors while moving (Todorov & Jordan, 2002). So how is this ability to quickly adapt in dynamic environments implemented in biological systems?

The following section reproduces a review paper related to this topic. Myself, Jason Gallivan, and Craig Chapman wrote the following review paper in part because of the surprising disconnect between the fields of decision making and motor control in cognitive neuroscience. We argue that there is significant evidence that in order to quickly adapt to a dynamic environment, there needs to be intimate connections between decision making and motor control in the brain. For example, animals must make rapid decisions on the fly about where to move when being chased by predators. They must continuously reevaluate information like their body positions, the states of the predator, and the benefits and drawbacks of locations to flee to as these factors change rapidly in time. These task requirements greatly benefit from fast mechanisms to integrate such cognitive and motor information to adapt movements. Despite this, many researchers regard perception, decision making, and motor control as distinct processes that operate in a serial loop—animals perceive the world, then make a decision, and only then plan a movement with respect to that decision. Below we argue against such a view, and review computational models, behavioural experiments, and neural recordings that point toward mechanisms that intimately tie together decision making and motor control in the service of adaptive behaviour.

1.5 Bridging the gap between decision making and action

The text in this section has been previously published as: Wispinski, N. J., Gallivan, J. P., & Chapman, C. S. (2018). Models, movements, and minds: Bridging the gap

between decision making and action. *Annals of the New York Academy of Sciences*. [The Year in Cognitive Neuroscience series]. doi:10.1111/nyas.13973

1.5.1 Abstract

Decision making is a fundamental cognitive function, which not only determines our day-to-day choices but also shapes the trajectories of our movements, our lives, and our societies. While immense progress has been made in recent years on our understanding of the mechanisms underlying decision making, research on this topic is still largely split into two halves. Good-based models largely state that decisions are made between representations of abstract value associated with available options; while action-based models largely state that decisions are made at the level of action representations. These models are further divided between those that state that a decision is made before an action is specified, and those that regard decision making as an evolving process that continues until movement completion. Here, we review computational models, behavioural findings, and results from neural recordings associated with these frameworks. In synthesizing this literature, we submit that decision making is best understood as a continuous, graded, and distributed process that traverses a landscape of behaviourally relevant options, from their presentation until movement completion. Identifying and understanding the intimate links between decision making and action processing has important implications for the study of complex, goal-directed behaviours such as social communication, and for elucidating the underlying mechanisms by which decisions are formed.

1.5.2 Introduction: Two halves of a whole decision

Imagine opening a refrigerator on a hot summer's day. To the left, a pitcher of iced tea, and to the right, a bottle of sparkling water with lemon. How do you decide between these two options, and how does this decision result in the movement required enacting it? As with all decisions, this involves reducing many choice options to only

one goal. And, as with the types of decisions we focus on in our review, most goals require a physical action.

The first question at the heart of our review is, In what representational space is an option selected? According to good-based decision theories (Padoa-Schioppa, 2011; Padoa-Schioppa & Assad, 2006; Pastor-Bernier et al., 2012) choice options are represented and are selected in an abstract value space where multiple sources of information are combined to construct a single subjective value for each of the options in a common neural currency (Levy & Glimcher, 2012). In the most rigid of these theories, the highest value, winning option is selected and only then is an action planned to enact the decision (Padoa-Schioppa, 2011). At the other end of the spectrum are action-based decision theories (Cisek, 2007; Dorris & Glimcher, 2004; Pezzulo & Cisek, 2016; Rangel & Hare, 2010; Shadlen et al., 2006), wherein choice options are represented and selected within sensorimotor maps of space that directly reflect how each option is physically situated in the environment. Under this view, the representation of every option is sensorimotor in nature, reflecting details of the movement associated with acting on each alternative.

Following from this distinction, the second question central to our review is: When does a decision get made? Returning to the cold-drink-on-a-hot-day conundrum, good-based theories generally argue that you first choose the drink and then you plan the movement toward it; that is, the decision is made before the associated movement is specified. Conversely, action-based theories generally argue that movement representations toward both drinks are maintained in parallel; that is, the decision does not end until the movement is complete (Fig. 1.1). However, while good-based theories are often implicitly associated with serial processing and action-based theories are often implicitly associated with parallel processing, neither framework strictly requires that they conform to these specific temporal sequences of decision making.

When decisions are made and at which level options are selected has a profound impact on understanding the underlying neural architecture involved, why we choose

certain options over others, and how we behave in between. For example, most action-based models would predict diverse and intimate neural connections between motor and perceptual systems for sensory decision information to shape motor representations, whereas good-based models would instead predict that an abstract value space mediates many of these connections. Additionally, action-based models would predict quicker responses after target selection, as the motor response associated with the selected option has already been specified in the sensorimotor system. It would also predict, however, that unselected movements might seep into movement execution.

Our review is broadly structured into three parts. The first two parts address good- and action-based models each in turn and reviews formal models, and behavioural and neural data in support of the theory. In the last part, we review recent models, behaviour, and neural findings that have made progress in bridging the gap between good- and action-based decision-making models. As with most problems in cognitive neuroscience, what initially appears to be a stark divide in theory is likely a result of dichotomous thinking—the real answer likely lies somewhere in between. In this review, we do not attempt to put forth a unifying theory of decision making, but rather identify gaps in our understanding and aim to outline current evidence for the three lines of thought.

1.5.3 Good-based models: From stimulus to reaction time

Good-based models of decision-making state that the selection of available options occurs at the level of abstract value representations (Padoa-Schioppa, 2011; Padoa-Schioppa & Conen, 2017). In this review, our use of the term *value* is specific to the task at hand. For example, in a perceptual decision task, like deciding in which direction a pattern of noisy dots is moving (Britten et al., 1992, Fig. 1.2), value is derived from perceived motion direction. By comparison, in tasks like deciding between two chocolate bars (Krajbich et al., 2010), value is derived from subjective preference. One of the most parsimonious features of a good-based model is that it

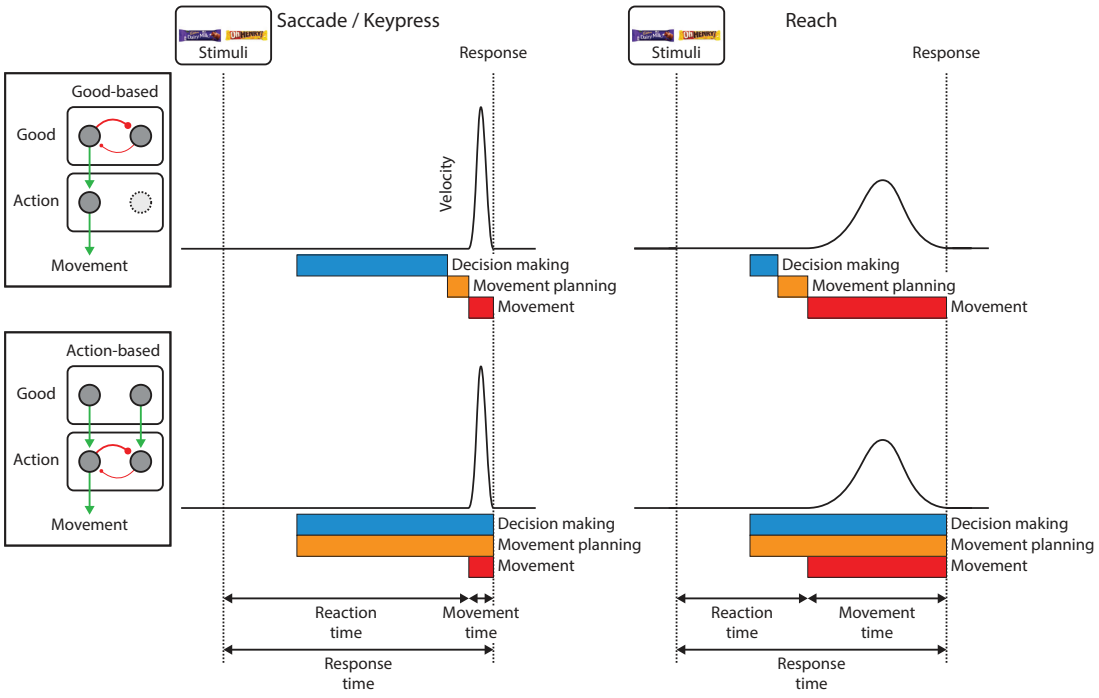

Figure 1.1: (a) Good-based models (top) state that a decision is made at the level of abstract values before an associated action is planned, while action-based models (bottom) state that a decision is made at the level of physical movements. (b) Serial and parallel processing timelines of decision making (blue), movement planning (orange), and movement (red) during a simple decision like choosing between chocolate bars. Choices requiring ballistic, short duration responses like saccades or key presses are shown on the left and those requiring longer duration responses like reaches are shown on the right. Good- and action-based schematics adapted from Chen and Stuphorn (2015)

is agnostic with respect to what information is used to construct value, since there is an abstract common currency on which an arbiter can judge. With respect to the relationship between a decision and the action associated with enacting it, a good-based model treats the decision process as a distinct module in a serial process (Fig. 1.1). Once perceptual processes have delivered a representation of the choice options, and decision processes have selected one for action, only then is the corresponding movement planned. To expand upon this serial architecture of decision making and movement, we first describe bounded evidence accumulation decision-making models. These models are distinct from good-based decision-making models since bounded accumulator models are typically agnostic regarding the level at which options are selected, and instead focus on how selection occurs. However, most bounded accumulation and good-based models share the common assumption that decision making is complete before movement processes that enact the decision begin.

Model

The most prominent (though by no means only) class of decision-making models are those based around the accumulation of evidence to a threshold (Bogacz, 2007; Gold & Shadlen, 2007). Here, the information relevant to a decision (i.e., evidence) is repeatedly sampled from the external world, or from internal sources such as memories (Shadlen & Shohamy, 2016). Evidence for or against a particular option is added over time. When this accumulated evidence in support of a particular option crosses some threshold, the decision is made (Fig. 1.2b). For example, in the random dot motion (RDM) task (Britten et al., 1992), moving dots are presented to a participant who is asked to discriminate the net direction of the dots (e.g., left versus right; Fig. 1.2a). Decision difficulty is manipulated by the amount of dots moving in the same direction on each trial (i.e., coherence). Additionally, the stimulus is noisy, as the remaining "noncoherent" dots move in random directions. Evidence accumulation models argue that subjects arrive at decisions in this task by sequentially sampling small portions

of the motion stimuli. This information is processed to extract whether, and how much, the motion sample favors responding left or right. This evidence is then added to left and right accumulators in the brain and sampling continues until some decision criterion based on accumulated evidence is met (Fig. 1.2b).

Two of the most widely used evidence accumulation models are deemed race (Smith & Vickers, 1988) and drift diffusion/random walk models (Link, 1975; Ratcliff & Rouder, 1998). Race models state that a decision is made the first time any one of multiple independent accumulators crosses some fixed decision threshold. In contrast, drift diffusion models state that decisions are made based on relative evidence—the difference in evidence between options is accumulated until reaching an upper or lower bound corresponding to the two options under consideration (Fig. 1.2b). It is beyond the scope of our review to summarize the support for and against the many kinds of bounded evidence accumulation models, but in general, these models account for behaviour in a wide range of tasks (Ratcliff & Smith, 2004; Ratcliff et al., 2016, Fig. 1.2c). Indeed, there is a mathematical elegance in this approach—the process of evidence accumulation is intimately related to signal detection theory (Green & Swets, 1966) and Bayesian inference (Beck et al., 2008), and is regarded as optimal in some sense (Bogacz et al., 2006).

While broadly successful, recent research shows decision making is more complex than simple evidence accumulation models can describe. Models need significant elaboration to account for very early responses (Noorani & Carpenter, 2016), the dynamic cost of accumulating evidence for a decision (Drugowitsch et al., 2012; Hawkins et al., 2015), the growing urgency to make a response (Thura et al., 2012; Thura & Cisek, 2016), or, on the other end of the spectrum, the ability to refrain from making a decision (often accomplished via "leaky" accumulators; Usher & McClelland, 2001).

Regarding the two central questions of our review, bounded evidence accumulation models are largely agnostic to what exactly is being represented during choice competition but are fairly committed to the position that decisions are made before

movement execution (i.e., the crossing of a decision threshold). That is, evidence accumulation can as easily be applied when what is being accumulated is a representation of abstract value (good-based) as it can if what is being accumulated reflects the value of specific planned movements (action-based). However, when it comes to the timing of decisions, bounded evidence accumulation models generally assume that the decision process is complete before a movement is initiated, mirroring a serial process of perception, decision making, and finally movement planning.

Behaviour

Recent reviews identify the "three pillars of choice behaviour" as accuracy, reaction time, and confidence (Pleskac & Busemeyer, 2010; Shadlen & Kiani, 2013), and it is the measurement and explanation of these three outcomes across a variety of tasks that grants bounded evidence accumulation models their status as one of the best theoretical accounts of decision making. Classically, signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2004) can explain choice accuracy (and perhaps confidence) but not choice timing. However, evidence accumulation with recent additions can account for all three (Kiani, Corthell, et al., 2014). While it is beyond the scope of our review to exhaustively describe these additions, it is useful to highlight that a successful decision theory should be able to account for how accuracy, reaction time, and confidence vary as a function of decision difficulty.

In a good-based framework, decisions vary in the degree of value similarity between available choice options. Two options that have very similar value will be harder to decide between, while two options that have disparate value will result in easier decisions. Classically, easier decisions are resolved more quickly and more accurately, leading to faster reaction times and more correct responses. Conversely, hard decisions take longer to make and result in more errors.

The effects of decision difficulty are particularly evident in psychophysical tasks of perceptual discrimination like the RDM task (Fig. 1.2a). The decision difficulty

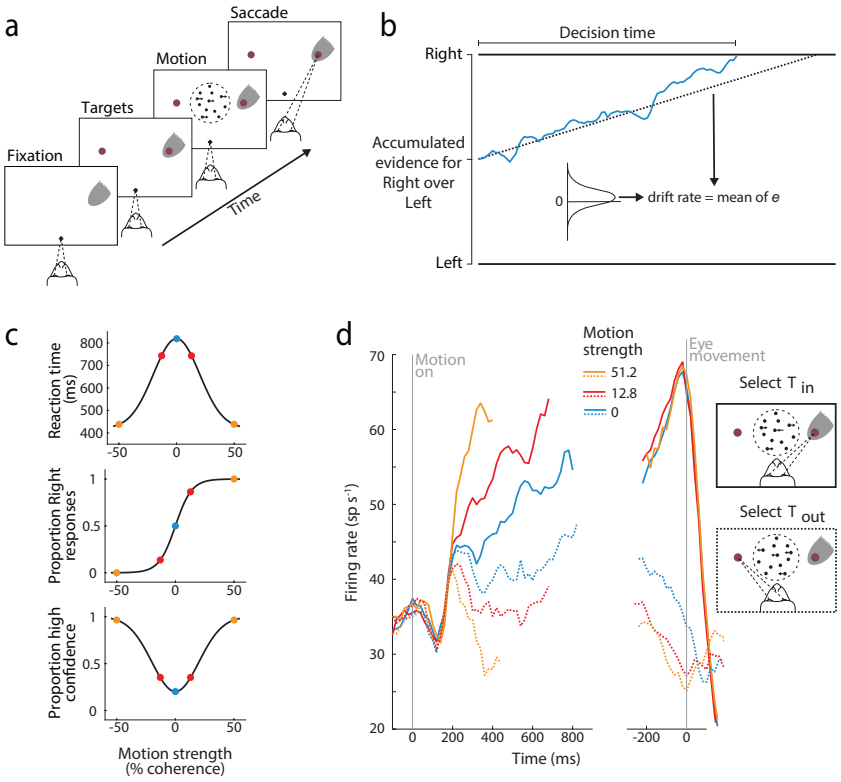

Figure 1.2: Findings from investigations on the mechanisms underlying perceptual decision making. (a) Trial structure of a typical random dot motion (RDM) task. After maintaining fixation, left and right targets appear (dark red). Animals are trained to execute a saccade to the target corresponding to the net direction of the moving dots. Moving dots appear within a central region of the screen, some of which move coherently left or right, while the remaining dots move in a random direction. Gray region indicates the receptive field of neurons typically recorded in this task. (b) Schematic of a drift-diffusion model applied to decision making on a single trial. Motion from the stimulus is sampled, and the difference in motion evidence for left and right target choices is accumulated (typically modeled as a normal distribution; see inset). When the accumulated evidence (blue trace) crosses a specified bound, the respective response is executed. (c) Example behavioural results in the RDM task. As the stimulus includes more coherently moving dots, reaction time decreases, accuracy increases, and confidence in the executed choice increases. (d) Average firing rates of recorded LIP neurons in the RDM task. Average firing rates of LIP neurons increase (or decrease) in proportion to the motion evidence favoring a saccade toward the receptive field of the recorded neuron. Average firing rates of LIP neurons reach a common firing rate "threshold" before a saccade is executed in the direction of the corresponding receptive field. Panels a, b, and c are reproduced and adapted from Gold and Shadlen (2007).

is manipulated by changing the number of dots moving coherently from very easy (100%) to very hard (<5%). The general finding in these tasks is that as coherence is reduced, accuracy decreases and reaction times get longer (Palmer et al., 2005, Fig. 1.2c). Similar results on accuracy and reaction time are abundant, even in the less-represented domain of decisions based on subjective value (Glimcher, 2010; Rangel & Hare, 2010; Sugrue et al., 2005).

Recent additions (Pleskac & Busemeyer, 2010) or extensions (Kiani & Shadlen, 2009) to evidence accumulation models can account for reductions in confidence with increases in decision difficulty. In a modified RDM task, in addition to the usual left and right choice options, Kiani and Shadlen (2009) presented monkeys a third "safe bet" option on some trials which gave a smaller "sure" reward. This allowed the monkeys to opt out of making a decision and instead take a small certain gain. Consistent with predictions from an evidence accumulation model when the trials were more difficult, monkeys more often opted for the safe bet.

Here, we define the three key behavioural outcomes that decision-making models must account for when decision difficulty is varied: as choice options become more similar, reaction times increase, while accuracy and confidence decrease (Fig. 1.2c). These features are well accounted for by robust models within a bounded evidence accumulator framework or the broader good-based theory, which state how the values of options are constructed, represented, and compared in order to ultimately select an action. Of note, these are all behavioural features that occur up to and including reaction time, but not after.

Neural

Strong neural evidence for both good-based competition and bounded evidence accumulation models in part affords the high status they enjoy within cognitive neuroscience and beyond.

In an exemplary group of studies, researchers recorded from the nonhuman pri-

mate (NHP) lateral intraparietal cortex (LIP), a brain area involved in oculomotor control (Andersen et al., 1985), while monkeys perform the RDM task. The neural responses are strikingly consistent with bounded evidence accumulation models (Platt & Glimcher, 1999; Roitman & Shadlen, 2002; Shadlen & Newsome, 1996, 2001, Fig. 1.2d). Firing rates of LIP neurons increase (or decrease) over time proportional to the amount of motion evidence favoring a saccade into the preferred direction of the recorded neuron. When the firing rate reaches some fixed threshold, a saccade is generated in that direction. This pattern is exactly consistent with that predicted by bounded evidence accumulation models—an accumulation of evidence in favor of each option until a threshold is crossed to execute an action.

Further experiments have provided even stronger support for this hypothesis. Short pulses of background motion during the RDM task briefly enhance or suppress the increase in firing rate for neurons associated with the correct response (Huk & Shadlen, 2005). Subthreshold stimulation of LIP neurons increases the proportion of saccades in the stimulated direction, and decreases their reaction times, as does subthreshold stimulation of the earlier motion-sensitive middle temporal visual area (MT), through which LIP receives significant input (Hanks et al., 2006). Together, this suggests that momentary motion evidence in the RDM task is computed within MT, and the accumulation of this evidence occurs downstream within LIP. Studies in humans using magnetoencephalography (Donner et al., 2009), electroencephalography (O'Connell et al., 2012), and functional magnetic resonance imaging (fMRI; Krueger et al., 2017) have also shown support for bounded evidence accumulation. While these experiments might seem to support choice selection at the level of saccades tuned to specific directions (action-based), others have argued that these patterns might instead simply reflect the motion direction of the random dot stimulus (Freedman & Assad, 2011). Furthermore, the idea that LIP plays a causal role in evidence accumulation is being reevaluated in light of recent experiments implementing pharmacological or optogenetic LIP inactivation, which fail to show corresponding deficits in decision

making (Katz et al., 2016; Licata et al., 2017; Yates et al., 2017). These challenges to LIP-based decision models give rise to the idea that perhaps options are represented and selected elsewhere in the brain, but at the same time do not invalidate the bounded evidence accumulation mechanism.

In support of a good-based decision-making model, studies have found evidence that the abstract value of available options is represented by orbitofrontal cortex (OFC) neurons integral in option selection. In a seminal experiment, NHPs made saccades to a left or right target offering different amounts of different kinds of juices, which the NHP would then receive (Padoa-Schioppa & Assad, 2006). The overwhelming majority of recorded neurons in the OFC were either sensitive to the amount of a particular type of juice offered, the type of juice the NHP was about to select, or the amount of juice the NHP was expecting to receive. In this task, the types and amounts of each juice option are sufficient for a rich representation of value, and the presence of neurons specifically encoding what option was to be chosen suggests the OFC may have a critical role in option selection. Importantly, neuronal responses were not found to vary with the spatial configuration of the options, nor with the direction of the upcoming saccade, suggesting the representations of value in the OFC were truly abstract. Other studies in NHPs have also shown little sensitivity to motor properties in OFC neurons despite significant sensitivity to aspects of subjective value (Grattan & Glimcher, 2014; Kennerley & Wallis, 2009). Furthermore, while associations to specific actions are difficult to parse with human fMRI data, value signals in a wide range of tasks and contexts have likewise been reported in the OFC (De Martino et al., 2009; Kable & Glimcher, 2007; Plassmann et al., 2007).

Discussion

The frameworks reviewed above are well supported and, as we will argue, are necessary for a complete understanding of the mechanisms underlying decision making. The bounded evidence accumulation framework provides an elegant explanation regarding

how options are selected, and a good-based theory provides a convincing solution as to how value is constructed and represented in the first place. However, they share a common limitation in that they generally argue a decision is complete before a movement is initiated.

In the vast majority of tasks to which these frameworks have been applied, the decision to move (measured via reaction time) is temporally bound with the movement outcome (measured via response time; Fig. 1.1). That is, saccadic eye movements (Padoa-Schioppa & Assad, 2006; Roitman & Shadlen, 2002), button presses (O'Connell et al., 2012), and verbal responses are essentially ballistic—reaction and response time are treated as the same value. Thus, it is perhaps unsurprising that many models do not account for decision making *after* movement initiation—indeed, in most tasks, this does not even exist. However, in the real world, the execution of most decisions takes time. The temporal protraction of movement has important implications for decision making; if an animal moving through a dynamic world wants to be optimally responsive to their environment (Cisek, 2007), it would be maladaptive to wait for one movement to complete before initiating a new decision process. Such models may also be a byproduct of the tasks used—decisions are studied in sequential isolation, deliberately separated by intertrial intervals. However, in the real world, new decision alternatives are constantly appearing or shifting, and require constant updating. This need to account for more ecologically relevant scenarios brings us to discuss another framework—action-based models, which may be particularly suited to account for decision making after movement initiation.

1.5.4 Action-based models: From reaction time to end of movement

Action-based models of decision-making state that available options are represented and selected in sensorimotor maps (Cisek, 2007; Cisek & Pastor-Bernier, 2014; Gallivan et al., 2018; Rangel & Hare, 2010), where options preserve their relative spatial

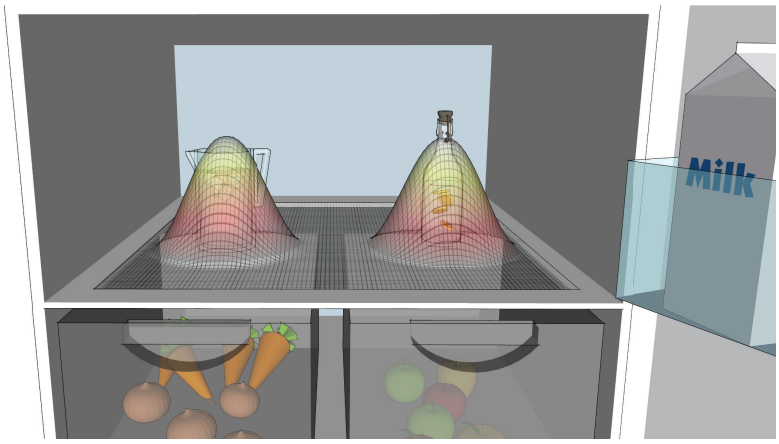

Figure 1.3: A visual depiction of how the brain might use a relevance landscape to represent the value of actions associated with real objects. Here, both a pitcher of iced tea (left) and bottle of sparkling water (right) are desirable drink options on a hot day. According to action-based models, both would have positive neural representations (e.g., hills of activity) in brain areas involved in specifying hand actions.

relation to the deciding agent. For example, when reaching for the pitcher of iced tea or the bottle of sparkling water (Fig. 1.3), both candidate objects would become activated in a map (or, more likely, multiple maps) of space preserving their relationship on the refrigerator shelf. Moreover, this topographic representation, at least in some maps, would contain information about the movements associated with successfully interacting with the object—also known as the object affordance (Gibson, 1986). These affordance competition maps (Cisek, 2007; Cisek & Kalaska, 2010) give rise to what have been called attentional landscapes (Baldauf & Deubel, 2010) or desirability density functions (Cisek & Pastor-Bernier, 2014; Dorris & Glimcher, 2004; Pezzulo & Cisek, 2016). According to action-based models, when one object is chosen, what is being selected is not some abstract representation to which an action then needs to be planned, but rather, some aspect of the action itself. Thus, decision making from an action-based framework can be viewed as representing the value of available actions, which shifts the body through the real world as a means of traversing a landscape of behavioural relevance specified in a neural map (Pezzulo & Cisek, 2016).

Model

In contrast to good-based decision-making models, few computational models within an action-based framework have been proposed. One of the earliest computational models, aimed at explaining observed neural data, consists of layers of simulated neurons in a frontoparietal network (Cisek, 2006). In this model, multiple competing actions are represented in parallel at several levels of the network and compete for selection. This model, while successful at recreating observed neural patterns, does not predict any specifics of how a movement is enacted (Cisek, 2006). A more recent action-based computational model accounts for both neural data and observed reaching movements during decision making (Christopoulos et al., 2015; Christopoulos & Schrater, 2015). This model integrates the value of options and goal-relevant information into a dynamic neural field, which simulates the activity of hundreds of neural populations each tuned to a different direction in space. These directionally tuned neuronal populations compete, and if any population reaches a specified activity threshold, an optimal control policy (Todorov & Jordan, 2002) for reaching in that direction is activated. A weighted average of active policies then determines how the hand moves before the process is updated by a new state of the hand in space. In essence, this model specifies that value influences representations for specific actions, and that option selection, rather than being specified solely before movement, is an outcome of a process that evolves during movement (Christopoulos et al., 2015; Cisek, 2007; Cisek & Kalaska, 2010).

Other models have mostly ignored accounting for neural data and instead, focus on accounting for specific movement features. For example, movement trajectories toward targets and away from obstacles are strikingly similar to a model of attractors and repellers in a dynamical system (Fajen & Warren, 2003). Similarly, attractor landscape models provide an appealing account of how the hand (or computer mouse) moves through space in decision-making tasks (Spivey & Dale, 2006; van der Wel et

al., 2009; Zgonnikov et al., 2017). These models show that action-based mechanisms provide many convincing frameworks to explain how animals decide when moving and move when deciding.

Regarding our two key questions, action-based models postulate that decisions are made through the competition at the level of actions and most of these models state that decisions are complete only when the movement enacting the decision is finished (cf. Erlhagen & Schöner, 2002). By and large, however, these action-based models suffer the opposite problem from that of many good-based models: they rarely explain or even attempt to explain the three classic behavioural hallmarks of decision making—reaction time, accuracy, and confidence—and instead, focus almost entirely on explaining what happens after a movement has been initiated.

Behaviour

In the set of behaviours a decision theory should account for, changes of mind have emerged as a fourth alongside choice accuracy, reaction time, and confidence. A change of mind refers to the infrequent (e.g., 5%) but reliable observation that individuals will sometimes initiate an action toward one choice option, but then switch to another choice option before the action is completed (Burk et al., 2014; Resulaj et al., 2009; Selen et al., 2012; van den Berg et al., 2016, Fig. 1.4a). These changes of mind are overwhelmingly corrective (i.e., they shift actions from incorrect to correct targets), indicating that they are based on a decision process that continues throughout movement (Resulaj et al., 2009)—something outside the scope of most good-based models.

Further evidence for the continued access to, and influence of, decision information on in-flight movements can be seen in what is a now a long list of studies showing the influence of multiple potential targets on both eye (Doyle & Walker, 2001; Ludwig & Gilchrist, 2003) and hand movements (J.-H. Song & Nakayama, 2009; Trommershäuser et al., 2008; Welsh & Elliott, 2004; Welsh et al., 1999). A partic-

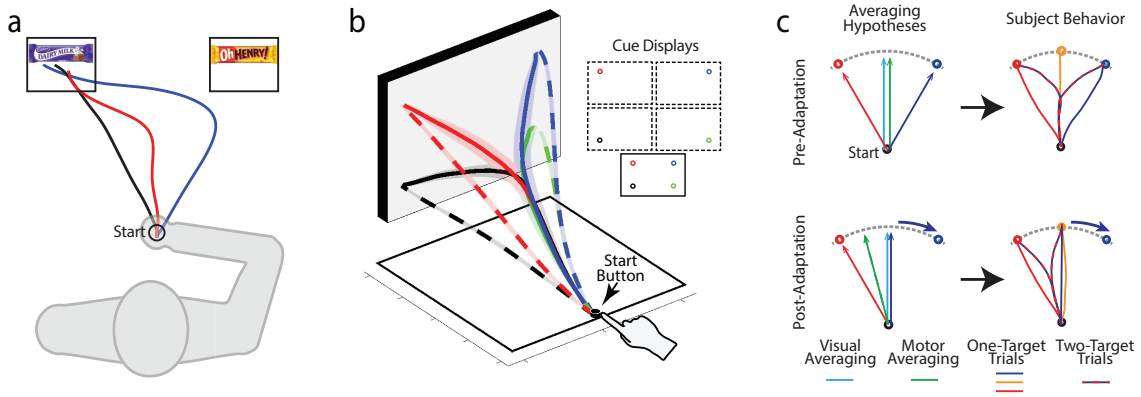

Figure 1.4: Behavioural evidence for an action-based framework of decision making. (a) When choosing between chocolate bars, participants sometimes move straight toward the chosen option (black), sometimes move on an intermediate path between options before committing (red), and sometimes move toward one option before changing their mind and switching to the other option (blue). From unpublished data. (b) In a go-before-you-know task, participants are required to initiate a reach movement toward a cue-display before the final target is revealed (after movement onset). When there is only one potential target presented (dashed traces), the hand moves straight toward its location. When four potential targets are presented (solid traces), the hand initially moves midway between all targets (spatial averaging) before correcting to the cued final target. Adapted from Gallivan and Chapman (2014). (c) To test whether average movements are visual (predicted, blue, left) or motor (predicted, green, left) in nature, participants were gradually adapted to visuomotor rotations, which shifted their hand from a distinct reach to a single right target (blue, top) to a central reach for a single right target (blue, bottom). When simultaneously presented with the left and right targets (dashed, right) after adaptation (bottom), participants reached in a direction that averaged between movements, not visual, directions. Adapted from Gallivan et al. (2017).

ularly acute demonstration can be found in so-called *go-before-you-know* tasks (Fig. 1.4b), wherein participants are required to move before knowing which of several potential targets is the final option (Chapman et al., 2010; Hudson et al., 2007). Under these conditions, participants initially execute an averaged movement between both options before ultimately selecting one. Furthermore, it has been shown that the probability (Hudson et al., 2007), number (Chapman et al., 2010; Gallivan et al., 2011; Milne et al., 2013), spatial arrangement (Chapman et al., 2010; Gallivan & Chapman, 2014), luminance (Wood et al., 2011), reward-association (Chapman, Gallivan, & Enns, 2015; Chapman, Gallivan, Wong, et al., 2015), and symbolic representation (Chapman et al., 2014) of targets all impact rapid reach trajectories.

But what exactly is the nature of option representations that give rise to this behaviour according to action-based theory? Some researchers argue that curved or averaged reach trajectories reveal value represented at the level of possible states associated with movements, which then are used to optimize a single movement control policy (A. L. Wong & Haith, 2017). One such model proposes that effort, accuracy, and evidence for each available option act as inputs along with the state of the arm to form a single optimal control policy (Haith et al., 2015). Other researchers argue that these trajectories reveal the representation of multiple competing motor plans (Chapman et al., 2010; Cisek, 2012). While distinct, both views largely acknowledge that (1) decisions move from a space with many options to an action space with only one eventual movement, (2) ultimate movement output is largely based on the optimization of a single action and not a literal average of simultaneously executed movements, and (3) that fluctuations in the value of multiple and simultaneously held motor representations can influence the single resultant movement.

Thus, one of the most pressing questions facing action-based models is What information is available in parallel motor representations? The above studies show this information reflects both bottom-up (e.g., luminance; Wood et al., 2011) and top-down (e.g., learned reward; Chapman, Gallivan, Wong, et al., 2015) factors. But are

there properties about the details of the movement beyond the spatial (and usually visual) endpoint? To directly dissociate visual target from reach directions, Gallivan et al. (2017) used a visuomotor adaptation task (Fig. 1.4c). Over a series of trials, a gradual, imperceptible rotation of reach direction was applied such that eventually, two targets separated visually by 30° required identical straight-ahead movements. Critically, in go-before-you-know trials toward one adapted target and one nonadapted target, the hand direction followed the motor midpoint (e.g., was shifted by the adapted target's rotation) and not the visual midpoint. Consistent with other studies (Pearce & Moran, 2012; Stewart et al., 2014), these findings support the notion that the brain directly maps visual target locations onto associated motor representations, and uses these to compute initial movements in cases of competing targets. Such a mechanism might support the specification of initial movement directions that minimize the cost of corrected movements to the targets once selected (Haith et al., 2015), thereby reconciling the optimization of motor goals with the averaging of motor representations (Chapman et al., 2010).

A recent study has extended these findings to a go-*after*-you-know task (Gallivan et al., 2015). Here, participants viewed two targets of varying orientation and, when one of the targets was cued, were required to rapidly orient and place the tip of a handheld tool on that target. Movements toward an ambiguously oriented target (i.e., one that could equally be reached via wrist pronation or wrist supination) were biased by the noncued target, more often matching its orientation, even though it was never an explicit movement target. The fact that this "co-optimization" effect emerged in a go-*after*-you-know task suggests that multiple movements (in this case, wrist orientations) were specified in advance of target cueing. This raises the important question: Why would the brain expend its limited resources to directly map competing visual targets onto associated motor representations? According to action-based models, the preparation of multiple potential movement representations might support the rapid execution of any one of the possible movements if required (Cisek, 2007). Results

from the co-optimization experiments support this claim since individuals exhibited faster reaction and movement times on trials in which the co-optimized wrist posture was selected versus trials in which it was not selected (Gallivan et al., 2015; Gallivan et al., 2016).

Action-based models are also consistent with many experiments regarding how motor-related costs factor into decision making. Cos and colleagues provide compelling support for action-based models by showing that when individuals make free choices between two potential reaching movements, which vary in motor-related costs (e.g., energy, stability, distance, etc.), they tend to choose the movements that are biomechanically easiest (Cos et al., 2011) and simplest to control (Cos et al., 2012). Importantly, this indicates that information about the predicted biomechanical costs of both candidate movements is available to the decision making process. Going further, neurostimulation within 200 ms of target presentation suggests a causal role of motor cortex in these rapid, automatic predictions (Cos et al., 2014). Other recent work further shows that the costs associated with motor control bias decision making between actions (Manohar et al., 2015; Morel et al., 2017; Shadmehr et al., 2016). While the impact of motor costs on decision making is not limited to action-based models—for instance, good-based models can account for motor costs through learning or association—the representation of options in a sensorimotor space provides a convincing and direct way for motor information to influence value.

The role of biomechanical costs in decision making has also been extended to changes of mind. Studies show that when the motor costs associated with redirection are increased (through distance, Moher and Song, 2014; or force fields, Burk et al., 2014), changes of mind become more infrequent. Motor costs can even affect perceptual decision making when participants are unaware of them (i.e., when they are introduced very gradually), and these can bias verbal reports of perceptual discriminations, even when they are conveyed through a completely different effector system (Hagura et al., 2017). Together, this work indicates that the motor system,

rather than merely reflecting the output of upstream perceptual processing, can itself influence perceptual processes and the transformation into decision space.

Neural

Unlike the predictions from a good-based framework which argues that option selection precedes action specification (McClelland, 1979; Miller et al., 1960; Padoa-Schioppa, 2011), neural recordings often show parallel action-based representations throughout the decision process (Cisek, 2012; Cisek & Kalaska, 2005, 2010). Several studies of neural responses have shown that before a decision is made, value is not only represented abstractly (Padoa-Schioppa & Assad, 2006), but also with associations to specific actions (often called action-value responses; Sugrue et al., 2005). For example, studies have documented neurons whose firing rates are sensitive to the value of a leftward saccade on each trial (Roitman & Shadlen, 2002). Human fMRI, and NHP and rodent electrophysiological recordings have observed action-value responses in several brain areas including the anterior cingulate cortex (Croxson et al., 2009; Matsumoto et al., 2003; Rushworth et al., 2004), frontal eye fields (FEFs; Gold & Shadlen, 2000), LIP (Dorris & Glimcher, 2004; Louie et al., 2011), striatum (Lau & Glimcher, 2008; Samejima et al., 2005), basal ganglia (Hikosaka et al., 2006), dorsolateral prefrontal cortex (S. Kim et al., 2008), superior colliculus (SC; Horwitz & Newsome, 1999; Ikeda & Hikosaka, 2003; Mysore & Knudsen, 2011), and the supplementary motor area (Sohn & Lee, 2007; Wunderlich et al., 2009).

In a seminal study, when NHPs had to hold in mind two possible reach targets, dorsal premotor neural population activity increased in the directions of both potential targets (Cisek & Kalaska, 2005, Fig. 1.5a). In a more recent extension, it was shown that this activity was also evident if the possible reach directions were specified by rules, rather than spatial targets (Klaes et al., 2011). This activity reflecting multiple motor representations was observed even though the NHPs could have simply waited for the correct option to be cued before representing the single corresponding

movement.

Similar neurophysiological results have also been observed in oculomotor tasks (Basso & Wurtz, 1997; Costello et al., 2013; Keller & McPeek, 2002; McPeek et al., 2003; Munoz & Wurtz, 1995; Platt & Glimcher, 1997; Port & Wurtz, 2003). For example, simultaneous recordings from SC (an oculomotor structure only a few synapses removed from the eye muscles) neurons with nonoverlapping receptive fields mapped the competition between targets and distractors (B. Kim & Basso, 2008, Fig. 1.5b). Specifically, the difference in simultaneous activity between target and distractor-related neurons predicted task accuracy of the NHP. The link between multiple eye movement representations and decision making is even clearer in a study showing that subthreshold stimulation to SC neurons influenced the eventual choice (Thevarajah et al., 2009). These findings appear to directly refute the good-based account (Padoa-Schioppa, 2011) wherein value—the determinant factor of a choice outcome—should not be altered by neural stimulation of a putative motor structure.

Discussion

In many ways, action-based models are the mirror image of good-based models with the reflection point occurring at the moment of movement initiation—they are two halves of the same decision. That is, whereas good-based theories provide convincing mechanisms for decision making during reaction time but lack explanations for much of movement time behaviour, action-based theories tend to lack explanatory richness for reaction time mechanisms but offer compelling explanations for behaviour during movement time. This is highlighted in the key experiments discussed above. Experiments that use ballistic responses such as keypresses or eye movements, which either do not allow for or mitigate post movement decision processes, ultimately force a decision to be resolved entirely during the reaction time period. In our view, this scenario does not reflect the vast majority of evolutionarily old and ecologically valid decisions for which the primate brain is organized—for example, moving through the

world when deciding where to forage. However, in the same way, experiments that force movement initiation so that decisions are resolved entirely during the movement time period are similarly lacking, and again do not reflect the majority of decisions for which the primate brain is organized. Framed this way, it should be clear that regarding the competition between choice options, there is nothing particularly special about the time of movement onset—competition occurs before and continues after movement initiation. Granted, if a movement is very brief (e.g., as in a keypress or eye movement), and reaction time and response time collapse, then the movement is the end of the decision. However, in the real world, where actions enacting a choice are often voluntary and evolve over several hundreds of milliseconds or more, decision making that was initiated when options were presented evolves through reaction time and can continue to unfold during movement. In particular, the sequential sampling of evidence for a decision after stimuli onset (good-based models), and the competition between multiple motor representations during movement (action-based models), while each with their own limitations, may reflect a continuation of the same process.

1.5.5 Bringing two halves together: Decision making as a continuous process

Theories of decision making that cross the boundary between reaction and movement times are beginning to be more prevalent. This shift has been necessitated, in part, by behavioural observations of competition and changes of mind during movement. In fact, a change of mind is precisely the case where a (mostly resolved) competition during reaction time leads to a movement being initiated toward one option, but then further competition during movement time leads to a revised decision. The majority of the models we review here have therefore been concerned with predicting the frequency of a change of mind given the competition evident during reaction time.

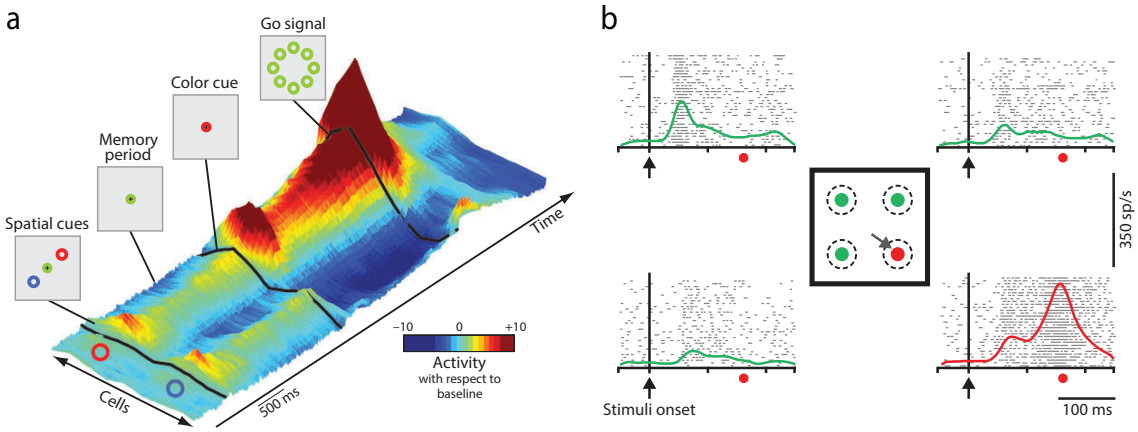

Figure 1.5: Neural evidence for the representation of competing movement options. (a) Population activity in dorsal premotor cortex while NHPs performed a delayed reaching task. Cells sorted by preferred direction along the bottom edge reveal sustained encoding of two potential reach directions when the final direction was unknown, even during a period when the potential targets were not visible. Reproduced from Cisek and Kalaska (2010). (b) Simultaneous recordings of four superior colliculus neurons with receptive fields (dashed circles) for three distractors (green) and one target (red). Each tick represents an action potential, and each row of ticks represents one trial (31 trials total, all correct). The black arrow represents stimuli onset, and the red circle represents the average saccade latency. Spike density functions for each neuron are overlaid on each raster plot. The discriminability of target and distractor neuronal activity was found to predict performance. Adapted from B. Kim and Basso (2008).

Model

Computational models aimed at bridging the gap between pre- and post-movement decision making are relatively recent and comparatively rare. The most prominent of these is the changes-of-mind model (CoMM) by Resulaj et al. (2009), which states that decision making both before and after movement initiation is based on a single, continuous process of evidence accumulation. Like a drift-diffusion model (Ratcliff & Rouder, 1998), the CoMM states that subjects base their decisions on the accumulated difference in evidence between options. When accumulated evidence crosses a decision threshold associated with one of the options, the subject initiates an "initial choice" movement straight toward that option. Unlike other bounded evidence accumulation models, however, evidence sampled just before an initial choice that has not yet been processed continues to accumulate, even during movement. If this post-initiation evidence causes the crossing of a new threshold, the subject changes their mind and begins moving straight toward the other option. This CoMM predicts and explains reaction times, accuracy, and the frequency of changes of mind. It is particularly powerful in that it can explain our flexibility to adapt actions as needed, all while preserving the elegant mathematics of an evidence accumulation process. Furthermore, a recent refinement (van den Berg et al., 2016), adapting a race model (Smith & Vickers, 1988), is able to explain changes in confidence as well.

However, in these models, threshold crossing determines one action, and if another threshold is crossed, another action is selected. This discrete switching between actions cannot explain several highly related behavioural phenomena reviewed above which support action-based models, such as intermediate movement trajectories (Chapman et al., 2010, Fig. 1.4a). Other models have likewise attempted to unify decision making before and after movement initiation by associating an evidence accumulation process with aspects of movement (Friedman et al., 2013; Haith et al., 2015; Lepora & Pezzulo, 2015). While this method has proven successful, several

important behavioural and neural phenomena remain unaccounted for.

Behaviour

Some of the best support for the idea that decision making is a single and continuous process that traverses stimulus presentation to movement completion comes from research that explicitly manipulates the amount of decision information prior to observable behaviour (early work reviewed by Meyer et al., 1988). Approaches to this problem have included fitting to reaction time distributions (as in evidence accumulation models; Ratcliff & Rouder, 1998), analyzing the conjunction of reaction time and kinematic parameters during response time (Abrams & Balota, 1991; Balota & Abrams, 1995), or looking for evidence of motor priming (Bub & Masson, 2012; Rosenbaum & Kornblum, 1982). Here, we focus on experiments that manipulate the speed–accuracy tradeoff. It is well known that increasing the speed of a movement also increases its variability (Fitts, 1966), and increasing the speed of a decision decreases its accuracy (Schouten & Bekker, 1967), suggesting that by forcing participants to respond faster than is natural, they are forced to act with less accumulated evidence.

Ultimately, however, even the analysis of speed–accuracy tradeoffs is somewhat impoverished since changes in accuracy or reaction time, while intimately linked to the amount and quality of evidence accumulated, can also arise for a variety of other reasons (A. L. Wong et al., 2017). To address this limitation, some research directly forces partial information by decoupling the stimulus cueing movement (the imperative stimulus) from the stimuli you are responding to (the test stimulus), thus varying the stimulus–response interval (SR interval; see Fig. 1.6). By using these timing techniques and observing changes in movements, researchers have access to a continuous measure that reflects ongoing decision making started during reaction time (J.-H. Song & Nakayama, 2009). Adapting a rhythmic responding task (Schouten & Bekker, 1967), Ghez et al. developed a *timed response task* where the imperative stimulus was the fourth of four repeated tones and the test stimuli were visual targets

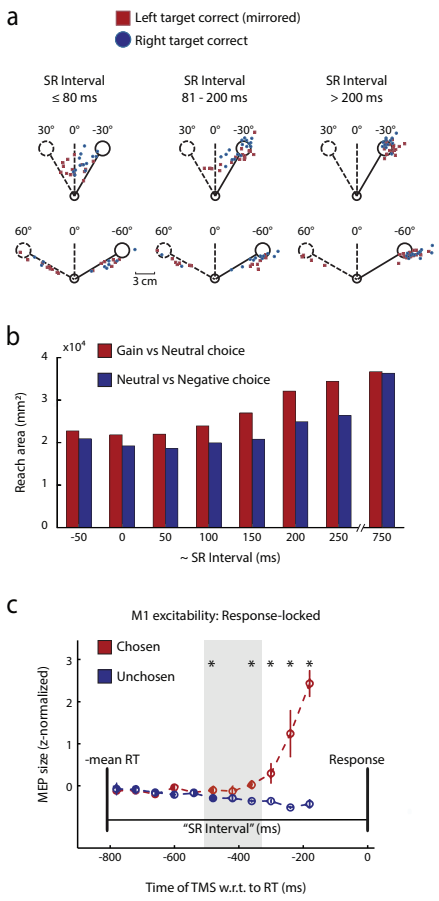

Figure 1.6: Behavioural evidence for the evolution of decision information both before and during movement. (a) Reaches toward unpredictable targets produce trajectories that average between directions when target separation is small (± 30°, top row) and the stimulus–response interval (SRI) is short (left). When the SRI is long, (right) reaches are directly toward the correct target. When target separation is large, (± 60°, bottom row), participants are forced to guess with short SRIs. Movement amplitude (not shown) was also varied, accounting for the observed undershoots. Adapted from Ghez et al. (1997). (b) Participants rapidly reached to choose positive over neutral targets (red) and neutral over negative targets (blue). Depicted is the area between reaches toward correct targets on the left and right (larger area corresponds to straighter reaches). The SRI was approximated by subtracting the average RT (250 ms) from the test stimuli presentation time. Short SRIs led to more competition (curved reaches) and choosing positive targets showed a consistent temporal advantage (straighter reaches) except for the shortest and longest SRIs. Adapted from Chapman, Gallivan, Wong, et al. (2015). (c) Response-locked analysis of changes in primary motor cortex (M1) excitability to transcranial magnetic stimulation (TMS) pulses, measured via normalized motor evoked potentials (MEPs) toward chosen (red) versus unchosen (blue) options. Since the MEP is not an explicit response, the term "SRI" is used, and shows that excitability rises well before response, and is significant (*) even for time windows (gray shading) which isolate decision processes. Adapted from Klein-Flügge and Bestmann (2012).

toward which restricted arm movement responses were required (Ghez et al., 1990). By varying the SR interval, Ghez and colleagues were able to map the evolution of this response (Fig. 1.6a). With less processing time (<80 ms), initial movement directions were averaged between unpredictable targets (akin to go-before-you-know tasks, see Chapman et al., 2010) but with more processing time (>200 ms), responses were more directed toward the correct target. However, these results also demonstrated that the decision between movement targets was influenced by spatial layout (Ghez et al., 1997). If targets were closer together, averaging was more evident and lasted for longer SR intervals. But, if they were further apart, intermediate movements were reduced and even eliminated (Ghez et al., 1997; A. L. Wong & Haith, 2017, Fig. 1.6a). This implies that ongoing decision making must be informed early on by the potential motor consequences for each available option.

Chapman et al. recently extended this technique to explore the temporal evolution of a higher order decision bias between options with positive and negative values. Participants made a rapid reach choice (average RT 250 ms) as soon as they heard an imperative auditory tone (Chapman, Gallivan, Wong, et al., 2015, Fig. 1.6b). Approximate SR intervals ranged from -50 ms (move before test stimuli appear) to 750 milliseconds. These results showed clear evidence of the evolution of a value-based decision—reaches were more curved with less time to process targets. They also demonstrated a clear temporal advantage for processing gains relative to losses. A recent follow-up study has shown that this competition revealed through reaching is prevalent even in relatively slow, self-initiated movements (Wispinski et al., 2017). Furthermore, other research shows that the instantaneous changes in movement angles can reveal how competition between options and sources of decision evidence evolves over time (Scherbaum et al., 2010; Sullivan et al., 2015). In sum, these studies are consistent with competition being initiated during reaction time, but now seeping into response time and affecting movement.

Another tool used to probe the evolving competition between options has been to

force movement initiation via the startle response (Carlsen et al., 2011). In general, a loud auditory tone will elicit an electromyogram (EMG) signal from upper arm muscles after 70 ms and an arm movement after 115 ms—much faster than normal reaction times. When multiple options are available for selection, the startle response reveals clear cases of the representation of multiple options at a motor level (Carlsen et al., 2009; Forgaard et al., 2011). In another line of work, the imperative stimulus was instead a mechanical perturbation causing an elbow extension and a resulting stretch reflex, and the key dependent measure was the EMG of the resulting reflex response as participants performed the RDM task (Selen et al., 2012). Critically, the strength of the reflex (within 75 ms) was sensitive to both the direction and strength of evidence.

Other novel techniques show just how far the competition between options during decision making flows downstream. For example, experimenters read out motor excitability during value-based decision making by applying transcranial magnetic stimulation (TMS) over the motor cortex and measuring the motor evoked potential (MEP; Klein-Flügge & Bestmann, 2012, Fig. 1.6c). By varying the timing of the TMS pulse, they were able to map the evolution of motor excitability during the decision process, concluding that motor excitability scales with decision competition before an action is selected. Finally, Wood et al. (2015) measured intramuscular EMG from pectoral muscles involved in making a reach response to visual targets. Surprisingly, there were spatially sensitive muscle responses less than 100 ms after visual onsets that responded to the luminance contrast of the stimuli. Across these three examples, we see two important points of convergence: first, the readout of the accumulation of evidence toward a decision (motion coherence, value difference, and luminance) was entirely motor (reflex gain, MEP, and muscle response) and second, these responses were graded across time, scaling with the quality of accumulated evidence.

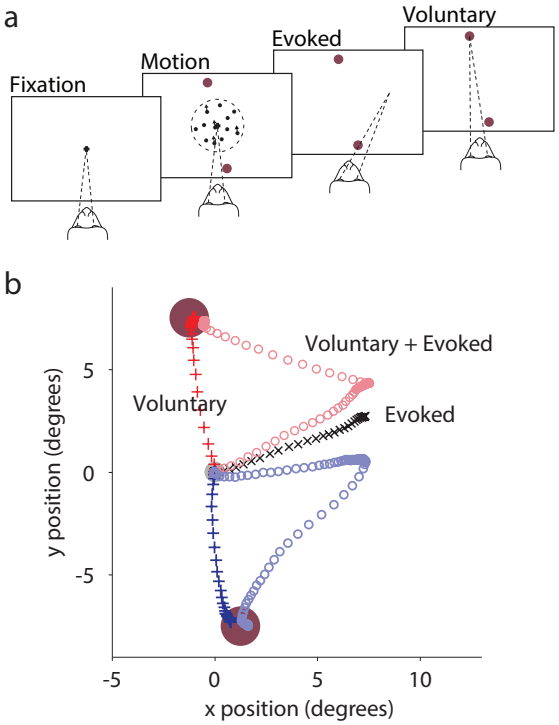

Figure 1.7: Neural evidence that decision formation and motor preparation use the same neural circuits. (a) NHPs performed the RDM task, and made saccades to the target associated with the greater net direction of dot motion (in this case, roughly up/down). (b) Voluntary saccades during this task were directly toward the target corresponding to the direction of perceived dot motion (blue and red crosses). Saccades evoked shortly after fixation using suprathreshold stimulation to FEF neurons resulted in saccades orthogonal to the two targets (black x). When suprathreshold stimulation was applied during dot motion discrimination but before a voluntary saccade, the evoked saccade deviated toward the direction with more dot motion evidence (light blue and red circles). Reproduced from Gold and Shadlen (2000).

Neural

Electrophysiological studies in NHPs parallel the behavioural results above. In one study, when NHPs self-initiated an eye-movement decision during the RDM task, saccades were straight toward the chosen target (Fig. 1.7). In separate trials, suprathreshold stimulation applied to oculomotor regions generated saccades orthogonal to the two targets. But when the same stimulation was applied during decision making but before self-initiation of a saccade, eye movements were a mixture of the orthogonal stimulated direction and the direction of the target with more dot motion evidence. These results have been shown in both the FEF (Gold & Shadlen, 2000) and SC (Thevarajah et al., 2009), and strongly suggest that information about the relative desirability of an option continuously updates circuits implicated in motor processes.

More recently, Kiani, Cueva, et al. (2014) recorded neural population activity from the prearcuate gyrus (a brain area involved in saccade planning; J.-N. Kim & Shadlen, 1999) while monkeys performed the RDM task (Fig. 1.8). By employing a sliding-temporal-window decoding approach prior to launching the decision, they were able to show that they could reliably predict the animal's decision before it was reported via a saccadic eye movement. Notably, by determining how far from the classification boundary (between choice options) the neural state is, the decoding approach can provide a moment-by-moment estimate of the competition between options. Consistent with evidence accumulation, this distance measure gradually increased from zero to large values over the course of the trial, with the rate of rise correlating to the strength of dot motion. Interestingly, however, on a minority of trials, the population response crossed from one side of the decision boundary to the other, suggesting a shift in the animal's choice from one target to the other. These internal changes of mind have the same features as their behavioural counterparts (Resulaj et al., 2009). That is, they were more likely: (1) to occur earlier, rather

than later; (2) for weak than strong stimuli; and (3) to shift from an incorrect to a correct choice.

Similar observations have recently been provided using a reaching task (Kaufman et al., 2015). NHPs were presented with two targets along with virtual barriers that could obstruct a nearby target. By varying when the barriers appeared in a trial, a continuum of situations was constructed ranging from complete free choice (no barriers) to forced choice (only one target was accessible). Of particular interest were cases where a barrier changed mid-trial, making a previously inaccessible target accessible. The investigators recorded neural population activity from the dorsal premotor (PMd) and primary motor (M1) cortices and trained a decoder to categorize the two different responses on "forced choice" trials. Not only could this decoder be used to predict reach direction on "free choice" trials, but also more interestingly, on the barrier-change trials, the decoder would sometimes initially indicate one choice, and then change to the opposite choice. Notably, these neural changes of mind were primarily observed when the animal was presented with free choices, and very rarely occurred on the forced choice trials. In other reaching tasks, neural activity in the PMd appears to represent the relative desirability of multiple potential actions simultaneously (Cisek & Kalaska, 2005; Pastor-Bernier & Cisek, 2011). Additionally, PMd neurons continue to be involved in action selection—if one of the potential options disappears when a "Go" signal is given, PMd activity predicts the switching of action before movement onset (Pastor-Bernier et al., 2012).

Taken together, these results show at the neural level what has been shown at the behavioural level—that the competition between options continuously evolves as a single process throughout decision making. However, these neural studies have yet to show the same results during movement itself, which reveals not only the relative infancy of this research area but also the exciting opportunities to come.

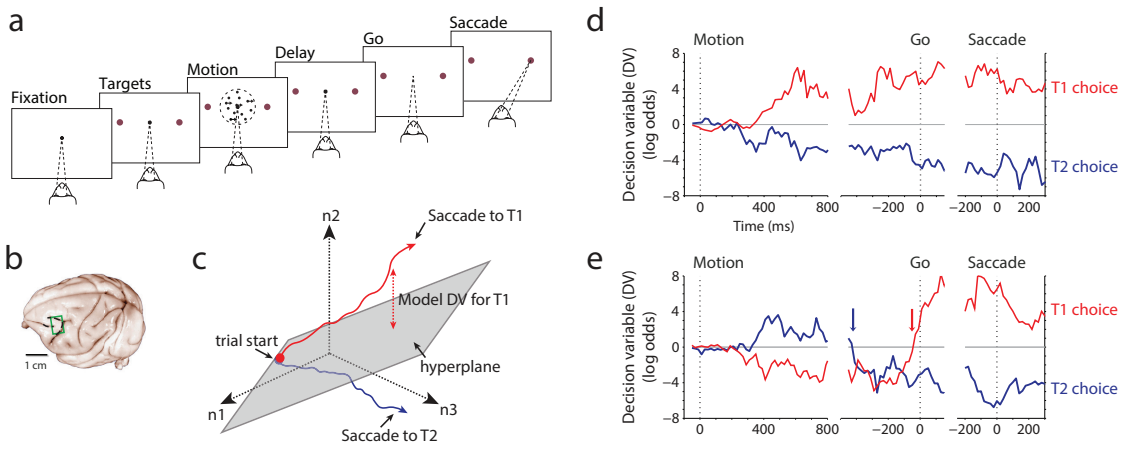

Figure 1.8: Neural evidence for changes of mind. (a) NHPs performed a delayed-response RDM task. (b) Multielectrode arrays recorded neural population activity from area 8Ar of the prearcuate gyrus. (c) The neural population response pattern at each time point can be envisioned as a point in high-dimensional space, whose axes correspond to the firing rates of individual neurons (shown for three hypothetical neurons, n1–n3). Logistic regression was used to find the hyperplane that best discriminated the neural population response patterns corresponding to leftward (T1) versus rightward (T2) choices. The distance of the population response pattern from the hyperplane, or decision variable (DV), indicates the certainty of the model's prediction about the upcoming choice. (d) Two sample trials in which the model DV maintained its sign throughout the trial, ending with T1 and T2 choices. (e) Two sample trials in which the sign of the model DV changed during the delay period (see arrows), indicative of a change in the model's predictions and suggestive of a change of mind in the animal. Adapted from Kiani, Cueva, et al. (2014).

1.5.6 Conclusions and extensions

At the start of our review, we presented two central questions. First, given that decision making is best conceptualized as a competition between choice options, In what representational space do these options compete? Second, especially with respect to the movements required to enact a choice, When are decisions made—before movement onset or at the time of movement completion?

Given the wealth of evidence reviewed here, the answer to the second question appears to strongly favor a decision process that does not end at movement onset. Rather, a convergence of modeling, behavioural, and neural evidence indicates that decision making is a single and gradual process that begins with the presentation (or consideration) of choice options and continues throughout movement execution. This is perhaps most evident in changes of mind or tasks where choice options are not completely defined before movement. In both of these cases and many others, there is clear evidence for the continued contribution of decision making during movement.

In contrast, the first question is still very much up for debate. Here, we presented two somewhat opposing views: first, good-based theories that advocate for the competition of abstract values and, in the extreme, completely separate this value competition from movement consequences, and second, action-based theories, which argue that decisions are fundamentally sensorimotor in nature and, in the extreme, believe decisions are always the resolution of competitions between actions. One alternative to both pure good- and action-based models is a distributed consensus decision-making model (Cisek, 2012). According to this theory, competition occurs at multiple levels of representation, and decisions emerge as the result of reciprocal connections between these distributed competitions. This theory can explain how decisions can be made between both actions and abstract values (Thevarajah et al., 2009). Furthermore, recent evidence suggests that competition can indeed occur at both good- and action-based representations (Chen & Stuphorn, 2015), and

that these might share strong reciprocal connections (Cos et al., 2011; Hagura et al., 2017)—but see Chen and Stuphorn (2015).

This resolution between good- and action-based theories leaves us with a picture of decision making as a dynamic, distributed system across the brain. Most often, choice options are presented via primary sensory inputs and the resolution of a decision results in a motor response. In these situations, cascades of sensory information (usually flowing *up* from sensory to movement/planning areas) flow together with the cascades of task goals (usually flowing *down* from movement/planning to sensory areas) to shape ongoing competition (Siegel et al., 2015). This idea is consistent with continuous cortical feedback where sensory areas are updated so that behaviourally relevant stimuli receive preferential processing as early as possible. For example, the activity in the primary visual cortex (V1) of rodents (Shuler & Bear, 2006), NHPs (Stănişor et al., 2013), and humans (Serences, 2008) is modulated by reward, and presents a likely candidate for the operation of selective attention (Desimone & Duncan, 1995; Stănişor et al., 2013).

In this framework, since most decisions ultimately lead to actions, action-related information and neural structures are usually involved in the milieu of biasing signals. Several researchers have argued against such an architecture, as it might be unnecessarily costly for the brain to continuously transmit such information, or to update motor plans (A. L. Wong & Haith, 2017; A. L. Wong et al., 2015). From these views, a single, central decision system may seem more resource efficient (Padoa-Schioppa, 2011). However, the cost of neural resources might well be worth the benefit of adaptive and flexible behaviour. If decision information is constantly ready to shift our actions, we are able to efficiently adapt to changes in our environments (Cisek, 2012). Ultimately, from many perspectives, the main goal of information processing within the brain is to guide action (Churchland et al., 1993; Cisek, 2007; Clark, 1997; Gibson, 1986; O'Regan & Noë, 2001; Rizzolatti et al., 1997).

The flexibility of such a decision network likely gives rise to its multiple character-

izations. For example, in tasks where a movement is not required, perhaps there is no need for the current decision-network configuration to include motor areas. If so, this could account for results of abstract value, divorced from movement. Similarly, in low-level perceptual decisions (e.g., RDM task) or even in nonconscious movement decisions (e.g., hand preshaping), it is not clear that abstract value is important, and there might be no need for the decision network to engage abstract value structures. Importantly, however, in all decisions, there is a requirement for the system to converge from a space of many options to a single choice, and this convergence evolves over time. Thus, more broadly, and more speculatively, it may be most accurate to say that the brain is a flexible conflict resolving machine, and decision making is one way of studying its capacities. One enticing theory that emerges from this framework is that all cognition is, at its core, reliant on the resolution of competition. We are by no means the first to articulate this kind of position (Shadlen & Kiani, 2013), and it is interesting to consider how memory recall, navigation, or even relevance determination can be conceptualized as the competition between options (with, respectively, candidate memories, possible routes, and decisions themselves being the options that are compared). This idea is not a new one and has some of its earliest origins in seminal writings of William James, who, back in 1890, wrote, "the mind is at every stage a theatre of simultaneous possibilities" (James, 1960).

If decision making is the central function of the brain, many lines of research emerge from its study. Here, we briefly consider two lines of work that appear poised to make real progress. First, by conceptualizing decision making as an evolving process that continues throughout movement, we can better account for sequential decisions. While the vast majority of decision-making tasks study single decisions in isolation, in the real world, the enacting of one decision invariably impacts and leads to subsequent decisions. For example, a prey fleeing from a predator may initially choose to flee toward a tree, but then must decide to climb it, to hide behind it, or to keep running. These subsequent decisions are directly impacted by both the current

environment and the animal's movement through it. Thus, truly sequential decision making appears to be an important next step in decision-making theories. Models like evidence accumulation can likely be extended to show how the evidence from one decision continues to affect not just the movement enacting that decision, but also remains available and biases the next decision (Doll et al., 2012; Murphy et al., 2015; Urai et al., 2019; Zylberberg et al., 2017). Second, social signaling appears to be significantly impacted by the results of the work reviewed. If movement is the result of competition between internally represented choice options, then our movements broadcast our evolving decision process to the world. Others are able to pick up on these decision-making signals simply by observing our movements and are able to use them to guide their own actions (Pesquita et al., 2016; Vaziri-Pashkam et al., 2017). This is a key aspect of body language, gesturing, and coordination, and might have been an important mechanism for the evolution of humans as a social species.

If sequential decision making and social signaling are two questions, we seem better equipped to address, countless other conundrums are enticing and unsolved. For example, How do we account for decisions that do not require an action? What competes during decisions that require the complex coordination of multiple actions and many effectors (Gottlieb, 2012)? What happens during decision making to inhibit an action or to move away from an object (Chapman et al., 2011; Chapman & Goodale, 2008, 2010a, 2010b; Striemer et al., 2009)? And how do we decide when to begin moving? Such questions pose great challenges to current decision-making models and ultimately speak to the difficulty of using neuroscientific techniques and approaches to understand the hidden inner workings of the human mind. Fortunately, this difficulty has only added to the adventure of the expedition.

1.6 A primer on reinforcement learning

We have now discussed adaptive behaviour in biological agents, and the mechanisms that underlie rapid adaptive behaviour—especially with respect to decision making

and motor control. Before turning to discuss adaptive behaviour and mechanisms in artificial agents, some readers may benefit from a short background on reinforcement learning (RL) aimed at biological scientists. Just as many papers in cognitive psychology begin with a quote from William James, many reinforcement learning papers begin with a variant of the following section:

The RL problem is typically illustrated in a figure similar to Fig 1.9 (Sutton & Barto, 2018). Starting on the far left of the figure at time t, the agent receives an observation from the state of the environment (S_t). This can look like an array of numbers that represent the agent's position in a board game. The agent then takes some action (A_t) within the environment. This can look like moving forward one position in the board game. The agent then receives a reward (R_{t+1}) and new observation (S_{t+1}) from the environment. Then the process repeats.

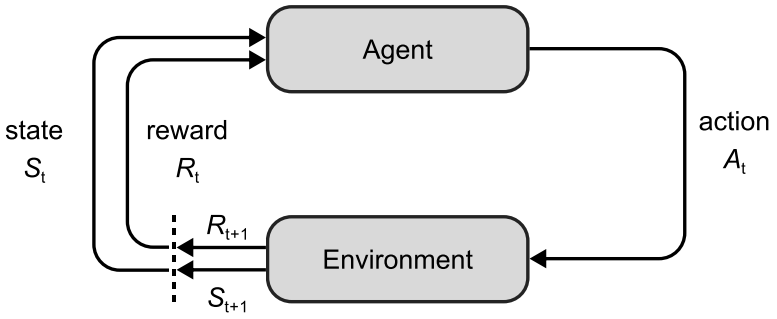

Figure 1.9: (a) Reinforcement learning framework.

In RL, the agent's task is to learn how to maximize the sum of future rewards:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + ... = \sum_{i=0}^{\infty} R_{t+1+i}$$

This is often called the *return* (G_t). In practice, learning to maximize the return is difficult. For instance, in environments with many time steps (i.e., a very long board game), the sum of rewards can reach some very large number. In environments that do not have a formal end point (i.e., a *continuing*, rather than *episodic* task), like managing the electric grid for a city, the return can reach $+\infty$ or $-\infty$. Therefore,

most RL work instead uses the *discounted* return:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{i=0}^{\infty} \gamma^i R_{t+1+i}$$

Where γ is often called the *discount rate*. Discounting applies an exponentially decreasing weight to future rewards, which prevents the sum of rewards from reaching a very large value (providing an easier return to learn to maximize). The discount rate can be set to any value from 0 to 1, with $\gamma = 0$ meaning that the agent would only concern itself with immediate rewards, and $\gamma = 1$ meaning that the agent concerns itself with all rewards in the future equally (i.e., the undiscounted return).

The agent is tasked with maximizing this return by learning an appropriate action policy (π). This policy is some function that gives a probability of choosing an action given the current state ($\pi(a|S_t)$). Ideally, the agent would learn the *optimal* policy (π^*), which when executed would give the maximum achievable return. This optimal policy is known in some reduced settings, like the shortest path through simple mazes or simplistic patch foraging (see Chapter 4), but is unknown in large and complex environments.

In order to learn good policies, many RL algorithms use *value estimates*. For example agents can learn the value of each state under the current policy (π):

$$V_\pi(S_t) = \mathbb{E}_\pi[G_t|S_t]$$

Where $\mathbb{E}$ is the expectation, or mean. In this way, the agent can learn to estimate the expected return under a given policy from a particular state. In other words, the agent can learn the answer to the question, "what is the average discounted return I would expect to get following my action policy from the current state?" Many agents are tasked with both learning to make good value estimates, and using these value estimates to improve their action policies.

In some RL algorithms, agents need to collect a full episode of experience (sometimes called a *trajectory*) from start to end in order to perform a learning update

(i.e., Monte Carlo methods). Biological scientists typically call these "trials" instead of "episodes". However, many researchers would like agents to learn from each new experience as it happens. In addition, some episodes are very long, or never end (i.e., continuing tasks). As such, we often desire agents that can perform *online* learning. Temporal difference (TD) methods offer a solution to this online learning problem by using a technique called bootstrapping. Instead of updating value estimates based on a full trajectory of experienced rewards, value estimates are updated with the current reward, and the future value estimate:

$$\delta_t = R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t)$$

Where δ_t is the TD-error used to update the agent's value estimate ($V_\pi(S_t)$). In other words, the new number that the current value estimate ($V_\pi(S_t)$) should move closer to is the reward that the agent actually received (R_{t+1}) from taking an action in the current state (S_t), plus the agent's expectation of the value on the next state that the agent found itself in after taking this action ($V_\pi(S_{t+1})$), discounted by some factor (γ). Bootstrapping involves updating an estimate based on an estimate, so while it allows for online updating, it introduces bias. However, there are many methods to help mitigate the bias involved in bootstrapping in practice (see Sutton & Barto, 2018).

In some environments, all relevant information about the environment state (S_t) is passed to the agent. For example, in chess, all piece positions on the board are observable by the agent. However, some environments are *partially observable*. That is, the agent is only given partial information about the true state. For example, in the Texas hold 'em variant of poker, your opponent's cards are part of the true and relevant environment state, but are hidden from view. Partial observability is a critical component of most biological tasks, where the true state of the environment is not available to an animal's sensory system or is corrupted by noise. For example, in the random dot motion discrimination task (Chapter 3), the true direction of the

moving dots is hidden from the decision maker, but is partially observable through motion signals in random noise. The agent's partial observation of a state is sometimes denoted by O_t to distinguish it from the environment state, S_t.

RL algorithms can be split into model-based and model-free algorithms. Model-based algorithms are given access to a model of the environment. That is, these models have a function which predicts the transitions between states and their associated rewards. These models are useful for asking questions like "what would happen if I do X?", which the agent can use to guide subsequent behaviour. For example, in robotics tasks, the agent might be given a dynamics model that it can use to simulate movement trajectories. In chess, model-based agents sometimes use tree search to look for the best series of moves, and update the value of states in this tree through experience. In contrast, model-free RL algorithms learn a policy directly without a model of the world. There are many trade offs to consider between model-based and model-free algorithms, which include learning speed (i.e., sample efficiency) and environmental complexity. In practice however, these distinctions between model-based and model-free can become blurred. Throughout this thesis, I deal exclusively with model-free RL algorithms.

Finally, reinforcement learning in artificial agents has always had strong links to biological learning. In fact, *the* book on RL has multiple chapters dedicated to inspirations and parallels with neuroscience and psychology (Sutton & Barto, 2018). As such, RL provides a stellar framework for both building artificial agents, and also better understanding biological systems.

1.7 Why deep reinforcement learning?

While this dissertation looks at adaptive decision making in biological agents and artificial agents, the artificial agents I focus on here are specifically deep reinforcement learning agents. In this section, I would like to take some time to justify this choice.

First, explicitly specifying all the rules for artificial agents is a daunting task,

especially given the incredibly complex, dynamic environments we would like them to operate in (Russell & Norvig, 2021). As such, I turn to popular machine learning approaches instead of symbolic approaches.

Second, we would like agents to learn themselves *how* to behave adaptively, rather than providing the agents many examples of desired adaptive behaviour as in supervised learning. Explicitly specifying desired behaviours potentially limits behavioural solutions to those that are known and able to be specified by human researchers. In addition, the agent's effectors or environment might differ in subtle ways that we as researchers might not anticipate. Agents trained using reinforcement learning are tasked with maximizing reward, and are able to discover novel behaviour in order to do so through trial-and-error interactions with the environment. Further, reinforcement learning has strong links to psychology and neuroscience (Botvinick et al., 2020; Sutton & Barto, 2018), which provides a nice foundation for comparing artificial and biological agents.

So why *deep* reinforcement learning? Reinforcement learning is often implemented using tabular methods, where each unique state (or state-action pair) is given a value within a table, and these values are updated using RL algorithms. Tabular reinforcement learning approaches are powerful, elegant, and have formal convergence guarantees (Sutton & Barto, 2018). However, in some environments, the state and/or action space is large enough for a table of unique states and/or actions to be computationally impractical or impossible. For example, simply representing each unique position in a 32 x 32 discrete grid requires a table of size 1024. Representing each unique position in a 2000 x 2000 discrete grid requires a table of size 4 million. And representing each unique position in a 32 meter x 32 meter continuous arena requires an infinitely large table. In Chapter 3, agents are given video input of size 55 x 55 x 4 on each time step, where most pixels can take either a 1 or 0 value. Some back-of-the-napkin math estimates that there are more than $4.9x10^{79}$ unique possible states in this task[1]. In

[1]Video frames are 55 x 55 pixels (3025 pixels total). Only an inscribed circle within this square

Chapter 4, part of an agent's sensory state are the continuous collision distances of 24 LIDAR rays. Even with floating-point arithmetic to discretize continuous values, there are even more unique sensory state combinations than in Chapter 3. To overcome this limitation, instead of tabular methods we can use function approximators. We could perhaps use linear function approximators, which have been successfully applied to solve complex tasks for which tabular methods are impractical, such as Atari games (Liang et al., 2015). However, linear methods often rely on experimenter-designed features to pre-process raw data. This feature engineering is a science and an art all to itself. In addition, we would like to have agents learn their own features from raw data, which may give some insight into the features that biological agents may use to produce adaptive behaviour in similar environments. Given these goals and restrictions, I use non-linear function approximators (e.g., deep neural networks) for the reinforcement learning agents investigated in this dissertation.

This dissertation also makes use of other artificial decision making agents, such as evidence accumulation models (Ratcliff & McKoon, 2008). These models are much easier to interpret and work with—most have fewer than 10 parameters to fit in comparison with the tens of thousands (Chapter 3), millions (Chapter 4), or even billions (Scao et al., 2022) of free parameters in modern deep learning research. However, while these relatively simple computational models are powerful decision making mechanisms, using these explicitly programmed decision rules to make decisions with large, raw, sensory data is cumbersome and an arguably impractical approach to developing adaptive artificial agents. As such, these models are treated as tools to understand cognition in this dissertation, rather than artificial agents themselves.

can display information in the task; $3025 \times \pi/4 \approx 2375$. On every frame, 7 pixels are white and the rest are black, which gives $2375!/(7! \times (2375-7)!) \approx 8.38 \times 10^{19}$ permutations. There are 4 video frames input to the agent per time step, which gives $(8.38 \times 10^{19})^4 \approx 4.9 \times 10^{79}$ unique combinations. These are only the pixel states of the 55 x 55 x 4 video input—the network also receives input of its previous reward and previous action at every step!

1.8 Adaptation in artificial systems

Recall that a core goal of this thesis is to develop ways in which artificial systems can learn to exhibit adaptive behaviour. The preceding section also outlined the purpose behind my specific focus on deep reinforcement learning agents. Here I would like to take some time to outline some of the ways in which deep reinforcement learning agents have already displayed adaptive behaviour, and the mechanisms by which this is typically achieved. However, this is a very active area of research with many more ideas and techniques than discussed here.

Returning to the discussion on the different forms of adaptive behaviour (see Adaptive behaviour), work in reinforcement learning typically revolves around the second form of adaptive behaviour—learning throughout a single task. Learning from new experiences within a task is a key part of adapting to large, dynamic environments. The process of continual learning (also sometimes called lifelong learning, never-ending learning, or other terms) is based on the idea that agents should be able to robustly add new knowledge and adapt their behaviour because the world is large, dynamic, and continuous. In this area, researchers have developed techniques such as periodic or automatic retraining, regularization, or dynamic network expansion to incorporate new experiences (Parisi et al., 2019).

Agents are also able to adapt to new environments using tools from the area of curriculum learning, where an agent's training experience is systematically varied to produce robust behaviour across a variety of environments. In this area, curricula can be researcher-defined (e.g., Leibo et al., 2019), or automatic (e.g., Open Ended Learning Team et al., 2021; Samvelyan et al., 2023), depending on the goals for the agent and the environment. For example, a curriculum can involve first learning to stand, then to walk, and then to kick a ball, building up to playing a complex game of soccer (Haarnoja et al., 2023).

Other ways of incorporating experience from large dynamic environments are by

learning from the observation of other, potentially more experienced, agents (e.g., as in cultural evolution; Cultural General Intelligence Team et al., 2022). Further, in order to gain more diverse experiences, researchers can develop ways to increase intrinsic motivation or curiosity into agents (Ady et al., 2022), which has been shown to increase metrics of adaptive behaviour (Kauvar et al., 2023). Through this work and more, agents have been able to navigate obstacles in procedurally generated 3D spaces (Cultural General Intelligence Team et al., 2022), adapt to barter with different agents in a marketplace environment (Johanson et al., 2022), and play real-world robotic soccer (Haarnoja et al., 2023). However, learning to adapt to new situations comes with its own host of problems. For example, deep neural networks are prone to catastrophic forgetting, where learning from new examples causes overwriting of old, potentially still-useful information (Kirkpatrick et al., 2017).

In dynamic reinforcement learning environments, such as an n-armed bandit task where the reward probabilities of actions change over time (Daw et al., 2006), continual learning is necessary for agents to continually update their action policy to a changing environment. In situations like these, it has been argued that standard reinforcement learning is too slow for many aspects of adaptive behaviour that we desire in artificial agents. Instead, researchers have argued in favour of alternate frameworks like Bayesian learning (Pearson et al., 2014). This is because standard model-free reinforcement learning agents (see A primer on reinforcement learning) have to update value estimates in a chain backwards in time through repeated experiences, which makes learning long and tedious in large environments. Afterall, if an agent adapts too slowly, it may miss out on exploiting the current environment before the world changes under it. However, advances in model-based reinforcement learning, deep learning, and meta-reinforcement learning allow for much more flexible and rapid updating of behaviour.

In deep neural networks, learning typically takes the form of changing model parameters—weights and biases between neurons in the network. Learning takes place

during training, and then testing or evaluation takes place after learning is stopped and model parameters are frozen (Patterson et al., 2023). However, the system can also learn without changing these parameters—solely through the dynamics of the network. A meta-reinforcement learning (meta-RL) framework allows learning to take place within the dynamics of a system through time without changing model parameters, and is related to dual learning systems in the brain (J. X. Wang et al., 2016). In addition to the usual sensory state, recurrent neural networks are also given access to their own rewards and actions from the previous time step. The idea here is that these are the key ingredients for a reinforcement learning algorithm—the network has access to a previous internal state, the action that it executed, and the reward received from that action execution. In meta-RL, work has shown that agents are able to learn their own reinforcement learning algorithms and adapt to environments even after model parameters are frozen (J. X. Wang et al., 2018; J. X. Wang et al., 2016).

So far, adaptive behaviour discussed in these artificial agents takes place through learning. Learning in meta-RL can be fast and take place within network dynamics, but the other form of learning discussed is slow and only takes place through periodically modifying network parameters. Returning to the discussion on adaptive behaviour (see Adaptive behaviour), agents can also learn adaptive mechanisms implemented in fixed network dynamics that more closely resemble rapid automatic changes in behaviour due to environmental changes. One example of this is training an agent to walk forward in a variety of environments where the agent is exposed to uneven ground or external perturbations. After training, agents are still able to walk forward even when exposed to novel ground conditions or perturbations (Heess et al., 2017). Such agents are able to perform rapid sensorimotor adaptation to novel states through their learned network dynamics. This is exactly the kind of adaptive property we desire—agents that can appropriately update their behaviour on the fly to environmental changes within a trial. We don't want agents that continue to walk

as if no perturbation had taken place; reminiscent of the fixed action patterns seen in greylag geese. While this kind of adaptive behaviour in deep neural networks is becoming more common, it is a significant endevour to evaluate exactly how robust these adaptive behaviours are to all possible environmental changes within a task (Weng et al., 2019). In addition, very little is known about how such behaviours actually work within these agents. Deep neural networks are often viewed as black boxes, which necessitates the field of explainable AI to try to understand these underlying emergent mechanisms (Montavon et al., 2018). Ironically, this begins to mirror problems faced in psychology and neuroscience—to understand the mechanisms underlying behaviour in a complex system (Kietzmann et al., 2017).

1.9 Pushing forward

To recap, this work has two goals—to understand the computational mechanisms of rapid sensorimotor adaptation in biological systems, and to develop ways in which artificial systems can learn these mechanisms themselves. Here in Chapter 1 I discussed how adaptive behaviour can take many forms in biological and artificial agents and some of the ways in which adaptive behaviour is achieved. In the following three chapters, I present work that advances the study of adaptive behaviour. First, I further document the ability of biological agents to make adaptive decisions. Second, I demonstrate two examples of artificial agents which learn to behave adaptively via trial-and-error. Through this work, I argue that the pressure of acting in a dynamic world is a critical factor for the emergence of adaptive mechanisms in artificial agents, and may have been a large factor in how biological systems are able to adapt as well.

1.10 References

Abrams, R. A., & Balota, D. A. (1991). Mental chronometry: Beyond reaction time. *Psychological Science*, *2*(3), 153–157.

Ady, N. M., Shariff, R., Günther, J., & Pilarski, P. M. (2022). Five properties of specific curiosity you didn't know curious machines should have. *arXiv preprint arXiv:2212.00187*.

Andersen, R. A., Essick, G. K., & Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, *230*(4724), 456–458.

Baldauf, D., & Deubel, H. (2010). Attentional landscapes in reaching and grasping. *Vision Research*, *50*(11), 999–1013.

Balota, D. A., & Abrams, R. A. (1995). Mental chronometry: Beyond onset latencies in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1289.

Basso, M. A., & Wurtz, R. H. (1997). Modulation of neuronal activity by target uncertainty. *Nature*, *389*(6646), 66–69.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Probabilistic population codes for bayesian decision making. *Neuron*, *60*(6), 1142–1152.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Bogacz, R. (2007). Optimal decision-making theories: Linking neurobiology with behaviour. *Trends in Cognitive Sciences*, *11*(3), 118–125.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*, *107*(4), 603–616.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, *12*(12), 4745–4765.

Bub, D. N., & Masson, M. E. (2012). On the dynamics of action representations evoked by names of manipulable objects. *Journal of Experimental Psychology: General*, *141*(3), 502.

Burk, D., Ingram, J. N., Franklin, D. W., Shadlen, M. N., & Wolpert, D. M. (2014). Motor effort alters changes of mind in sensorimotor decision making. *PloS One*, *9*(3), e92681.

Carlsen, A. N., Chua, R., Summers, J. J., Inglis, J. T., Sanderson, D. J., & Franks, I. M. (2009). Precues enable multiple response preprogramming: Evidence from startle. *Psychophysiology*, *46*(2), 241–251.

Carlsen, A. N., Maslovat, D., Lam, M. Y., Chua, R., & Franks, I. M. (2011). Considerations for the use of a startling acoustic stimulus in studies of motor

preparation in humans. *Neuroscience & Biobehavioral Reviews*, *35*(3), 366–376.

Chapman, C. S., Gallivan, J. P., Culham, J. C., & Goodale, M. A. (2011). Mental blocks: Fmri reveals top-down modulation of early visual cortex when obstacles interfere with grasp planning. *Neuropsychologia*, *49*(7), 1703–1717.

Chapman, C. S., Gallivan, J. P., & Enns, J. T. (2015). Separating value from selection frequency in rapid reaching biases to visual targets. *Visual Cognition*, *23*(1-2), 249–271.

Chapman, C. S., Gallivan, J. P., Wong, J. D., Wispinski, N. J., & Enns, J. T. (2015). The snooze of lose: Rapid reaching reveals that losses are processed more slowly than gains. *Journal of Experimental Psychology: General*, *144*(4), 844.

Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Ansari, D., Culham, J. C., & Goodale, M. A. (2014). Counting on the motor system: Rapid action planning reveals the format-and magnitude-dependent extraction of numerical quantity. *Journal of Vision*, *14*(3), 30–30.

Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Culham, J. C., & Goodale, M. A. (2010). Reaching for the unknown: Multiple target encoding and real-time decision-making in a rapid reach task. *Cognition*, *116*(2), 168–176.

Chapman, C. S., & Goodale, M. A. (2008). Missing in action: The effect of obstacle position and size on avoidance while reaching. *Experimental Brain Research*, *191*(1), 83–97.

Chapman, C. S., & Goodale, M. A. (2010a). Obstacle avoidance during online corrections. *Journal of Vision*, *10*(11), 17–17.

Chapman, C. S., & Goodale, M. A. (2010b). Seeing all the obstacles in your way: The effect of visual feedback and visual feedback schedule on obstacle avoidance while reaching. *Experimental Brain Research*, *202*(2), 363–375.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Charnov, E. L., & Orians, G. H. (2006). Optimal foraging: Some theoretical explorations.

Chen, X., & Stuphorn, V. (2015). Sequential selection of economic good and action in medial frontal cortex of macaques during value-based decisions. *eLife*, *4*, e09418.

Christopoulos, V., Bonaiuto, J., & Andersen, R. A. (2015). A biologically plausible computational theory for value integration and action selection in decisions with competing alternatives. *PLoS Computational Biology*, *11*(3), e1004104.

Christopoulos, V., & Schrater, P. R. (2015). Dynamic integration of value information into a common probability currency as a theory for flexible decision making. *PLoS Computational Biology*, *11*(9), e1004402.

Churchland, P. S., Ramachandran, V. S., & Sejnowski, T. J. (1993). A critique of pure vision.

Cisek, P. (2006). Integrated neural processes for defining potential actions and deciding between them: A computational model. *Journal of Neuroscience*, *26*(38), 9761–9770.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1585–1599.

Cisek, P. (2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology*, *22*(6), 927–936.

Cisek, P., & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron*, *45*(5), 801–814.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, *33*, 269–298.

Cisek, P., & Pastor-Bernier, A. (2014). On the challenges and mechanisms of embodied decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130479.

Cisek, P., Puskas, G. A., & El-Murr, S. (2009). Decisions in changing conditions: The urgency-gating model. *Journal of Neuroscience*, *29*(37), 11560–11571.

Clark, A. (1997). The dynamical challenge. *Cognitive Science*, *21*(4), 461–481.

Cos, I., Bélanger, N., & Cisek, P. (2011). The influence of predicted arm biomechanics on decision making. *Journal of Neurophysiology*, *105*(6), 3022–3033.

Cos, I., Duque, J., & Cisek, P. (2014). Rapid prediction of biomechanical costs during action decisions. *Journal of Neurophysiology*, *112*(6), 1256–1266.

Cos, I., Medleg, F., & Cisek, P. (2012). The modulatory influence of end-point controllability on decisions between actions. *Journal of Neurophysiology*, *108*(6), 1764–1780.

Costello, M. G., Zhu, D., Salinas, E., & Stanford, T. R. (2013). Perceptual modulation of motor—but not visual—responses in the frontal eye field during an urgent-decision task. *Journal of Neuroscience*, *33*(41), 16394–16408.

Cowie, R. J. (1977). Optimal foraging in great tits (Parus major). *Nature*, *268*(5616), 137–139.

Croxson, P. L., Walton, M. E., O'Reilly, J. X., Behrens, T. E., & Rushworth, M. F. (2009). Effort-based cost–benefit valuation and the human brain. *Journal of Neuroscience*, *29*(14), 4531–4541.

Cultural General Intelligence Team, Bhoopchand, A., Brownfield, B., Collister, A., Lago, A. D., Edwards, A., Everett, R., Frechette, A., Oliveira, Y. G., Hughes, E., Mathewson, K. W., Mendolicchio, P., Pawar, J., Pislar, M., Platonov, A., Senter, E., Singh, S., Zacherl, A., & Zhang, L. M. (2022). Learning robust real-time cultural transmission without human data. *arXiv preprint arXiv.2203.00715*.

Darwin, C. (1872). *The descent of man, and selection in relation to sex* (Vol. 2). D. Appleton.

Davidson, J. D., & El Hady, A. (2019). Foraging as an evidence accumulation process. *PLoS Computational Biology*, *15*(7), e1007060.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

De Martino, B., Kumaran, D., Holt, B., & Dolan, R. J. (2009). The neurobiology of reference-dependent value computation. *Journal of Neuroscience*, *29*(12), 3833–3842.

Delsuc, F. (2003). Army ants trapped by their evolutionary history. *PLoS Biology*, *1*(2), e37.

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

Devlin, S., Georgescu, R., Momennejad, I., Rzepecki, J., Zuniga, E., Costello, G., Leroy, G., Shaw, A., & Hofmann, K. (2021). Navigation turing test (NTT): Learning to evaluate human-like navigation, In *International Conference on Machine Learning*.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081.

Donner, T. H., Siegel, M., Fries, P., & Engel, A. K. (2009). Buildup of choice-predictive activity in human motor cortex during perceptual decision making. *Current Biology*, *19*(18), 1581–1585.

Dorris, M. C., & Glimcher, P. W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron*, *44*(2), 365–378.

Doyle, M., & Walker, R. (2001). Curved saccade trajectories: Voluntary and reflexive saccades curve away from irrelevant distractors. *Experimental Brain Research*, *139*(3), 333–344.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*(11), 3612–3628.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545.

Fajen, B. R., & Warren, W. H. (2003). Behavioral dynamics of steering, obstable avoidance, and route selection. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 343.

Fitts, P. M. (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Psychology*, *71*(6), 849.

Fooken, J., Yeo, S.-H., Pai, D. K., & Spering, M. (2016). Eye movement accuracy determines natural interception strategies. *Journal of Vision*, *16*(14), 1–1.

Forgaard, C. J., Maslovat, D., Carlsen, A. N., & Franks, I. M. (2011). Default motor preparation under conditions of response uncertainty. *Experimental Brain Research*, *215*, 235–245.

Freedman, D. J., & Assad, J. A. (2011). A proposed common neural mechanism for categorization and perceptual decisions. *Nature Neuroscience*, *14*(2), 143–146.

Friedman, J., Brown, S., & Finkbeiner, M. (2013). Linking cognitive and reaching trajectories via intermittent movement control. *Journal of Mathematical Psychology*, *57*(3-4), 140–151.

Gallivan, J. P., Barton, K. S., Chapman, C. S., Wolpert, D. M., & Randall Flanagan, J. (2015). Action plan co-optimization reveals the parallel encoding of competing reach movements. *Nature Communications*, *6*(1), 7428.

Gallivan, J. P., & Chapman, C. S. (2014). Three-dimensional reach trajectories as a probe of real-time decision-making between multiple competing targets. *Frontiers in Neuroscience*, *8*, 215.

Gallivan, J. P., Chapman, C. S., Wolpert, D. M., & Flanagan, J. R. (2018). Decision-making in sensorimotor control. *Nature Reviews Neuroscience*, *19*(9), 519–534.

Gallivan, J. P., Chapman, C. S., Wood, D. K., Milne, J. L., Ansari, D., Culham, J. C., & Goodale, M. A. (2011). One to four, and nothing more: Nonconscious parallel individuation of objects during action planning. *Psychological Science*, *22*(6), 803–811.

Gallivan, J. P., Logan, L., Wolpert, D. M., & Flanagan, J. R. (2016). Parallel specification of competing sensorimotor control policies for alternative action options. *Nature Neuroscience*, *19*(2), 320.

Gallivan, J. P., Stewart, B. M., Baugh, L. A., Wolpert, D. M., & Flanagan, J. R. (2017). Rapid automatic motor encoding of competing reach options. *Cell Reports*, *18*(7), 1619–1626.

Ghez, C., Favilla, M., Ghilardi, M., Gordon, J., Bermejo, R., & Pullman, S. (1997). Discrete and continuous planning of hand movements and isometric force trajectories. *Experimental Brain Research*, *115*, 217–233.

Ghez, C., Gordon, J., Ghilardi, M., Christakos, C., & Cooper, S. (1990). Roles of proprioceptive input in the programming of arm trajectories, In *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Psychology Press.

Glimcher, P. W. (2010). *Foundations of neuroeconomic analysis*. Oxford University Press.

Gold, J. I., & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, *404*(6776), 390–394.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

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

Grattan, L. E., & Glimcher, P. W. (2014). Absence of spatial tuning in the orbitofrontal cortex. *PLoS One*, *9*(11), e112750.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Haarnoja, T., Moran, B., Lever, G., Huang, S. H., Tirumala, D., Wulfmeier, M., Humplik, J., Tunyasuvunakool, S., Siegel, N. Y., Hafner, R., Et al. (2023). Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *arXiv preprint arXiv:2304.13653*.

Hagura, N., Haggard, P., & Diedrichsen, J. (2017). Perceptual decisions are biased by the cost to act. *eLife*, *6*, e18422.

Haith, A. M., Huberdeau, D. M., & Krakauer, J. W. (2015). Hedging your bets: Intermediate movements as optimal behavior in the context of an incomplete decision. *PLoS Computational Biology*, *11*(3), e1004171.

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience*, *9*(5), 682–689.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *Journal of Neuroscience*, *35*(6), 2476–2484.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*(7), 933–939.

Heess, N., Tb, D., Sriram, S., Lemmon, J., Merel, J., Wayne, G., Tassa, Y., Erez, T., Wang, Z., Eslami, S., Et al. (2017). Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*.

Hikosaka, O., Nakamura, K., & Nakahara, H. (2006). Basal ganglia orient eyes to reward. *Journal of Neurophysiology*, *95*(2), 567–584.

Horwitz, G. D., & Newsome, W. T. (1999). Separate signals for target selection and movement specification in the superior colliculus. *Science*, *284*(5417), 1158–1161.

Hudson, T. E., Maloney, L. T., & Landy, M. S. (2007). Movement planning with probabilistic target information. *Journal of Neurophysiology*, *98*(5), 3034–3046.

Huk, A. C., & Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *Journal of Neuroscience*, *25*(45), 10420–10436.

Ikeda, T., & Hikosaka, O. (2003). Reward-dependent gain and bias of visual responses in primate superior colliculus. *Neuron*, *39*(4), 693–700.

James, W. (1960). *The principles of psychology* (Vol. 1). Henry Holt and Company.

Johanson, M. B., Hughes, E., Timbers, F., & Leibo, J. Z. (2022). Emergent bartering behaviour in multi-agent reinforcement learning. *arXiv preprint arXiv:2205.06760*.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633.

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*(7611), 285–288.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2015). Vacillation, indecision and hesitation in moment-by-moment decoding of monkey motor cortex. *eLife*, *4*, e04677.

Kauvar, I., Doyle, C., Zhou, L., & Haber, N. (2023). Curious replay for model-based adaptation, In *International Conference on Machine Learning*.

Keller, E. L., & McPeek, R. M. (2002). Neural discharge in the superior colliculus during target search paradigms. *Annals of the New York Academy of Sciences*, *956*(1), 130–142.

Kennerley, S. W., & Wallis, J. D. (2009). Encoding of reward and space during a working memory task in the orbitofrontal cortex and anterior cingulate sulcus. *Journal of Neurophysiology*.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342.

Kiani, R., Cueva, C. J., Reppas, J. B., & Newsome, W. T. (2014). Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. *Current Biology*, *24*(13), 1542–1547.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017). Deep neural networks in computational neuroscience. *bioRxiv*, 133504.

Kim, B., & Basso, M. A. (2008). Saccade target selection in the superior colliculus: A signal detection theory approach. *Journal of Neuroscience*, *28*(12), 2991–3007.

Kim, J.-N., & Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, *2*(2), 176–185.

Kim, S., Hwang, J., & Lee, D. (2008). Prefrontal coding of temporally discounted values during intertemporal choice. *Neuron*, *59*(1), 161–172.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526.

Klaes, C., Westendorff, S., Chakrabarti, S., & Gail, A. (2011). Choosing goals, not rules: Deciding among rule-based action plans. *Neuron*, *70*(3), 536–548.

Klein-Flügge, M. C., & Bestmann, S. (2012). Time-dependent changes in human corticospinal excitability reveal value-based competition for action during decision processing. *Journal of Neuroscience*, *32*(24), 8373–8382.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857.

Krakauer, J. W., Hadjiosif, A. M., Xu, J., Wong, A. L., & Haith, A. M. (2019). Motor learning. *Comprehensive Physiology*, *9*(2), 613–663.

Krebs, J. R., Ryan, J. C., & Charnov, E. L. (1974). Hunting by expectation or optimal foraging? A study of patch use by chickadees. *Animal Behaviour*, *22*, 953–IN3.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Krotkov, E., Hackett, D., Jackel, L., Perschbacher, M., Pippine, J., Strauss, J., Pratt, G., & Orlowski, C. (2018). The darpa robotics challenge finals: Results and perspectives. *The DARPA robotics challenge finals: Humanoid robots to the rescue*, 1–26.

Krueger, P. M., van Vugt, M. K., Simen, P., Nystrom, L., Holmes, P., & Cohen, J. D. (2017). Evidence accumulation detected in bold signal using slow perceptual decision making. *Journal of Neuroscience Methods*, *281*, 21–32.

Lau, B., & Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, *58*(3), 451–463.

Leibo, J. Z., Hughes, E., Lanctot, M., & Graepel, T. (2019). Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*.

Lepora, N. F., & Pezzulo, G. (2015). Embodied choice: How action influences perceptual decision making. *PLoS Computational Biology*, *11*(4), e1004110.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038.

Liang, Y., Machado, M. C., Talvitie, E., & Bowling, M. (2015). State of the art control of atari games using shallow reinforcement learning. *arXiv preprint arXiv:1512.01563*.

Licata, A. M., Kaufman, M. T., Raposo, D., Ryan, M. B., Sheppard, J. P., & Churchland, A. K. (2017). Posterior parietal cortex guides visual decisions in rats. *Journal of Neuroscience*, *37*(19), 4954–4966.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Link, S. W. (1975). The relative judgment theory of two choice response time. *Journal of Mathematical Psychology*, *12*(1), 114–135.

Lorenz, K., & Tinbergen, N. (1938). Taxis und instinkthandlung in der eirollbewegung der graugans. *Zeitschrift für Tierpsychologie*.

Lottem, E., Banerjee, D., Vertechi, P., Sarra, D., oude Lohuis, M. N., & Mainen, Z. F. (2018). Activation of serotonin neurons promotes active persistence in a probabilistic foraging task. *Nature Communications*, *9*(1), 1–12.

Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward value-based gain control: Divisive normalization in parietal cortex. *Journal of Neuroscience*, *31*(29), 10627–10639.

Ludwig, C. J., & Gilchrist, I. D. (2003). Target similarity affects saccade curvature away from irrelevant onsets. *Experimental Brain Research*, *152*, 60–69.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

Manohar, S. G., Chong, T. T.-J., Apps, M. A., Batla, A., Stamelou, M., Jarman, P. R., Bhatia, K. P., & Husain, M. (2015). Reward pays the cost of noise reduction in motor and cognitive control. *Current Biology*, *25*(13), 1707–1716.

Matsumoto, K., Suzuki, W., & Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, *301*(5630), 229–232.

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*(4), 287.

McPeek, R. M., Han, J. H., & Keller, E. L. (2003). Competition between saccade goals in the superior colliculus produces saccade curvature. *Journal of Neurophysiology*, *89*(5), 2577–2590.

Megaw, E. (1974). Possible modification to a rapid on-going programmed manual response. *Brain Research*, *71*(2-3), 425–441.

Meyer, D. E., Irwin, D. E., Osman, A. M., & Kounois, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. *Psychological Review*, *95*(2), 183.

Milani, S., Juliani, A., Momennejad, I., Georgescu, R., Rzepecki, J., Shaw, A., Costello, G., Fang, F., Devlin, S., & Hofmann, K. (2023). Navigates like me: Under-

standing how people evaluate human-like AI in video games, In *Conference on Human Factors in Computing Systems*.

Miller, G. A., Eugene, G., & Pribram, K. H. (1960). *Plans and the structure of behaviour*. Henry Holt and Company.

Milne, J. L., Chapman, C. S., Gallivan, J. P., Wood, D. K., Culham, J. C., & Goodale, M. A. (2013). Connecting the dots: Object connectedness deceives perception but not movement planning. *Psychological Science*, *24*(8), 1456–1465.

Minsky, M. (1988). *Society of mind*. Simon and Schuster.

Moher, J., & Song, J.-H. (2014). Perceptual decision processes flexibly adapt to avoid change-of-mind motor costs. *Journal of Vision*, *14*(8), 1–1.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Morel, P., Ulbrich, P., & Gail, A. (2017). What makes a reach movement effortful? Physical effort discounting supports common minimization principles in decision making and motor control. *PLoS Biology*, *15*(6), e2001323.

Munoz, D. P., & Wurtz, R. H. (1995). Saccade-related activity in monkey superior colliculus. I. Characteristics of burst and buildup cells. *Journal of Neurophysiology*, *73*(6), 2313–2333.

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife*, *4*, e11946.

Mysore, S. P., & Knudsen, E. I. (2011). The role of a midbrain network in competitive stimulus selection. *Current Opinion in Neurobiology*, *21*(4), 653–660.

Newsome, W. T., & Paré, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience*, *8*(6), 2201–2211.

Noorani, I., & Carpenter, R. (2016). The LATER model of reaction time and decision. *Neuroscience & Biobehavioral Reviews*, *64*, 229–251.

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729–1735.

Open Ended Learning Team, Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., Et al. (2021). Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973.

Pacheco-Cobos, L., Winterhalder, B., Cuatianquiz-Lima, C., Rosetti, M. F., Hudson, R., & Ross, C. T. (2019). Nahua mushroom gatherers use area-restricted search strategies that conform to marginal value theorem predictions. *Proceedings of the National Academy of Sciences*, *116*(21), 10339–10347.

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: A good-based model. *Annual Review of Neuroscience*, *34*, 333–359.

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223–226.

Padoa-Schioppa, C., & Conen, K. E. (2017). Orbitofrontal cortex: A neural circuit for economic decisions. *Neuron*, *96*(4), 736–754.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 1–1.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, *113*, 54–71.

Pastor-Bernier, A., & Cisek, P. (2011). Neural correlates of biased competition in premotor cortex. *Journal of Neuroscience*, *31*(19), 7083–7088.

Pastor-Bernier, A., Tremblay, E., & Cisek, P. (2012). Dorsal premotor cortex is involved in switching motor plans. *Frontiers in Neuroengineering*, *5*, 5.

Patterson, A., Neumann, S., White, M., & White, A. (2023). Empirical design in reinforcement learning. *arXiv preprint arXiv:2304.01315*.

Pearce, T. M., & Moran, D. W. (2012). Strategy-dependent encoding of planned arm movements in the dorsal premotor cortex. *Science*, *337*(6097), 984–988.

Pearson, J. M., Watson, K. K., & Platt, M. L. (2014). Decision making: The neuroethological turn. *Neuron*, *82*(5), 950–965.

Pesquita, A., Chapman, C. S., & Enns, J. T. (2016). Humans are sensitive to attention control when predicting others' actions. *Proceedings of the National Academy of Sciences*, *113*(31), 8669–8674.

Pezzulo, G., & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences*, *20*(6), 414–424.

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience*, *27*(37), 9984–9988.

Platt, M. L., & Glimcher, P. W. (1997). Responses of intraparietal neurons to saccadic targets and visual distractors. *Journal of Neurophysiology*, *78*(3), 1574–1589.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864.

Port, N. L., & Wurtz, R. H. (2003). Sequential activity of simultaneously recorded neurons in the superior colliculus during curved saccades. *Journal of Neurophysiology*, *90*(3), 1887–1903.

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, *20*(2), 262–270.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266.

Rideaux, R., & Welchman, A. E. (2020). But still it moves: Static image statistics underlie how we see motion. *Journal of Neuroscience*, *40*(12), 2538–2552.

Rizzolatti, G., Fogassi, L., & Gallese, V. (1997). Parietal cortex: From sight to action. *Current Opinion in Neurobiology*, *7*(4), 562–567.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, *22*(21), 9475–9489.

Rosenbaum, D. A., & Kornblum, S. (1982). A priming method for investigating the selection of motor responses. *Acta Psychologica*, *51*(3), 223–243.

Rushworth, M., Walton, M. E., Kennerley, S. W., & Bannerman, D. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*(9), 410–417.

Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach. *University of California, Berkeley*.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337–1340.

Samvelyan, M., Khan, A., Dennis, M., Jiang, M., Parker-Holder, J., Foerster, J., Raileanu, R., & Rocktäschel, T. (2023). MAESTRO: Open-ended environment design for multi-agent reinforcement learning. *arXiv preprint arXiv:2303.03376*.

Sarlegna, F. R., & Mutha, P. K. (2015). The influence of visual target information on the online control of movements. *Vision Research*, *110*, 144–154.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Scherbaum, S., Dshemuchadse, M., Fischer, R., & Goschke, T. (2010). How decisions evolve: The temporal dynamics of action selection. *Cognition*, *115*(3), 407–416.

Schouten, J., & Bekker, J. (1967). Reaction time and accuracy. *Acta Psychologica*, *27*, 143–153.

Scialfa, C. T. (2002). The role of sensory factors in cognitive aging research. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *56*(3), 153.

Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience*, *5*(7), 532–545.

Selen, L. P., Shadlen, M. N., & Wolpert, D. M. (2012). Deliberation in the motor system: Reflex gains track evolving evidence leading to a decision. *Journal of Neuroscience*, *32*(7), 2276–2286.

Serences, J. T. (2008). Value-based modulations in human visual cortex. *Neuron*, *60*(6), 1169–1181.

Shadlen, M. N., Hanks, T. D., Churchland, A. K., Kiani, R., & Yang, T. (2006). The speed and accuracy of a simple perceptual decision: A mathematical primer. *Bayesian brain: Probabilistic approaches to neural coding*, 209–37.

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791–806.

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: Seeing and deciding. *Proceedings of the National Academy of Sciences*, *93*(2), 628–633.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939.

Shadmehr, R., Huang, H. J., & Ahmed, A. A. (2016). A representation of effort in decision-making and motor control. *Current Biology*, *26*(14), 1929–1934.

Shuler, M. G., & Bear, M. F. (2006). Reward timing in the primary visual cortex. *Science*, *311*(5767), 1606–1609.

Siegel, M., Buschman, T. J., & Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science*, *348*(6241), 1352–1355.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, *32*(2), 135–168.

Sohn, J.-W., & Lee, D. (2007). Order-dependent modulation of directional signals in the supplementary and presupplementary motor areas. *Journal of Neuroscience*, *27*(50), 13655–13666.

Song, H. F., Yang, G. R., & Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, *6*, e21492.

Song, J.-H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, *13*(8), 360–366.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*(5), 207–211.

Stănişor, L., van der Togt, C., Pennartz, C. M., & Roelfsema, P. R. (2013). A unified selection signal for attention and reward in primary visual cortex. *Proceedings of the National Academy of Sciences*, *110*(22), 9136–9141.

Stephens, D. W., & Krebs, J. R. (2019). *Foraging theory*. Princeton University Press.

Stewart, B. M., Gallivan, J. P., Baugh, L. A., & Flanagan, J. R. (2014). Motor, not visual, encoding of potential reach targets. *Current Biology*, *24*(19), R953–R954.

Striemer, C. L., Chapman, C. S., & Goodale, M. A. (2009). "Real-time" obstacle avoidance in the absence of primary visual cortex. *Proceedings of the National Academy of Sciences*, *106*(37), 15996–16001.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*(5), 363–375.

Sullivan, N., Hutcherson, C., Harris, A., & Rangel, A. (2015). Dietary self-control is related to the speed with which attributes of healthfulness and tastiness are processed. *Psychological Science*, *26*(2), 122–134.

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

Thevarajah, D., Mikulić, A., & Dorris, M. C. (2009). Role of the superior colliculus in choosing mixed-strategy saccades. *Journal of Neuroscience*, *29*(7), 1998–2008.

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: Theory and experimental support. *Journal of Neurophysiology*, *108*(11), 2912–2930.

Thura, D., & Cisek, P. (2014). Deliberation and commitment in the premotor and primary motor cortex during dynamic decision making. *Neuron*, *81*(6), 1401–1416.

Thura, D., & Cisek, P. (2016). On the difference between evidence accumulator models and the urgency gating model. *Journal of Neurophysiology*, *115*(1), 622–623.

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, *5*(11), 1226–1235.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, *12*(8), 291–297.

Tunyasuvunakool, S., Muldal, A., Doron, Y., Liu, S., Bohez, S., Merel, J., Erez, T., Lillicrap, T. P., Heess, N., & Tassa, Y. (2020). dm_control: Software and tasks for continuous control. *Software Impacts*, *6*, 100022.

Urai, A. E., De Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *eLife*, *8*, e46331.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, *5*, e12192.

van der Wel, R. P., Eder, J. R., Mitchel, A. D., Walsh, M. M., & Rosenbaum, D. A. (2009). Trajectories emerging from discrete versus continuous processing models in phonological competitor tasks: A commentary on Spivey, Grosjean, and Knoblich. *Journal of Experimental Psychology: Human Perception and Performance*.

Vaziri-Pashkam, M., Cormiea, S., & Nakayama, K. (2017). Predicting actions from subtle preparatory movements. *Cognition*, *168*, 65–75.

Vertechi, P., Lottem, E., Sarra, D., Godinho, B., Treves, I., Quendera, T., oude Lohuis, M. N., & Mainen, Z. F. (2020). Inference-based decisions in a hidden state foraging task: Differential contributions of prefrontal cortical areas. *Neuron*, *106*(1), 166–176.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Welsh, T. N., & Elliott, D. (2004). Movement trajectories in the presence of a distracting stimulus: Evidence for a response activation model of selective reaching. *The Quarterly Journal of Experimental Psychology Section A*, *57*(6), 1031–1057.

Welsh, T. N., Elliott, D., & Weeks, D. J. (1999). Hand deviations toward distractors evidence for response competition: Evidence for response competition. *Experimental Brain Research*, *127*, 207–212.

Weng, T.-W., Dvijotham, K. D., Uesato, J., Xiao, K., Gowal, S., Stanforth, R., & Kohli, P. (2019). Toward evaluating robustness of deep reinforcement learning with continuous control, In *International Conference on Learning Representations*.

Wispinski, N. J. (2017). Modelling movement as an ongoing decision.

Wispinski, N. J., Truong, G., Handy, T. C., & Chapman, C. S. (2017). Reaching reveals that best-versus-rest processing contributes to biased decision making. *Acta Psychologica*, *176*, 32–38.

Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences*, *5*(11), 487–494.

Wong, A. L., Goldsmith, J., Forrence, A. D., Haith, A. M., & Krakauer, J. W. (2017). Reaction times can reflect habits rather than computations. *eLife*, *6*, e28075.

Wong, A. L., & Haith, A. M. (2017). Motor planning flexibly optimizes performance under uncertainty about task goals. *Nature Communications*, *8*(1), 14624.

Wong, A. L., Haith, A. M., & Krakauer, J. W. (2015). Motor planning. *The Neuroscientist*, *21*(4), 385–398.

Wood, D. K., Gallivan, J. P., Chapman, C. S., Milne, J. L., Culham, J. C., & Goodale, M. A. (2011). Visual salience dominates early visuomotor competition in reaching behavior. *Journal of Vision*, *11*(10), 16–16.

Wood, D. K., Gu, C., Corneil, B. D., Gribble, P. L., & Goodale, M. A. (2015). Transient visual responses reset the phase of low-frequency oscillations in the skeletomotor periphery. *European Journal of Neuroscience*, *42*(3), 1919–1932.

Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2009). Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences*, *106*(40), 17199–17204.

Wurman, P. R., Barrett, S., Kawamoto, K., MacGlashan, J., Subramanian, K., Walsh, T. J., Capobianco, R., Devlic, A., Eckert, F., Fuchs, F., Et al. (2022). Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, *602*(7896), 223–228.

Yates, J. L., Katz, L. N., Levi, A. J., Pillow, J. W., & Huk, A. C. (2020). A simple linear readout of MT supports motion direction-discrimination performance. *Journal of Neurophysiology*.

Yates, J. L., Park, I. M., Katz, L. N., Pillow, J. W., & Huk, A. C. (2017). Functional dissection of signal and noise in MT and LIP during decision-making. *Nature Neuroscience*, *20*(9), 1285–1292.

Yoo, S. B. M., Tu, J. C., & Hayden, B. Y. (2021). Multicentric tracking of multiple agents by anterior cingulate cortex during pursuit and evasion. *Nature Communications*, *12*(1), 1985.

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., Et al. (2023). Catalyzing next-generation artificial intelligence through NeuroAI. *Nature Communications*, *14*(1), 1597.

Zgonnikov, A., Aleni, A., Piiroinen, P. T., O'Hora, D., & di Bernardo, M. (2017). Decision landscapes: Visualizing mouse-tracking data. *Royal Society Open Science*, *4*(11), 170482.

Zylberberg, A., Lorteije, J. A., Ouellette, B. G., De Zeeuw, C. I., Sigman, M., & Roelfsema, P. (2017). Serial, parallel and hierarchical decision making in primates. *eLife*, *6*, e17331.

Chapter 2

Reaching for known unknowns: Rapid reach decisions accurately reflect the future state of dynamic probabilistic information

2.1 Abstract

Everyday tasks such as catching a ball appear effortless, but in fact require complex interactions and tight temporal coordination between the brain's visual and motor systems. What makes such interceptive actions particularly impressive is the capacity of the brain to account for temporal delays in the central nervous system—a limitation that can be mitigated by making predictions about the environment as well as one's own actions. Here, we wanted to assess how well human participants can plan an upcoming movement based on a dynamic, predictable stimulus that is not the target of action. A central stationary or rotating stimulus determined the probability that each of two potential targets would be the eventual target of a rapid reach-to-touch movement. We examined the extent to which reach movement trajectories convey

A version of this work was previously published as: Wispinski, N. J., Stone, S. A., Bertrand, J. K., Ouellette Zuk, A. A., Lavoie, E. B., Gallivan, J. P., & Chapman, C. S. (2021). Reaching for the known unknowns: Rapid reach decisions accurately reflect the future state of dynamic probabilistic information. *Cortex*, *138*, 253-265. doi:10.1016/j.cortex.2021.02.010. This work has been reproduced with permission.

internal predictions about the future state of dynamic probabilistic information conveyed by the rotating stimulus. We show that movement trajectories reflect the target probabilities determined at movement onset, suggesting that humans rapidly and accurately integrate visuospatial predictions and estimates of their own reaction times to effectively guide action.

2.2 Introduction

Humans exist in a dynamic world. Everyday tasks such as walking onto a moving escalator or catching a ball appear simple, but require tight temporally-coupled communication between visual and motor areas of the brain to ensure the action is successful. A key aspect of both of these tasks is that they require interception—demanding that the person get their body to the right place at the right time. To have this kind of successful interaction with the environment, predictions about the future state of moving objects must be computed by the brain and transformed into action. Catching a ball, for example, requires that the visual representation of the ball and its likely trajectory be transformed into the appropriate arm and hand movements, ultimately producing an anticipatory interceptive movement based on predictive internal models of object acceleration and gravity (Brenner et al., 2014; Zago et al., 2004; Zago et al., 2008).

What makes such interceptive actions particularly impressive is the capacity of the brain to account for the various temporal limitations of the central nervous system. Visuomotor processes, involving sensory evidence integration, action planning, and movement initiation, are subject to neurophysiological transmission delays ranging from 100 to 450 msec (Resulaj et al., 2009; van den Berg et al., 2016; Zago et al., 2009). Given the natural aptitude to intercept moving objects (Brenner et al., 2014; Brenner & Smeets, 2013; Brenner & Smeets, 2010; Fooken et al., 2016; Gellman & Carl, 1991) even when their motion cannot be fully observed (Fooken et al., 2016; Mazyn et al., 2007; Sharp & Whiting, 1975), theories articulate that humans must pre-plan

(Tyldesley & Whiting, 1975; Zago et al., 2009), adjust on the fly (Dessing et al., 2002), or mix planning and adjustment (Katsumata & Russell, 2012) to overcome these delays and produce successful interception actions. Empirically, there is evidence that the brain predicts the delays of sensory inputs in visual illusions. For example, in the flash-lag effect (Nijhawan, 2002), a predictably moving object is perceived as occupying its future location. Likewise, there is evidence that the brain predicts the delays of motor outputs during decision making. For example, during a random dot motion task, neuronal activity thought to reflect the decision variable terminates 50 msec before movement initiation (Roitman & Shadlen, 2002).

Studies of interception tasks have shown that humans are adept at predicting the future location of an object based on the movement of that object, e.g., by continuing smooth pursuit of an object through a period of occlusion (Fooken et al., 2016), fixating on the object intended for interception (Brenner & Smeets, 2013; Brenner & Smeets, 2010), or hitting a ball with a bat (Brenner et al., 2014). Yet, anecdotally, we also know that humans can make predictions about where to move based on other objects in the environment (e.g., obstacles; Chapman and Goodale, 2008, 2010), and plan actions toward locations where the eyes are *not* fixated (e.g., anti-pointing tasks; Johnson et al., 2002; Knights et al., 2015; Verneau et al., 2016). An intuitive example is a hockey forward who shoots opposite the position of the goalie to score a goal. Here, the already complex sensory-to-motor transformation must introduce yet another mediating cognitive variable—the representation of where the goalie *will not be* based on where the goalie will be. Here, we wanted to assess this particular capacity—how well can participants plan an upcoming movement based on a dynamic, predictable stimulus that is not the target of action.

One tool for assessing dynamic cognitive states is to analyze the shape of movement trajectories (Chapman et al., 2010a; Freeman et al., 2011; Gallivan et al., 2019; Resulaj et al., 2009; Scherbaum et al., 2010; Song & Nakayama, 2009; Spivey et al., 2005; Trommershäuser et al., 2008; Welsh & Elliott, 2004; Wispinski et al., 2020).

Hand or computer mouse trajectories can reflect the deliberation of external information, such as random dot motion stimuli (Resulaj et al., 2009; van den Berg et al., 2016), number magnitude (Chapman et al., 2014; Faulkenberry et al., 2016), or word processing (Spivey et al., 2005). Fluctuations in movements toward a final choice can also reflect internal information, such as the subjective value of snack foods (Sullivan et al., 2015). Typically, movement trajectories that curve between potential targets suggest conflict or indecision, while trajectories relatively straight toward a target reflect less competition between alternatives (Cisek & Kalaska, 2010; Song & Nakayama, 2009; Wispinski et al., 2020).

Of particular note, movements can reflect static or changing probabilities of multiple potential targets in space. When required to reach toward one of many potential targets on a screen, movement trajectories are sensitive to target number, suggesting a rapid integration of static probabilistic information during movement planning (Chapman et al., 2010a; Gallivan et al., 2011; Hudson et al., 2007). This information can bias movement trajectories even when movements need to be initiated less than 325 msec after stimuli onset (Gallivan et al., 2011). Others have shown that movement trajectory planning can also incorporate changing probabilistic information over time (Resulaj et al., 2009). In one group of studies, subjects were asked to reach toward a left or right target to indicate whether a group of dots on a screen are moving left or right. In this task, dot motion is noisy, so motion information fluctuates from timepoint-to-timepoint. In these studies, initial movement trajectories reflect fluctuating dot motion information that occurred roughly 350 msec in the past (Resulaj et al., 2009; van den Berg et al., 2016).

Here we questioned the extent to which movement trajectories also convey internal predictions about the future state of dynamic probabilistic information. To examine this, we manipulated probabilistic information dynamically between two potential targets. Participants were presented with a stimulus that rotated in a circle and were required to launch a movement towards the potential targets prior to the final target

being cued (i.e., go-before-you-know task). Critically, the position of the stimulus at movement onset determined how likely each of the two potential targets would be selected as the ultimate target of action on that trial. Previous work has shown that central (endogenous) versus peripheral (exogenous) cue-stimuli elicit different patterns of prediction and evolve over different time courses (Berger et al., 2005). To examine if this affected the dynamic prediction task, we collected data from two groups of participants—one where the rotation stimulus was an arrow that rotated about the central fixation and one where the rotation stimulus was a box that moved on a more peripheral path adjacent to the potential targets. We show that, across both groups, movement trajectories reflect target probabilistic information determined at movement onset, suggesting that humans rapidly and accurately integrate visuospatial predictions and estimates of their own reaction times to effectively guide action.

2.3 Materials and methods

2.3.1 Overview of procedure

Humans use predictions to overcome sensorimotor delays such that successful actions are generated to intercept moving objects. Here we use an analysis of behaviour (accuracy and movement trajectories) in a rapid reach task to test whether these same predictive capacities extend to movement planning based on predictable, but dynamic sensory evidence. The stimulus in the current experiments is separate from the target and conveys information about the probability of the final target location, rather than cueing location directly. In this task, we extend a previous go-before-you know paradigm (Chapman et al., 2010a, 2010b; Gallivan et al., 2017; Milne et al., 2013; Wood et al., 2011), which requires participants to initiate a movement in response to a go-signal before one of two potential target locations is revealed as the final target. Here, the probability of the upcoming target location was conveyed to participants

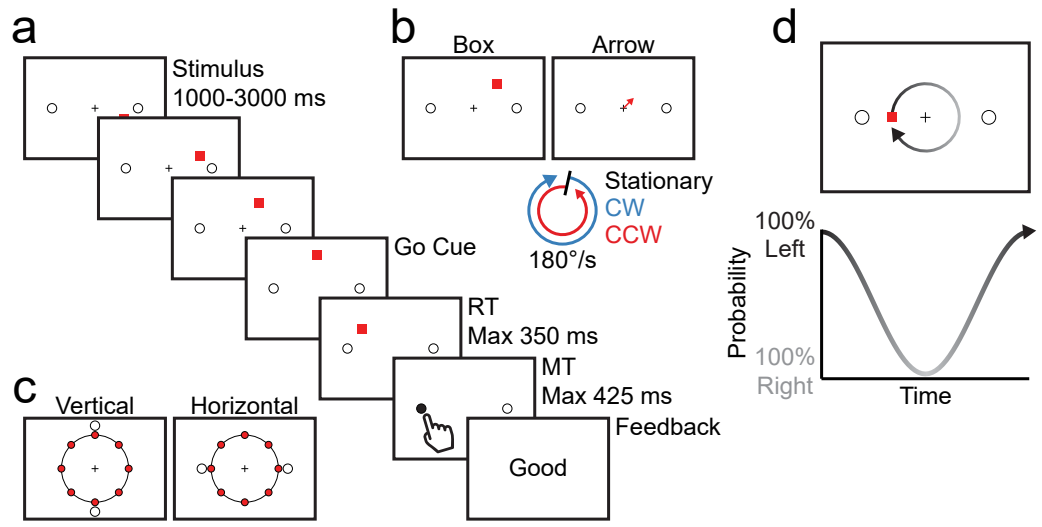

Figure 2.1: Stimuli and trial sequence. Also see videos linked in Open practices. (a) Example trial where a stimulus rotates counterclockwise (CCW) around the fixation. Once the stimulus hits a predetermined point along the circle (e.g., at the very top of the circle), the fixation disappears and a 'beep' plays. The stimulus continues rotating until the participant lifts their finger off the start button, after which the stimulus disappears and the final target is cued. Participants must begin their rapid reach before knowing which of the two targets will be cued. (b) The stimulus that determined target probability was a box or an arrow for different sets of participants. (c) Targets were arranged vertically or horizontally on each trial. One of eight equally-spaced points along the circle was pre-determined for each trial at which the go-signal would occur (depicted here as red circles). (d) As the stimulus moves (top panel, shown clockwise), the probability of target location oscillates (closer to left, black, left target; or closer to right, grey, right target).

via a stimulus that rotated at a fixed rate, either clockwise (CW), counter-clockwise (CCW), or, on baseline/control conditions, remained stationary (Fig. 2.1; see videos linked in Open practices). To test for possible differences in endogenous versus exogenous cueing (Berger et al., 2005), the stimuli conveying probability used in the present study differed across two groups: one group saw a central red arrow, and the other group saw a more peripheral red box, both of which rotated around the central fixation cross (Fig. 2.1b). The position of the probability-stimulus at movement onset dictated the probability with which one of two targets was selected as the final target for action (Fig. 2.1d). Thus, to have the highest chance for success participants needed to be monitoring and predicting from the rotating probability stimulus before

the go-signal *and* during reaction time.

Since the stimulus conveying target probability moves in a circle, the probability of any one target being selected varies sinusoidally (Fig. 2.1d). We capitalized on this sinusoidal feature of target probability to test for sinusoidal characteristics of behaviour in our key dependent measures—choice accuracy and reach curvature (indexed by reach area; Fig. 2.2). The trials in which the probability stimulus remained stationary serve as the starting point for our analysis (black curves in Fig. 2.3). In these stationary trials, we predict and find that participants are most accurate and reach trajectories are most straight (low area, e.g., grey trajectory in Fig. 2.2b) when the probability-stimulus perfectly predicts the target location (i.e., 100% probability), and least accurate and least straight (high area, e.g., green trajectory in Fig. 2.2b) when the probability stimulus is ambiguous with respect to final target location (i.e., 50% probability). To test for participants' ability to use the rotating probability-stimulus to guide action planning we compare the rotating trials (CW in blue, CCW in red; Fig. 2.3) to these stationary trials.

Our results test between three patterns of hypothesized behaviour (Fig. 2.3). First, as a baseline, we show what we would expect to see if participants were not predicting the future location of the rotating stimulus, but rather, "living in the past" (Delayed, Fig. 2.3a). Here, the CW and CCW data would show a shifted sinusoid whose phase (polar plots; Fig. 2.3) is out of alignment (specifically, delayed in time) for both accuracy and reach area. In this case, accuracy and reach area would reflect a temporally-outdated location of the probability-stimulus. Under this prediction, behavioural measures could reflect target probability determined at a salient event like the go-signal, or at a constant delay reflecting computation and transmission delays. This result would be consistent with data from unpredictable stimuli like in a random dot motion task, where movement trajectories and accuracy reflect the status of a decision variable several hundred milliseconds in the past (Resulaj et al., 2009).

Second, if we imagine that participants are living in the past at the onset of move-

ment but use the time available during the executed movement to make online corrections (e.g., changes of mind; Resulaj et al., 2009), we would predict a "catch-up" pattern of results (Fig. 2.3b). Here, the phase of the sinusoid for the reach area of CW and CCW trials lags the phase of the reach area across static trials, but the phase of the sinusoid for accuracy "catches up" such that all across-trial phases align. In this case, participants initially aim toward an outdated probabilistic location, but successfully correct their movements in flight to reach and touch the final target. Finally, third, if we imagine that participants are successfully predicting the future probability at the moment of movement onset (and thus, accurately accounting for sensorimotor processing delays and being unbiased by other factors), we would expect to observe a "complete" pattern of results (Fig. 2.3c). Here the CW and CCW data would match the stationary data. That is, even though rotation trials are dynamic, the prediction is accurate, rapid, and updated in real-time such that participants both aim toward an up-to-date probabilistic location and correctly touch the final target.

While the measures of reach behaviour described above are the focus of this study, we can also test how the sinusoidal nature of target probability might induce sinusoidal changes in reaction time. In tasks requiring action in response to targets of varying uncertainty, participants have been shown to adjust movement and reaction times to improve visuomotor accuracy in trials with greater uncertainty (Battaglia & Schrater, 2007). We would therefore predict reaction times to fluctuate sinusoidally with target probability. Specifically, when anchored to the go-signal, we would predict trials where target uncertainty is high (probability $\approx 50\%$) to result in longer reaction times, with participants maximizing the amount of visual evidence accumulated in support of final target probability. In contrast, we would predict trials where the target uncertainty is low (probability $\approx 100\%$) to result in shorter reaction times, potentially allowing participants to decrease motor errors by increasing movement time.

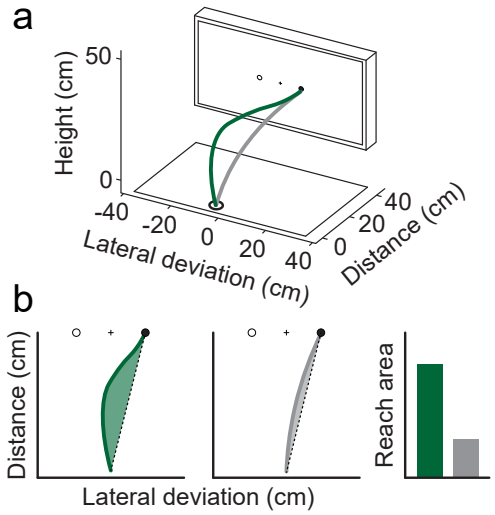

Figure 2.2: (a) Example of the three-dimensional reach trajectories collected. (b) Examples of two reach trajectories on trials where the right target is cued. The area between the reach trajectory, and a straight line from the start position to this participant's mean endpoint for right target trials, is used to index reach curvature. When reach trajectories travel between the two targets the reach area is larger (green), and when trajectories travel straight to one target the area is smaller (grey).

2.3.2 Participants

Twenty-seven participants (19 women; Age: M = 22.78, SD = 4.19) took part in the arrow experiment, while twenty-eight participants (13 women; Age: M = 22.96, SD = 3.53) took part in the box experiment. Sample size was determined based on recommendations from previous research with similar experimental paradigms (Gallivan & Chapman, 2014). All participants provided written consent before the experiment, and were compensated with course credit for participation. Experimental procedures were approved by Western University's Research Ethics Board. Only data from right-handed participants with normal or corrected-to-normal vision were analyzed.

2.3.3 Equipment and stimuli

Participants sat in front of a 40" touchscreen (NEC MultiSync© LCD4020 refresh rate 60 Hz; Fig. 2.2a), and made rapid reaching movements to targets on the screen

(see videos linked in Open practices). Two active infrared markers were taped to the participant's right index finger, and tracked reaching movements throughout the experiment (Optotrak, 150 Hz). All stimuli presentation and data collection were controlled with MATLAB (The Mathworks, Natick, MA) using Psychtoolbox (Version 3; Brainard and Vision, 1997; Kleiner et al., 2007; Pelli, 1997).

2.3.4 Trial sequence and procedure

Participants performed a variant of a go-before-you-know task (e.g., Chapman et al., 2010a; Gallivan and Chapman, 2014), requiring them to initiate a rapid reach movement before they knew which of two potential targets would be cued as the final target. The current study involved the presentation of a box (box experiment) or arrow (arrow experiment) stimulus that could either rotate around a central fixation (clockwise or counterclockwise) or remain in a fixed position (stationary). Two potential targets were presented (placed horizontally or vertically), and after a variable delay, an auditory beep would signal the participant to begin their reaching movement. At movement onset, one of the two targets was cued as the final target—the probability of which was determined by the location of the probability-stimulus at movement onset (Fig. 2.1d). Participants were informed that the final location of the stimulus dictated target probability prior to commencing the task, and were given practice trials until they reported feeling comfortable with the experimental procedure (e.g., timing constraints).

Trials began with the participant holding down the start button (Fig. 2.2a, positioned 5 cm from the front edge of the table) with their right index finger. The start button was placed so that participants would need to reach forward 40 cm and up 25 cm to touch the center of the screen in front of them.

With the start button held down, a central fixation cross would appear with two targets on a screen with a white background (Fig. 2.1a). The targets on each trial were arranged either horizontally or vertically, evenly counterbalanced across all trials

(Fig. 2.1c). Potential targets were black outlines of circles 2 cm in diameter, and located 9 cm from the fixation cross at the center of the screen. Participants were instructed to maintain central fixation at all times during the experiment.

Next, a stimulus would appear. For participants in the box experiment, this stimulus was a red square 2 cm wide (Fig. 2.1b). For participants in the arrow experiment, this stimulus was a red arrow with its base at the fixation, and extending ∼2.2 cm outward. On stationary trials, the stimulus would appear located at, or pointing toward, one of 8 evenly-spaced locations 7 cm from the origin (0°, 45°, 90°, 135°, etc.; evenly counterbalanced across trials, Fig. 2.1c) and not move throughout the trial. On non-stationary trials, this stimulus would appear at, or point toward, one of 120 evenly-spaced points centered 7 cm from the origin, with the start location of the stimulus chosen from a random uniform distribution. During these trials, the stimulus would rotate either clockwise or counterclockwise about the fixation along (box experiment), or pointing toward (arrow experiment), an invisible circle with 7 cm radius at a constant angular velocity of 180°/s. For both trial types (stationary and non-stationary), the stimulus remained on the display until participants initiated their reaching movements in response to a go-signal. The go-signal consisted of an auditory beep paired with the simultaneous disappearance of the central fixation cross. To clarify, the box or arrow stimulus continued to rotate after the go-signal until the participant had lifted their finger off the start button. This meant that the box or arrow stimulus rotated for roughly 51.30 additional degrees after its location at the go-signal, depending on the participant's reaction time on that trial (285 msec average RT in non-stationary trials across participants).

On stationary trials, this go-signal always occurred one second after the onset of the stimulus. On non-stationary trials, the stimulus would rotate around the origin for a minimum of one second, but would continue moving until it had reached one of the eight predetermined locations on the circle (0°, 45°, 90°, 135°, etc.; evenly counterbalanced across trials; Fig. 2.1c). Once the box or arrow stimulus had reached

this specified location, the participant would be signalled (via fixation disappearance and the coincident beep) to initiate their movement.

Participants had 350 msec (box experiment) or 325 msec (arrow experiment) after the go-signal to lift their finger off the start button. Upon successful button release, the stimulus disappeared and one of the two circles was filled in. Participants then had 425 msec to touch the cued final target on the screen (i.e., reach movements were required to be ballistic). The probability of a target filling in was based on the location of the stimulus when the start button was released.

For example, a target had a 100% probability of being cued as the final target if the probability-stimulus was located directly next to it (box) or pointed directly towards it (arrow) at reach onset. If the probability-stimulus was halfway between the targets when the reach was initiated, both targets had a 50% chance of being filled in (Fig. 2.1d). At the end of each trial, participants received feedback on their performance. If participants lifted their finger earlier than 100 msec after the go-signal (i.e., the reach movement was anticipatory), a "Too Early" error message would be presented after trial completion. If participants exceeded the reaction time limit, or the 425 msec movement time limit, the trial would similarly end with a "Time Out" or "Too Slow" error message, respectively. Participant accuracy was denoted by either a "Miss" message should they have touched the screen outside of a 6 cm x 6 cm invisible box centered on the correct final target, or a "Good" message should they successfully complete the trial without any errors.

Trials were equally counterbalanced for target arrangements (horizontal or vertical), stimulus motion (stationary, clockwise, or counterclockwise), and stimulus position at the time of the go-signal (eight equally-spaced positions around the origin; Fig. 2.1c). As such, there were 48 unique conditions (2 target arrangements x 8 trigger positions x 3 rotations), each repeated 12 times for a total of 576 trials. Trial order was fully randomized for each participant.

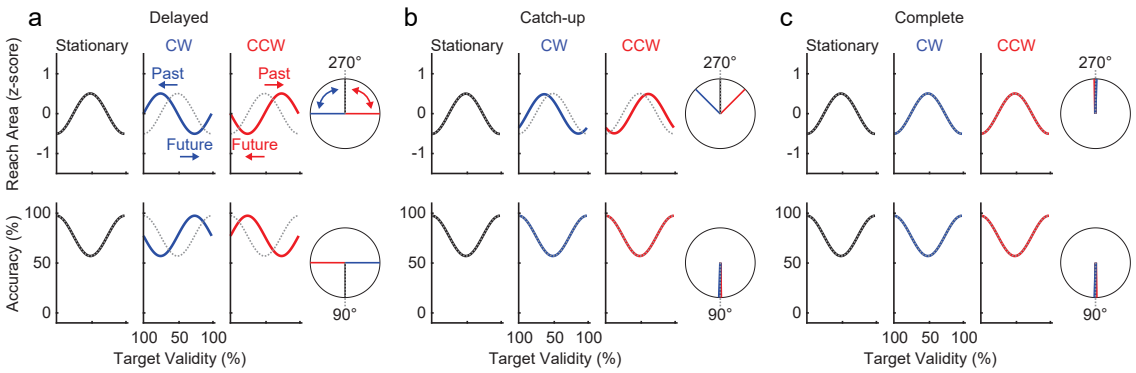

Figure 2.3: Predictions from left to right. Sinusoids for stationary (black), clockwise (blue), and counterclockwise (red) trials show predictions for reach area (top) and accuracy (bottom). Circles next to sinusoid panels are polar plots, with condition-coloured lines showing the predicted phase offsets relative to an expected phase (grey dotted lines). (a) Delayed. No predictive processing, or insufficient prediction. Accuracy and reach area reflect target probability in the past. (b) Catch-up. Reach area reflects target probability in the past, but information is used during the movement to correct the reach so that accuracy reflects the final target probabilities. (c) Complete. Prediction is accurate and fast. Information about target probability is able to be used at the time of movement onset, when target probability is actually determined, and is not biased by other factors. Reach area and accuracy both reflect the final target probability.

2.3.5 Pre-processing

Trials were deemed as useable for analysis if they were not "Too Early" or "Time Out" trials, did not contain movement recording errors, or did not contain "out of bounds" start or end positions. Additionally, participants were rejected for analysis if they had 25% or fewer useable trials in 8 or more of the 48 unique conditions. These rejected participants are not discussed further. This criterion was enforced so that participants had at least three trials in most conditions for analysis. Three subjects were rejected from the box experiment, while six subjects were rejected from the arrow experiment. One subject was also rejected from each experiment for initially reaching backward off the start position in the majority of trials, leaving n = 24 and n = 20 for the box and arrow experiments, respectively.

Data cleaning and trial rejection were conducted following the recommendations in Gallivan and Chapman (2014) for rapid reaching experiments. In brief, reach

trajectories were space-normalized to 200 equally-spaced points along the ∼40 cm distance from the start position to the screen (Gallivan & Chapman, 2014). Reach area was calculated as the approximate area between a reach trajectory on a correct trial and the straight line between the start position and average endpoint for that corresponding target (left, right, up, down) calculated for each subject (Fig. 2.2b; for previous use see Chapman, Gallivan, and Enns, 2015; Chapman, Gallivan, Wong, et al., 2015). Area was calculated in two-dimensional space along the axis of interest on that trial (e.g., horizontal axis for horizontal target trials). Reach areas were then z-scored for each subject within each target orientation condition (left, right, up, down). Reach area normalization was performed because biomechanical differences within and between subjects created differences in reach area for different reach directions that were not of interest in this study. Larger normalized reach areas correspond to trajectories that move more in between the two targets, whereas smaller reach areas correspond to trajectories that more closely follow a straight line path to the correct, filled-in target. As such, reach area can be used to estimate the level of competition or indecision between several potential targets in space (Gallivan & Chapman, 2014; Gallivan et al., 2018; Wispinski et al., 2020).

Reaction time was calculated as the time from the go-signal auditory beep to the release of the start button. Movement time was calculated as the elapsed time between button release and when a touch was detected on the touchscreen. Unlike "Too Early", "Time Out", and "Miss" trials, we did not automatically reject "Too Slow" trials. Instead, "Too Slow" trials >2 SD above a participant's mean (after excluding all trials with a movement time >850 msec) were rejected for analysis.

Errors on each trial could be a combination of "Too Early" (M = 1.02%, Range: 0% - 5.21%), "Miss" (M = 8.08%, Range: .52% - 20.31%), "Time Out" (M = 12.02%, Range: .35% - 32.12%), >2 SD of mean movement time (M = 4.06%, Range: 1.22% - 7.64%), reaches with recording errors (M = 1.21%, Range: 0% - 9.72%), and reaches with "out of bounds" start or end positions (M = 6.49%, Range: .17% - 32.29%). In

total, participants whose data was analyzed had a mean of 86.02% useable trials for analysis (Range: 57.81% - 98.96%), and of those trials a mean of 86.06% were correct (Range: 72.59% - 96.97%). These trial rejection numbers are generally in line with recommendations for rapid reach experiments (Gallivan & Chapman, 2014).

2.3.6 Model

To overcome sparsity of sampling (there were 120 possible stimulus locations) and to directly test for the predicted sinusoidal patterns of data (Fig. 2.3), we reduced the data collected in this experiment by fitting a sine wave model to each condition for each subject. The sine wave model consisted of a fixed period equal to the rate of stimulus rotation (180°/s), and three free parameters: mean shift (μ), amplitude (A), and phase shift (ϕ).

$$y = \mu + Asin(\phi + x)$$

To fit data to this sine wave model, circle positions on vertical target trials were rotated 90° so that they would line up with horizontal target trials (i.e., 100% target probability occurred at the same circle location for horizontal and vertical trials). Circle positions were then collapsed so that positions started at 100% target probability of left targets, decreased to 50% target probability, and then ended at 100% probability for right targets (Fig. 2.1d). For each subject and for each condition (e.g., Fig. 2.4a shows a subject in the box experiment, horizontal targets, and clockwise stimulus rotation), single trial data were fit to the sine wave model using a least squares cost function. One-hundred fits were performed using the `fminsearchbnd` function in MATLAB with random initial parameters, and the fit with the lowest cost was taken as the final parameter estimate. The amplitude parameter was constrained to be higher than zero for all fits, as it caused inaccurate phase parameter estimates if amplitude was too low.

By fitting sine waves to each condition for each participant, these data were reduced

to three parameters (mean, amplitude, and phase), which were used for statistical comparisons (Fig. 2.4b). Overall, these sine waves are reasonable descriptors of the data and provided useful data reduction. First, the model period corresponds directly to the independent variables of stimulus motion and changes in target probability with location (i.e., 180°/s). Second, the fitted models describe the dependent variables in different target probability locations reasonably well, given that the sine wave model is fit to single-trial data (reaction time, mean R^2 = .09, range: -.26 - .45; accuracy, mean R^2 = .08, range: 0 - .37; reach area, mean R^2 = .13, range: -.01 - .45). Reach area and reaction time were only calculated for correct trials.

2.3.7 Statistical analysis

Phase

Our primary theoretical motivation was to test whether prediction of probability would be evident in our dependent measures. As such, of our model-fitted dependent measures, the phase parameter is of the most theoretical importance. However, estimated phase parameters reasonably match a circular normal distribution, which violates assumptions of many statistical tests, such as a linear repeated-measures ANOVA. Therefore, the phase parameters for the sine waves fit to each of reaction time, reach area, and accuracy were compared using circular statistics (Berens, 2009). In particular, we were interested if estimated phases in each condition were significantly different from an expected phase. For instance, in the stationary stimulus condition, we would expect reaction times to be the fastest, reach area to be the smallest (reaches most straight), and accuracy to be highest when target probability was 100%. We expect the reverse pattern when the probability was 50% (slow reaction times, large reach areas, and low accuracy). Below we compare whether the observed phase estimates in each condition were significantly different from the expected phase using one-sample circular t-tests. In addition, we wanted to know how each of our stimulus conditions differed from one another. So, we also ran all

possible circular paired t-tests of stationary versus clockwise versus counterclockwise stimulus conditions. This led to 18 total circular t-tests (3 dependent measures x (3 one-sample + 3 paired)), which were Bonferroni-corrected to a statistical threshold of .0028 (i.e., .05/18). Our investigation of phase collapses across the other factors in our experiment (Experiment: Box or Arrow, and Target Arrangement: Vertical or Horizontal) because our main theoretical questions are driven by Rotation.

Mean and amplitude

For mean and amplitude parameters estimated from each dependent variable (reaction time, reach area, accuracy), we conducted a 2 (target arrangement: horizontal vs vertical) x 3 (rotation: stationary vs clockwise vs counterclockwise) x 2 (experiment: box vs arrow) mixed ANOVA. All main effects and interactions were Greenhouse-Geisser corrected, and also corrected using the sequential Bonferroni-Holm procedure (remedy 2; described in Cramer et al., 2016) to control for the familywise error rate of all the mixed ANOVA tests together.

2.4 Results

Accuracy and reach area were analyzed relative to the final target probabilities on each trial (i.e., when the probability-stimulus disappeared at the beginning of a movement). However, reaction time was analyzed relative to the target probabilities at the go beep. Locking reaction times to the go beep can give us a picture of how target probabilities influence movement onset times.

2.4.1 Effects of rotation on phase

As articulated in our Methods, our primary motivation was to analyze the effect of rotation condition on the estimated phase parameters of the data. These analyses speak to whether the sinusoidal pattern of the dependent measures are shifted depending on whether the stimulus was stationary or rotating, and should indicate whether reach

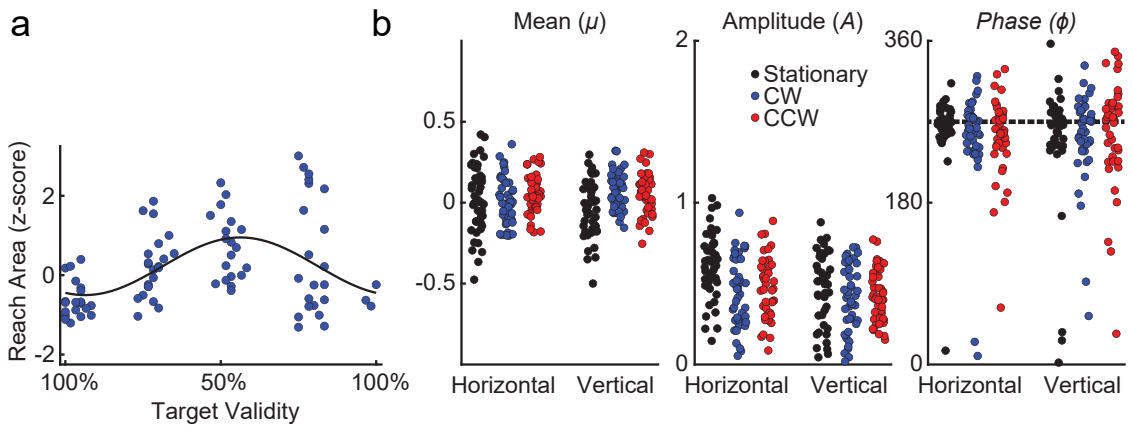

Figure 2.4: Fitting sine wave models to reduce data. (a) Example of a single subject, single condition sine-wave fit where each data point is a single trial. Sine waves with fixed period (matching the probability profile of the rotating stimulus), variable mean shift, amplitude, and phase shift were fit to single-trial data. Shown is normalized reach area by target validity locations for a single participant in the box experiment when targets were arranged horizontally and the probability stimulus was rotating clockwise, $R^2 = .23$. (b) Sine wave parameter fits to normalized reach area where each data point is one subject's data for each parameter (μ - left panel, A - middle panel, ϕ - right panel) in each condition (Stationary - black, CW - blue, CCW - red) and experiment (Box and Arrows). Dashed line in the phase panel represents the expected phase for normalized reach area (lowest at 100% target validity, highest at 50% target validity).

behaviour reflects a delayed, catch-up, or complete sensorimotor prediction process based on dynamic target probability (Fig. 2.3). For the following tests, the corrected statistical threshold was $p = .0028$ (see 2.3.7 Phase).

For reaction time data (Fig. 2.5), we find that the distribution of estimated phases when the stimulus is stationary is not different from the expected phase (phase difference = 2.65°, $p = .70$, 99% CI [21.06°, -15.76°]). Here we test against an expected phase where fastest reaction times occur when probability is 100% and slowest reaction times occur when probability is 50%. However, estimated phases in conditions where the targets are moving clockwise (phase difference = 51.16°, $p = .000013$, 99% CI [75.29°, 23.94°]) or counterclockwise (phase difference = -63.94°, $p < .000001$, 99% CI [-32.39°, -97.33°]) are significantly different from the expected phase. Pairwise comparisons indicate that both the stationary ($p = .0014$) and clockwise ($p =$

.00011) phases are significantly different from the counterclockwise phase. However, stationary and clockwise phases are not significantly different from each other at the corrected statistical threshold (Fig. 2.5; $p = .0051$). This pattern of results suggests that when the probability-stimulus is rotating, participants are reacting to the probability state that the stimulus is approaching, rather than reacting to where the probability-stimulus is actually located at the time of the go-signal.

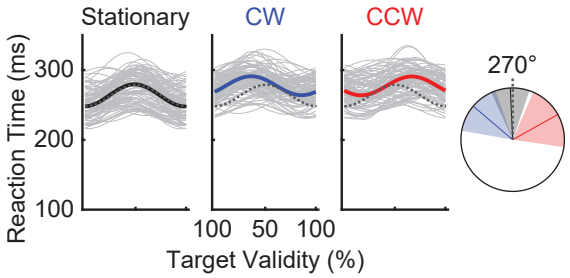

Figure 2.5: Sine wave models fit to reaction time data over target validity positions, time-locked to the go beep. Sine waves from each individual condition (e.g., a subject in the box experiment, horizontal targets, clockwise rotation) are in light grey and a sine wave with average mean, amplitude, and phase parameters are in solid colours (black for stationary, blue for clockwise, red for counterclockwise). On the right is a circle showing a polar plot describing the phase parameters and their confidence intervals (Bonferroni-corrected 95%). The sinusoidal pattern of results corresponding to the expected phase are plotted dashed grey lines. Average stationary phase was not significantly different from the expected phase, while phases in the rotating conditions were significantly different from the expected phase.

For reach area data (Fig. 2.6), the expected phase is that reach area would be smallest when probability was high, and largest when probability was low. All estimated reach area phases do not differ from the expected phase regardless of if the stimulus was stationary (phase difference = 2.39°, $p = .58$, 99% CI [-11.06°, 16.79°]), moving clockwise (phase difference = -5.61°, $p = .30$, 99% CI [-20.04°, 8.82°]), or moving counterclockwise (phase difference = -5.11°, $p = .34$, 99% CI [-22.36°, 10.50°]). For accuracy data (Fig. 2.6), the expected phase is that accuracy would be highest when probability was high, and lowest when probability was low. All estimated accuracy phases do not differ from the expected phase regardless of if the stimulus was stationary (phase difference = -8.85°, $p = .14$, 99% CI [-22.71°, 6.31°]), moving

clockwise (phase difference = 3.91°, p = .54, 99% CI [-13.15°, 20.96°]), or moving counterclockwise (phase difference = -.15°, p = .99, 99% CI [-22.34°, 22.64°]). Pairwise comparisons indicate that stationary, clockwise, and counter-clockwise phases are not significantly different from each other (ps [.11, .96]) for both reach area and accuracy data (Fig. 2.6). This pattern of results shows that people were accounting for sensorimotor delays and building those sensorimotor delays into their reach planning. This aligns with our Complete prediction hypothesis (Fig. 2.3) and demonstrates that, in this task, predictive mechanisms were being successfully deployed based on a probability-stimulus that was separate from the actual target location.

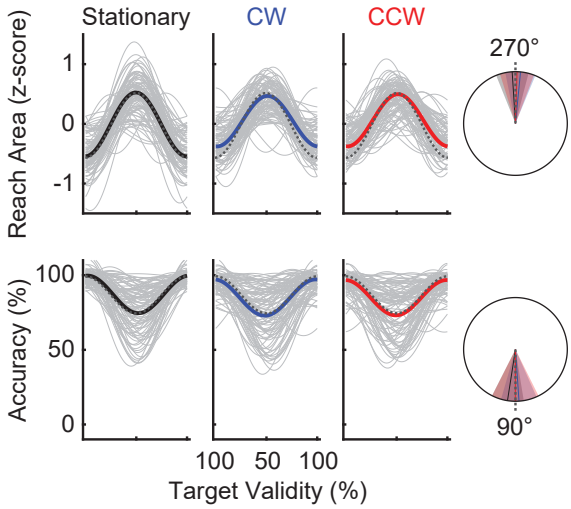

Figure 2.6: Sine wave models fit to reach area (top) and accuracy (bottom) data over target validity positions. As in Fig. 2.5, sine waves from each individual condition are in light grey, while a sine wave with average mean, amplitude, and phase parameters are in solid colours (black for stationary, blue for clockwise, red for counterclockwise). The average phase in each rotation condition and Bonferroni-corrected 95% confidence intervals are plotted along a circle, with the expected phase as a dashed line.

2.4.2 Additional main effects

Beyond the theoretically-motivated exploration of Phase parameters, we also examined differences in Mean and Amplitude for our sinusoidal parameter fits using a 3-factor mixed ANOVA applied to each of reaction time, reach area, and accuracy. After correcting for the number of statistical tests (Cramer et al., 2016), we found

no significant effects of Experiment, nor any significant interactions in these data (ps [.018, .97]). Five main effects passed the adjusted significance threshold and are described below. Again, for the following tests, the corrected statistical threshold varied per test between $p = .00119$ and $p = .05$ (Cramer et al., 2016).

Analyses showed a main effect of rotation for the mean parameters estimated from reaction time data, $F(1.21, 47.11) = 54.86$, $p = 5.8\text{e-}9$, $\eta_p^2 = .52$. Bonferroni-corrected post-hoc comparisons showed mean parameters were lower in the stationary condition relative to the clockwise ($t(42) = 8.38$, $p = 3.14\text{e-}12$) or counterclockwise ($t(42) = 8.21$, $p = 7.05\text{e-}12$) conditions, and that the clockwise and counterclockwise conditions did not differ ($t(42) = .17$, $p = 1.00$). In other words, participants were faster to start moving when the stimulus was stationary relative to when it was moving.

Analyses also showed a main effect of rotation for the amplitude parameters estimated from normalized reach area data, $F(1.61, 67.45) = 9.98$, $p = 4.39\text{e-}4$, $\eta_p^2 = .19$. Bonferroni-corrected post-hoc comparisons showed amplitude parameters were higher in the stationary condition relative to the clockwise ($t(42) = 4.12$, $p = .00026$) or counterclockwise conditions ($t(42) = 3.55$, $p = .0019$), and that the clockwise and counterclockwise conditions did not differ ($t(42) = .57$, $p = 1.00$). In other words, the difference between straight, confident reaches and indirect, conflicted reaches was larger for the stationary trials than the moving trials. This likely reflects that stationary trials' probabilities were more discernible than rotating trials. Analyses revealed a main effect of target arrangement on the mean parameters estimated from accuracy data, $F(1, 42) = 86.93$, $p = 8.69\text{e-}12$, $\eta_p^2 = .67$. Post-hoc comparisons showed mean parameters were higher for horizontal targets relative to vertical targets ($t(42) = 9.44$, $p = 4.83\text{e-}12$), suggesting participants found horizontal trials easier than vertical trials.

Finally, analyses revealed a main effect of target arrangement for the amplitude parameters estimated from normalized reach area, $F(1, 42) = 18.94$, $p = 8.46\text{e-}5$, $\eta_p^2 = .31$, and accuracy data, $F(1, 42) = 12.39$, $p = .001$, $\eta_p^2 = .23$. Post-hoc comparisons

showed amplitude parameters were higher for horizontal targets relative to vertical targets for reach area data ($t(42) = 4.35$, $p = .000085$). These results indicate that the change from straight to indirect reaches was larger for horizontal trials, likely because the hand started between the two targets for horizontal trials, but below the two targets for vertical trials. Conversely, amplitude parameters were higher for vertical targets relative to horizontal targets for accuracy data ($t(42) = 3.52$, $p = .0011$). In other words, participants found horizontal trials easier than vertical trials. Essentially, accuracy was near 100% when probability was high for both horizontal and vertical trials, but vertical trials' accuracy was much lower when probabilities neared 50%. This means that vertical trials have a larger amplitude to account for the decrease at 50% probability and subsequently have a lower mean.

2.5 Discussion

Here we assessed how well participants can plan an upcoming movement based on a dynamic, predictable stimulus that is not the target of action. A stationary or rotating stimulus determined the probability that each of two potential targets would be the ultimate target of a rapid reach-to-touch movement. Further, we used two different stimuli (box and arrow) to investigate processing differences in exogenous and endogenous attention systems. We questioned the extent to which the sensorimotor system integrates predictions about the future state of dynamic probabilistic information by examining movement trajectories.

We tested whether the sinusoidal pattern of reach area and accuracy was shifted in time relative to the rotation of the stimulus that determined target probability. We tested between three possible patterns of results (Fig. 2.3). According to the "delayed" prediction, transmission delays in the central nervous system would mean that reach area and accuracy would reflect the target probability at some time in the past. According to the "catch-up" prediction, information about target probability would be similarly delayed, but could still be used to correct online reach trajectories

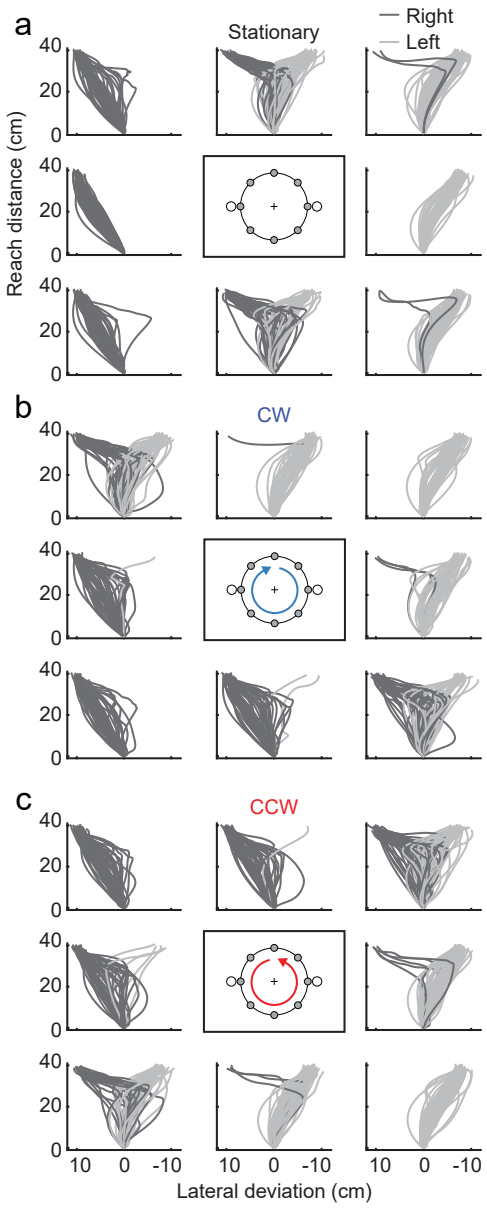

Figure 2.7: Average correct reach trajectories for each participant when reaches ended left (dark grey) and right (light grey) on trials where targets were arranged horizontally. Reaches are plotted in one of the eight trigger positions along the circle corresponding to the probability-stimulus location when the go-signal was presented. These data are intended to be qualitative descriptors of the overall reaching patterns by condition. For (a) Stationary trials, reaches tended to be straightest when the probability-stimulus is at the left or rightmost trigger position at the go-signal, and more curved when at the top and bottommost trigger position. For (b) Clockwise and (c) Counterclockwise trials, this pattern is shifted indicating that participants were anticipating the future location of the probability-stimulus.

toward the final target more often. This catch-up prediction would appear as a delayed offset in the sinusoidal pattern of reach area relative to the probability-stimulus, but with less temporal offset for the sinusoidal pattern of accuracy. Finally, according to a "complete" prediction, participants would be able to successfully predict the future location of the probability-stimulus while accounting for their own reaction time, ultimately producing a sinusoidal pattern for both reach area and accuracy in lock-step with information about the final target probabilities. These results support the notion of "complete" prediction (Fig. 2.3), wherein there is no temporal offset for patterns of reach area and accuracy between stationary, clockwise, and counter-clockwise conditions. Overall, we show that, despite sensory and motor delays in the central nervous system, movement trajectories reflect target probability determined at movement onset. This was true for both the box and arrow experiment, suggesting that the prediction of probability from a non-target stimulus is not subject to changes due to a central versus more peripheral focus. This suggests that humans rapidly and accurately integrate visuospatial predictions from various non-target stimuli and can estimate their own reaction times to effectively guide action.

It has long been argued that one of the major roles of the brain is to produce movement (Cisek & Kalaska, 2010; Gallivan et al., 2018; Hommel et al., 2019; Wolpert et al., 2001) and that this capacity, among others, involves prediction (e.g., Clark, 2013; Von Helmholtz, 2013). In short, several theories posit that the brain, rather than using the accumulation of bottom-up sensory cues to build a model of the world, instead builds predictions about the current state of the world and compares these predictions to incoming sensory information. The difference between the predicted sensory input and the actual sensory input—termed the "prediction error"—is used to continually update internal models of the world (Clark, 2013). Evidence for such predictive coding has been found for low-level sensory input (Hosoya et al., 2005; Rao & Ballard, 1999), as well as higher order cognitive functions (Spratling, 2008, 2016).

In addition to perception and cognition, the fundamental capacity for prediction

is required for effective motor control, where appropriate motor commands are computed through the use of internal forward models (Wolpert et al., 2001). Forward models are a theoretical construct that can be used to predict, given a particular motor command, the sensory consequences of executing the action. Such prediction allows the brain to account for transmission and computational delays in the central and peripheral nervous systems, effectively providing for robustness in both real-time control and perception. There is good behavioural and neural evidence that the brain contains such internal models (Blakemore et al., 1998; Schneider & Mooney, 2018; Wolpert et al., 1995). For example, with respect to perception, humans are unable to tickle themselves because forward models can be used to inhibit sensations arising from self-motion (Blakemore et al., 1999). Likewise, with respect to control, the prediction of the sensory consequences of action can allow the brain to rapidly detect performance errors, and rapidly launch effective corrective actions as needed. A forward model is useful especially when generating interceptive actions. How humans use an internal prediction model for interceptive actions was tested by Soechting et al. (2009) using a model that explained finger movements during interception of a randomly moving target on a screen. They found that the finger's position within 100 msec of movement onset reflected anticipatory predictions in advance of the target's location, similar to the current reach area results. However, Soechting et al. (2009) conclude that only "directly observable quantities" like target position and velocity are integrated into an internal prediction model, while higher order properties like statistical features of motion (i.e., sinusoidal motion laws) are not dynamically refined. In contrast, the current results suggest that some unobservable quantities, in this case target probability derived from a rotating stimulus, do indeed directly impact real-time predictions.

In this study, we used movement trajectories to reveal the sensitivity to changes in target probability. Previous work has shown that trajectories are thought to be a real-time readout of several cognitive variables, shown in behaviours such as changes

of mind (Resulaj et al., 2009; van den Berg et al., 2016), or moment-to-moment fluctuations throughout movement (Dshemuchadse et al., 2013; Freeman et al., 2011). Here, we show that curved reach trajectories (i.e., those with large reach areas) reflect uncertainty about the predicted target position, while relatively straighter reach trajectories (i.e., those with smaller reach areas) reflect more certainty about target predictions. We provide Fig. 2.7 as a useful descriptive tool demonstrating the effect of the position of our dynamic probability stimulus at the time of the go-signal on average participant trajectories. For stationary stimuli, reach trajectories are most curved when the stimulus is positioned half-way between the two targets (50% target probability, Fig. 2.7a, middle panels of top and bottom rows), whereas for rotating stimuli reach trajectories are most curved when the stimulus is moving toward 50% probability at the go-cue (top left and bottom right corners for Fig. 2.7b, CW; top right and bottom left corners for Fig. 2.7c, CCW). Overall, this suggests that participants are successfully predicting the future probability of both potential targets, and planning their movements accordingly.

Behavioural measures such as reaction time, accuracy, and movement trajectories are often thought to index the same internal cognitive processes (Wispinski et al., 2020). However, these behaviours are measured at different times. For instance, reaction time may reveal cognitive variables several hundred milliseconds before accuracy, particularly when a reaching movement separates the two. Differences between these measures may reveal the evolution of cognition over the course of a trial, especially in dynamic environments. Here we see a dissociation between reaction time and measures of reach area and accuracy. The pattern of reaction time on clockwise and counterclockwise trials is out of phase with the pattern of reaction time on stationary trials—results not observed for reach area and accuracy. On one hand, this difference could reflect that reaction time is indexed at an earlier point in the trial than accuracy and most of the movement trajectory. This might suggest that the internal prediction of target probability is still evolving when reaction time is measured, while

predictions of target probability are accurate at the time movement and accuracy are measured. Some models explicitly theorize that reaction time, movement trajectories, and accuracy can be explained in some tasks from the same internal decision variable (Resulaj et al., 2009). However, these measures, while similar in some tasks, may arise from distinct computation. Such differences may also explain the discrepancy between reaction time, reach area, and accuracy in the current results. Finally, it is also possible participants are adjusting reaction times to improve visuomotor accuracy in trials with greater uncertainty (Battaglia & Schrater, 2007). When trials are uncertain, longer reaction times may be used to accumulate more sensory evidence to guide their decision. Overall, however, these results suggest more work needs to be done to determine if reaction time and movement variables in a reach decision task reflect common or separate cognitive processes.

In the present study, we demonstrated that humans are able to accurately predict future states from a predictable, dynamic, non-target object and account for sensorimotor delays to guide rapid reaching movements. Such predictions are likely a key part of neural computation within and between different systems of the brain. The results of this study speak to one key part of how humans are able to carry out actions in complex and dynamic environments.

2.6 Open practices

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. No part of the study procedures or analyses were pre-registered prior to the research being conducted.

Videos of the task, data, analysis code, and digital study materials are publicly available at the following website: `https://osf.io/rt5xv/`.

2.7 Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Killam Trusts. We thank Jody Culham, Mel Goodale, and their labs at Western University for providing resources and supervision for data collection.

2.8 References

Battaglia, P. W., & Schrater, P. R. (2007). Humans trade off viewing time and movement duration to improve visuomotor accuracy in a fast reaching task. *Journal of Neuroscience*, *27*(26), 6984–6994.

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, *31*, 1–21.

Berger, A., Henik, A., & Rafal, R. (2005). Competition between endogenous and exogenous orienting of visual attention. *Journal of Experimental Psychology: General*, *134*(2), 207.

Blakemore, S.-J., Frith, C. D., & Wolpert, D. M. (1999). Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience*, *11*(5), 551–559.

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, *1*(7), 635–640.

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

Brenner, E., Driesen, B., & Smeets, J. B. (2014). Precise timing when hitting falling balls. *Frontiers in Human Neuroscience*, *8*, 342.

Brenner, E., & Smeets, J. (2013). Introduction to active vision: The complexities of continuous visual control. *Journal of Vision*, *13*(9), 1375–1375.

Brenner, E., & Smeets, J. B. (2010). Why we need continuous visual control to intercept a moving target. *Journal of Vision*, *10*(7), 1081–1081.

Chapman, C. S., Gallivan, J. P., & Enns, J. T. (2015). Separating value from selection frequency in rapid reaching biases to visual targets. *Visual Cognition*, *23*(1-2), 249–271.

Chapman, C. S., Gallivan, J. P., Wong, J. D., Wispinski, N. J., & Enns, J. T. (2015). The snooze of lose: Rapid reaching reveals that losses are processed more slowly than gains. *Journal of Experimental Psychology: General*, *144*(4), 844.

Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Ansari, D., Culham, J. C., & Goodale, M. A. (2014). Counting on the motor system: Rapid action planning reveals the format-and magnitude-dependent extraction of numerical quantity. *Journal of Vision*, *14*(3), 30–30.

Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Culham, J. C., & Goodale, M. A. (2010a). Reaching for the unknown: Multiple target encoding and real-time decision-making in a rapid reach task. *Cognition*, *116*(2), 168–176.

Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Culham, J. C., & Goodale, M. A. (2010b). Short-term motor plasticity revealed in a visuomotor decision-making task. *Behavioural Brain Research*, *214*(1), 130–134.

Chapman, C. S., & Goodale, M. A. (2008). Missing in action: The effect of obstacle position and size on avoidance while reaching. *Experimental Brain Research*, *191*(1), 83–97.

Chapman, C. S., & Goodale, M. A. (2010). Seeing all the obstacles in your way: The effect of visual feedback and visual feedback schedule on obstacle avoidance while reaching. *Experimental Brain Research*, *202*(2), 363–375.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, *33*, 269–298.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., Waldorp, L. J., & Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2), 640–647.

Dessing, J. C., Bullock, D., Peper, C. L. E., & Beek, P. J. (2002). Prospective control of manual interceptive actions: Comparative simulations of extant and new model constructs. *Neural Networks*, *15*(2), 163–179.

Dshemuchadse, M., Scherbaum, S., & Goschke, T. (2013). How decisions emerge: Action dynamics in intertemporal decision making. *Journal of Experimental Psychology: General*, *142*(1), 93.

Faulkenberry, T. J., Cruise, A., Lavro, D., & Shaki, S. (2016). Response trajectories capture the continuous dynamics of the size congruity effect. *Acta Psychologica*, *163*, 114–123.

Fooken, J., Yeo, S.-H., Pai, D. K., & Spering, M. (2016). Eye movement accuracy determines natural interception strategies. *Journal of Vision*, *16*(14), 1–1.

Freeman, J., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 59.

Gallivan, J. P., & Chapman, C. S. (2014). Three-dimensional reach trajectories as a probe of real-time decision-making between multiple competing targets. *Frontiers in Neuroscience*, *8*, 215.

Gallivan, J. P., Chapman, C. S., Gale, D. J., Flanagan, J. R., & Culham, J. C. (2019). Selective modulation of early visual cortical activity by movement intention. *Cerebral Cortex*, *29*(11), 4662–4678.

Gallivan, J. P., Chapman, C. S., Wolpert, D. M., & Flanagan, J. R. (2018). Decision-making in sensorimotor control. *Nature Reviews Neuroscience*, *19*(9), 519–534.

Gallivan, J. P., McLean, D. A., Valyear, K. F., Pettypiece, C. E., & Culham, J. C. (2011). Decoding action intentions from preparatory brain activity in human parieto-frontal networks. *Journal of Neuroscience*, *31*(26), 9599–9610.

Gallivan, J. P., Stewart, B. M., Baugh, L. A., Wolpert, D. M., & Flanagan, J. R. (2017). Rapid automatic motor encoding of competing reach options. *Cell Reports*, *18*(7), 1619–1626.

Gellman, R., & Carl, J. (1991). Motion processing for saccadic eye movements in humans. *Experimental Brain Research*, *84*(3), 660–667.

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics*, *81*(7), 2288–2303.

Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, *436*(7047), 71–77.

Hudson, T. E., Maloney, L. T., & Landy, M. S. (2007). Movement planning with probabilistic target information. *Journal of Neurophysiology*, *98*(5), 3034–3046.

Johnson, H., Van Beers, R. J., & Haggard, P. (2002). Action and awareness in pointing tasks. *Experimental Brain Research*, *146*(4), 451–459.

Katsumata, H., & Russell, D. M. (2012). Prospective versus predictive control in timing of hitting a falling ball. *Experimental Brain Research*, *216*(4), 499–514.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?

Knights, E., Bultitude, J., & Rossit, S. (2015). Prism adaptation effects are not limited to dorsal visual processing: Evidence from pro-pointing and anti-pointing, In *British association for cognitive neuroscience*.

Mazyn, L. I., Savelsbergh, G. J., Montagne, G., & Lenoir, M. (2007). Planning and on-line control of catching as a function of perceptual-motor constraints. *Acta Psychologica*, *126*(1), 59–78.

Milne, J. L., Chapman, C. S., Gallivan, J. P., Wood, D. K., Culham, J. C., & Goodale, M. A. (2013). Connecting the dots: Object connectedness deceives perception but not movement planning. *Psychological Science*, *24*(8), 1456–1465.

Nijhawan, R. (2002). Neural delays, visual motion and the flash-lag effect. *Trends in Cognitive Sciences*, *6*(9), 387–393.

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

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, *22*(21), 9475–9489.

Scherbaum, S., Dshemuchadse, M., Fischer, R., & Goschke, T. (2010). How decisions evolve: The temporal dynamics of action selection. *Cognition*, *115*(3), 407–416.

Schneider, D. M., & Mooney, R. (2018). How movement modulates hearing. *Annual Review of Neuroscience*, *41*, 553–572.

Sharp, R., & Whiting, H. (1975). Information-processing and eye movement behaviour in a ball catching skill. *Journal of Human Movement Studies*.

Soechting, J. F., Juveli, J. Z., & Rao, H. M. (2009). Models for the extrapolation of target motion for manual interception. *Journal of Neurophysiology*, *102*(3), 1491–1502.

Song, J.-H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, *13*(8), 360–366.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398.

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, *48*(12), 1391–1408.

Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing*, *17*(3), 279–305.

Sullivan, N., Hutcherson, C., Harris, A., & Rangel, A. (2015). Dietary self-control is related to the speed with which attributes of healthfulness and tastiness are processed. *Psychological Science*, *26*(2), 122–134.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, *12*(8), 291–297.

Tyldesley, D., & Whiting, H. (1975). Operational timing. *Journal of Human Movement Studies*.

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, *5*, e12192.

Verneau, M., van der Kamp, J., de Looze, M. P., & Savelsbergh, G. J. (2016). Age effects on voluntary and automatic adjustments in anti-pointing tasks. *Experimental Brain Research*, *234*(2), 419–428.

Von Helmholtz, H. (2013). *Treatise on physiological optics* (Vol. 3). Courier Corporation.

Welsh, T. N., & Elliott, D. (2004). Movement trajectories in the presence of a distracting stimulus: Evidence for a response activation model of selective reaching. *The Quarterly Journal of Experimental Psychology Section A*, *57*(6), 1031–1057.

Wispinski, N. J., Gallivan, J. P., & Chapman, C. S. (2020). Models, movements, and minds: Bridging the gap between decision making and action. *Annals of the New York Academy of Sciences*, *1464*(1), 30–51.

Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences*, *5*(11), 487–494.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

Wood, D. K., Gallivan, J. P., Chapman, C. S., Milne, J. L., Culham, J. C., & Goodale, M. A. (2011). Visual salience dominates early visuomotor competition in reaching behavior. *Journal of Vision*, *11*(10), 16–16.

Zago, M., Bosco, G., Maffei, V., Iosa, M., Ivanenko, Y. P., & Lacquaniti, F. (2004). Internal models of target motion: Expected dynamics overrides measured kinematics in timing manual interceptions. *Journal of Neurophysiology*, *91*(4), 1620–1634.

Zago, M., McIntyre, J., Senot, P., & Lacquaniti, F. (2008). Internal models and prediction of visual gravitational motion. *Vision Research*, *48*(14), 1532–1538.

Zago, M., McIntyre, J., Senot, P., & Lacquaniti, F. (2009). Visuo-motor coordination and internal models for object interception. *Experimental Brain Research*, *192*(4), 571–604.

Chapter 3

Primate-like perceptual decision making through deep recurrent reinforcement learning

3.1 Abstract

Progress has led to a detailed understanding of the neural mechanisms that underlie decision making in primates. However, less is known about why such mechanisms are present in the first place. Theory suggests that primate decision making mechanisms, and their resultant behavioural abilities, emerged to maximize reward in the face of noisy, temporally evolving information. To test this theory, we trained an end-to-end deep recurrent neural network using reinforcement learning on a noisy perceptual discrimination task. Networks learned several key abilities of primate-like decision making including trading off speed for accuracy, and flexibly changing their mind in the face of new information. Correlational and causal analysis showed that these abilities were supported by similar decision mechanisms as those observed in primate neurophysiological studies. These results provide experimental support for

Chapter 3 of this thesis has not been previously published. A version of this chapter has been presented as a talk at the 2022 Society for Neuroscience Conference as Wispinski, N. J., Stone, S. A., Singhal, A., Pilarski, P. M., & Chapman, C. S. (2021). Primate-like perceptual decision making through deep recurrent reinforcement learning.

key pressures that gave rise to the primate ability to make flexible decisions, and provide a model to further investigate biological cognition.

3.2 Introduction

The process of decision making determines how people choose between entrées at a restaurant, strategies in a competitive game, or votes between political candidates. Decision making in humans and non-human primates is well-studied and its neural mechanisms are increasingly well-understood. Behaviour (Ratcliff & McKoon, 2008), neural recordings (Shadlen & Newsome, 2001), and causal neural perturbations (Hanks et al., 2006; Salzman et al., 1992) all strongly support the idea that primates make decisions using a common mechanism termed evidence accumulation. Evidence accumulation models state that internal or external information is converted to momentary evidence in support of a decision. Momentary evidence is then accumulated over sequential samples in time to a decision threshold, which determines both choices and response times (Gold & Shadlen, 2007). Extensions to evidence accumulation models are also able to explain complex phenomena such as changes of mind, where primates flexibly change their commitment from one option to an alternative after considering new information (Atiya et al., 2020; Resulaj et al., 2009).

Theory suggests that primate decision making mechanisms, and their resultant behavioural abilities, emerged via the biological need to act in noisy, temporally uncertain environments (Cisek, 2012; Wispinski et al., 2020). Recent advances in artificial neural network research afford the ability to experimentally test such theories about emergence—by asking if networks optimized to perform a task given particular constraints and assumptions develop similar properties as the biological systems under investigation (Kanwisher et al., 2023; Kell & McDermott, 2019). For example, deep neural networks have provided compelling accounts for how primate-like image recognition (Kanwisher et al., 2023; Lindsay, 2021) and motion processing (Rideaux & Welchman, 2020) emerge when networks are trained to classify natural images via

supervised learning. Other work has shown that biological-like mechanisms to solve detour or motor control problems emerge when trained to reproduce behaviour from animals (Banino et al., 2018; Sussillo et al., 2015).

Here we ask if primate-like decision making emerges in artificial agents trained to maximize reward in a noisy, temporally uncertain environment. Specifically, we train agents via reinforcement learning to solve the random dot motion discrimination task (Shadlen & Newsome, 2001). In this task, decision makers are shown dots that move to the left or to the right with some level of random noise (termed coherence), and are asked to report in which direction the dots are moving. Because dot motion is noisy, decision makers need to consider multiple time samples of motion to make a good decision. The random dot motion task is widely used in perceptual decision making research in part because it acts as a proxy for an uncertain and dynamic world and allows for experimental control over environmental noise (Fig 3.1a). Agents were trained to complete this task using either a simulated saccadic response (as in many non-human primate studies; Britten et al., 1992; Roitman and Shadlen, 2002), or by controlling a two-degree-of-freedom arm (similar to collected human data; Resulaj et al., 2009; van den Berg et al., 2016).

Below, we identify five key properties of primate-like decision making that we aim to observe in these trained agents. Here we focus on algorithmic-level properties given the high level of abstraction in modeling primate brains with deep neural networks. As such, we do not consider several other important properties closer to the implementational-level of primate decision making like spike count variance (Churchland et al., 2011).

First, agents need to display stereotyped behavioural signatures. Animals tend to respond faster, more accurately, and with higher confidence during easy relative to hard decisions across a wide array of decision making tasks (Gold & Shadlen, 2007; Roitman & Shadlen, 2002; Wispinski et al., 2020)—also known as the three pillars of choice behaviour (Shadlen & Kiani, 2013). Agents also need to be able to

trade off increases in accuracy at the expense of decision speed; in humans, speed-accuracy trade-offs vary naturally between individuals, but can also be influenced via instructions or reward structures (Heitz, 2014; Palmer et al., 2005).

Second, agent internal dynamics should mirror those found in biological agents. Specifically, agent dynamics should match two functions identified from studies on the neural basis of decision making in primates: learned representations of relevant decision evidence, and the accumulation of this evidence (Gold & Shadlen, 2007). For instance, when making saccadic responses during the random dot motion discrimination task, recordings from primate medial temporal (MT) cortex suggest this area encodes the direction and magnitude of momentary motion on the screen (Britten et al., 1992). Specifically, MT cells selective for motion display tonic firing rates proportional to dot motion direction and coherence. Downstream, the lateral intraparietal area (LIP) is thought to accumulate this momentary motion evidence over time to a decision threshold, which determines in what direction the animal responds with and when (Roitman & Shadlen, 2002; Shadlen & Newsome, 2001). Specifically, selective LIP cells display a *change* in firing rate proportional to dot motion and coherence.

Third, targeted causal perturbation should predictably alter artificial agent behaviour in line with primate experiments. That is, while similar agent internal dynamics (the second key property) provide only correlational evidence for primate-like mechanisms, causal manipulations provide much stronger support for mechanistic understanding. Specifically, in the context of primate decision making, microstimulation in areas MT and LIP have differential effects on behaviour consistent with their proposed functions from neural correlate studies (Hanks et al., 2006; Salzman et al., 1992).

Fourth, agents should display distinguishing characteristics of flexible primate decision making. Specifically, primates are able to change their mind regarding which decision option they prefer in the face of new information—a hallmark of cognitive flexibility (Atiya et al., 2020; Resulaj et al., 2009). Changes of mind during deci-

sion making have been decoded from neural activity in primates (Kiani et al., 2014; Peixoto et al., 2018), even in real-time (Peixoto et al., 2021). Therefore, agent internal dynamics should suggest the same ability.

Fifth and finally, changes of mind have most notably been observed via movement trajectories in humans (Resulaj et al., 2009; van den Berg et al., 2016), where humans initially reach toward one target before switching in-flight to ultimately choose an alternative target. These changes often correct for initial mistakes, suggesting that they arise from the consideration of additional information to continuously improve accuracy while decision information is available. Changes of mind should not only be inferred from neural dynamics (as in the fourth key property), but also overtly observed through behaviour as an agent interacts with the environment.

Below, we investigate each of these five key properties in turn, and show that the trained agents meet all of the above criteria for primate-like decision making. We additionally describe two simple changes to agent architecture and training environment in which primate-like decision making does *not* emerge, suggesting critical pressures that contributed to the emergence of flexible decision making in biological agents. Overall, we argue these results provide unique insight into the origins of primate decision mechanisms, and additionally offer a computational model with which to complement future biological research.

3.3 Results

3.3.1 Agents learn stereotyped decision making behaviours

During perceptual decision making, animals tend to respond faster and more accurately during easy (high absolute coherence) relative to hard (low absolute coherence) decisions (Gold & Shadlen, 2007; Roitman & Shadlen, 2002; Wispinski et al., 2020). Here, trained artificial agents replicated these results with accuracy and response times that similarly varied with dot motion coherence and direction (Fig 3.2b). This

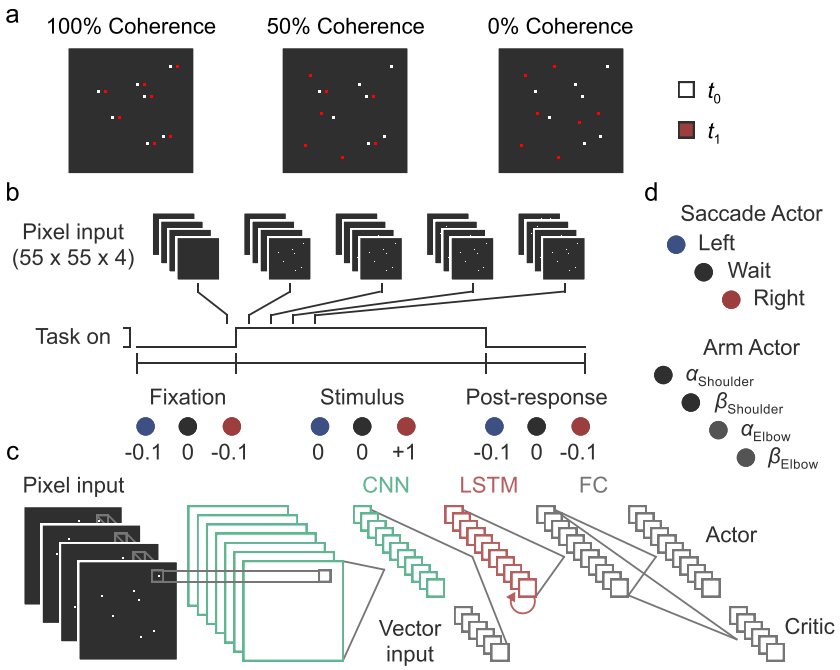

Figure 3.1: Task and agent network architecture. (a) Random dot motion discrimination stimuli with varying levels of noise. Dots at t_0 (white) either move coherently in one direction, or are replaced randomly at t_1 (red for illustration only) with an independent probability determined by coherence. All examples are from trials where the correct response is "right". (b) Saccadic random dot motion task structure. Agents are rewarded with +1 for responding correctly during the stimulus period, -0.1 for responding before or after the stimulus period, and 0 otherwise. (c) Actor-critic agent architecture. Pixel and vector input are passed left to right through layers of a deep neural network consisting of convolutional and sum pool operations (CNN), a recurrent layer (LSTM), and fully-connected (FC) layers. Critic output is a single linear unit. Actor output depends on the action space of the task. (d) Agents responding with saccades have access to discrete "left", "wait", and "right" actions. Agents responding with continuous arm movements control shoulder and elbow joint forces, parameterized by alpha (α) and beta (β) parameters for independent beta distributions (see Methods). After training, network parameters were frozen and performance was analyzed.

pattern of accuracy over coherence levels was well described by a logistic function for all agents ($R^2 = 0.996 \pm 0.0004$). Trained agents rarely failed to respond by the end of the trial, and successfully withheld responses outside of the dot motion period (mean trials without responses: $0.35\% \pm 0.11\%$).

In the random dot motion task, individual decision makers can trade off increases

in accuracy at the expense of decision speed (Palmer et al., 2005). Similarly, differences in the reinforcement learning discount rate hyperparameter (γ; see Methods) during training altered the speed-accuracy tradeoff between individual trained agents at evaluation time (Fig 3.2b, c, d). There was a significant change in the slope (linear mixed effects regression; $b_1 = 2184.68 \pm 199.21, p = 5.51e\text{-}28$; Fig 3.2c) of logistic functions fit to choices, and in mean reaction time (linear mixed effects regression; $b_1 = 11335.65 \pm 3162.19, p = 3.47e\text{-}4$; Fig 3.2d) between models trained with discount rates of 0.96, 0.98, 0.99, and 1.0, respectively. No group's indifference points significantly differed from 0% coherence (one-sample t-tests, ps > 0.05; Fig 3.2c). Overall, trained agents displayed stereotyped speed and accuracy signatures of primate-like decision making.

The level of accuracy given the time agents took to respond suggests that these agents considered multiple samples of motion in support of their decision. Simulating an evidence accumulation model shows that agents exceeded the maximum accuracy achievable from considering only a single time step of dot motion, $t(9) = 33.34, p = 4.84e\text{-}11$ (80.3% line in Fig 3.2a; see Methods). Accuracy during training suggests that agents started to consider multiple steps of motion in support of their decision after roughly one million steps of experience. In contrast, agents that were instead trained on noiseless motion (i.e., only coherences of 100%), failed to exceed this one-sample accuracy threshold on average (see Appendix A). Consistent with theory, this suggests that environmental noise was a critical factor in the emergence of primate decision making abilities.

However, patterns of response speed and accuracy alone are not enough to support claims of primate-like decision making. It is possible for similar speed and accuracy patterns to arise from decision making mechanisms other than evidence accumulation—some with limited support in primates (Stine et al., 2020, see Appendix A). For example, an extrema detection mechanism compares individual samples of evidence against response thresholds *without* accumulation. With this mechanism, a decision

maker waits until an individual sample is large enough to trigger a response. Therefore, to distinguish between primate-like and non-primate-like decision mechanisms, we now turn to the internal dynamics of these trained agents.

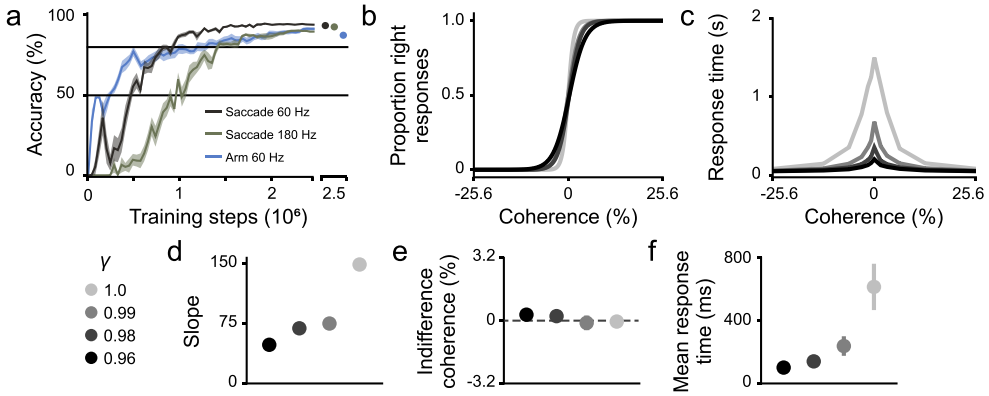

Figure 3.2: Behavioural performance of agents during and after training. Agents were trained on the random dot motion task to respond either with discrete actions (left-/right saccades), or continuous actions (controlling the shoulder and elbow forces of a simulated arm; in blue). Saccade agents were trained at either 60 Hz (black) or 180 Hz (green), as the 180 Hz agents allow for higher temporal resolution to investigate dynamics (see Methods). Results did not differ between 60 and 180 Hz agents. (a) Periodic evaluation of agents during training (coloured traces), and final evaluation after training (dots). Shaded areas (traces) and vertical bars (dots) represent standard errors across 10 random seeds. Horizontal line at 50% accuracy represents chance performance on the random dot motion task for an agent that always responds within the stimulus period. Horizontal line at 80.3% represents maximum accuracy achievable from considering only a single time step of dot motion by a hand-constructed evidence accumulation model (see Methods). (b) Average accuracy by agent discount rate (γ; N = 15 per γ). Agents trained with higher discount rates (light gray) were more accurate than those trained with lower discount rates (dark gray). Models are 60 Hz saccade agents. (c) Average response time by agent discount rate. Agents trained with higher discount rates (light gray) were slower to respond than those trained with lower discount rates (dark gray). (d) Slopes were extracted from logistic functions fit to each agent's final evaluation accuracy (in b). (e) Indifference points were extracted from logistic functions fit to each agent's final evaluation accuracy (in b). (f) Mean response time by discount rate condition (extracted from c).

3.3.2 Dynamics reflect momentary and accumulated decision evidence

To support a claim of primate-like decision making, agent internal dynamics should be selective for both momentary and accumulated motion direction and coherence, consistent with an evidence accumulation framework and neuronal recordings from primate areas MT and LIP (Britten et al., 1992; Roitman & Shadlen, 2002; Shadlen & Newsome, 2001). Here we investigate individual unit activations from the sum pool layer of the network (CNN; Fig 3.1b), and the recurrent layer of the network (LSTM; Fig 3.1b) as analogues for areas MT and LIP in primates during saccadic decision making (see Methods).

Several CNN units ($75\% \pm 1.4\%$) consistently showed a sustained response proportional or inversely proportional to momentary motion strength and direction (Fig 3.3a; see Appendix A). These results are consistent with the idea that CNN units allow the agent to compute a measure related to momentary motion evidence on every time step, similar to the proposed function of primate area MT in this task (Gold & Shadlen, 2007). In addition, several LSTM units ($72.9\% \pm 1.3\%$) showed a significant relationship between buildup slope and coherence (Fig 3.3c; see Appendix A), as in primate area LIP (O'Connell et al., 2018; Roitman & Shadlen, 2002). When LSTM dynamics were aligned to response, the activity of a subset of units appeared to meet or exceed a threshold level, mimicking the proposed decision threshold gating saccadic responses (Gold & Shadlen, 2007). Specifically, these response-aligned dynamics were most consistent with a decision threshold that collapsed over time (e.g., Drugowitsch et al., 2012). LSTM and CNN dynamics were both robust to the training of several agents that differed only in initial random seeds.

Together, these dynamics suggest agents learned a two-part, primate-like evidence accumulation mechanism. If agents instead learned an alternative, non-accumulation mechanism such as extrema detection (Stine et al., 2020), units selective for accumulated decision evidence would not be predicted to emerge. Indeed, when agents were

trained without recurrence, behaviour was similar to recurrent agents, and agents showed several units selective for momentary decision evidence (see Appendix A). However, LSTM dynamics strongly suggest that these non-recurrent agents make non-primate-like decisions, based on a single, extreme sample of motion.

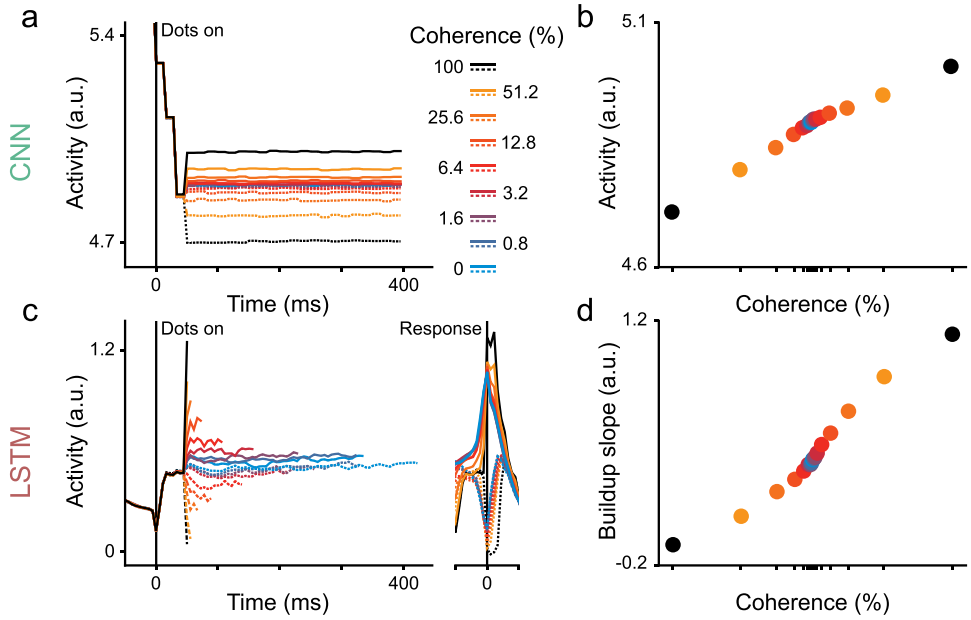

Figure 3.3: Internal dynamics of a representative agent in the 180 Hz saccade task. Solid lines indicate rightward motion trials, and dashed lines indicate leftward motion trials. Data are from correct trials only. For dynamics on the left, individual trials making up the plotted averages are considered up until the first of: response time or median response time. (a) An example CNN unit with average activity proportional or inversely proportional to momentary motion coherence and direction, similar to primate area MT. (b) The average activity of this CNN unit is roughly linearly related to motion coherence. (c) An example LSTM unit with activity proportional or inversely proportional to accumulated momentary motion coherence and direction, similar to primate area LIP. Note the blue (0%) traces, where solid lines indicate the average of trials that ended in a rightward response, and dashed lines indicate the average of trials that ended in a leftward response. In the middle panel, the same data is time-locked to response time. Dynamics increase or decrease roughly proportional to dot motion coherence. (d) The average buildup slope, from dot motion onset to the first step of full motion, of this LSTM unit is roughly linearly related to coherence. To compare with non-human primate neurophysiological recordings, see Gold and Shadlen (2007).

3.3.3 Causal stimulation predictably alters behaviour

Causal manipulations, such as microstimulation, in tandem with modeling have shown that MT and LIP have differential effects on behaviour that align with their function from theory (Hanks et al., 2006; Salzman et al., 1992). Here we increase or decrease the activation of single CNN or LSTM units over the course of the motion stimulus on a trial and observe similar causal changes in behaviour as microstimulation in primates.

Example CNN and LSTM units were identified based on their selectivity—half for leftward and half for rightward motion (see Methods). Individual unit activations during motion were increased or decreased throughout the motion stimulus, and agent evaluation was repeated independently for each individual unit being stimulated. Consistent with theoretical predictions, and primate experiments, stimulation of motion-selective units corresponded to increases in response frequency and decreases in response time in the preferred direction of the unit, relative to no stimulation (Fig 3.4). These effects were observed for both CNN and LSTM unit stimulation.

In primates, there are subtle differences between the impact that MT and LIP stimulation has on choices and response times (Hanks et al., 2006). Here we adapt the drift diffusion model described in Hanks et al. (2006) to illustrate these differences. MT stimulation, thought to change *momentary* evidence, causes a lateral shift in behaviour with respect to coherence (i.e., a translation; Fig 3.4a, b). Conversely, LIP stimulation, thought to change *accumulated* evidence, causes changes in choices and response times reflected about the 0% coherence point (i.e., a transformation; Fig 3.4c, d). That is, choices are more likely and response times are quicker for the stimulated direction, and choices are less likely and response times are slower for the un-stimulated direction. Here we observe that stimulating single units in the CNN and LSTM layers of the network qualitatively align with these subtle differences as predicted by drift diffusion modeling. Overall, these results suggest that the CNN

and LSTM layers in the current agents perform similar functions in support of solving the dot motion task as primate areas MT and LIP, respectively. Namely, the CNN layer encodes the momentary motion energy on every step, and a subset of units in the LSTM layer encode the accumulated evidence in support of each response.

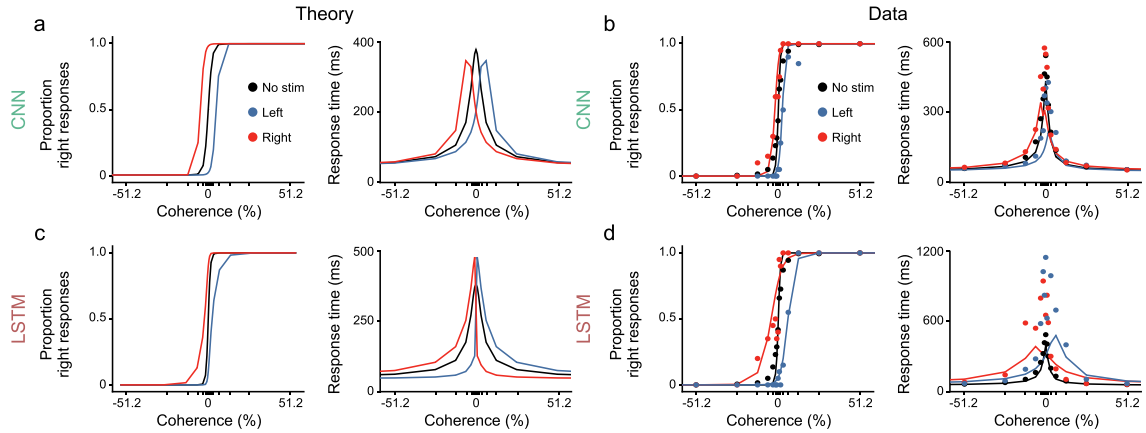

Figure 3.4: Behavioural effects of targeted microstimulation in a representative agent trained in the 180 Hz saccade task. On the left, theoretical behavioural results generated from a modified drift diffusion model (see Methods). Under these predictions, CNN and LSTM stimulation have the same impact as the theorized functions of primate areas MT and LIP, respectively (Hanks et al., 2006). On the right, observed behavioural results from stimulating example leftward and rightward selective CNN and LSTM units (dots), and drift diffusion model fits to these observed data (lines). (a) Theory predicts that changing *momentary* evidence shifts choices and response times laterally with respect to coherence. Right stimulation (red) and left stimulation (blue), relative to no stimulation (black). (b) Observed results from stimulating an example rightward selective CNN unit (red), an example leftward selective CNN unit (blue), or no stimulation (black). (c) Theory predicts that changing *accumulated* evidence causes choices to be more likely and response times to be quicker in the stimulated motion direction, and choices to be less likely and response times to be slower in the un-stimulated motion direction (about the 0% coherence point). (d) Observed results from stimulating an example rightward selective LSTM unit (red), an example leftward selective LSTM unit (blue), or no stimulation (black).

3.3.4 Changes of mind decoded from dynamics

We next look at the adaptive behaviour of these trained agents by investigating changes of mind—a phenomenon where a decision maker revises their decision online in the face of new information (Resulaj et al., 2009), and a behavioural hallmark

of an animal that can best adapt to a noisy and unpredictable world. Changes of mind (CoMs) have been decoded from neural activity in primates (Kiani et al., 2014; Peixoto et al., 2018), even in real-time (Peixoto et al., 2021), and have their own key properties (Peixoto et al., 2021). First, changes of mind are more frequent when decisions are more difficult (i.e., during low relative to high coherence trials). Second, changes of mind are more likely to be corrective than erroneous. In other words, changes more likely move from an initially incorrect choice to an ultimately correct one rather than vice versa, indicating that these changes are based on additional information to improve the accuracy of a decision.

Similar to primate studies (Kiani et al., 2014; Peixoto et al., 2018; Peixoto et al., 2021), we train a linear decoder to predict the choices of a decision maker throughout a trial based on evolving neural activity. Specifically, we train a logistic regression classifier, which decodes a decision variable (DV) to predict left/right choices based on LSTM layer activity at every time step. Changes in the decoder's prediction (i.e., a change of sign in the DV) before a response suggest a neural change of mind (Fig 3.5).

Consistent with non-human primate experiments, changes of mind were more frequent when decisions were more difficult (linear regression of the proportion of CoMs by log coherence; $b_1 = -0.33 \pm 0.006, p = 0.001$). Importantly, changes of mind were more likely to be corrective than erroneous (chi-squared goodness-of-fit test; $\chi^2 = 43.28, p = 4.74e\text{-}11$). Overall, these results show that the trained artificial agents considered here are able to learn highly flexible, error-correcting behaviour similar to primates.

3.3.5 Changes of mind in movements

Ballistic tasks such as those requiring a saccadic response are the exception rather than the rule of decision making in animals. More typically, animals must execute temporally and spatially extended movements to interact with the world (Wispinski

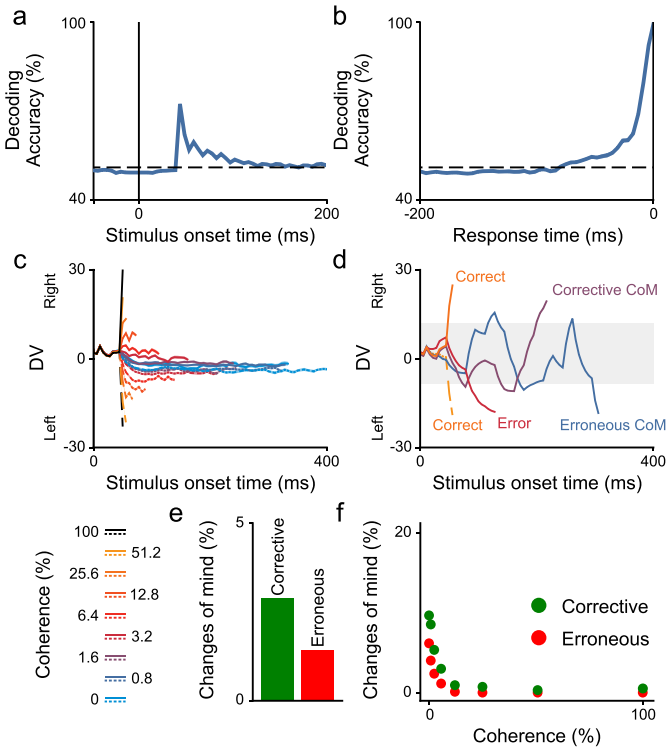

Figure 3.5: Decoding of neural changes of mind in a representative agent trained in the 180 Hz saccadic task. (a) 10-fold cross validation accuracy for a logistic decoder trained on LSTM layer activity. Dashed line indicates chance accuracy. Decoding time-locked to dot motion stimulus onset. Standard errors are smaller than the decoding line widths. (b) Decoding time-locked to response. (c) Logistic decoder decision variable (DV) by coherence condition. Shaded regions denote standard errors. Traces plotted until median response time. Solid traces indicate rightward motion trials, and dashed traces indicate leftward motion trials. (d) Example of single trial DV traces. Line colours correspond to the coherence conditions of each single trial. Shaded region denotes change of mind threshold (see Methods). (e) Decoded changes of mind. Agents displayed more corrective (green) than erroneous (red) changes of mind. f) Decoded changes of mind by coherence. Corrective changes (green) are those where the model DV at some point predicts an incorrect choice before an ultimately correct choice is made. Erroneous changes (red) are those where the model DV predicts a correct choice before an ultimately incorrect choice is made.

et al., 2020). When biological agents move to make a decision, valuable time elapses, and the consideration of decision information extends into movement (Michalski et al., 2020; Resulaj et al., 2009; Wispinski et al., 2020). This continuous consideration of information throughout a movement leads to the final key property of primate-like decision making we consider—decision-related movement fluctuations (J.-H. Song &

Nakayama, 2009; Wispinski et al., 2020), and changes of mind observable in movement trajectories (Resulaj et al., 2009; van den Berg et al., 2016).

We trained a new set of end-to-end agents that instead responded by controlling the joint forces of a simulated two-degree-of-freedom planar arm to move a fingertip toward a left or a right target (see Methods). We compared this agent behaviour to collected data from 13 humans in a similar reaching task, where participants would perform the random dot motion discrimination task by reaching to one of two targets on an upright screen. Human participants also rated their confidence after every decision, and feedback was withheld.

With the saccade agents above, decoding an agent's internal state reveals the evolution of an agent's decision state between left and right options throughout a trial (Fig 3.5). In this continuous control task, decision states can instead be inferred from a decision maker's movements in physical space between left and right options. Looking at these movements, both humans and artificial agents displayed more curved movement trajectories on hard trials, and changed their mind while moving to correct for initial errors. Similar to those decoded from neural activity, behavioural changes of mind were more frequent when decisions were more difficult (linear mixed effects regression; Humans: $b_1 = -0.055 \pm 0.017, p = 0.0015$; Artificial agents: $b_1 = -0.20 \pm 0.087, p = 0.024$). Importantly, these behavioural changes of mind tended to be more corrective than erroneous (paired one-tailed t-tests; Humans: $t(12) = 6.11, p = 2.61e\text{-}5$; Artificial agents: $t(9) = 1.86, p = 0.048$), suggesting that both the humans and artificial agents flexibly altered movements online as a result of incoming information. In addition to behavioural similarities, the trained agents in this continuous control task developed similar internal dynamics as the trained agents in the saccadic response task, suggesting that these agents also continuously integrate evidence over time in support of their decision (see Appendix A).

Up to now, we have only considered speed and accuracy behaviour, but not the third pillar of choice behaviour—confidence (Shadlen & Kiani, 2013). Animals tend

to respond with higher confidence during easy relative to hard decisions across a wide array of tasks. Results show human post-decision confidence judgements follow these established patterns (Fig 3.6e). However, decision confidence is difficult to determine without language. For example, without the ability to directly ask about the confidence of non-human primate decision makers, researchers employ alternative tasks such as post-decision wagering (Kepecs & Mainen, 2014), or analyze neural patterns (Kepecs et al., 2008).

While we cannot ask the current reinforcement learning agents to verbalize metacognitive judgments about their confidence after each decision, we can leverage their architecture to answer similar questions. The current agents use an actor-critic reinforcement learning method (Sutton & Barto, 2018, see Methods). In short, part of the agent's output represents its policy—the probabilities of each action to be selected on each time step (i.e., the actor; Fig 3.1b). The other part of the agent's output represents the agent's estimate about future cumulative discounted rewards (i.e., the critic; Fig 3.1b). In brief, actor-critic methods in reinforcement learning work by improving the critic's estimate about rewards based on experiences, and changing the probabilities of actions relative to the critic's estimates (Sutton & Barto, 2018). Here we can query the critic's reward estimate on the final step of each trial, just before the artificial agent touches one of the two targets. This value output approximates the agent's prediction of a correct response (+1) over an incorrect response (0; see Methods). The pattern of agent critic output emerged to closely match the pattern of human post-decision confidence judgements—both consistent with theory (Wispinski et al., 2020) and primate experiments (Kiani & Shadlen, 2009). Overall, these patterns show a proxy for emergent primate-like confidence in artificial agents, completing the three pillars of choice behaviour (Shadlen & Kiani, 2013).

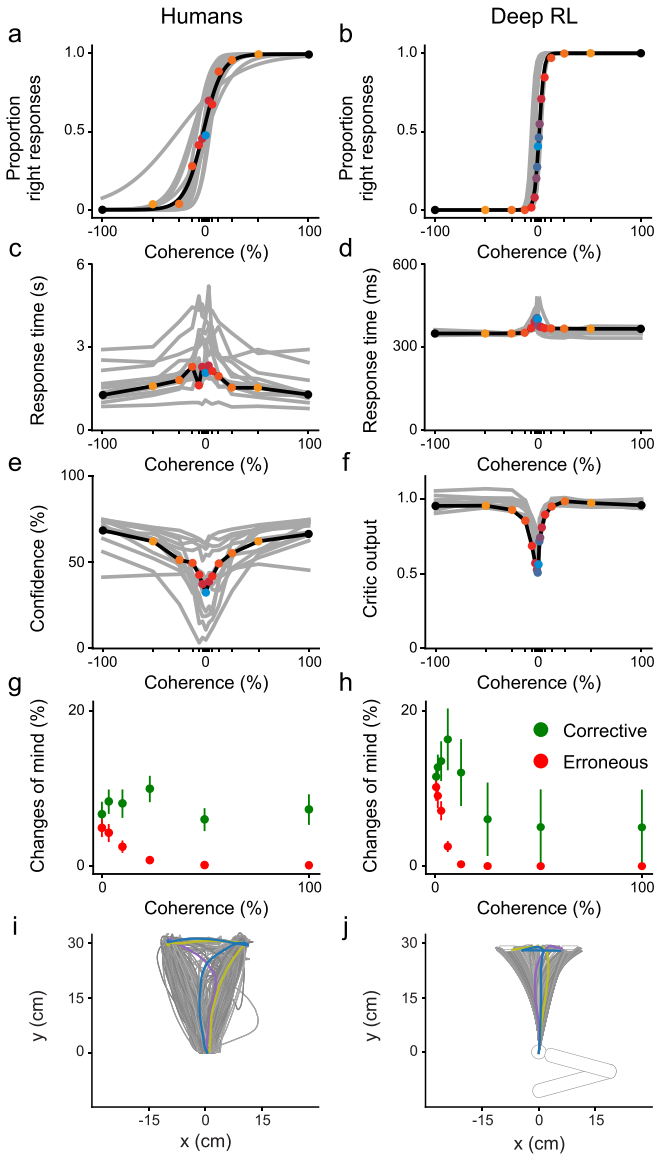

Figure 3.6: Comparison between human reaching performance (top row; N = 13) and trained deep reinforcement learning agents controlling a 2DOF arm (bottom row; N = 10). Individual representative human and trained agent highlighted (black line, coloured dots), and all other individuals (gray lines). (a, b) Accuracy. Logistic function fits; humans: $R^2 = 0.957 \pm 0.019$; artificial trained agents: $R^2 = 0.998 \pm 0.001$. (c, d) Response time. Note differences in scaling between humans and artificial agents. (e, f) Human post-decision confidence ratings, and network critic output at response time. (g, h) Mean changes of mind across all subjects. Vertical lines indicate standard errors. (i, j) Movement trajectories from an individual representative human and artificial agent (2DOF arm shown), with selected changes of mind highlighted.

3.4 Discussion

Here we show that artificial agents trained to maximize reward in the face of noisy, temporally evolving information learn behaviours and internal dynamics similar to primate decision makers. Agents learned to trade off speed and accuracy, to estimate future rewards which closely aligned with human post-decision confidence judgements, and to flexibly change their mind in the face of new information. These results also suggest an underlying decision making mechanism similar to evidence accumulation, which is thought to underlie many perceptual- and value-based decisions in biological agents (Shadlen & Shohamy, 2016).

That these phenomena presented here emerge in artificial agents support theory that these mechanisms emerged similarly in biological systems—via the need to act in noisy, temporally uncertain environments (Cisek, 2012; Wispinski et al., 2020). It may seem unsurprising that the agents considered here discover a similar mechanism to evidence accumulation, given that evidence accumulation models are thought to be the optimal solution to several perceptual decision making tasks (Ratcliff et al., 2016). However, we identify two simple changes where primate-like decision making does *not* emerge. Agents without recurrence, and agents trained on noiseless motion were both able to complete the task. However, both failed to reliably display key properties of primate-like decision making, suggesting at least two critical pressures for primate-like decision making to emerge—recurrence and environmental noise. The current results fit with other work showing that training artificial agents on biological tasks encourages biological-like solutions, for example motion processing (Rideaux & Welchman, 2020), detour problems (Banino et al., 2018), and patch foraging (Wispinski et al., 2022). Overall, these results support long-standing theory that biological intelligence (and biological-like artificial intelligence) may emerge through optimization in environments that mirror developmental (Smith & Gasser, 2005; Turing, 1950) or evolutionary pressures (Cisek, 2019) faced by biological systems (Kanwisher et al.,

2023; Kell & McDermott, 2019). Further, agents learned these abilities solely through reinforcement learning, providing additional experimental support for the idea that reward signals may be enough for the development of intelligent, human-like artificial systems (Silver et al., 2021).

The agents considered here are simulated at a high level of biological abstraction and do not consider several components relevant to biological researchers. Future work may further consider the simulation of decision making using spiking neural networks (Lillicrap et al., 2016), biologically plausible methods of network weight updating (Bengio et al., 2016), or considering architectures with long-range recurrence. The current agents are also simulated with no neural or motor delays or noise, which is apparent when looking at the low response times and low behavioural and neural variability of these agents relative to biological decision makers. Future work may consider integrating these limitations into artificial agents to better understand biological solutions via deep reinforcement learning.

Here we describe agents that meet all of the above criteria for primate-like decision making. However, it should be noted that the majority of the above criteria are also well-captured by extensions to evidence accumulation models (Kiani & Shadlen, 2009; Ratcliff & McKoon, 2008; Resulaj et al., 2009; Shadlen & Newsome, 2001). Additionally, evidence accumulation mechanisms have similarly emerged in artificial agents trained to produce discrete actions via supervised learning (Mante et al., 2013) or reinforcement learning (H. F. Song et al., 2017) given noisy numerical input. Why then use models with hundreds of thousands of parameters when hundreds (H. F. Song et al., 2017), or fewer than ten (Resulaj et al., 2009), can capture many of the same results? In contrast to rule-based models, the agents described here discover similar mechanisms on their own solely from rewards. Second, agents do not simply reduce to evidence accumulation models—these agents at the same time also learn relevant sensory representations from raw pixels, and map learned internal states onto joint forces in order to effect decisions in a simulated environment. We argue

that end-to-end learning is important when testing emergence hypotheses, as it also considers the biological problems of simultaneously learning sensory representations and motor control policies that may have competing objectives.

The agents here are not intended to replace other computational models of decision making. Instead, we argue they complement existing models. The deep learning agents here would be much more difficult to explain without the use of an evidence accumulation framework. Conversely, we argue these agents can provide fully accessible systems to better predict and explain biological behaviour and computation. For example, these trained agents provide a model with which to further investigate the computations underlying dynamic decision making and motor control—especially in motor control tasks where biological data is difficult and time-consuming to collect, and neural activity is often more difficult to interpret relative to stationary, in-lab tasks (Musall et al., 2019). Deep learning models of decision making also potentially provide ways to investigate other aspects of biological cognition such as social decisions (Rorie & Newsome, 2005; Shadlen & Kiani, 2013), ecological foraging (Davidson & El Hady, 2019; Wispinski et al., 2022), and even consciousness (Kang et al., 2017), which are all thought to rely to varying degrees on similar evidence accumulation processes (Shadlen & Kiani, 2013).

3.5 Methods

3.5.1 Motion discrimination task

Agents completed a reaction time version of the random dot motion discrimination task commonly used in studies of decision making in humans and non-human primates (Roitman & Shadlen, 2002; Shadlen & Newsome, 2001). Agents were shown a video of noisy dots moving to the left or to the right, and guessed in which direction the dots were moving and when to respond.

On each frame, seven dots were drawn within a circular aperture within a 55 x 55

pixel image. Each dot had an independent probability of moving at a fixed angle and distance from its current location on the next frame, or being drawn at a new random location. This probability of dot motion, also known as coherence, determined the strength of motion on each trial. Although the expectation of motion strength in each coherence condition over time is constant, motion strength is noisy on single trials. During training, dots could move either left or right, and dot motion strength on each trial was selected from seven coherence values used in primate decision research: [0%, 3.2%, 6.4%, 12.8%, 25.6%, 51.2%, 100%]. Motion direction and coherence varied randomly from trial-to-trial, but remained fixed within each trial. Responses on zero coherence trials were rewarded with 50% probability since there was no correct response on these trials. The dot motion stimulus was simulated at 60 Hz, in line with refresh rates of motion stimuli presented to mammals (Katz et al., 2016; Shadlen & Newsome, 2001).

Dot motion discrimination stimuli are often presented with three interleaved sets of motion, such that dots on frame one are moved or randomly replaced on frame four, while an independent set of dots are presented on frames two and five, etc. (Shadlen & Newsome, 2001). During training, agents were presented either dot motion stimuli with three interleaved sets of motion (as in a typical biological experiment), or stimuli with no interleaved frames to approximate natural consistent motion (as in Rideaux & Welchman, 2020), randomly determined at the start of each training trial with equal probability. After training, evaluation and all subsequent analyses were completed using motion stimuli with three interleaved frames. Dot speed was kept the same regardless of the number of interleaved frames—in other words, coherently moving dots were displaced one pixel horizontally for stimuli with no interleaved frames, and three pixels horizontally for stimuli with three interleaved frames.

Agents responded via a simulated saccadic response as in most non-human primate decision research, or a simulated reaching movement to one of two targets modeled after human data (see Human data). In the saccadic task, agents responded with

discrete “left”, “right”, or “wait” actions. In the reaching task, agents controlled the shoulder and elbow forces (continuous, [-1, +1]) of a two-degree-of-freedom planar arm in the MuJoCo physics engine (Todorov et al., 2012).

Correct responses during dot motion presentation corresponded to a reward of +1, while incorrect responses corresponded to a reward of 0. If an agent responded before stimulus onset or after stimulus offset, it received a reward of -0.1. Wait actions on every time step before, during, and after stimulus presentation corresponded to a reward of 0. In the reaching version of the task, agents were additionally rewarded on each time step based on the forward distance (in meters) of the simulated fingertip (multiplied by a scaling coefficient) to encourage forward reaching movements. Specifically, the fingertip of the agent started at a distance of 0 m on every trial, and the two targets and screen were located 0.3 m forward from this start position. With a scaling coefficient of 0.005 (see Supplementary Table A.1), this meant that agents would receive an additional reward of 0 on every step at the start position, and a maximum additional reward of 0.0015 on every time step when the finger was touching the screen. Agents with this small reward for moving forward learned the task quicker and more reliably than those without the reward, as it encouraged the agents to explore states further from the start position early on in training.

The saccadic random dots task was simulated as trials that ended 5 steps after a response, or after 3 seconds without a response. For the saccadic agent simulated at 180 Hz, the trial ended 15 steps after a response, or after 2 seconds without a response. For the reaching task, trials always ended when the fingertip made contact with the target, or after 3 seconds without a response. Stimuli onset times on each trial were drawn from a random uniform distribution (see Supplementary Table A.1; Fig 3.1b), and dot motion was extinguished immediately after a response for the remainder of the trial. For the reaching task, the agent’s arm was held in place until 4 frames of dot motion had been input to the network.

Saccadic agents were simulated at both 60 Hz and at 180 Hz. 60 Hz agents per-

formed one forward pass of information through the neural network and selected actions in sync with new dot motion frames. In contrast, 180 Hz agents performed three forward passes, and corresponding action selections, per new dot motion frame. While slower and more difficult to train, 180 Hz agents allow for investigation of internal dynamics at a higher temporal resolution than 60 Hz agents. Results did not qualitatively differ between 60 Hz and 180 Hz agents.

3.5.2 Network architecture

All networks were implemented in Python version 3.9 (`https://python.org`) using The DeepMind JAX Ecosystem (Babuschkin et al., 2020).

The neural network described below accepted 55 x 55 x 4 pixel input, corresponding to the most recent four frames of the dot motion stimulus. Stacked frames as input have been used in deep reinforcement learning agents for playing Atari games (Mnih et al., 2015), and in supervised learning networks to recreate several properties of primate motion-selective area MT (Rideaux & Welchman, 2020).

At each time step, input of shape 55 x 55 x 4 was convolved with 64 3D kernels each with a shape of 5 x 5 x 4 to produce 64 convolutional output maps. Input was padded with zeros so that each convolutional output map was of shape 55 x 55. Units used a rectified linear (ReLU) activation function to model neurophysiological data (Rideaux & Welchman, 2020). Each convolutional output map was then summed so that network output was reduced to 64 values (one for each map). Maps were summed across space as the dot motion discrimination task relies on global, rather than local, perception of motion. Convolution as a first operation was chosen because of the sharing of parameters across image space as in many deep learning models of biological vision (Lindsay, 2021), and because the convolution of stacked frames of moving images has been shown to approximate key properties observed in primate area MT (Rideaux & Welchman, 2020). These 64 sum-pooled outputs are referred to as "CNN" throughout.

The 64 CNN outputs were then concatenated with a vector of task-relevant inputs. In the saccadic version of the task, these were a binary task off signal, a binary task on signal, the agent's action on the previous step, and the reward on the previous step. In the reaching version of the task, vector inputs additionally included the sine and cosine of shoulder and elbow angles, the velocity of the shoulder and elbow, and the x and y distances of the fingertip to both left and right targets.

After concatenation, inputs were fed to a layer of 128 LSTM units (Hochreiter & Schmidhuber, 1997). LSTM units introduce recurrence by copying their internal 'cell state' between time steps. LSTM units are also gated by 'forget', 'input', and 'output' gates, which allow these units to choose to forget information, allow new information to enter, and contextually output memory contents at each time step. The dynamics of these units are governed by the standard equations:

$$i_t = \sigma(W_{ii}x_t + W_{hi}h_{t-1} + b_i)$$

$$f_t = \sigma(W_{if}x_t + W_{hf}h_{t-1} + b_f)$$

$$g_t = \tanh(W_{ig}x_t + W_{hg}h_{t-1} + b_g)$$

$$o_t = \sigma(W_{io}x_t + W_{ho}h_{t-1} + b_o)$$

$$c_t = f_t \cdot c_{t-1} + i_t g_t$$

$$h_t = o_t \cdot \tanh(c_t)$$

Where x_t is the LSTM input at time t, h_t is the hidden state, i_t, f_t and o_t are the input, forget, and output gate activations, c_t is the cell state, g_t is a vector of cell state updates, and σ is the sigmoid function. All LSTM states were initialized to zero at the beginning of each trial. Recurrent units, such as LSTMs, were chosen because they have been shown in cases to approximate the accumulation of evidence in favor of a decision given noisy numerical input (H. F. Song et al., 2017). LSTM layer output was finally fed through a fully-connected layer with 128 ReLU units.

Output was then fully-connected to independent actor and critic network heads,

consistent with many deep reinforcement learning architectures (Sutton & Barto, 2018). The critic network head consisted of 64 fully-connected units with ReLU activations connected to a single linear unit. The critic head acts to estimate the expected return of the agent's policy given the current state, s (see Training).

In the saccadic network, the actor network head consisted of 64 fully-connected units with ReLU activations connected to three linear output units, one for each action available at every time step (left, wait, right). A softmax operation was performed on the three output units so that their sum was equal to one, and an action at each time step was randomly selected from these probabilities. This part of the network determined the agent's action policy, π.

For the reaching network, the final layer of the actor head consisted of four output units with a shifted softplus activation function ($[1, \infty]$). These outputs corresponded to alpha and beta parameters for a beta distribution (Chou et al., 2017)—one for shoulder and one for elbow forces. During training, joint forces were randomly sampled from beta distributions parameterised by the network at each time step.

3.5.3 Training

Reinforcement learning was implemented by the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017). In brief, the full objective function is a weighted sum of a clipped policy gradient loss (L_π), a state-value function loss (L_V), and an entropy regularization term (L_H):

$$L = L_\pi + \beta_V L_V + \beta_H L_H$$

Where β_V and β_H are coefficients (see Supplementary Table A.1). The clipped policy gradient loss (L_π) is defined by:

$$L_\pi(\theta) = \hat{E}_t[\min(\rho_t(\theta)\hat{A}_t, \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]$$
$$\rho_t = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

Where $\hat{E}_t$ denotes the empirical expectation over time steps, ρ_t is the ratio of state-action probabilities under the new and old policies, respectively, and ϵ is a clipping hyperparameter. $\hat{A}_t$ is the estimated advantage at time t, defined by the truncated generalized advantage estimator from Schulman et al. (2017):

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + ... + (\gamma\lambda)^{T-t+1}\delta_{T-1}$$
$$\delta_t = r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)$$

Where γ is a discount factor $\in$ [0, 1], λ is a mixing parameter $\in$ [0, 1], r_t is the experienced reward at time t, $V(s_t)$ is the critic head output given state s_t, and t specifies the time index in [0, T] within a given length-T trajectory segment. The state-value function loss (L_V) is defined by (Huang et al., 2022; Schulman et al., 2017):

$$L_V(\theta) = \hat{E}_t[(R_t - V_\theta(s_t))^2]$$

Where R_t is the TD(λ) return (Sutton & Barto, 2018). Finally, the entropy regularization term is the mean entropy of the policy distribution given all states in the batch:

$$L_H(\theta) = \hat{E}_t[H(\pi_\theta(\cdot \mid s_t))]$$

Network parameters were updated after collecting each batch of experiences using backpropagation through time and the Adam method for gradient-based optimization (Kingma & Ba, 2014, see Supplementary Table A.1). Unless stated, all agents

were run with the same hyperparameters and different random seeds initializing network parameters and the starting environment state. Hyperparameters were chosen through an informal search.

All 10 random seeds of the 60 Hz saccade agents, 180 Hz saccade agents, and 60 Hz reaching agents exceeded 80.3% accuracy at final evaluation, and so were not rejected from analysis (Range: 81.7% by a 60 Hz reaching agent, and 95.5% by a 60 Hz saccade agent). Distributed training used 20 parallel CPU cores, and took approximately 10, 20, and 12 hours for the 60 Hz saccade, 180 Hz saccade, and 60 Hz reaching agents, respectively (including online and post-training evaluation).

After training, network parameters were frozen and agents were evaluated on a version of the task with three interleaved sets of motion. 500 trials were simulated in each coherence and direction condition. Consistent with many deep reinforcement learning experiments (e.g., Lillicrap et al., 2016), agents were evaluated with deterministic actions: in the saccadic network, the most probable of all 3 actions at every step; in the reaching network, the mode of each beta distribution at every step. A single representative model was selected for stimulation and decoding experiments, however results extend to all models analyzed.

3.5.4 Evidence accumulation model

We evaluated agent performance against a lossless accumulation model with *a priori* information about the dot motion stimulus to estimate the degree of temporal information used by the agent (see Fig A.3).

First we constructed two 5 x 5 x 4 convolutional kernels for left and right motion at the ground truth of the dot motion stimulus (i.e., left or right motion with three interleaved frames at a speed of 1 pixel per frame). The stimulus was convolved with these kernels, output was summed across space (i.e., from local activation to global activation), and the difference between these left and right motion kernels was taken to approximate net motion energy (see Adelson & Bergen, 1985; Waskom et al., 2018).

Approximated net motion energy was then accumulated for N steps of the motion stimulus for all coherence levels and directions. Accumulated values in each condition were normally distributed (KS tests; ps > 0.05). The mean accumulated value across all conditions was chosen as the signal detection theory threshold for left/right choices (Green & Swets, 1966). Each condition for each N-step accumulation was simulated for 10,000 trials.

Results showed that the accuracy for this model was 50%, 80.3%, 86.2%, and 89.1% for 0, 1, 2, and 3 steps of accumulation, respectively. Trained 60 Hz saccadic reinforcement learning agents achieved an average of 93.30% ± 0.37% accuracy on the task across the same conditions—significantly higher than the hand-constructed accumulation model for 1 step of motion energy (80.3%); $t(9) = 33.34, p = 4.84e\text{-}11$. These results suggest that the trained agents consider multiple samples of evidence over time in support of a decision, rather than considering single samples in isolation.

3.5.5 Microstimulation experiments

We employed a drift diffusion model from Hanks et al. (2006) to describe the impact on choices and response times from CNN and LSTM stimulation. The model accumulates momentary evidence to an upper (A) or lower bound (B), corresponding to left and right choices. The momentary evidence at every time step is modeled as draws from a unit-variance Gaussian with mean $\mu = kC$, where C is motion strength (e.g., 0.512) and k is a free parameter that scales motion sensitivity. This momentary evidence is accumulated over time to one of the two bounds, which determines both choice and response time. Consistent with Hanks et al. (2006), we modeled the bounds as symmetric (i.e., $B = -A$) when there was no stimulation. Choice probabilities and response times under this model can be solved analytically. The probability that accumulated evidence reaches bound A first is:

$$P_a(\mu, A, B) = \frac{e^{2\mu B} - 1}{e^{2\mu B} - e^{-2\mu A}}$$

And as μ approaches 0, this converges to:

$$\lim_{\mu \to 0} P_a(\mu, A, B) = \frac{B}{A+B}$$

The mean time to bound A is:

$$T_a(\mu, A, B) = \frac{A+B}{\mu}\coth((A+B)\mu) - \frac{B}{\mu}\coth(B\mu)$$

And as μ approaches 0, this converges to:

$$\lim_{\mu \to 0} T_a(\mu, A, B) = \frac{1}{3}(A^2 + 2AB)$$

To find the mean time to bound B instead of A, exchange the two in the above equations. While drift diffusion models often consider non-decision time (t_{nd}) as a free parameter, here we know the non-decision time for the artificial agents beforehand, and so can set it appropriately as a fixed parameter. To incorporate effects of microstimulation into the model, Hanks et al. (2006) add two additional free parameters. First, to simulate an increase or decrease in momentary evidence, consistent with MT microstimulation, an additional parameter is used to increase or decrease the motion strength (ΔC). Second, to simulate an increase or decrease in the accumulated evidence, consistent with LIP microstimulation, an additional parameter is used to shift the accumulated evidence toward one bound and away from the other (here implemented as an asymmetric change in bounds, ΔA and $-\Delta B$). Overall, the model had four free parameters, k, A, the change in motion strength, and the change in bounds.

3.5.6 Decoding

Decoding analyses were inspired by studies decoding decisions in non-human primates based on neural activity (Kiani et al., 2014; Peixoto et al., 2018; Peixoto et al., 2021). We trained a logistic regression classifier to predict the choices of a trained agent

based on LSTM layer activity. We defined the decision variable (DV) as the log odds ratio of observing a particular choice (T_1: rightward, T_2: leftward) given the activity of all units considered (r):

$$DV = log\frac{P(T_1|r)}{P(T_2|r)} = \beta_0 + \sum_{i=1}^{n} \beta_i r_i(t)$$

Where $r_i(t)$ are the z-scored activations for each unit, β_0 is an intercept term, and β_i are the classifier weights. Classifier training finds a set of linear weights (β_i) on the activity of each unit that maximizes the probability of correctly predicting the agent's choices. This can be viewed as finding the hyperplane that best separates neural activity according to observed choices. The distance of a point in high-dimensional activity space from this discriminant hyperplane can be viewed as the classifier's degree of belief about the agent's choice (i.e., the DV). Changes in the decoder's prediction (i.e., a change of sign in the DV) before a response suggests a neural change of mind (Kiani et al., 2014; Peixoto et al., 2018; Peixoto et al., 2021).

We trained a logistic regression classifier exclusively on the time step where agents made a response, and applied this trained classifier to all other time steps (both locked to dot motion stimulus onset, and response time). Trials with no responses were rejected, and data was randomly sampled so that an equal number of trials in each coherence condition remained. We performed 10-fold cross validation, stratified by response so that each fold had an equal number of left and right response trials for classification training. Chance decoding accuracy was defined as the response bias for the agent considered, which was slightly above 50% for all agents (e.g., 50.9% for the agent considered in Fig 3.5).

Neural changes of mind are typically identified as the instance when a DV changes sign before a response on a single trial. In practice, DVs can be biased and noisy, requiring the addition of restrictions (Peixoto et al., 2021). Here we defined the change of mind threshold as the mean DV value at dot motion stimulus onset, instead of a change in sign (i.e., 2.2 instead of 0). We also required DVs to have exceeded this

threshold by at least 10 DV units on the opposite side of space corresponding to the chosen side. In other words, if the agent ultimately chose the leftward option (corresponding to a negative DV), then a change of mind requires a DV that had at some point exceeded 12.2 (i.e., 10 + 2.2). Finally, we additionally required this threshold crossing to occur for at least 4 consecutive time steps to reduce changes of mind attributable to temporal noise. Decoded changes of mind results were qualitatively similar for the majority of these specific restriction values.

3.5.7 Human data

13 participants (7 women; Age: M = 19.29, SD = 1.44) took part in the experiment. All participants gave written consent prior to the experiment, which was approved by the University of Alberta's Research Ethics Board. All participants were right-handed, had normal or corrected-to-normal vision, and did not know the purpose of the study. Participants were compensated with course credit.

Participants' movements were recorded at 60 Hz using six Optitrak Flex 13 cameras (NaturalPoint, Inc., Corvallis, Oregon) mounted on two tripods, which tracked one passive, reflective motion-tracking marker placed near the tip of each participant's right index finger. Stimuli were presented at 60 Hz (synchronized with the motion capture rate) on a vertical, table-mounted monitor (VIEWPixx/EEG; Saint-Bruno, Quebec). The position of the finger marker was co-registered in space with the monitor so the tabletop and monitor could be used as touch-interactive surfaces. Stimuli presentation and data collection were controlled with MATLAB using Psychtoolbox (Kleiner et al., 2007).

Participants were seated in a semi-dark room with a computer monitor at a viewing distance of 57 cm. Participants were instructed to maintain their gaze on the central red fixation square during every trial. Stimuli were presented on a black background.

On each trial, participants saw a random dot motion stimulus and were asked to report the net motion of the left/right moving dots as quickly and accurately as

possible. White dots (3 x 3 pixel squares) were presented within a 5° circular aperture at the center of the screen. Dots moved at a speed of 5 deg/s, and average dot density was set at 16.8 dots/degree^2/s. As in Kiani et al. (2013), the stimulus consisted of three independent and interleaved sets of dots presented on successive video frames. Motion direction (left, right) and coherence (0%, 3.2%, 6.4%, 12.8%, 25.6%, 51.2%, and 100%) varied pseudorandomly from trial-to-trial. Dot motion within each trial continued from stimulus onset until one of the two targets had been selected.

Throughout the trial, two gray circles (3 degrees^2) to the left and right of center fixation (10 degrees) were visible at a height of ≈30 cm relative to the tabletop. Participants responded by reaching to touch one of the two circles with their right index finger. Reaching distance from start position to targets was approximately 30 cm forward, and 10 cm laterally. No feedback was given on dot motion trials so post-decision confidence could be rated.

At the end of each trial, participants returned their finger to the start position to rate their confidence in their decision. A white horizontal line (30 cm long) with the text, "How confident?" appeared on the screen, and participants were asked to reach and touch a point on the line corresponding to their confidence in the previous decision (left, 0% confident; right, 100% confident). Once participants had finished adjusting the confidence slider, they could return to the start position to begin the next trial.

Intertrial intervals were randomly drawn from a truncated exponential distribution (range: 700-1000 ms; mean: 820 ms; as in Resulaj et al., 2009). Participants completed 10 practice trials, followed by 10 blocks of 39 trials for a total of 400 trials per participant. Trials were excluded for reaction times less than 150 ms or greater than 6 s, movement times greater than 2 s, confidence rating times greater than 5 s, or for motion tracking recording errors (Gallivan and Chapman, 2014; trial rejection: M = 15.61%, SD = 7.64%).

3.6 Acknowledgements

This research was supported by the National Science and Engineering Research Council of Canada (NSERC) and the Killam Trusts. This research was enabled in part by computational support provided by the BC and the Prairie DRI Groups, and the Digital Research Alliance of Canada (`alliancecan.ca`). We would like to thank Garett Motley for assisting in human data collection, and Andrew Butcher for helpful discussions.

3.7 References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*(2), 284–299.

Atiya, N. A., Zgonnikov, A., O'Hora, D., Schoemann, M., Scherbaum, S., & Wong-Lin, K. (2020). Changes-of-mind in the absence of new post-decision evidence. *PLOS Computational Biology*, *16*(2), e1007149.

Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Dedieu, A., Fantacci, C., Godwin, J., Jones, C., Hemsley, R., Hennigan, T., Hessel, M., Hou, S., Kapturowski, S., ... Viola, F. (2020). *The DeepMind JAX Ecosystem*. http://github.com/deepmind

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T. P., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433.

Bengio, Y., Scellier, B., Bilaniuk, O., Sacramento, J., & Senn, W. (2016). Feedforward initialization for fast inference of deep generative networks is biologically plausible. *arXiv preprint arXiv:1606.01651*.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, *12*(12), 4745–4765.

Chou, P.-W., Maturana, D., & Scherer, S. (2017). Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution, In *International Conference on Machine Learning*.

Churchland, A. K., Kiani, R., Chaudhuri, R., Wang, X.-J., Pouget, A., & Shadlen, M. N. (2011). Variance as a signature of neural computations during decision making. *Neuron*, *69*(4), 818–831.

Cisek, P. (2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology*, *22*(6), 927–936.

Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*, *81*(7), 2265–2287.

Davidson, J. D., & El Hady, A. (2019). Foraging as an evidence accumulation process. *PLoS Computational Biology*, *15*(7), e1007060.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*(11), 3612–3628.

Gallivan, J. P., & Chapman, C. S. (2014). Three-dimensional reach trajectories as a probe of real-time decision-making between multiple competing targets. *Frontiers in Neuroscience*, *8*, 215.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience*, *9*(5), 682–689.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huang, S., Dossa, R. F. J., Ye, C., Braga, J., Chakraborty, D., Mehta, K., & Araújo, J. G. (2022). Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *The Journal of Machine Learning Research*, *23*(1), 12585–12602.

Kang, Y. H., Petzschner, F. H., Wolpert, D. M., & Shadlen, M. N. (2017). Piercing of consciousness as a threshold-crossing operation. *Current Biology*, *27*(15), 2285–2295.

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254.

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*(7611), 285–288.

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132.

Kepecs, A., & Mainen, Z. F. (2014). A computational framework for the study of confidence across species. *The Cognitive Neuroscience of Metacognition*, 115–145.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Kiani, R., Churchland, A. K., & Shadlen, M. N. (2013). Integration of direction cues is invariant to the temporal gap between them. *Journal of Neuroscience*, *33*(42), 16483–16489.

Kiani, R., Cueva, C. J., Reppas, J. B., & Newsome, W. T. (2014). Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. *Current Biology*, *24*(13), 1542–1547.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, *7*(1), 13276.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84.

Michalski, J., Green, A. M., & Cisek, P. (2020). Reaching decisions during ongoing movements. *Journal of Neurophysiology*, *123*(3), 1090–1102.

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

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, *22*(10), 1677–1686.

O'Connell, R. G., Shadlen, M. N., Wong-Lin, K., & Kelly, S. P. (2018). Bridging neural and computational viewpoints on perceptual decision-making. *Trends in Neurosciences*, *41*(11), 838–852.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 1–1.

Peixoto, D., Kiani, R., Chandrasekaran, C., Ryu, S. I., Shenoy, K. V., & Newsome, W. T. (2018). Population dynamics of choice representation in dorsal premotor and primary motor cortex. *bioRxiv*, 283960.

Peixoto, D., Verhein, J. R., Kiani, R., Kao, J. C., Nuyujukian, P., Chandrasekaran, C., Brown, J., Fong, S., Ryu, S. I., Shenoy, K. V., & Newsome, W. T. (2021). Decoding and perturbing decision states in real time. *Nature*, *591*(7851), 604–609.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266.

Rideaux, R., & Welchman, A. E. (2020). But still it moves: Static image statistics underlie how we see motion. *Journal of Neuroscience*, *40*(12), 2538–2552.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, *22*(21), 9475–9489.

Rorie, A. E., & Newsome, W. T. (2005). A general mechanism for decision-making in the human brain? *Trends in Cognitive Sciences*, *9*(2), 41–43.

Salzman, C. D., Murasugi, C. M., Britten, K. H., & Newsome, W. T. (1992). Microstimulation in visual area MT: Effects on direction discrimination performance. *Journal of Neuroscience*, *12*(6), 2331–2355.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791–806.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535.

Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life*, *11*(1-2), 13–29.

Song, H. F., Yang, G. R., & Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, *6*, e21492.

Song, J.-H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, *13*(8), 360–366.

Stine, G. M., Zylberberg, A., Ditterich, J., & Shadlen, M. N. (2020). Differentiating between integration and non-integration strategies in perceptual decision making. *eLife*, *9*, e55365.

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, *18*(7), 1025–1033.

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

Todorov, E., Erez, T., & Tassa, Y. (2012). MuJoCo: A physics engine for model-based control, In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Turing, A. M. (1950). Mind. *Mind*, *59*(236), 433–460.

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, *5*, e12192.

Waskom, M. L., Asfour, J., & Kiani, R. (2018). Perceptual insensitivity to higher-order statistical moments of coherent random dot motion. *Journal of Vision*, *18*(6), 9–9.

Wispinski, N. J., Butcher, A., Mathewson, K. W., Chapman, C. S., Botvinick, M. M., & Pilarski, P. M. (2022). Adaptive patch foraging in deep reinforcement learning agents. *Transactions on Machine Learning Research*.

Wispinski, N. J., Gallivan, J. P., & Chapman, C. S. (2020). Models, movements, and minds: Bridging the gap between decision making and action. *Annals of the New York Academy of Sciences*, *1464*(1), 30–51.

Chapter 4

Adaptive patch foraging in deep reinforcement learning agents

4.1 Abstract

Patch foraging is one of the most heavily studied behavioural optimization challenges in biology. However, despite its importance to biological intelligence, this behavioural optimization problem is understudied in artificial intelligence research. Patch foraging is especially amenable to study given that it has a known optimal solution, which may be difficult to discover given current techniques in deep reinforcement learning. Here, we investigate deep reinforcement learning agents in an ecological patch foraging task. For the first time, we show that machine learning agents can learn to patch forage adaptively in patterns similar to biological foragers, and approach optimal patch foraging behaviour when accounting for temporal discounting. Finally, we show emergent internal dynamics in these agents that resemble single-cell recordings from foraging non-human primates, which complements experimental and theoretical work on the neural mechanisms of biological foraging. This work suggests that agents interacting in complex environments with ecologically valid pressures arrive

A version of this work was previously published as: Wispinski, N. J., Butcher, A., Mathewson, K. W., Chapman, C. S., Botvinick, M. M., & Pilarski, P. M. (2023). Adaptive patch foraging in deep reinforcement learning agents. *Transactions on Machine Learning Research*.

at common solutions, suggesting the emergence of foundational computations behind adaptive, intelligent behaviour in both biological and artificial agents.

4.2 Introduction

Patch foraging is one of the most critical behavioural optimization problems that biological agents encounter in nature. Almost all animals forage, and must do so effectively to survive. In patch foraging theory, spatial patches are frequently modeled as exponentially decaying in resources, with areas outside of patches as having no resources (Charnov, 1976). Agents are faced with a decision about when to cease foraging in a depleting patch in order to begin travelling some distance to a richer patch. Research has shown that animals are adaptive patch foragers in this context, intelligently staying in patches for longer when the environment is resource-scarce, and staying a shorter time in patches when the environment is resource-rich (Cowie, 1977; Hayden et al., 2011; Kacelnik, 1984; Krebs et al., 1974).

Despite its importance to biological intelligence, patch foraging is understudied in artificial intelligence research. Computational models of patch foraging are often agent-based models with fixed decision rules (Pleasants, 1989; Tang & Bennett, 2010), although recent work has involved the use of tabular reinforcement learning models (Constantino & Daw, 2015; Goldshtein et al., 2020; Miller et al., 2017; Morimoto, 2019). Additionally, neural networks trained using methods such as Hebbian learning have displayed foraging behaviour in separate ecological tasks such as patch selection (Coleman et al., 2005; Montague et al., 1995; Niv et al., 2002). Many deep reinforcement learning agents are able to successfully search environments for rewarding collectibles like apples while avoiding obstacles and/or enemies (as in gridworlds or popular video games; e.g., Lin, 1991; Mnih et al., 2015; Platanios et al., 2020), and even to stay or switch in the face of decaying rewards (Shuvaev et al., 2020). However, these environments significantly differ from core principles of theoretical and experimental foraging research. Namely, many environments lack the repeating, tem-

porally evolving tradeoff between immediate, decaying resources and delayed, richer ones (Stephens & Krebs, 2019).

Patch foraging is especially feasible to study computationally given that it has a known optimal solution—the marginal value theorem (MVT; Charnov, 1976). In short, the MVT states that the optimal solution is to cease foraging within a patch and begin traveling toward a new patch when the reward rate of the current patch drops below the average reward rate of the environment. Foraging is so important to the survival of biological agents that theorists argue that the foraging behaviour of animals is not only adaptive, but approaches this MVT optimal behaviour in natural environments because of strong selective pressures (Charnov, 1976; Pearson et al., 2014; Stephens & Krebs, 2019). Many animals, including humans, have been shown to behave optimally in patch foraging tasks in the wild and in the laboratory. For example, human mushroom foragers (Pacheco-Cobos et al., 2019), rodents (Lottem et al., 2018; Vertechi et al., 2020), and birds, fish, and bees (Cowie, 1977; Krebs et al., 1974; Stephens & Krebs, 2019) all behave consistent with the MVT solution of optimal patch foraging (although many examples of suboptimal patch over-staying exist; see: Nonacs, 2001).

However, the optimal patch foraging solution may be difficult to discover using many frequently used deep reinforcement learning approaches. The MVT dictates a comparison between the long-run average reward rate of the environment with the instantaneous reward rate of the current patch. This value comparison requires multiscale temporal resolutions, both local and global, which can be difficult to represent using model-free reinforcement learning, but are seen in animal cognition and neural dynamics (Badman et al., 2020). Further, model-free reinforcement learning potentially requires significant temporal exploration involved with the trial-and-error learning of leaving patches at different depletion times. Finally, changes in the resource-richness of the environment or the agent's movement time between patches impact the optimal patch time, as prescribed by the MVT.

Given the potential difficulty for deep reinforcement learning agents to learn to patch forage, how do biological agents solve the same problem? Theoretical work suggests that a simple evidence accumulation mechanism, similar to one commonly proposed for perceptual and value-based decision making in mammals (Brunton et al., 2013; Gold & Shadlen, 2007; Hanks & Summerfield, 2017; Wispinski et al., 2020), can approximate MVT solutions in complex environments (Davidson & El Hady, 2019). Neural recordings from non-human primate cingulate cortex suggests that such a mechanism may underlie decisions in a computerized patch foraging task (Hayden et al., 2011). Past work in deep reinforcement learning has used task paradigms from animal research to probe the abilities of agents to learn abstract rule structures (Wang et al., 2016), detour problems (Banino et al., 2018), and perceptual decision making (Song et al., 2017). In these studies, internal dynamics show patterns similar to those recorded from animals completing the same behavioural task, suggesting the discovery of similar mechanisms solely through reward learning (Silver et al., 2021). Further, studying animal intelligence, the behavioural pressures that shaped biological intelligence (Cisek, 2019; Stephens & Krebs, 2019), and their neural implementation is likely to expedite progress in artificial intelligence (Hassabis et al., 2017).

Here, we first ask if deep reinforcement learning agents can learn to forage in a 3D patch foraging environment inspired by experiments from behavioural ecology. Second, we ask whether these agents forage intelligently—adapting their behaviour to the environment in which they find themselves. Next, we investigate if agent foraging behaviour in these environments approaches the known optimal solution. Finally, we investigate how these agents solve the patch foraging problem by interrogating internal dynamics in comparison to theory and neural dynamics recorded from non-human primates.

This paper offers a number of novel contributions. We demonstrate:

1. The first investigation of deep recurrent reinforcement learning agents in a com-

plex ecological patch foraging task;

2. Deep reinforcement learning agents that learn to adaptively trade off travel time and patch reward in patterns similar to biological foragers;

3. That these agents approach optimal patch foraging behaviour when accounting for temporal discounting;

4. That the internal dynamics of these agents show key patterns that are similar to neural recordings from foraging non-human primates and patterns predicted by foraging theory.

This paper is an empirical investigation into the emergence of complex patch foraging behaviour, and offers a model to the biology community with which to study patch foraging. Additionally, these results add to the neuroscientific literature on how biological agents approximate the optimal solution using a general decision mechanism, and how adaptive behaviour arises from simple changes in this mechanism.

4.3 Experiments

4.3.1 Environment

A continuous 3D environment was selected to approximate the rich sensorimotor experience involved in ecological foraging experiments. The environment consisted of a 32 x 32 m flat world with two patches (i.e., half spheres) equidistant from the center of the world (Fig. 4.1a; see Cultural General Intelligence Team et al., 2022). Patches always had a diameter of 4 m. Agents started each episode at the middle of the world, facing perpendicular to the direction of the patches. Each episode terminated after 3600 steps. Agents received a reward of zero on each step they were outside of both patches. When an agent was within a patch, it received reward according to the exponentially decaying function, $r(n) = N_0 e^{-\lambda n}$, where n is the number of non-consecutive steps the agent has been inside a patch without being inside the

alternative patch. In this way, as soon as an agent entered a patch, the alternative patch was refreshed to its initial reward state (i.e., $n = 0$). As such, agents are faced with a decision about how long to deplete the current patch before traveling toward a newly refreshed patch. For all experiments, the initial patch reward, N_0, was set to 1/30, and the patch reward decay rate, λ, was set to 0.01 (Fig. 4.1c). The surface colour of each patch changed proportional to the reward state of the patch in RGB space. Patches changed colour from white (i.e., [1, 1, 1]) to black (i.e., [0, 0, 0]) following the function, $r(n)/N_0$. In this way, agents had access to the instantaneous reward rate of the patch through patch colour, rather than having to estimate patch reward rate by estimating the decay function and keeping track of steps spent within a patch. We chose to make instantaneous patch reward information available in the environment as biological agents often have complete or partial sensory information regarding the current reward state of a patch (e.g., visual input of apple density on a tree).

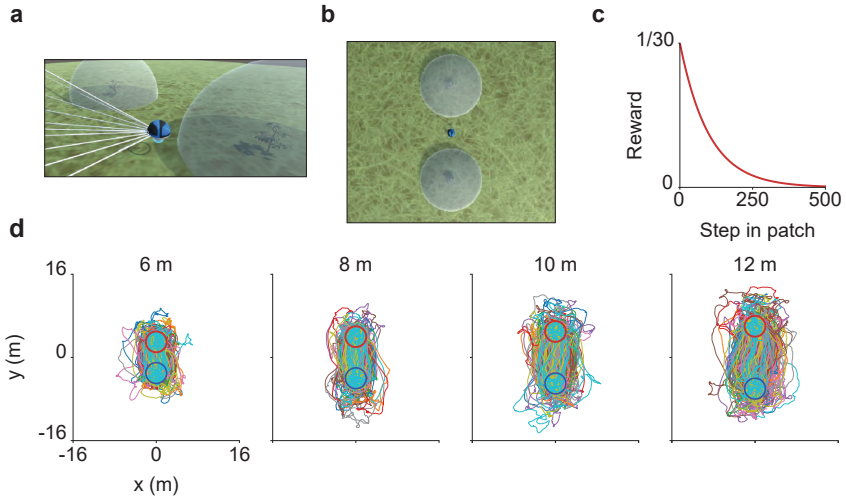

Figure 4.1: Task. (a) Mock-up of the 3D foraging environment and agent with LIDAR rays. (b) Overhead view. An agent starts each episode between two equidistant patches. (c) The agent receives exponentially decreasing reward on every step it is within a patch. When the agent enters one patch, the opposite patch is immediately refreshed to its starting reward state. (d) Overhead spatial trajectories of a representative trained agent in each evaluation environment.

4.3.2 Agents

The agents observe the environment through a LIDAR sensor-based observation sensor. LIDAR-based sensing is common for physical robots (Malavazi et al., 2018), and agents in other simulated environments (e.g., Baker et al., 2019; Cultural General Intelligence Team et al., 2022). The LIDAR sensor casts uniformly across 3 rows of 8 rays. Rays are evenly spaced, with azimuth ranging from -45° to +45°, and altitude ranging from -30° to +30°. Each ray returns an encoding of the first object it intersects with which includes object type, colour, and distance of the object from the agent. Object type is encoded as a one-hot vector for each valid object type (ground plane, patch sphere, no-intersection; e.g., [0, 1, 0]). Colour is encoded for patch spheres only, as [R,G,B] where each is a float [0, 1]. Distance is encoded as a float [0, 1] normalized to the maximum LIDAR distance (128 m). This corresponds to a LIDAR space with 24 channels. Agents analyzed in the dynamics results section below instead had 14 rows of 14 rays with azimuth ranging from 0° to 360°, and altitude ranging from -90° to +90°, consistent with Cultural General Intelligence Team et al. (2022). LIDAR differences did not substantially impact agent behaviour. LIDAR inputs were convolved (24 output channels, 2x2 kernel shape), before they were concatenated with the reward and action taken on the previous step. These values were then passed through a MLP (3 layers of 128, 256, and 256 units) and a LSTM layer (256 units). Finally, LSTM outputs were passed to an actor and a critic network head.

The action space is 5-dimensional and continuous (adapted from Cultural General Intelligence Team et al., 2022). Each action dimension takes a value in $[-1, 1]$. The dimensions correspond to 1) moving forward and backwards, 2) moving left and right, 3) rotating left and right, 4) rotating up and down, and 5) jumping or crouching. Agents can take any combination of actions simultaneously. Movement dynamics are subject to the inertia of the environment. Actions were taken by sampling from Gaussians

parameterized by the policy network head output for each action dimension.

We use a state-of-the-art continuous control deep reinforcement learning (RL) algorithm for training our agents: maximum a posteriori policy optimization (MPO; Abdolmaleki et al., 2018). MPO is an actor-critic, model-free RL algorithm which leverages samples to compare different actions in a particular state and then updates the policy to ensure that better actions have a high probability of being sampled. MPO alternates between policy improvement which updates the policy π using a fixed Q-function and policy evaluation which updates the estimate of the Q-function. Following previously described methods, agents were trained in a distributed manner with each agent interacting with 16 environments in parallel (Cultural General Intelligence Team et al., 2022). Agent architecture and training hyperparameters were the same as in Cultural General Intelligence Team et al. (2022), unless otherwise stated. Experience was saved in an experience buffer. Agent parameter updates were accomplished by sampling batches and using MPO to fit the mean and covariance of the Gaussian distribution over the joint action space (Abdolmaleki et al., 2018). Agent architecture, observation space, and the MPO algorithm were selected in part because of the success of these implementational choices in more complex environments (see Cultural General Intelligence Team et al., 2022).

Three agents were trained in each of four discount rate treatments ($N = 12$), selected on the basis of MVT simulations (Fig. 4.3d). Agents were each initialized with a different random seed, and trained for $12e^7$ steps using the Adam optimizer (Kingma & Ba, 2014) and a learning rate of $3e^{-4}$. On each training episode, patch distance was drawn from a random uniform distribution between 5 m and 12 m, and held constant for each episode. Trained agents were evaluated on 50 episodes of each evaluation patch distance (i.e., 6, 8, 10, and 12 m). By manipulating patch distance, we vary the amount of time it takes agents to travel between patches, which in turn varies the resource-richness of the environment—similar to many animal experiments on adaptive patch foraging behaviour (Cowie, 1977; Kacelnik, 1984). If an agent was

within a patch at the end of an evaluation episode (e.g., Fig. 4.2a), this final patch encounter was excluded from all analyses, as no distinct patch leave behaviour could be verified to determine the total steps in this patch.

Videos of a representative agent during training and evaluation are available in the supplementary material.

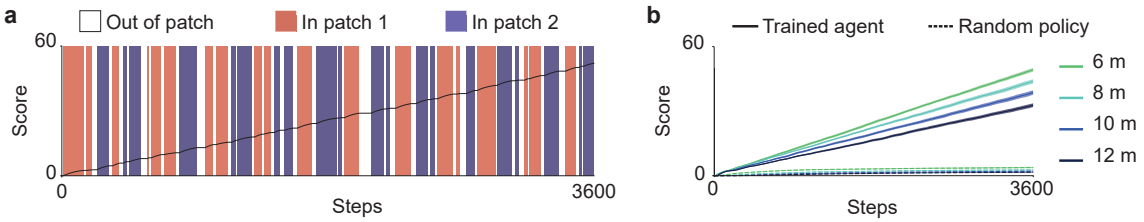

Figure 4.2: Performance. (a) Agent behaviour from a representative evaluation episode. Shaded regions define when the agent is outside of any patch (white), inside patch 1 (red), or inside patch 2 (blue). (b) Episode score for a representative trained agent (solid lines), and a representative agent with a random uniform action policy (dashed lines) in each evaluation environment. Shaded regions denote standard errors across evaluation episodes.

4.4 Results

For several results below, we fit data using a linear regression of the form, $y = bx + a$, where free parameter b is the slope of the fitted line, and a is a constant. b references the predicted relationship, negative or positive, of the rate of change in y (e.g., steps in patch; Fig. 4.3a) per unit change in x (e.g., patch distance in meters). In text, we report b, along with its standard error (e.g., $b = -1.50 \pm 0.50$). For full statistical reporting, see Appendix B.1.

4.4.1 Environment adaptation

Trained agents displayed behaviour consistent with successful patch foraging—agents learned to leave patches before they were fully depleted of reward (mean leaving step = 121.7), and traveled for several steps without reward in order to reach a refreshed patch (mean travel steps between patches = 57.7; e.g., Fig. 4.2a). We computed each

agent's mean final score in each patch distance evaluation environment (e.g., single representative agent in Fig. 4.2b) to analyze whether agent score varied systematically with patch distance. Analysis showed agents achieved a higher score on episodes where patches were closer together ($p < 0.05$, linear regression slope $b = -5.82 \pm 0.21$).

In patch foraging theory and experimental animal research, animals intelligently adapt patch leaving times to the environment in which they find themselves. Theory predicts agents monotonically increase their steps in patch when the distance between patches increases (Charnov, 1976; Stephens & Krebs, 2019). Similarly, in biological data, agents stay a longer time in patches when alternative patches are further away (Cowie, 1977; Hayden et al., 2011; Kacelnik, 1984; Krebs et al., 1974). This adaptive behaviour is present in the current agents—trained agents increased their patch leaving times when patch distance increased, leaving patches later when travel distance was higher ($p < 0.05$, $b = 9.60 \pm 0.87$; Fig. 4.3a). These results support our contribution that the current deep reinforcement learning agents learn to adaptively trade off travel time and patch reward in patterns similar to biological foragers.

4.4.2 Optimality

Above we show that trained agents are able to successfully forage, and intelligently adapt their foraging behaviour to the environment in accordance with patch foraging theory (Charnov, 1976; Stephens & Krebs, 2019), and animal behaviour (e.g., Cowie, 1977). However, do these agents adapt optimally according to the marginal value theorem (MVT), like many animals (Stephens & Krebs, 2019)? As stated above, the MVT provides a simple rule for when to leave patches optimally (Charnov, 1976; Shuvaev et al., 2020). That is, an agent should leave a patch when the reward rate of the patch drops below the average reward rate of the environment (Fig. 4.3c).

The current agents however use temporal discounting methods, which exponentially diminish rewards in the future. In other words, agents are tasked with solving a *discounted* patch foraging problem. Given that agents are effectively asked to compare

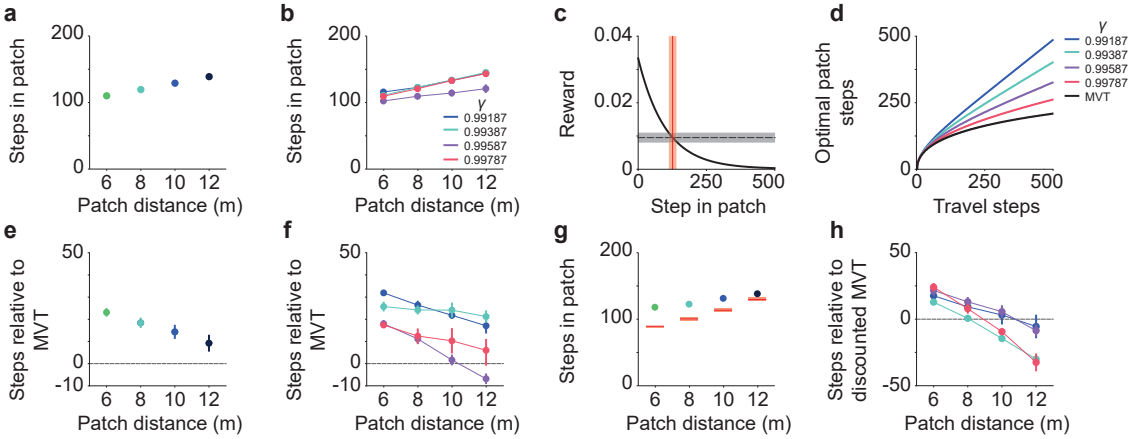

Figure 4.3: Patch leaving times. (a) Average of all agents. (b) Average of agents grouped by discount rate. (c) Graphical model of the MVT solution. Where the patch reward rate (solid black line) intersects the observed average reward rate of the environment as determined by the agent's behaviour (dashed horizontal black line), determines the MVT optimal average patch leaving time (solid vertical red line). (d) Patch leaving times prescribed by the MVT (black), and simulation results for the MVT considering discount rates. (e) Mean difference between the observed and optimal patch leaving time for all agents, and (f) agents grouped by discount rate. (g) Representative single trained agent patch leaving times (dots) against the MVT solution (red lines). (h) Mean difference between the observed and discounted MVT patch leaving time for agents grouped by discount rate. All vertical lines and/or shaded regions denote standard errors over agent means in each patch distance evaluation environment.

the values between the current patch reward on the next step relative to a refreshed patch reward after several travel steps, temporal discounting encourages longer patch leaving times relative to predictions from the MVT. Below, we evaluate agents relative to the MVT, which animals are typically compared against (Charnov & Parker, 1995; Cowie, 1977; Hayden et al., 2011; Krebs et al., 1974; Pacheco-Cobos et al., 2019; Stephens & Krebs, 2019). We also attempt to account for temporal discounting in the MVT to compare agents against the task that they are given.

For each agent and evaluation environment (e.g., 6 m), we can estimate the average reward rate of the environment by calculating the average reward per step for each evaluation episode. Over all evaluation episodes for each agent and environment, this provides an estimate of the optimal patch leaving step (see Fig. 4.3c and Fig. 4.3g).

Here, we take each agent's mean patch leaving time relative to the MVT across patch distance environments, and test if these 12 values (one for each agent) are significantly different from zero (Fig. 4.3e) using a one-sample t-test. Comparing the difference between average observed and MVT optimal patch leaving times, agents tend to overstay in patches relative to the MVT optimal solution ($p < 0.05$; mean = 15.8 steps above MVT optimal, standard error = 2.7; Fig. 4.3e). These results are consistent with the patch foraging behaviour of humans and other animals in computerized laboratory tasks (Cash-Padgett & Hayden, 2020; Constantino & Daw, 2015; Hutchinson et al., 2008; Kane et al., 2017; Nonacs, 2001), and are also consistent with temporal discounting predictions.

We now ask if agents trained with higher temporal discounting rates behave closer to MVT optimal (Fig. 4.3f). Here, we take the mean patch leaving time relative to the MVT across patch distance environments, and group these values by agent discount rate, leaving 3 agents per discount rate. Using linear regression, we test whether the difference between observed and MVT optimal patch leaving times decreases with discount rate. As expected, agents trained with higher temporal discounting rates tend to behave closer to MVT optimal ($p < 0.05$, $b = -2784.01 \pm 992.19$; Fig. 4.3f; for details see Appendix B.1).

Are agents then optimal after accounting for temporal discounting rates in the MVT solution? We accounted for the temporal discounting rate by simulating individual stay and leave decisions at many patch leaving steps (for details, see Appendix B.1). Agents could either stay for an additional step of reward before leaving a patch, or immediately leave the patch, where the subsequent 5000 steps were simulated as alternating between a fixed number of steps in a patch and a fixed number of steps traveling between patches. For example, on step 42 within a patch, and given a future fixed travel time of 50 steps and a future fixed patch time of 100 steps, is it more beneficial to leave immediately or stay in the patch for an additional step of reward? Over a grid of subsequent fixed patch and travel steps, the difference in

the discounted return (sum of discounted rewards) between each stay/leave decision provided an indifference curve, where the 5000-step discounted return was equal for staying relative to leaving. Where this stay/leave indifference step matched the fixed patch steps provided an approximation of an average patch time where the value of leaving is about to exceed the value of staying.

After accounting for each agent's temporal discounting rate in the MVT (Fig. 4.3d), we ask if agent patch leaving times approach the optimal solution (Fig. 4.3h). Similar to above, we take each agent's mean patch leaving time relative to the discounted MVT solution across patch distance environments, and test if these 12 values are significantly different from zero (Fig. 4.3h) using a one-sample t-test. Comparing the difference between average observed and discounted MVT optimal patch leaving times, agents were not significantly different from the optimal solution ($p = 0.74$; mean = 0.9 steps above discounted MVT optimal, standard error = 2.8; Fig. 4.3h)—although agents appear to slightly understay or overstay in patches depending on the environment (see Appendix B.1). Overall, these results support our contribution that the current agents approach optimal patch foraging behaviour when accounting for temporal discounting.

4.4.3 Dynamics and patch leaving time variability

Above we show that trained agents approach optimal patch leaving time behaviour when accounting for temporal discounting. Here, we investigate how these agents decide to leave a patch by interrogating internal dynamics in comparison to theory and neural dynamics recorded from non-human primates.

Theory and modeling work have shown that patch foraging decisions can be made using a simple evidence accumulation mechanism, which can approximate MVT solutions in complex environments (Davidson & El Hady, 2019). Evidence accumulation is a general decision making mechanism, which has been proposed to underlie many perceptual- and value-based decisions in animals (Brunton et al., 2013; Gold

& Shadlen, 2007; Wispinski et al., 2020). In brief, a decision variable representing the degree of evidence in support of a decision (i.e., leaving the current patch) accumulates over multiple time steps until it reaches a threshold (Fig. 4.4a). When a decision variable reaches threshold dictates when a decision is made (i.e., when the agent commits to leaving a patch). Evidence accumulation is especially useful when considering evidence that is delivered over multiple time points (e.g., rewards in a patch, Davidson and El Hady, 2019; Hayden et al., 2011; or noisy visual input, Gold and Shadlen, 2007; Ratcliff et al., 2016; Wispinski et al., 2020). Key mechanisms in this model typically include the slope with which evidence accumulates, and the distance that evidence needs to travel from baseline to threshold for a decision to be made (Gold & Shadlen, 2007; Ratcliff et al., 2016). For example, when evidence for a decision is stronger, the decision variable tends to accumulate with a higher slope, reaching threshold sooner (Fig. 4.4a). Similarly, when a decision requires less evidence, the distance between baseline and threshold can be decreased so decisions are made sooner. Both of these mechanisms are typically subject to variability in the internal state of the decision maker (Gold & Shadlen, 2007; Ratcliff et al., 2016).

Neural recordings in primate cingulate cortex similar to an evidence accumulation mechanism have been observed during a computerized patch foraging task, suggesting that such a mechanism may underlie biological solutions to the patch foraging problem (Hayden et al., 2011). Here we investigate LSTM layer activity in a single trained agent, and show several emergent similarities with neural data in foraging primates. The results described below are consistent with several other agents we investigated.

In this section we ask if variability in the internal state of the agent can explain earlier or later patch leaving times within the same patch distance environment, as in biological agents (Hayden et al., 2011). We took patch encounters where an agent had first entered a newly refreshed patch until it first left that patch. In these encounters, agents experience the same exponential decrease in reward within a patch but displayed variable patch leaving times. As in Hayden et al. (2011), we divided

these data into quartiles based on patch leaving times. Patch encounters were split into exclusive groups named: Earliest (25th percentile), Early (50th percentile), Late (75th percentile), and Latest (100th percentile). These groups were taken equally from each evaluation environment (6, 8, 10, 12 m) so results were independent of patch distance (Fig. 4.4b).

Using these groups, we can investigate if earlier patch leaving times in each evaluation environment had a higher slope of rising activity, as predicted by theory and biological data (Fig. 4.4a). We perform a linear regression of the slope of LSTM unit change against patch leaving time quartile (e.g., Fig. 4.4e) at every time step, using data from all individual patch encounters. We performed these sliding regressions on both patch entry-aligned data (left panel Fig. 4.4c), and patch exit-aligned data (right panel Fig. 4.4c). Primate data show that cingulate neurons have a higher slope of activity before the Earliest patch leaving times, and a lower slope of activity before the Latest patch leaving times (inset of Fig. 4.4a). Here we find the same pattern in a number of LSTM units for several steps after patch entry, but before patch exit (blue shaded bars; left panel of Fig. 4.4c). In the example unit shown, there are 19 consecutive steps where there is a significantly higher slope for encounters that had shorter patch leaving times (e.g., step 20: $p < 0.001$, $b = -0.0174 \pm 0.002$; Fig. 4.4e). In other words, the pattern of results in Fig. 4.4e match the idealized results consistent with primate neural data and theory in the inset of Fig. 4.4a. Additionally, only 1/10 steps had a significant relationship just before and after the patch encounter.

A principal components dimensionality reduction was performed on the LSTM data, and most results for the example unit can also be seen in a component accounting for roughly 14% of LSTM layer variability (Fig. 4.4f). The expected negative relationship between activity slope and patch leaving quartile (e.g., Fig. 4.4a and Fig. 4.4e) was statistically significant for 15 consecutive steps after patch entry (blue shaded bars; left panel of Fig. 4.4g).

Additionally, activity traces appear to begin at a similar level at patch entry,

and appear to converge just before patch exit (Fig. 4.4c and g), suggesting that changes in the distance activity must travel from baseline to threshold do not underlie patch leaving time variability (although not significantly; see Appendix B.1). Overall, this suggests that variability in the slope of rising activity can explain variability in the agent's patch leaving times, similar to primate neural recordings during patch foraging (Hayden et al., 2011). These results support the first half of our contribution that internal dynamics of these agents show key patterns that are similar to neural recordings from foraging non-human primates and patterns predicted by foraging theory.

4.4.4 Dynamics and environment adaptation

Above we show that within a patch distance environment, variability in the slope of LSTM layer activity predicts patch leaving time. We now turn to investigate the same relationships, but *between* patch distance environments. The behavioural analysis above showed that agents adapt patch leaving times to the resource-richness of the environment (Fig. 4.3a). Agents stayed in patches for more steps when travel time to a new patch was longer, and stayed in patches for fewer steps when travel time to a new patch was shorter.

Under an evidence accumulation mechanism, patch leaving times can be adapted by changing the slope of activity, or by changing the distance a decision variable needs to accumulate from baseline to threshold in each evaluation environment. Primate neural data suggests that when travel times between patches increase, the average slope of neural activity decreases, and the distance between baseline and threshold increases—both acting to prolong the time the animals stay in a patch (Hayden et al., 2011). We investigated which of these changes, if any, may drive the increase of patch leaving time when patches are further apart.

Decreasing the slope with which a decision variable accumulates toward a threshold prolongs patch leaving times. Here we took the average slope of activity during

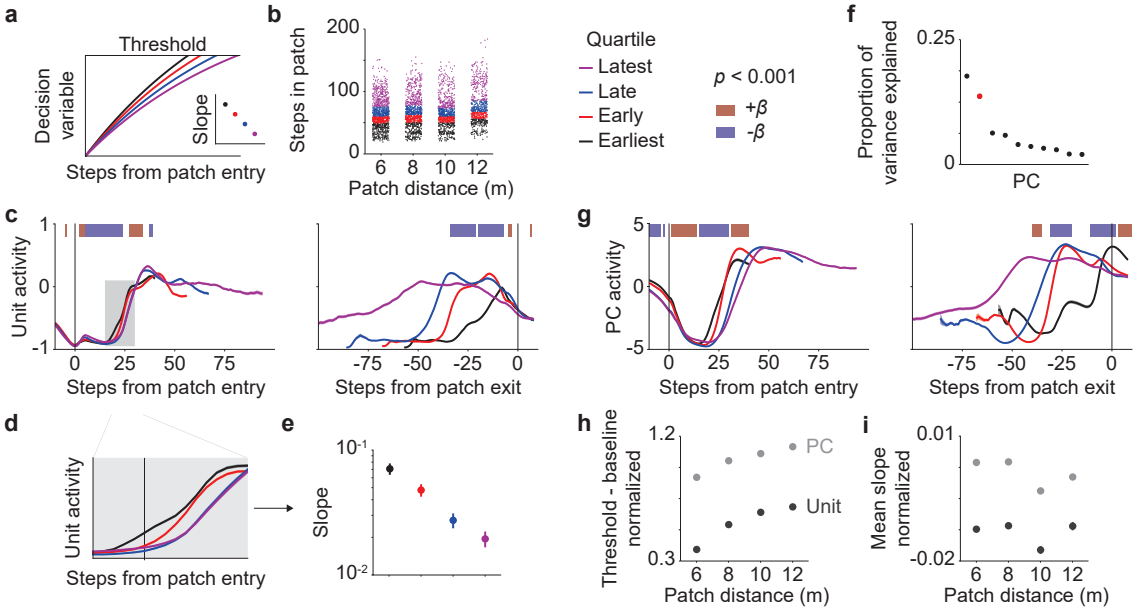

Figure 4.4: Dynamics from a single trained agent. (a) Example of an idealized evidence accumulation process. A decision variable accumulates to a threshold, which determines patch leaving time. Here, variability in the slope of the decision variable explains patch leaving time differences. (b) Patch encounters for a single agent split into patch leaving time quartiles in each evaluation environment. (c) Mean activity of an example LSTM unit aligned to patch entry (left) and patch exit (right), for each quartile of patch leaving times. Shaded regions along activity traces denote standard errors across patch encounters. Shaded bars at the top of the plot indicate steps where there is a significant slope-quartile relationship (negative in blue, positive in red). For patch entry-aligned data, traces in each quartile are plotted until median patch leaving time. (d) Zoomed-in region of c. (e) Mean slope by patch leaving time quartile for step 20 in d (log-scale). (f) Principal components decomposition of LSTM layer dynamics. Selected principal component (PC) in red. (g) Activation of the selected principal component. (h) Mean difference in example unit and example PC activity from patch exit to patch entry in each evaluation environment. (i) Mean slope of example unit and example PC in each evaluation environment. All vertical lines and/or shaded regions denote standard errors across patch encounters.

significant time steps (i.e., blue bars; Fig. 4.4c and g) for every patch encounter, and separated these data by patch distance environment. Using a linear regression, we find no significant relationship between the average slope of activity and patch distance environment in the example unit ($p = 0.18$, $b = -0.00026 \pm 0.00020$; Fig. 4.4i), nor for the selected PC ($p = 0.13$, $b = -0.0034 \pm 0.002$; Fig. 4.4i).

Increasing the distance a decision variable needs to travel from baseline to threshold

is another mechanism which prolongs patch leaving times (Davidson & El Hady, 2019). Here we took the difference between activity at patch exit and activity at patch entry for every patch encounter, and separated these data by patch distance. This difference in activity approximates the range a decision variable may need to span to complete a patch leaving decision (e.g., Fig. 4.4a). Using a linear regression, we find a positive relationship between activity range and patch distance environment for both the example unit ($p < 0.05$, $b = 0.053 \pm 0.003$; Fig. 4.4h), and for the selected PC ($p < 0.05$, $b = 0.20 \pm 0.020$; Fig. 4.4h).

Through an evidence accumulation framework, these results suggest that adaptive behaviour across environments is accomplished via changes in the distance a decision variable must travel between baseline and threshold. In other words, when it takes less time to travel to a new patch, distance between baseline and threshold decrease, shortening the agent's time in the current patch. When it takes more time to travel to a new patch, distance between baseline and threshold increase, prolonging the agent's time in the current patch. This mechanism is similar to work which suggests that decision thresholds during foraging may be estimated by an exponential moving average of past rewards (Constantino & Daw, 2015; Davidson & El Hady, 2019; Shuvaev et al., 2020) to approximate the average reward rate of the environment in the MVT (Charnov, 1976). In contrast, primate neural recordings suggest that adaptation between environments with short or long travel times between patches is also driven by changes in the slope of activity (Hayden et al., 2011)—a different mechanism within an evidence accumulation framework, albeit with very similar behavioural results in this context. These results support the second half of our contribution that internal dynamics of these agents show key patterns that are similar to neural recordings from foraging non-human primates and patterns predicted by foraging theory—although in this case results are different but are consistent with the same underlying mechanism.

4.5 Discussion

Here we tested deep reinforcement learning agents in a foundational decision problem facing biological agents—patch foraging. We find that these agents successfully learn to forage in a 3D patch foraging environment. Further, these agents intelligently adapt their foraging behaviour to the resource-richness of the environment in a pattern similar to many biological agents (Cowie, 1977; Hayden et al., 2011; Kacelnik, 1984; Krebs et al., 1974; Stephens & Krebs, 2019).

Many animals (Cowie, 1977; Krebs et al., 1974; Stephens & Krebs, 2019), including humans in the wild (Pacheco-Cobos et al., 2019), have been shown to be optimal patch foragers. The deep reinforcement learning agents investigated here also approach optimal foraging behaviour after accounting for discount rate. These results are similar to those from humans and non-human primates in computerized patch foraging tasks—participants tended to overstay relative to the MVT solution, but this discrepancy was significantly reduced after accounting for discount rate or risk sensitivity (Cash-Padgett & Hayden, 2020; Constantino & Daw, 2015). Together, these results raise questions as to how and why humans seemingly discount future rewards in lab foraging (Constantino & Daw, 2015; Hutchinson et al., 2008), but show undiscounted patch leaving times in nature (Pacheco-Cobos et al., 2019). Although outside the scope of this paper, these issues may be reconciled with further work on foraging using different reinforcement learning approaches, such as average reward reinforcement learning (Kolling & Akam, 2017; Sutton & Barto, 2018), especially in deep reinforcement learning models (Shuvaev et al., 2020; Zhang & Ross, 2021), to potentially better model ecological decision making in biological agents. While the current paper considers only the MPO reinforcement learning algorithm due to its success in a similar environment (Cultural General Intelligence Team et al., 2022), future work may find improvements in learning to solve the patch foraging problem with alternative methods.

In several agents that learned to adaptively patch forage, internal dynamics emerged to resemble several key properties of single-cell neural recordings from foraging non-human primates completing a computerized patch foraging task (Hayden et al., 2011). These results complement experimental and theoretical work on how adaptive patch foraging behaviour may be accomplished in biological agents by a general decision making mechanism—evidence accumulation (Davidson & El Hady, 2019; Gold & Shadlen, 2007; Wispinski et al., 2020). These results however do not necessarily mean that the agents investigated here, nor biological agents, use an evidence accumulation mechanism to solve the patch foraging problem (Blanchard & Hayden, 2014; Kane et al., 2021). Other strategies, or variations on accumulation models have also been proposed to underlie foraging behaviour (Cazettes et al., 2022; Davidson & El Hady, 2019; Kilpatrick et al., 2021). As in biological research, decision making frameworks provide one way to interpret and test predictions about behavioural and neural data. Emergent patterns in artificial agents completing similar tasks to biological agents provide concrete predictions for neural data and aid in interpretability (Banino et al., 2018; Hassabis et al., 2017; Song et al., 2017; Wang et al., 2016)—especially in ecological tasks like patch foraging where neural data is often more difficult to interpret relative to in-lab, stationary, computerized tasks (Pearson et al., 2014).

In this paper we demonstrate agents capable of patch foraging in a complex environment using a continuous action space. Future experiments may build on the ecological complexity of the environment and interesting agent behaviour may arise in environments where the assumptions and predictions of the MVT start to break down (Davidson & El Hady, 2019; Stephens & Krebs, 2019). For example, biological realism may be increased by adding multiple patches, modeling a biologically realistic patch refresh rate, or adding competitive or collaborative agents to the environment (Bidari et al., 2022). Other biological research argues that there are optimal movement strategies for finding unknown patch locations, and that many animals abide by these movement policies (Calhoun et al., 2014; Cisek, 2019; Sims et al., 2008;

Tello-Ramos et al., 2015; Woodgate et al., 2017). Given that the current agents generate continuous and complex movement trajectories (Fig. 4.1d), future work may investigate situations in which these optimal movement policies may emerge in artificial agents. The current movement trajectories may also be improved in future experiments by using alternative RL methods or modifying the sensory inputs available to the agents. Finally, foraging frameworks have been successfully extended to explain other aspects of intelligence, such as visual search (Wolfe et al., 2018), or human memory (Hills et al., 2012), which provide another avenue where reinforcement learning models of foraging may aid in intelligence research.

In conclusion, we have trained deep reinforcement learning agents on a complex patch foraging task and for the first time observed the emergence of adaptive, optimal behaviour, and neural dynamics that resembled those of biological agents. This paper contributes a model with which biological and artificial intelligence researchers may further understand patch foraging (Frankenhuis et al., 2019)—a fundamental decision problem that strongly guided the evolution of biological intelligence (Cisek, 2019; Stephens & Krebs, 2019). Such paradigms have been, and may continue to be, critical in the continued development of artificial intelligence (Hassabis et al., 2017; Lindsay, 2021).

4.6 Acknowledgements

I am deeply indebted to colleagues Leslie Acker, Andrew Bolt, Michael Bowling, Dylan Brenneis, Adrian Collister, Elnaz Davoodi, Richard Everett, Arne Olav Hallingstad, Nik Hemmings, Edward Hughes, Michael Johanson, Marlos Machado, Kory Mathewson, Drew Purves, Kimberly Stachenfeld, Richard Sutton, Jane Wang, Alexander Zacherl, and the entire DeepMind Cultural General Intelligence Team for their support, suggestions, and insight regarding this work. I also thank Ben Hayden and Eric Charnov for their insightful comments on this work. This work was funded solely by DeepMind.

4.7 References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., & Riedmiller, M. (2018). Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*.

Badman, R. P., Hills, T. T., & Akaishi, R. (2020). Multiscale computation and dynamic attention in biological and artificial intelligence. *Brain Sciences*, *10*(6), 396.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T. P., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433.

Bidari, S., El Hady, A., Davidson, J. D., & Kilpatrick, Z. P. (2022). Stochastic dynamics of social patch foraging decisions. *Physical Review Research*, *4*(3), 033128.

Blanchard, T. C., & Hayden, B. Y. (2014). Neurons in dorsal anterior cingulate cortex signal postdecisional variables in a foraging task. *Journal of Neuroscience*, *34*(2), 646–655.

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, *340*(6128), 95–98.

Calhoun, A. J., Chalasani, S. H., & Sharpee, T. O. (2014). Maximally informative foraging by caenorhabditis elegans. *eLife*, *3*, e04220.

Cash-Padgett, T., & Hayden, B. (2020). Behavioural variability contributes to overstaying in patchy foraging. *Biology Letters*, *16*(3), 20190915.

Cazettes, F., Mazzucato, L., Murakami, M., Morais, J. P., Renart, A., & Mainen, Z. F. (2022). A repertoire of foraging decision variables in the mouse brain. *bioRxiv*.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Charnov, E. L., & Parker, G. A. (1995). Dimensionless invariants from foraging theory's marginal value theorem. *Proceedings of the National Academy of Sciences*, *92*(5), 1446–1450.

Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*, *81*(7), 2265–2287.

Coleman, S. L., Brown, V. R., Levine, D. S., & Mellgren, R. L. (2005). A neural network model of foraging decisions made under predation risk. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(4), 434–451.

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(4), 837–853.

Cowie, R. J. (1977). Optimal foraging in great tits (Parus major). *Nature*, *268*(5616), 137–139.

Cultural General Intelligence Team, Bhoopchand, A., Brownfield, B., Collister, A., Lago, A. D., Edwards, A., Everett, R., Frechette, A., Oliveira, Y. G., Hughes, E., Mathewson, K. W., Mendolicchio, P., Pawar, J., Pislar, M., Platonov, A., Senter, E., Singh, S., Zacherl, A., & Zhang, L. M. (2022). Learning robust real-time cultural transmission without human data. *arXiv preprint arXiv.2203.00715*.

Davidson, J. D., & El Hady, A. (2019). Foraging as an evidence accumulation process. *PLoS Computational Biology*, *15*(7), e1007060.

Frankenhuis, W. E., Panchanathan, K., & Barto, A. G. (2019). Enriching behavioral ecology with reinforcement learning methods. *Behavioural Processes*, *161*, 94–100.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Goldshtein, A., Handel, M., Eitan, O., Bonstein, A., Shaler, T., Collet, S., Greif, S., Medellín, R. A., Emek, Y., Korman, A., Et al. (2020). Reinforcement learning enables resource partitioning in foraging bats. *Current Biology*, *30*(20), 4096–4102.

Hanks, T. D., & Summerfield, C. (2017). Perceptual decision making in rodents, monkeys, and humans. *Neuron*, *93*(1), 15–31.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*(7), 933–939.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431.

Hutchinson, J. M., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: Can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour*, *75*(4), 1331–1349.

Kacelnik, A. (1984). Central place foraging in starlings (Sturnus vulgaris). I. Patch residence time. *The Journal of Animal Ecology*, 283–299.

Kane, G. A., James, M. H., Shenhav, A., Daw, N. D., Cohen, J. D., & Aston-Jones, G. (2021). Rat anterior cingulate cortex continuously signals decision variables in a patch foraging task. *bioRxiv*.

Kane, G. A., Vazey, E. M., Wilson, R. C., Shenhav, A., Daw, N. D., Aston-Jones, G., & Cohen, J. D. (2017). Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *17*(6), 1073–1083.

Kilpatrick, Z. P., Davidson, J. D., & El Hady, A. (2021). Uncertainty drives deviations in normative foraging decision strategies. *Journal of the Royal Society Interface*, *18*(180), 20210337.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolling, N., & Akam, T. (2017). (Reinforcement?) Learning to forage optimally. *Current Opinion in Neurobiology*, *46*, 162–169.

Krebs, J. R., Ryan, J. C., & Charnov, E. L. (1974). Hunting by expectation or optimal foraging? A study of patch use by chickadees. *Animal Behaviour*, *22*, 953–IN3.

Lin, L. J. (1991). Self-improvement based on reinforcement learning, planning and teaching, In *International Conference on Machine Learning*.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031.

Lottem, E., Banerjee, D., Vertechi, P., Sarra, D., oude Lohuis, M. N., & Mainen, Z. F. (2018). Activation of serotonin neurons promotes active persistence in a probabilistic foraging task. *Nature Communications*, *9*(1), 1–12.

Malavazi, F. B., Guyonneau, R., Fasquel, J.-B., Lagrange, S., & Mercier, F. (2018). Lidar-only based navigation algorithm for an autonomous agricultural robot. *Computers and Electronics in Agriculture*, *154*, 71–79.

Miller, M. L., Ringelman, K. M., Eadie, J. M., & Schank, J. C. (2017). Time to fly: A comparison of marginal value theorem approximations in an agent-based model of foraging waterfowl. *Ecological Modelling*, *351*, 77–86.

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

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, *377*(6551), 725–728.

Morimoto, J. (2019). Foraging decisions as multi-armed bandit problems: Applying reinforcement learning algorithms to foraging data. *Journal of Theoretical Biology*, *467*, 48–56.

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior*.

Nonacs, P. (2001). State dependent behavior and the marginal value theorem. *Behavioral Ecology*, *12*(1), 71–83.

Pacheco-Cobos, L., Winterhalder, B., Cuatianquiz-Lima, C., Rosetti, M. F., Hudson, R., & Ross, C. T. (2019). Nahua mushroom gatherers use area-restricted search strategies that conform to marginal value theorem predictions. *Proceedings of the National Academy of Sciences*, *116*(21), 10339–10347.

Pearson, J. M., Watson, K. K., & Platt, M. L. (2014). Decision making: The neuroethological turn. *Neuron*, *82*(5), 950–965.

Platanios, E. A., Saparov, A., & Mitchell, T. (2020). Jelly bean world: A testbed for never-ending learning. *arXiv preprint arXiv:2002.06306*.

Pleasants, J. M. (1989). Optimal foraging by nectarivores: A test of the marginal-value theorem. *The American Naturalist*, *134*(1), 51–71.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Shuvaev, S., Starosta, S., Kvitsiani, D., Kepecs, A., & Koulakov, A. (2020). R-learning in actor-critic model offers a biologically relevant mechanism for sequential decision-making. *Advances in Neural Information Processing Systems*.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535.

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J., Pitchford, J. W., James, A., Ahmed, M. Z., Brierley, A. S., Hindell, M. A., Et al. (2008). Scaling laws of marine predator search behaviour. *Nature*, *451*(7182), 1098–1102.

Song, H. F., Yang, G. R., & Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, *6*, e21492.

Stephens, D. W., & Krebs, J. R. (2019). *Foraging theory*. Princeton University Press.

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

Tang, W., & Bennett, D. A. (2010). Agent-based modeling of animal movement: A review. *Geography Compass*, *4*(7), 682–700.

Tello-Ramos, M. C., Hurly, T. A., & Healy, S. D. (2015). Traplining in hummingbirds: Flying short-distance sequences among several locations. *Behavioral Ecology*, *26*(3), 812–819.

Vertechi, P., Lottem, E., Sarra, D., Godinho, B., Treves, I., Quendera, T., oude Lohuis, M. N., & Mainen, Z. F. (2020). Inference-based decisions in a hidden state foraging task: Differential contributions of prefrontal cortical areas. *Neuron*, *106*(1), 166–176.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Wispinski, N. J., Gallivan, J. P., & Chapman, C. S. (2020). Models, movements, and minds: Bridging the gap between decision making and action. *Annals of the New York Academy of Sciences*, *1464*(1), 30–51.

Wolfe, J. M., Cain, M. S., & Alaoui-Soce, A. (2018). Hybrid value foraging: How the value of targets shapes human foraging behavior. *Attention, Perception, & Psychophysics*, *80*(3), 609–621.

Woodgate, J. L., Makinson, J. C., Lim, K. S., Reynolds, A. M., & Chittka, L. (2017). Continuous radar tracking illustrates the development of multi-destination routes of bumblebees. *Scientific Reports*, *7*(1), 1–15.

Zhang, Y., & Ross, K. W. (2021). On-policy deep reinforcement learning for the average-reward criterion. *arXiv preprint arXiv:2106.07329*.

Chapter 5

General Discussion

Adaptive behaviour is critical for effectively acting in a dynamic world. As the environment and our place in it changes over time, so too must we adapt how we engage with our environment and other agents within it. The purpose of this thesis was twofold—to understand the computational mechanisms of rapid sensorimotor adaptation in biological systems, and to develop ways in which artificial systems can learn these mechanisms themselves. Below, I will discuss each of the three studies presented above individually before concluding with a general discussion.

5.1 On Chapter 2: Humans acting in dynamic environments

In Chapter 2, I pushed forward our understanding of rapid sensorimotor adaptation in biological systems. In this study, we asked human participants to make a rapid reaching movement to one of two targets. The correct target was determined by the location of a central square or arrow, which dictated the probability that each of the two targets would be the correct target of action. Analyzing human accuracy, reaction times, and movement trajectories, we found that participants moved and selected targets in sync with the stimulus probability in the moment, rather than delayed by some time. This optimal behaviour was observed despite known perceptual and motor delays in the human central and peripheral nervous systems.

The human participants here were able to rapidly, accurately, and continuously make visuomotor predictions about the probability of both targets and their own temporal limitations to guide action. This behaviour is even more impressive when we consider that human participants were only given a limited number of practice trials before the experiment began, only had 1 second of stimulus motion before the movement signal, 350 ms to begin a movement, and 425 ms to touch one of the two targets after movement initiation. In other words, participants were able to leverage their vast sensorimotor experience to perform a novel rapid reaching task optimally. If participants learned *during* the task we would expect to see behavioural patterns different from optimal performance, with a trend throughout the experiment. This was not the case—participants on average were not different from optimal performance when considering data throughout the whole experiment. As such, the adaptive behaviour observed in this experiment is likely the result of repurposing already-learned neural mechanisms for sensorimotor adaptation, rather than learning and refining new policies.

As stated in the paper, theorists argue that one of the major roles of the brain is to produce movement (Cisek & Kalaska, 2010; Gallivan et al., 2018; Hommel et al., 2019; Wolpert et al., 2001). Because of the biological delays involved in motor control, this requires a significant amount of temporal prediction. These predictions likely take the form of internal forward models which can be used to predict, given a particular motor command, the sensory consequences of action (Wolpert et al., 2001). Forward models can be used to correct behaviour online by computing "prediction errors", where low-level incoming sensory input is compared to forward model predictions of what sensory input *should* be given recent motor control signals. This prediction error framework is exceptionally useful, and is present in many other frameworks such as the "reward prediction error" in reinforcement learning (Sutton & Barto, 2018), prediction error in general value functions (GVFs; Sutton et al., 2011), and the concepts of "nexting" in psychology and computing science (White, 2015).

This particular study could have gone further by increasing the temporal restrictions on the participant (e.g., as in Chapman, Gallivan, Wong, et al., 2015). For example, we could have systematically varied the length of the stimulus period to investigate how much stimulus motion is needed for a good sensorimotor prediction to be computed. The current results show that participants are able to compute and use accurate predictions within just over 1 second, but sheds no light on the temporal limits of the cognitive processes under consideration. It may be useful to also systematically shorten reaction and movement time restrictions to investigate how long good predictions take to compute, and how performance is impacted with less processing time. However, it is difficult to shorten reaction time and movement time limits, as this rapid reaching task already starts to approach comfortable limits for humans without significant practice.

This study adds to a rich literature on the ability of humans to adapt to dynamic environments. For example, humans optimally adapt learning rates to the statistics of a dynamic environment (Behrens et al., 2007), and human motor behaviour aligns well with optimal control models (Scott, 2004; Todorov, 2004). In this study, humans were able to continuously and optimally adapt their motor behaviour to fluctuating target probabilities as they changed over time. The dynamic target probabilities in this task can be compared to n-armed bandit tasks with dynamic probabilities (e.g., Behrens et al., 2007; Daw et al., 2006; Pearson et al., 2009; J. X. Wang et al., 2018). In these tasks, the optimal action (i.e., arm with the highest win probability) varies randomly over time. In this scenario, agents need to adapt their actions dynamically, but instead have limited ability to predict these changes. In Chapter 2, probabilities vary predictably over time according to a sinusoidal function. In addition, adaptation in bandit tasks takes place over the course of a task, whereas the Chapter 2 task requires dynamically updating continuous actions in real time within a single decision. Interpolating between these two dynamic tasks gets us tasks with relatively unpredictable changes that should impact continuous real-time actions. Such tasks more

closely resemble the complex pressures of the real world, and have some examples in recent biological research such as dynamic pursuit and avoidance tasks (Yoo et al., 2021).

Tying this work together with the goals of developing adaptive artificial agents, how might insights from these studies advance adaptive artificial behaviour? For artificial agents, we might expect predictive abilities to emerge from training agents on tasks that have statistical relationships between past and future states. In some tasks such as chess, all information is present in the state space, and the task is a Markov process (i.e., the probability of each event depends only on the state attained in the previous event). In contrast, many tasks are partially observable and have long temporal delays between cause and effect, which require acting with respect to future rather than present states. For example, racecar driving (Wurman et al., 2022), patch foraging (Wispinski et al., 2022), some Atari video games (Mnih et al., 2015), and robotic control tasks (Beattie et al., 2016) all have situations in which present actions impact (partially observable) states in the future more than a handful of time steps away. Especially in robotic control tasks, agents may be able to discover their own forward models and prediction errors if given the right ingredients (Feulner et al., 2022). After all, meta-RL models are able to learn their own reward prediction errors to rapidly adapt their own behaviour (J. X. Wang et al., 2018). Artificial agents that learn their own forward models for motor control might better integrate the nuances of their effectors and environments into their movements, relative to forward models that are explicitly programmed into these agents. As in meta-RL, these forward models could also be fine-tuned through the slower process of network parameter updates to gradually adapt to slower changes like drift in motors or sensors. Overall, taking inspiration from how biological agents learned to make predictions and use them to guide movement may provide one way to advance the challenging task of robust motor control in artificial agents.

5.2 On Chapter 3: Artificial agents that make decisions like humans

Primates are able to make highly flexible decisions and can adapt their actions to dynamic sensory information on the fly. Much (but far from everything) is known about the neural mechanisms underlying these abilities, but less is known about how they came about in the first place. Some preliminary theory suggests that these decision mechanisms emerged because of the environmental pressures to act in a noisy, dynamic world (Cisek, 2012; Wispinski et al., 2020).

In this study, we trained deep recurrent reinforcement learning agents to complete a noisy perceptual decision task used in primate research. We found that agent behaviour and neural dynamics emerged to match those of primates completing a similar task. In addition, targeted causal perturbations inspired by primate microstimulation experiments impacted the behaviour of artificial agents in a similar pattern to their biological counterparts. We show that agents learned to change their minds on the fly—a hallmark of cognitive flexibility (Atiya et al., 2020; Resulaj et al., 2009). We also found that agent value estimates closely aligned with human post-decision confidence judgments. Finally, we showed that primate-like decision making does *not* emerge in agents trained without recurrence, nor in the absence of environmental noise. Interestingly, these alternate networks sometimes behaved like primates, but employed dissimilar decision making mechanisms. This distinction is important because theorized primate decision mechanisms are known for their adaptability and generalizability to many tasks, which these alternative decision mechanisms cannot match.

Overall, we argue that these results support the theory that primate decision mechanisms emerged via evolutionary and developmental pressures to act in noisy, dynamic environments. These results align well with other work showing that training artificial agents on biological tasks encourages biological-like solutions, for example motion

processing (Rideaux & Welchman, 2020), detour problems (Banino et al., 2018), and patch foraging (Wispinski et al., 2022, Chapter 4). Later, I will elaborate more on how biologically inspired evolutionary or developmental pressures can be a powerful tool for developing adaptive artificial agents (see How can biological systems aid artificial intelligence research?).

There are many limitations to consider with respect to this work. First, these agents are simulated at a high level of abstraction—we do not consider spiking neural networks (Lillicrap et al., 2016), biologically plausible methods of network weight updating (Bengio et al., 2016), cell types, or long-range recurrence (among many, many other things). Most notably, we do not consider internal neural noise or biologically plausible motor noise, which increases the interpretability and reliability of the agents within the current task, but may be important for robust representation learning (McDonnell & Ward, 2011). Related to Chapter 2, these agents also have no perceptual or motor delays, which is apparent when looking at their low response times. The action space of these artificial agents is also heavily abstracted relative to our knowledge about biological systems. In the saccadic version of the task, agents respond with three discrete actions, rather than controlling oculomotor muscles (e.g., Enderle & Sierra, 2013; Zhou et al., 2009). In the reaching version of the task, agents respond by controlling the joint forces of a two-degree-of-freedom planar arm, which is a significantly easier problem relative to controlling biomechanical upper arm models (Schaffelhofer et al., 2014; Seth et al., 2018), and the real-world limb control problem that primates need to solve.

Despite these limitations, I argue the agents developed in this work can have several potential uses. First, these trained agents provide a model with which to further investigate the computations underlying dynamic decision making and motor control in biological agents. This is especially true in motor control tasks where biological data is difficult and time-consuming to collect, and neural activity is often more difficult to interpret relative to stationary, in-lab tasks (Musall et al., 2019). For

example, most non-human primate tasks involve head-fixed saccadic responses instead of reaching because it significantly reduces motor confounds and variability. However, we as researchers should want to investigate ecologically valid tasks that involve limb movements, as this is the primary way that we as humans engage with the world (Stone, 2023). Computational models provide one way to mitigate these concerns, by studying these tasks and systems, but with the ability to control or remove noise. Overall, these techniques provide one way to better understand biological systems in messy, naturalistic environments.

Second, the agents here develop value estimates that closely align with human confidence judgments in the same task. These results fit well with some ideas about the purpose of confidence in biological agents being partially used to adapt behaviour in a graded way after being compared to decision outcomes (Boldt et al., 2019; Kepecs et al., 2008; Peters, 2022). Future work in this area may shed light on the function and neural basis of confidence in humans (Peters, 2022), but also other actor-critic reinforcement learning algorithms (Sutton & Barto, 2018). For instance, actor-critic models of reinforcement learning have experimental support from neural data, specifically in the basal ganglia (Bogacz & Larsen, 2011; O'Doherty et al., 2004). This work leads us to investigate specific brain areas with concrete theories about what kind of confidence information should be fed back to change action policies in the brain. Further, when neuroscientists are able to make high density recordings of these areas, dimensionality reduction techniques may lead to clues regarding useful variations on actor-critic reinforcement learning algorithms.

Third, this work provides agents that we can now use and extend to make primate-like decisions instead of requiring the use of humans. For example, these agents can potentially be adapted to assist with or perform tasks like autonomous driving, air traffic control, and maritime surveillance, which partially rely on flexibly integrating noisy sensory information over time (Boag et al., 2022). Such agents provide the additional benefit that they have no internal neural noise or nervous system delays,

making them both faster and more consistent than biological decision makers. In addition, we demonstrate that it is possible to predict and explain the decisions of these artificial agents to a similar degree that we can predict and explain primate decisions. As such, these agents potentially provide a path toward addressing explainability and safety concerns regarding artificial agents that are deployed to make autonomous decisions in real or virtual environments.

Finally, I argue this work has implications for the emergence of human-readable movements in artificial agents. Here we show that the reaching movements of trained agents vary with decision difficulty in similar patterns to those observed in humans—movements are slower, more curved, and indicate more changes of mind when decisions are difficult, and movements are faster, straighter, and indicate fewer changes of mind when decisions are easy (Gallivan et al., 2018; J.-H. Song & Nakayama, 2009; Wispinski et al., 2020). Research has shown that humans can observe the movements of other humans in order to "read" their internal cognitive states (Barrett et al., 2019; De Gelder, 2006; Pesquita et al., 2016; Zhu & Thagard, 2002). Here I propose that humans may be able to similarly "read" the internal states of these artificial agents simply by observing their movements. Specifically, humans may be able to rapidly and intuitively determine how confident the artificial agent is in its action through observation. If this idea holds true, then it suggests that a large part of body language may not have developed socially, but rather emerges through the individual need to continuously interact within noisy environments. Further, this information channel for human-computer interaction can potentially bypass other forms of information like text or speech, which can be difficult or relatively slow. This would open several doors for applied AI research. First, artificial agents that can display their internal states to humans through movement may be able to efficiently alert humans of instances when they are not confident in their actions to signal for help or for safety concerns. For example, a delivery robot that is unsure of where to go next may look confused to human bystanders based on its movements, potentially allowing humans

to help with some directions. Second, this work could potentially provide artificial agents that move in ways that are compelling and intuitively understandable to humans. In this realm, I imagine video game characters that automatically move with respect to their confidence in human-readable form.

Overall, this work advances both goals of this thesis at the same time. I show that humans are able to rapidly adapt their movements online to improve their decisions. And I show that we can leverage our knowledge of biological environmental pressures so that agents learn adaptive primate-like decision mechanisms.

5.3 On Chapter 4: Foraging

In Chapter 4, I further investigate this idea that artificial agents might learn adaptive mechanisms on their own via interacting within biologically inspired environments. Specifically, I looked at patch foraging. Patch foraging is one of the most critical behavioural problems that biological agents encounter in nature. Almost all animals forage, and must do so effectively to survive. Patch foraging is a sequential explore/exploit decision problem, where agents are faced with dynamic decisions about whether to continue to exploit depleting resources within the current patch or to explore the environment for alternative patches with more resources. The marginal value theorem (MVT) dictates an elegant optimal solution to this problem—to leave the current patch when its instantaneous reward rate drops below the average reward rate of the environment (Charnov, 1976). In addition, the MVT prescribes the optimal adaptive behaviour to environmental change. Specifically, when the environment becomes relatively scarce (i.e., the average reward rate of the environment decreases), agents should increase their patch residence times. Conversely, when the environment becomes relatively plentiful (i.e., the average reward rate of the environment increases), agents should decrease their patch residence times. A wide variety of experimental and observational studies across several species show that animals adapt their patch foraging behaviour optimally to environmental changes (Cowie, 1977; Krebs

et al., 1974; Lottem et al., 2018; Pacheco-Cobos et al., 2019; Stephens & Krebs, 2019; Vertechi et al., 2020), although many exceptions exist (Nonacs, 2001).

In this study, we trained deep recurrent reinforcement learning agents on a simplistic version of the patch foraging task that adhered to assumptions made by the MVT. After agents were trained and parameters were frozen, they were evaluated in four different environments. Agents adapted their patch leaving times to the environment even after their parameters were frozen—staying in patches longer in scarce environments and staying in patches for a shorter amount of time in plentiful environments. Agents approached the optimal patch leaving time dictated by the MVT, but only after accounting for temporal discounting in each agent. Finally, an analysis of agent recurrent dynamics revealed several similarities to theorized evidence accumulation mechanisms underlying patch foraging behaviour in biological agents (Davidson & El Hady, 2019), and single-cell recordings from foraging primates (Hayden et al., 2011).

The patch foraging work described in this thesis can be viewed as a proof of concept, and is limited in several ways. Agents used the maximum a posteriori policy optimization (MPO) reinforcement learning algorithm (Abdolmaleki et al., 2018), and a LIDAR-based sensory system. These were both choices made based on the success of previous work (Cultural General Intelligence Team et al., 2022), but future work may explore alternative learning algorithms and sensory spaces. Further, the neural analyses employed in this work were only correlational. Results in Chapter 3, along with decades of work in neuroscience, show that these correlational analyses are not enough to draw conclusions about the mechanisms underlying the behaviour of interest. Other strategies, or variations on accumulation models have also been proposed to underlie foraging behaviour (Cazettes et al., 2021; Davidson & El Hady, 2019; Kilpatrick et al., 2021), which this work does not distinguish between. Additional analyses, simulations, behavioural experiments, and causal manipulations are needed to draw stronger conclusions about the neural mechanisms underlying patch foraging behaviour.

While these results could have been demonstrated in a 1D discrete environment, the 3D continuous environment used allows for immediate extensions. For example, the foraging environment can be easily extended to investigate behaviours like traplining between multiple patches in 2D or 3D space, where agents learn repeatable sequences of patch-checking (Woodgate et al., 2017). Varying the environmental statistics of patch generation may allow researchers to test under which situations adaptive behaviour does and does not emerge. The agents and environments used also allow for immediate extensions into multi-agent foraging to investigate emergent cooperative and competitive behaviours during foraging. Finally, theory suggests that animal movements during foraging should adhere to a Lévy distribution, as it may be an optimal way to search for unknown patch locations (i.e., the Lévy flight foraging hypothesis; Viswanathan et al., 1996). Lévy distributions are characterised by a high density near small values, but a large tail that extends infinitely. In foraging navigation, this looks like many steps that are small, broken by infrequent steps that are very large. The idea is that this allows animals to perform local exploration of an area before travelling a long distance to a completely new area. This is related to the idea of dual motor systems in the early evolutionary stages of the brain, where locomotion policies are thought to have alternated between local and long-range exploration via dopamine (Cisek, 2019). This is also thought to be the original purpose of dopamine in the brain (Hills et al., 2015). Direct future work can test if Lévy-like movement patterns emerge in deep reinforcement learning agents optimized to maximize resource intake in environments where patch locations are unknown. For example, agents tasked with maximizing resource intake in a patchy environment with no (or very impoverished) sensors would still need to learn locomotion policies to maximize resource intake. These agents might learn to alternate between local and long-range locomotion to explore the world more effectively than a random policy. These results would provide further support for the Lévy flight foraging hypothesis, give insight into the environmental statistics that shaped the evolution of the brain,

and quantify locomotion policies that we desire of artificial agents in the real world.

The current work also raises questions about temporal discounting in deep reinforcement learning models of biological agents. The agents here approached optimal patch foraging behaviour, but only *after* accounting for temporal discounting. These results are at odds with many biological experiments that show optimal foraging behaviour without the need to account for any temporal discounting (Cowie, 1977; Krebs et al., 1974; Lottem et al., 2018; Pacheco-Cobos et al., 2019; Stephens & Krebs, 2019; Vertechi et al., 2020). Future work may consider reinforcement learning methods that optimize the same objective as the MVT—the average reward rate (e.g., R-learning; Schwartz, 1993). However, average reward methods for deep reinforcement learning are not nearly as prominent as their discounted counterparts (c.f., Zhang & Ross, 2021). Other research has shown that human in-lab foraging behaviour is better explained by an explicit MVT model compared to both average reward learning and discounted reinforcement learning (Constantino & Daw, 2015). However it is possible that an MVT-like adaptive mechanism may emerge via meta-RL if trained to maximize the average reward rate instead of a cumulative sum of discounted rewards.

Foraging is a fundamental decision problem, and has strong links to many real-world problems such as information search and resource management. Foraging frameworks have also been successfully used to explain other aspects of cognition and neural phenomena, such as visual search (Wolfe et al., 2018), human memory (Hills et al., 2012), and hippocampal reactivation patterns (McNamee et al., 2021), which provide another avenue where models of foraging may aid in biological research. Overall, I argue that foraging provides an exciting framework for testing and developing adaptive artificial agents, and also represents an important unsolved problem in biology.

5.4 General discussion

Through these three studies, I contribute to our understanding of rapid decision-related sensorimotor adaptation and its underlying mechanisms in biological agents. Specifically, I show that humans are able to optimally perform a novel rapid reaching task, and change their mind in the face of new information. I provide support that these abilities are achieved through cognitive mechanisms such as forward visuomotor models and evidence accumulation. In addition, I contribute two examples of artificial agents that converge toward rapid sensorimotor adaptive behaviours and mechanisms similar to biological agents in both perceptual decision making and in patch foraging. These studies provide compelling evidence that environmental pressures from biology offer one path toward adaptive artificial agents. In the sections above, I discussed issues related to single chapters in isolation. In the remainder of this section, I discuss issues that pertain to two or more of the chapters presented, before turning to discuss the mutualistic relationship of artificial and biological intelligence research.

In both Chapter 3 and Chapter 4, I present artificial agents that discovered mechanisms similar to the evidence accumulation processes theorized to underlie perceptual and foraging decisions in mammals. In Chapter 3, I showed that agents trained without environmental noise do *not* learn primate-like evidence accumulation mechanisms, and argued that environmental noise is a critical factor in discovering accumulation-like mechanisms. However agents in Chapter 4 were trained *without* environmental noise (albeit in a different task), and yet there I argue that these agents *did* discover accumulation like-mechanisms. We can perhaps reconcile these conflicting results by speculating that environmental complexity can serve a similar function to environmental noise. Agents acting in a continuous 3D environment are presented with a diversity of sensory experiences, even in very similar situations. For instance, even if the agent is in a slightly different position in 3D space, many LIDAR sensor values change despite having little relevance for the agent's action policy. In this way, very

large environments and fine sensors may provide one way to skip the need for external (or internal) noise to learn adaptive mechanisms. On the other hand, environmental noise in small environments may prepare agents to make robust decisions in more complex environments, by learning robust adaptive mechanisms. While environmental, neural, or motor noise is often regarded as a limitation, some argue that it serves a purpose for learning, exploration, or developing robust representations (Faisal et al., 2008; McDonnell & Ward, 2011). Despite these ideas and the ubiquity of noise in biological systems, noise is still understudied in artificial intelligence research, which suggests an important path for future work.

Evidence accumulation mechanisms are incredibly powerful and elegant. Many complex and adaptive behaviours fall out of these mechanisms, such as changes of mind. Further, they present a general decision mechanism, which can adapt to different contexts (e.g., Mante et al., 2013). The results in this thesis suggest two examples of agents that discover evidence accumulation-like mechanisms to make adaptive decisions. I find this result endlessly exciting, as many other aspects of biological cognition such as value-based decision making (Shadlen & Shohamy, 2016), social decisions (Rorie & Newsome, 2005; Shadlen & Kiani, 2013), and even consciousness (Kang et al., 2017), are all thought to rely to varying degrees on similar evidence accumulation processes (Shadlen & Kiani, 2013). Artificial agents that can reliably discover a fundamental decision making mechanism in biological agents potentially provide a rich testing ground for many aspects of cognition. In particular, I am excited about the idea that evidence accumulation mechanisms underlie value-based decisions, where sensory information is largely static (e.g., pictures of two snack food items). This work suggests that recalling relevant memories as evidence for or against different value-based options is a noisy, sequential process (Shadlen & Shohamy, 2016). Is this due to a biological limitation in how internal, value-based memories are recalled? Or is this a feature of a process traversing structured knowledge within connections of a network? Investigating how networks learn by themselves to store and recall

memories within a neural network in the service of action may shed light on this issue.

In Chapter 3, I also present results that recurrence is needed for agents to learn an evidence accumulation mechanism. This result may seem intuitive to biological researchers. However, many deep learning models still use networks that contain feedforward connections only, rather than including recurrent or backward connections (Spoerer et al., 2020). This is in part because of the difficulty of training recurrent networks relative to feedforward networks (Hochreiter, 1998), and in part because many tasks of interest have been solved using feedforward networks in the past. However, the real-world is complex and dynamic in time, necessitating processes like recurrence and long-term memory recall. Purely feedforward agents that act irrespective of temporal context seem like an ill-suited direction to achieving adaptive artificial agents.

In this vein, when constructing a deep reinforcement learning agent, it is critical to consider the temporal resolution in which they are simulated. In tasks like chess, each discrete move is encoded as a single time step. However, these clean choices break down when enacting movements in a real or simulated world. How frequently should agents be able to control their actions? Too slow, and agents are unable to adapt to changes rapidly. Too fast, and agents face the problem of learning when many timesteps have no relevant information. As in a bandit task, deep neural networks are typically given one feedforward pass through the network to make a decision about what arm to choose. But studies on biological agents show that a significant amount of neural processing happens over time when making similar bandit decisions (Costa & Averbeck, 2020). In addition, visual object processing is typically implemented in deep neural networks as a single feedforward pass, yet performance and biological similarities both increase when networks are given recurrence and processing time (Kietzmann et al., 2019; Spoerer et al., 2020). It is unlikely that deep neural networks can perform all the computations that biological agents do with a single forward pass

through a network, given that biological computation is highly recurrent. It is interesting to speculate then what biological agents do with this additional processing time and recurrence. The answers are likely other processes in the service of learning, planning, and adaptive behaviour, such as recalling past experiences, dynamic programming in the service of motor control (Todorov & Jordan, 2002), simulating potential behaviours (e.g., forward models, as in Chapter 2), and integrating evidence. Some deep reinforcement learning research already approaches this problem in its own way. For example, some model-based reinforcement learning agents are explicitly programmed to take several steps of simulation to update internal models of the world between making each new decision (e.g., Dreamer; Hafner et al., 2019). Others approach this problem by abstracting time using multiple temporal resolutions or hierarchical learning (Precup & Sutton, 1997; Singh, 1992). Conversely, much of deep learning research in 2023 relies on methods like the Transformer (Vaswani et al., 2017), which often uses a feedforward architecture but includes many time steps of input at once (termed context windows). While extremely impactful, these architectures face the same problem—if simulated steps are too slow, behaviour cannot adapt quickly, but if simulated steps are too fast, relevant information may fall outside the scope of the context window and be forgotten. Context windows are getting much bigger (Yu et al., 2023), but this ignores the fact that recurrence gives researchers agents with potentially infinite context. Taking a long-term view from an ivory tower, methods looking at increasing the reliability of recurrence in deep learning research (potentially inspired by biological agents) may provide lasting solutions to many of the above problems.

Throughout this thesis, I use the term "optimally" to describe certain behaviours. The use of this term may seem intuitive to those in machine learning. Machine learning has intimate ties to mathematical optimization, where learning is posed as the gradual maximization (or minimization) of some objective function. In many simulated environments, there is a known global minimum or maximum—in reinforcement

learning, an optimal policy that maximizes reward (π^*; see A primer on reinforcement learning). These claims of optimality are possible because the problems under consideration are significantly smaller than the vast, real world. In contrast, researchers in the biological sciences might recoil at claims of optimal behaviour. There are many examples of both optimal behaviour (e.g., Körding & Wolpert, 2006), and suboptimal behaviour (e.g., Tversky & Kahneman, 1992) observed in humans. However, it is important to consider the big picture outside of individual tasks when discussing optimality in biological systems. For example, humans are poor at estimating magnitudes (Tversky & Kahneman, 1992), but this could be viewed as optimal with respect to neural constraints (Summerfield & Parpart, 2022; Woodford, 2020). Others argue that optimality is often poorly defined across fields like psychology, behavioural economics, and biology (Schoemaker, 1991), and that researchers should move away from making claims of optimal behaviour in biological studies (Rahnev & Denison, 2018). One important consideration is in comparing artificial agents to human performance. Superhuman performance within a particular task is sometimes claimed by machine learning researchers. However, human suboptimality on a given task may be the result of a larger optimization for many tasks, for which an artificial agent would dramatically fail. If trained to complete a wider variety of tasks, artificial agents might develop similar suboptimalities, despite having much more impressive abilities than agents optimized for a single task. Overall, researchers (myself included) should be more careful and nuanced with their definitions of optimality—especially when conducting interdisciplinary work, as it can be harmful for the development of artificial *general* intelligence.

Finally, in the biological cognitive sciences, we often make the convenient and knowingly-incorrect assumption that the subjects in our experiments don't learn. For example during the random dot motion task, we model these decision processes as if they are stable and frozen at the beginning of the task, much like an evaluation phase in machine learning research. But work has shown that this assumption is

hiding results in plain sight. For example, lapses during decision making can be explained by an overzealous learning rate operating throughout the course of a task (Gupta et al., 2023). That is, errors made by animals on very simple decisions are often attributed to cognitive limitations. However, these results can be explained by animals learning throughout an experiment. If an animal sees four left-correct trials in a row on the random dot motion task (Chapter 3), it might erroneously learn that the task has changed to be left-biased, rather than attributing this pattern to random noise (Gupta et al., 2023). As such, these animals that appear to be performing poorly in a static task are actually employing adaptive mechanisms under the assumption that the task is dynamic. In a similar vein, hidden issues can arise when using the typical two-stage approach of training and testing in machine learning research (Dulac-Arnold et al., 2019). This is especially true if we want artificial agents that continuously learn as biological agents do. Overall, more dynamic methods are likely needed to help analyze complex systems that change under our feet while we are trying to understand them.

5.5 The mutualistic relationship of artificial and biological intelligence research

A mutualistic relationship is a concept from biology, which describes the ecological interaction between two or more species where each species has a net benefit (Bronstein, 2015). A popular example is from clownfish and sea anemones, which both protect each other from their respective predators ("Finding Nemo", 2003). Similarly, there have been many instances of net benefit interactions between biological and artificial intelligence research in the past. Here I conclude by arguing that these two fields still have much to benefit from closely interacting with each other.

5.5.1 How can artificial intelligence aid biological research?

The dizzying pace of artificial intelligence research has provided a wealth of valuable tools and techniques—some of which have direct applications to understanding biological function. First, stellar recent work has leveraged training recurrent neural networks using supervised learning to reproduce animal behaviour. Through this process, researchers can then investigate the emergent dynamics of these networks trained to reproduce behaviour for clues about the underlying cognitive processes which generate such behaviour. This technique replaces one black box with another, but with the benefit that the *artificial* black box is significantly more interpretable and accessible than its biological archetype (Barak, 2017). In addition, these artificial black boxes are significantly easier to manipulate for functional investigations. For example, the microstimulation experiment from Chapter 3 is significantly easier to perform compared to its primate predecessor (Hanks et al., 2006). Another example is a "teleportation" experiment that extends Chapter 3. We can induce some level of decision preference in the agent, and then teleport the agent's arm to different positions between the two targets before letting the agent enact its decision. In this situation, we would expect the agent to choose its preferred target when released in many positions, but choose the unpreferred target when the preferred target is relatively far away. As such, we can map the agent's decision-space tradeoff over many arm positions to provide insight into how primates consider both decision preference, time, and motor costs. This experiment is presently impossible to run in the lab, and illustrates a unique way for artificial neural networks to contribute to biological research.

Deep neural network techniques in neuroscience and psychology are close cousins to traditional cognitive modeling. However, whereas cognitive models need to be explicitly defined by the researcher, recurrent neural networks can perform model discovery themselves. This brings about a tension—hand-specified cognitive models may be too

restrictive, but recurrent neural networks can potentially specify any function that can fit data (i.e., they are thought to be Turing complete; Sontag and Siegelmann, 1995). Work has shown insight into biological cognition by training recurrent neural networks on behavioural data like discrete choices (Driscoll et al., 2022; K. J. Miller et al., 2023), or relatively dense data like movement trajectories (M. M. Churchland et al., 2012; Mante et al., 2013; Michaels et al., 2020; Sussillo et al., 2015). For example, researchers have found that motor cortex embeds muscle commands in a way that disentangles them from other similar states (Russo et al., 2018). Other work has shown that the brain might contextually reuse and combine learned dynamic motifs to solve a variety of tasks (Driscoll et al., 2022). Further, exciting new techniques have been developed to encourage parsimony in these recurrent neural networks, so that their discovered solutions are human-interpretable (K. J. Miller et al., 2023). These methods provide compelling tools to understanding biological cognition, and may rapidly advance as machine learning techniques improve, and neural (e.g., Steinmetz et al., 2021) and behavioural (e.g., Nath et al., 2019) recordings in biological agents become much richer and easier to collect.

Another application of artificial intelligence to the study of biological cognition, on display in this thesis, is training deep neural networks to optimize a behaviour given constraints inspired by biological agents. Here the idea is that we can infer the pressures that gave rise to biological cognitive mechanisms if artificial neural networks do or do not converge to a solution with similarities to biology. In this area, much work has been performed on biological vision with the supervised training of convolutional neural networks (Kanwisher et al., 2023). For example, researchers have trained convolutional neural networks (CNNs) to maximize performance in an object recognition task. In this work, psychophysical phenomena such as set size effects (Nicholson & Prinz, 2021) and scene incongruence effects (Jacob et al., 2021) emerge in these CNNs trained for object recognition. Conversely, other phenomena in human vision such as 3D processing or part-based processing do not emerge in these systems

(Jacob et al., 2021). Together, these results give us clues that the importance of recognizing objects for biological systems may have given rise to some behaviours and mechanisms, but not others (Kanwisher et al., 2023). Other work has emphasized architectural constraints, rather than objective functions. For example, CNNs with a split architecture, intended to mirror the separation of higher-order visual processing into dorsal and ventral pathways, showed functional specialization in each split pathway emerged when trained on multiple, dissimilar tasks (Scholte et al., 2018). Such results can shed light on "why" questions regarding human brain organization. Finally, similar questions have been asked in deep reinforcement learning work. In one study, deep reinforcement learning agents were trained to maximize reward on a series of navigation tasks. Agents that were given a grid cell-like mechanism outperformed agents with place cell-like mechanisms, agents with no pre-existing mechanisms, and even human experts (Banino et al., 2018). These grid cell-like agents further learned to take shortcuts in navigation tasks where appropriate. Such results suggest the "why" of grid cells observed in the brains of animals—that grid cells offer an extremely effective basis for navigation. However, this line of work relies on providing a rough parallel to specific ethologically important objective functions for which biological agents are under pressure to optimize (Marblestone et al., 2016). Overall, the emergence of similar behaviours and dynamics in deep neural networks to those of biological agents provides unique evidence for the purpose and origin of phenomena in biology.

This kind of optimization research at the intersection of artificial and biological agents may also provide insight into the reward hypothesis. The reward hypothesis is the idea that intelligence, and its associated abilities, can be understood as subserving the maximization of reward (Bowling et al., 2023; Silver et al., 2021). In short, if we could define the correct reward function for a flexible enough agent, all the properties we can think of as intelligent would emerge. The reward hypothesis stands in contrast to other approaches, like a modular approach where specialized

problem formulations are needed for each ability, each with their own optimization process. Investigating this hypothesis is a daunting task, and requires that we identify compelling, intelligent behaviours across the animal kingdom to test if they emerge through the maximization of reward. Such work runs into several problems. For instance, how is this grand reward function specified in the first place (Summerfield, 2022)? Some argue that such a reward function would be too complex to specify with the diversity of goals and abilities observed in intelligent biological agents (Ringstrom, 2022). The reward hypothesis of course also has compelling parallels to the evolutionary process of fitness maximization in biology, especially the quantitative measure of individual reproductive success in population genetics (w or ω in population genetics models). That we observe emergent adaptive patch foraging and several primate-like decision making phenomena through reinforcement learning with relatively simple reward structures provides support that perhaps these grand reward functions are relatively simple, but it is the world that is extremely complex. This remains to be seen and would require training agents in very large and dynamic environments to observe emergent behaviours.

Those in the biological sciences may shy away from artificial intelligence work because of its mathematical or technical requirements, or its perception as seeking engineering solutions instead of truth about the natural world. However, tools from AI provide incredible untapped potential to understanding complex, adaptive biological systems. Further, biological researchers may find much more common ground with AI researchers than they expected when tackling the big questions—and may find many unintended benefits in the diversity of perspectives.

5.5.2 How can biological systems aid artificial intelligence research?

Despite some claims otherwise, biological systems have continuously provided inspiration and useful analogies for the study of artificial intelligence (Summerfield, 2022;

Zador et al., 2023). In addition, we desire intelligent artificial systems not for themselves in isolation, but with respect to how they can work with humans or within natural environments. Some of the work in artificial intelligence can even be conceptualized as "part of theoretical psychology" (Ady, 2023; Moore & Newell, 1974). In this final section, I outline some thoughts on how AI research may still find use in looking at adaptive biological agents.

A core message in this thesis is that environmental pressures from biology offer one path toward adaptive artificial agents. I provide two concrete computational experiments that show the potential of this approach. Artificial agents learned to rapidly adapt in complex foraging and perceptual decision making environments similar to biological agents. These results fit with other work showing that training artificial agents on biological tasks encourages biological-like solutions, for example motion processing (Rideaux & Welchman, 2020), detour problems (Banino et al., 2018), and visual object detection (Jacob et al., 2021). If we desire artificial agents that can adapt as well as biological agents, I argue we should push this path forward. This might look like creating extremely large and dynamic simulated environments inspired by biology. Many current experiments rightfully focus on one specific problem like patch foraging. However, needing to solve many biological problems in a large world simultaneously may provide unintended benefits, as different processes and experiences may be able to naturally scaffold learning. There are many theories about what major factors gave rise to the emergence of natural intelligence, such as fundamental decision problems, social collaboration (the social brain hypothesis; Dunbar, 1998), or motor control (motor chauvinism; Wolpert et al., 2001). Optimizing artificial agents to complete these tasks can help us answer questions about the origins of biological intelligence. However, complex environments that involve all of these tasks at once provide a potential path for robust and general artificial agents to emerge. In this line of thinking, research may be able to identify a list of critical tasks and pressures that encourage the adaptive behaviours we desire of artificial agents. This

might give rise to some kind of standard curriculum inspired by biology, which would encourage the development of all the necessary building blocks (or motifs) for agents to compositionally adapt to many novel tasks (e.g., Driscoll et al., 2022). Speculating, this line of work could give rise to a biological foundation model (Bommasani et al., 2021), which could be fine-tuned for specific uses.

The idea that human-level adaptive artificial agents may emerge through optimization in environments that mirror developmental (L. Smith & Gasser, 2005; Turing, 1950) or evolutionary pressures (Cisek, 2019) faced by biological systems is not a new idea (Kanwisher et al., 2023; Kell & McDermott, 2019). This idea simply translates theory about how biological agents developed intelligent adaptive behaviours in the first place to artificial agents. Stellar work provides experimental evidence for this idea that environmental statistics have strongly influenced the organization of perceptual, cognitive, and motor systems in humans (Behrens et al., 2007; Ernst & Banks, 2002; Körding & Wolpert, 2004). In this thesis, I provide two concrete examples, which show that a parallel to this process is a fruitful endeavour for developing adaptive artificial agents that should be seriously explored further.

A main goal of AI research is to make agents that behave like humans (Summerfield, 2022). However, one concern is that such agents modeled after humans would be limited in the same way that humans are limited. Humans are suboptimal decision makers in many contexts when considered in isolation (Kahneman, 2011), and agents that develop as humans do may similarly inherit these suboptimalities. It could be argued that biological suboptimalities in specific tasks arise because humans are optimized to perform many tasks in the real world. However, some goals of artificial agents are to act in specific situations rather than the world as a whole. For these and many other reasons, people should be very clear with their intended goals and limitations for artificial agents. There are of course many safety and ethics issues in this area outside the scope of this thesis (see Christian, 2020). Conversely, training AI agents to develop as humans do may be one approach to AI that aligns with human

values. If artificial agents need to cooperate and interact with humans to achieve their goals, and integrate experiences within human society, then it seems likely that such interaction may encourage abilities like social perception and human-machine cooperation. Many non-human animals work in concert with humans (e.g., working dogs), despite communication barriers. Further, productive biological interspecies collaboration is possible despite these animals having significantly different sensors, effectors, and experiences possibly out of the realm of human understanding—could we possibly understand what it is like to be a bat, dog, octopus, or an artificial neural network (Carls-Diamante, 2022; Godfrey-Smith, 2016; Horowitz, 2002; Nagel, 1980)? I argue that it is impossible to truly understand what it is like to be another agent, human or otherwise. However, this has not deterred humans in the past from developing net benefit relationships with other agents, and I believe that this factor should not deter us from developing these relationships in the future as well.

Overall, this framework for AI means that artificial intelligence researchers need to help discover much more about biology, evolution, and development, and work to leverage what we already do know. Research into biological adaptive behaviour not only helps us understand pressures to use for AI research, but helps us better understand ourselves, provide ways to improve our decisions, and sheds insight into clinical treatments for neurological disorders. Research into artificial adaptive behaviour not only helps us better understand biological systems, but helps us build better virtual assistants that adapt to our needs, autonomously drive vehicles, or support humans in health care settings. While humans possess the incredible ability to rapidly adapt to a changing world without effortful thinking, it may require slow, deliberate, and effortful actions to work on these challenges as they should be faced—together.

5.6 References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., & Riedmiller, M. (2018). Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*.

Ady, N. M. (2023). Specific machine curiosity.

Atiya, N. A., Zgonnikov, A., O'Hora, D., Schoemann, M., Scherbaum, S., & Wong-Lin, K. (2020). Changes-of-mind in the absence of new post-decision evidence. *PLOS Computational Biology*, *16*(2), e1007149.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T. P., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433.

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, *46*, 1–6.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1–68.

Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., Lefrancq, A., Green, S., Valdés, V., Sadik, A., Et al. (2016). DeepMind Lab. *arXiv preprint arXiv:1612.03801*.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Bengio, Y., Scellier, B., Bilaniuk, O., Sacramento, J., & Senn, W. (2016). Feedforward initialization for fast inference of deep generative networks is biologically plausible. *arXiv preprint arXiv:1606.01651*.

Boag, R. J., Strickland, L., Heathcote, A., Neal, A., Palada, H., & Loft, S. (2022). Evidence accumulation modelling in the wild: Understanding safety-critical decisions. *Trends in Cognitive Sciences*.

Bogacz, R., & Larsen, T. (2011). Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. *Neural Computation*, *23*(4), 817–851.

Boldt, A., Schiffer, A.-M., Waszak, F., & Yeung, N. (2019). Confidence predictions affect performance confidence and neural preparation in perceptual decision making. *Scientific Reports*, *9*(1), 4031.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bowling, M., Martin, J. D., Abel, D., & Dabney, W. (2023). Settling the reward hypothesis, In *International Conference on Machine Learning*.

Bronstein, J. L. (2015). *Mutualism*. Oxford University Press.

Carls-Diamante, S. (2022). Where is it like to be an octopus? *Frontiers in Systems Neuroscience*, *16*, 840022.

Cazettes, F., Murakami, M., Renart, A., & Mainen, Z. (2021). Reservoir of decision strategies in the mouse brain. *bioRxiv*.

Chapman, C. S., Gallivan, J. P., Wong, J. D., Wispinski, N. J., & Enns, J. T. (2015). The snooze of lose: Rapid reaching reveals that losses are processed more slowly than gains. *Journal of Experimental Psychology: General*, *144*(4), 844.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, *487*(7405), 51–56.

Cisek, P. (2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology*, *22*(6), 927–936.

Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*, *81*(7), 2265–2287.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, *33*, 269–298.

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(4), 837–853.

Costa, V. D., & Averbeck, B. B. (2020). Primate orbitofrontal cortex codes information relevant for managing explore–exploit tradeoffs. *Journal of Neuroscience*, *40*(12), 2553–2561.

Cowie, R. J. (1977). Optimal foraging in great tits (Parus major). *Nature*, *268*(5616), 137–139.

Cultural General Intelligence Team, Bhoopchand, A., Brownfield, B., Collister, A., Lago, A. D., Edwards, A., Everett, R., Frechette, A., Oliveira, Y. G., Hughes, E., Mathewson, K. W., Mendolicchio, P., Pawar, J., Pislar, M., Platonov, A., Senter, E., Singh, S., Zacherl, A., & Zhang, L. M. (2022). Learning robust real-time cultural transmission without human data. *arXiv preprint arXiv.2203.00715*.

Davidson, J. D., & El Hady, A. (2019). Foraging as an evidence accumulation process. *PLoS Computational Biology*, *15*(7), e1007060.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

De Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, *7*(3), 242–249.

Driscoll, L., Shenoy, K., & Sussillo, D. (2022). Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *bioRxiv*, 2022–08.

Dulac-Arnold, G., Mankowitz, D., & Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, *6*(5), 178–190.

Enderle, J. D., & Sierra, D. A. (2013). A new linear muscle fiber model for neural control of saccades. *International Journal of Neural Systems*, *23*(02), 1350002.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303.

Feulner, B., Perich, M. G., Miller, L. E., Clopath, C., & Gallego, J. A. (2022). Feedback-based motor control can guide plasticity and drive rapid learning. *bioRxiv*.

Finding Nemo. (2003). Pixar Animation Studios.

Gallivan, J. P., Chapman, C. S., Wolpert, D. M., & Flanagan, J. R. (2018). Decision-making in sensorimotor control. *Nature Reviews Neuroscience*, *19*(9), 519–534.

Godfrey-Smith, P. (2016). *Other minds: The octopus, the sea, and the deep origins of consciousness*. William Collins.

Gupta, D., DePasquale, B., Kopec, C. D., & Brody, C. D. (2023). Trial-history biases in evidence accumulation can give rise to apparent lapses. *bioRxiv*, 2023–01.

Hafner, D., Lillicrap, T. P., Ba, J., & Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience*, *9*(5), 682–689.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*(7), 933–939.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116.

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics*, *81*(7), 2288–2303.

Horowitz, A. C. (2002). *The behaviors of theories of mind, and a case study of dogs at play*. University of California, San Diego.

Jacob, G., Pramod, R., Katti, H., & Arun, S. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, *12*(1), 1872.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kang, Y. H., Petzschner, F. H., Wolpert, D. M., & Shadlen, M. N. (2017). Piercing of consciousness as a threshold-crossing operation. *Current Biology*, *27*(15), 2285–2295.

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254.

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Kilpatrick, Z. P., Davidson, J. D., & El Hady, A. (2021). Uncertainty drives deviations in normative foraging decision strategies. *Journal of the Royal Society Interface*, *18*(180), 20210337.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Krebs, J. R., Ryan, J. C., & Charnov, E. L. (1974). Hunting by expectation or optimal foraging? A study of patch use by chickadees. *Animal Behaviour*, *22*, 953–IN3.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, *7*(1), 13276.

Lottem, E., Banerjee, D., Vertechi, P., Sarra, D., oude Lohuis, M. N., & Mainen, Z. F. (2018). Activation of serotonin neurons promotes active persistence in a probabilistic foraging task. *Nature Communications*, *9*(1), 1–12.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 94.

McDonnell, M. D., & Ward, L. M. (2011). The benefits of noise in neural systems: Bridging theory and experiment. *Nature Reviews Neuroscience*, *12*(7), 415–425.

McNamee, D. C., Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2021). Flexible modulation of sequence generation in the entorhinal–hippocampal system. *Nature Neuroscience*, *24*(6), 851–862.

Michaels, J. A., Schaffelhofer, S., Agudelo-Toro, A., & Scherberger, H. (2020). A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *Proceedings of the National Academy of Sciences*, *117*(50), 32124–32135.

Miller, K. J., Eckstein, M., Botvinick, M. M., & Kurth-Nelson, Z. (2023). Cognitive model discovery via disentangled RNNs. *bioRxiv*, 2023–06.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Et al. (2015).

Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Moore, J., & Newell, A. (1974). How can MERLIN understand? Hillsdale, NJ: Erlbaum Assoc.

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, *22*(10), 1677–1686.

Nagel, T. (1980). What is it like to be a bat?, In *The language and thought series*. Harvard University Press.

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., & Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, *14*(7), 2152–2176.

Nicholson, D. A., & Prinz, A. A. (2021). Deep neural network models of object recognition exhibit human-like limitations when performing visual search tasks. *bioRxiv*, 2020–10.

Nonacs, P. (2001). State dependent behavior and the marginal value theorem. *Behavioral Ecology*, *12*(1), 71–83.

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.

Pacheco-Cobos, L., Winterhalder, B., Cuatianquiz-Lima, C., Rosetti, M. F., Hudson, R., & Ross, C. T. (2019). Nahua mushroom gatherers use area-restricted search strategies that conform to marginal value theorem predictions. *Proceedings of the National Academy of Sciences*, *116*(21), 10339–10347.

Pearson, J. M., Hayden, B. Y., Raghavachari, S., & Platt, M. L. (2009). Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. *Current Biology*, *19*(18), 1532–1537.

Pesquita, A., Chapman, C. S., & Enns, J. T. (2016). Humans are sensitive to attention control when predicting others' actions. *Proceedings of the National Academy of Sciences*, *113*(31), 8669–8674.

Peters, M. A. (2022). Confidence in decision-making, In *Oxford research encyclopedia of neuroscience*.

Precup, D., & Sutton, R. S. (1997). Multi-time models for temporally abstract planning. *Advances in Neural Information Processing Systems*, *10*.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*, e223.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266.

Rideaux, R., & Welchman, A. E. (2020). But still it moves: Static image statistics underlie how we see motion. *Journal of Neuroscience*, *40*(12), 2538–2552.

Ringstrom, T. J. (2022). Reward is not necessary: How to create a compositional self-preserving agent for life-long learning. *arXiv preprint arXiv:2211.10851*.

Rorie, A. E., & Newsome, W. T. (2005). A general mechanism for decision-making in the human brain? *Trends in Cognitive Sciences*, *9*(2), 41–43.

Russo, A. A., Bittner, S. R., Perkins, S. M., Seely, J. S., London, B. M., Lara, A. H., Miri, A., Marshall, N. J., Kohn, A., Jessell, T. M., Et al. (2018). Motor cortex embeds muscle-like commands in an untangled population response. *Neuron*, *97*(4), 953–966.

Schaffelhofer, S., Sartori, M., Scherberger, H., & Farina, D. (2014). Musculoskeletal representation of a large repertoire of hand grasping actions in primates. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *23*(2), 210–220.

Schoemaker, P. J. (1991). The quest for optimality: A positive heuristic of science? *Behavioral and Brain Sciences*, *14*(2), 205–215.

Scholte, H. S., Losch, M. M., Ramakrishnan, K., de Haan, E. H., & Bohte, S. M. (2018). Visual pathways from the perspective of cost functions and multi-task deep neural networks. *Cortex*, *98*, 249–261.

Schwartz, A. (1993). A reinforcement learning method for maximizing undiscounted rewards, In *International Conference on Machine Learning*.

Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience*, *5*(7), 532–545.

Seth, A., Hicks, J. L., Uchida, T. K., Habib, A., Dembia, C. L., Dunne, J. J., Ong, C. F., DeMers, M. S., Rajagopal, A., Millard, M., Et al. (2018). OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLoS Computational Biology*, *14*(7), e1006223.

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791–806.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535.

Singh, S. P. (1992). Reinforcement learning with a hierarchy of abstract models, In *Proceedings of the National Conference on Artificial Intelligence*.

Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life*, *11*(1-2), 13–29.

Song, J.-H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, *13*(8), 360–366.

Sontag, E. D., & Siegelmann, H. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences*, *50*, 132–150.

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Computational Biology*, *16*(10), e1008215.

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., Et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, *372*(6539), eabf4588.

Stephens, D. W., & Krebs, J. R. (2019). *Foraging theory*. Princeton University Press.

Stone, S. A. (2023). Eye and body tracking in the lab, in the wild, and in the clinic.

Summerfield, C. (2022). *Natural general intelligence: How understanding the brain can help us build AI*. Oxford University Press.

Summerfield, C., & Parpart, P. (2022). Normative principles for decision-making in natural environments. *Annual Review of Psychology*, *73*, 53–77.

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, *18*(7), 1025–1033.

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

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., & Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction, In *International Conference on Autonomous Agents and Multiagent Systems*.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*(9), 907–915.

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, *5*(11), 1226–1235.

Turing, A. M. (1950). Mind. *Mind*, *59*(236), 433–460.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vertechi, P., Lottem, E., Sarra, D., Godinho, B., Treves, I., Quendera, T., oude Lohuis, M. N., & Mainen, Z. F. (2020). Inference-based decisions in a hidden state foraging task: Differential contributions of prefrontal cortical areas. *Neuron*, *106*(1), 166–176.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., & Stanley, H. E. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, *381*(6581), 413–415.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868.

White, A. (2015). Developing a predictive approach to knowledge.

Wispinski, N. J., Butcher, A., Mathewson, K. W., Chapman, C. S., Botvinick, M. M., & Pilarski, P. M. (2022). Adaptive patch foraging in deep reinforcement learning agents. *Transactions on Machine Learning Research*.

Wispinski, N. J., Gallivan, J. P., & Chapman, C. S. (2020). Models, movements, and minds: Bridging the gap between decision making and action. *Annals of the New York Academy of Sciences*, *1464*(1), 30–51.

Wolfe, J. M., Cain, M. S., & Alaoui-Soce, A. (2018). Hybrid value foraging: How the value of targets shapes human foraging behavior. *Attention, Perception, & Psychophysics*, *80*(3), 609–621.

Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences*, *5*(11), 487–494.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, *12*, 579–601.

Woodgate, J. L., Makinson, J. C., Lim, K. S., Reynolds, A. M., & Chittka, L. (2017). Continuous radar tracking illustrates the development of multi-destination routes of bumblebees. *Scientific Reports*, *7*(1), 1–15.

Wurman, P. R., Barrett, S., Kawamoto, K., MacGlashan, J., Subramanian, K., Walsh, T. J., Capobianco, R., Devlic, A., Eckert, F., Fuchs, F., Et al. (2022). Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, *602*(7896), 223–228.

Yoo, S. B. M., Tu, J. C., & Hayden, B. Y. (2021). Multicentric tracking of multiple agents by anterior cingulate cortex during pursuit and evasion. *Nature Communications*, *12*(1), 1985.

Yu, L., Simig, D., Flaherty, C., Aghajanyan, A., Zettlemoyer, L., & Lewis, M. (2023). MEGABYTE: Predicting million-byte sequences with multiscale transformers. *arXiv preprint arXiv:2305.07185*.

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., Et al. (2023). Catalyzing next-generation artificial intelligence through NeuroAI. *Nature Communications*, *14*(1), 1597.

Zhang, Y., & Ross, K. W. (2021). On-policy deep reinforcement learning for the average-reward criterion. *arXiv preprint arXiv:2106.07329*.

Zhou, W., Chen, X., & Enderle, J. (2009). An updated time-optimal 3rd-order linear saccadic eye plant model. *International Journal of Neural Systems*, *19*(05), 309–330.

Zhu, J., & Thagard, P. (2002). Emotion and action. *Philosophical Psychology*, *15*(1), 19–36.

Bibliography

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., & Riedmiller, M. (2018). Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*.

Abrams, R. A., & Balota, D. A. (1991). Mental chronometry: Beyond reaction time. *Psychological Science*, *2*(3), 153–157.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*(2), 284–299.

Ady, N. M. (2023). Specific machine curiosity.

Ady, N. M., Shariff, R., Günther, J., & Pilarski, P. M. (2022). Five properties of specific curiosity you didn't know curious machines should have. *arXiv preprint arXiv:2212.00187*.

Andersen, R. A., Essick, G. K., & Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, *230*(4724), 456–458.

Atiya, N. A., Zgonnikov, A., O'Hora, D., Schoemann, M., Scherbaum, S., & Wong-Lin, K. (2020). Changes-of-mind in the absence of new post-decision evidence. *PLOS Computational Biology*, *16*(2), e1007149.

Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Dedieu, A., Fantacci, C., Godwin, J., Jones, C., Hemsley, R., Hennigan, T., Hessel, M., Hou, S., Kapturowski, S., ... Viola, F. (2020). *The DeepMind JAX Ecosystem*. http://github.com/deepmind

Badman, R. P., Hills, T. T., & Akaishi, R. (2020). Multiscale computation and dynamic attention in biological and artificial intelligence. *Brain Sciences*, *10*(6), 396.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*.

Baldauf, D., & Deubel, H. (2010). Attentional landscapes in reaching and grasping. *Vision Research*, *50*(11), 999–1013.

Balota, D. A., & Abrams, R. A. (1995). Mental chronometry: Beyond onset latencies in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1289.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T. P., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Et al. (2018). Vector-based nav-

igation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433.

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, *46*, 1–6.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1–68.

Basso, M. A., & Wurtz, R. H. (1997). Modulation of neuronal activity by target uncertainty. *Nature*, *389*(6646), 66–69.

Battaglia, P. W., & Schrater, P. R. (2007). Humans trade off viewing time and movement duration to improve visuomotor accuracy in a fast reaching task. *Journal of Neuroscience*, *27*(26), 6984–6994.

Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., Lefrancq, A., Green, S., Valdés, V., Sadik, A., Et al. (2016). DeepMind Lab. *arXiv preprint arXiv:1612.03801*.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Probabilistic population codes for bayesian decision making. *Neuron*, *60*(6), 1142–1152.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Bengio, Y., Scellier, B., Bilaniuk, O., Sacramento, J., & Senn, W. (2016). Feedforward initialization for fast inference of deep generative networks is biologically plausible. *arXiv preprint arXiv:1606.01651*.

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, *31*, 1–21.

Berger, A., Henik, A., & Rafal, R. (2005). Competition between endogenous and exogenous orienting of visual attention. *Journal of Experimental Psychology: General*, *134*(2), 207.

Bidari, S., El Hady, A., Davidson, J. D., & Kilpatrick, Z. P. (2022). Stochastic dynamics of social patch foraging decisions. *Physical Review Research*, *4*(3), 033128.

Blakemore, S.-J., Frith, C. D., & Wolpert, D. M. (1999). Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience*, *11*(5), 551–559.

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, *1*(7), 635–640.

Blanchard, T. C., & Hayden, B. Y. (2014). Neurons in dorsal anterior cingulate cortex signal postdecisional variables in a foraging task. *Journal of Neuroscience*, *34*(2), 646–655.

Boag, R. J., Strickland, L., Heathcote, A., Neal, A., Palada, H., & Loft, S. (2022). Evidence accumulation modelling in the wild: Understanding safety-critical decisions. *Trends in Cognitive Sciences*.

Bogacz, R. (2007). Optimal decision-making theories: Linking neurobiology with behaviour. *Trends in Cognitive Sciences*, *11*(3), 118–125.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700.

Bogacz, R., & Larsen, T. (2011). Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. *Neural Computation*, *23*(4), 817–851.

Boldt, A., Schiffer, A.-M., Waszak, F., & Yeung, N. (2019). Confidence predictions affect performance confidence and neural preparation in perceptual decision making. *Scientific Reports*, *9*(1), 4031.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*, *107*(4), 603–616.

Bowling, M., Martin, J. D., Abel, D., & Dabney, W. (2023). Settling the reward hypothesis, In *International Conference on Machine Learning*.

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

Brenner, E., Driesen, B., & Smeets, J. B. (2014). Precise timing when hitting falling balls. *Frontiers in Human Neuroscience*, *8*, 342.

Brenner, E., & Smeets, J. (2013). Introduction to active vision: The complexities of continuous visual control. *Journal of Vision*, *13*(9), 1375–1375.

Brenner, E., & Smeets, J. B. (2010). Why we need continuous visual control to intercept a moving target. *Journal of Vision*, *10*(7), 1081–1081.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, *12*(12), 4745–4765.

Bronstein, J. L. (2015). *Mutualism*. Oxford University Press.

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, *340*(6128), 95–98.

Bub, D. N., & Masson, M. E. (2012). On the dynamics of action representations evoked by names of manipulable objects. *Journal of Experimental Psychology: General*, *141*(3), 502.

Burk, D., Ingram, J. N., Franklin, D. W., Shadlen, M. N., & Wolpert, D. M. (2014). Motor effort alters changes of mind in sensorimotor decision making. *PloS One*, *9*(3), e92681.

Calhoun, A. J., Chalasani, S. H., & Sharpee, T. O. (2014). Maximally informative foraging by caenorhabditis elegans. *eLife*, *3*, e04220.

Carls-Diamante, S. (2022). Where is it like to be an octopus? *Frontiers in Systems Neuroscience*, *16*, 840022.

Carlsen, A. N., Chua, R., Summers, J. J., Inglis, J. T., Sanderson, D. J., & Franks, I. M. (2009). Precues enable multiple response preprogramming: Evidence from startle. *Psychophysiology*, *46*(2), 241–251.

Carlsen, A. N., Maslovat, D., Lam, M. Y., Chua, R., & Franks, I. M. (2011). Considerations for the use of a startling acoustic stimulus in studies of motor preparation in humans. *Neuroscience & Biobehavioral Reviews*, *35*(3), 366–376.

Cash-Padgett, T., & Hayden, B. (2020). Behavioural variability contributes to overstaying in patchy foraging. *Biology Letters*, *16*(3), 20190915.

Cazettes, F., Mazzucato, L., Murakami, M., Morais, J. P., Renart, A., & Mainen, Z. F. (2022). A repertoire of foraging decision variables in the mouse brain. *bioRxiv*.

Cazettes, F., Murakami, M., Renart, A., & Mainen, Z. (2021). Reservoir of decision strategies in the mouse brain. *bioRxiv*.

Chapman, C. S., Gallivan, J. P., Culham, J. C., & Goodale, M. A. (2011). Mental blocks: Fmri reveals top-down modulation of early visual cortex when obstacles interfere with grasp planning. *Neuropsychologia*, *49*(7), 1703–1717.

Chapman, C. S., Gallivan, J. P., & Enns, J. T. (2015). Separating value from selection frequency in rapid reaching biases to visual targets. *Visual Cognition*, *23*(1-2), 249–271.

Chapman, C. S., Gallivan, J. P., Wong, J. D., Wispinski, N. J., & Enns, J. T. (2015). The snooze of lose: Rapid reaching reveals that losses are processed more slowly than gains. *Journal of Experimental Psychology: General*, *144*(4), 844.

Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Ansari, D., Culham, J. C., & Goodale, M. A. (2014). Counting on the motor system: Rapid action planning reveals the format-and magnitude-dependent extraction of numerical quantity. *Journal of Vision*, *14*(3), 30–30.

Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Culham, J. C., & Goodale, M. A. (2010a). Reaching for the unknown: Multiple target encoding and real-time decision-making in a rapid reach task. *Cognition*, *116*(2), 168–176.

Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Culham, J. C., & Goodale, M. A. (2010b). Short-term motor plasticity revealed in a visuomotor decision-making task. *Behavioural Brain Research*, *214*(1), 130–134.

Chapman, C. S., & Goodale, M. A. (2008). Missing in action: The effect of obstacle position and size on avoidance while reaching. *Experimental Brain Research*, *191*(1), 83–97.

Chapman, C. S., & Goodale, M. A. (2010a). Obstacle avoidance during online corrections. *Journal of Vision*, *10*(11), 17–17.

Chapman, C. S., & Goodale, M. A. (2010b). Seeing all the obstacles in your way: The effect of visual feedback and visual feedback schedule on obstacle avoidance while reaching. *Experimental Brain Research*, *202*(2), 363–375.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Charnov, E. L., & Orians, G. H. (2006). Optimal foraging: Some theoretical explorations.

Charnov, E. L., & Parker, G. A. (1995). Dimensionless invariants from foraging theory's marginal value theorem. *Proceedings of the National Academy of Sciences*, *92*(5), 1446–1450.

Chen, X., & Stuphorn, V. (2015). Sequential selection of economic good and action in medial frontal cortex of macaques during value-based decisions. *eLife*, *4*, e09418.

Chou, P.-W., Maturana, D., & Scherer, S. (2017). Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution, In *International Conference on Machine Learning*.

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Christopoulos, V., Bonaiuto, J., & Andersen, R. A. (2015). A biologically plausible computational theory for value integration and action selection in decisions with competing alternatives. *PLoS Computational Biology*, *11*(3), e1004104.

Christopoulos, V., & Schrater, P. R. (2015). Dynamic integration of value information into a common probability currency as a theory for flexible decision making. *PLoS Computational Biology*, *11*(9), e1004402.

Churchland, A. K., Kiani, R., Chaudhuri, R., Wang, X.-J., Pouget, A., & Shadlen, M. N. (2011). Variance as a signature of neural computations during decision making. *Neuron*, *69*(4), 818–831.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, *487*(7405), 51–56.

Churchland, P. S., Ramachandran, V. S., & Sejnowski, T. J. (1993). A critique of pure vision.

Cisek, P. (2006). Integrated neural processes for defining potential actions and deciding between them: A computational model. *Journal of Neuroscience*, *26*(38), 9761–9770.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1585–1599.

Cisek, P. (2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology*, *22*(6), 927–936.

Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*, *81*(7), 2265–2287.

Cisek, P., & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron*, *45*(5), 801–814.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, *33*, 269–298.

Cisek, P., & Pastor-Bernier, A. (2014). On the challenges and mechanisms of embodied decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130479.

Cisek, P., Puskas, G. A., & El-Murr, S. (2009). Decisions in changing conditions: The urgency-gating model. *Journal of Neuroscience*, *29*(37), 11560–11571.

Clark, A. (1997). The dynamical challenge. *Cognitive Science*, *21*(4), 461–481.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Coleman, S. L., Brown, V. R., Levine, D. S., & Mellgren, R. L. (2005). A neural network model of foraging decisions made under predation risk. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(4), 434–451.

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(4), 837–853.

Cos, I., Bélanger, N., & Cisek, P. (2011). The influence of predicted arm biomechanics on decision making. *Journal of Neurophysiology*, *105*(6), 3022–3033.

Cos, I., Duque, J., & Cisek, P. (2014). Rapid prediction of biomechanical costs during action decisions. *Journal of Neurophysiology*, *112*(6), 1256–1266.

Cos, I., Medleg, F., & Cisek, P. (2012). The modulatory influence of end-point controllability on decisions between actions. *Journal of Neurophysiology*, *108*(6), 1764–1780.

Costa, V. D., & Averbeck, B. B. (2020). Primate orbitofrontal cortex codes information relevant for managing explore–exploit tradeoffs. *Journal of Neuroscience*, *40*(12), 2553–2561.

Costello, M. G., Zhu, D., Salinas, E., & Stanford, T. R. (2013). Perceptual modulation of motor—but not visual—responses in the frontal eye field during an urgent-decision task. *Journal of Neuroscience*, *33*(41), 16394–16408.

Cowie, R. J. (1977). Optimal foraging in great tits (Parus major). *Nature*, *268*(5616), 137–139.

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., Waldorp, L. J., & Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2), 640–647.

Croxson, P. L., Walton, M. E., O'Reilly, J. X., Behrens, T. E., & Rushworth, M. F. (2009). Effort-based cost–benefit valuation and the human brain. *Journal of Neuroscience*, *29*(14), 4531–4541.

Cultural General Intelligence Team, Bhoopchand, A., Brownfield, B., Collister, A., Lago, A. D., Edwards, A., Everett, R., Frechette, A., Oliveira, Y. G., Hughes, E., Mathewson, K. W., Mendolicchio, P., Pawar, J., Pislar, M., Platonov, A., Senter, E., Singh, S., Zacherl, A., & Zhang, L. M. (2022). Learning robust real-time cultural transmission without human data. *arXiv preprint arXiv.2203.00715*.

Darwin, C. (1872). *The descent of man, and selection in relation to sex* (Vol. 2). D. Appleton.

Davidson, J. D., & El Hady, A. (2019). Foraging as an evidence accumulation process. *PLoS Computational Biology*, *15*(7), e1007060.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

De Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, *7*(3), 242–249.

De Martino, B., Kumaran, D., Holt, B., & Dolan, R. J. (2009). The neurobiology of reference-dependent value computation. *Journal of Neuroscience*, *29*(12), 3833–3842.

Delsuc, F. (2003). Army ants trapped by their evolutionary history. *PLoS Biology*, *1*(2), e37.

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

Dessing, J. C., Bullock, D., Peper, C. L. E., & Beek, P. J. (2002). Prospective control of manual interceptive actions: Comparative simulations of extant and new model constructs. *Neural Networks*, *15*(2), 163–179.

Devlin, S., Georgescu, R., Momennejad, I., Rzepecki, J., Zuniga, E., Costello, G., Leroy, G., Shaw, A., & Hofmann, K. (2021). Navigation turing test (NTT): Learning to evaluate human-like navigation, In *International Conference on Machine Learning*.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081.

Donner, T. H., Siegel, M., Fries, P., & Engel, A. K. (2009). Buildup of choice-predictive activity in human motor cortex during perceptual decision making. *Current Biology*, *19*(18), 1581–1585.

Dorris, M. C., & Glimcher, P. W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron*, *44*(2), 365–378.

Doyle, M., & Walker, R. (2001). Curved saccade trajectories: Voluntary and reflexive saccades curve away from irrelevant distractors. *Experimental Brain Research*, *139*(3), 333–344.

Driscoll, L., Shenoy, K., & Sussillo, D. (2022). Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *bioRxiv*, 2022–08.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*(11), 3612–3628.

Dshemuchadse, M., Scherbaum, S., & Goschke, T. (2013). How decisions emerge: Action dynamics in intertemporal decision making. *Journal of Experimental Psychology: General*, *142*(1), 93.

Dulac-Arnold, G., Mankowitz, D., & Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, *6*(5), 178–190.

Enderle, J. D., & Sierra, D. A. (2013). A new linear muscle fiber model for neural control of saccades. *International Journal of Neural Systems*, *23*(02), 1350002.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303.

Fajen, B. R., & Warren, W. H. (2003). Behavioral dynamics of steering, obstable avoidance, and route selection. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 343.

Faulkenberry, T. J., Cruise, A., Lavro, D., & Shaki, S. (2016). Response trajectories capture the continuous dynamics of the size congruity effect. *Acta Psychologica*, *163*, 114–123.

Feulner, B., Perich, M. G., Miller, L. E., Clopath, C., & Gallego, J. A. (2022). Feedback-based motor control can guide plasticity and drive rapid learning. *bioRxiv*.

Finding Nemo. (2003). Pixar Animation Studios.

Fitts, P. M. (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Psychology*, *71*(6), 849.

Fooken, J., Yeo, S.-H., Pai, D. K., & Spering, M. (2016). Eye movement accuracy determines natural interception strategies. *Journal of Vision*, *16*(14), 1–1.

Forgaard, C. J., Maslovat, D., Carlsen, A. N., & Franks, I. M. (2011). Default motor preparation under conditions of response uncertainty. *Experimental Brain Research*, *215*, 235–245.

Frankenhuis, W. E., Panchanathan, K., & Barto, A. G. (2019). Enriching behavioral ecology with reinforcement learning methods. *Behavioural Processes*, *161*, 94–100.

Freedman, D. J., & Assad, J. A. (2011). A proposed common neural mechanism for categorization and perceptual decisions. *Nature Neuroscience*, *14*(2), 143–146.

Freeman, J., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 59.

Friedman, J., Brown, S., & Finkbeiner, M. (2013). Linking cognitive and reaching trajectories via intermittent movement control. *Journal of Mathematical Psychology*, *57*(3-4), 140–151.

Gallivan, J. P., Barton, K. S., Chapman, C. S., Wolpert, D. M., & Randall Flanagan, J. (2015). Action plan co-optimization reveals the parallel encoding of competing reach movements. *Nature Communications*, *6*(1), 7428.

Gallivan, J. P., & Chapman, C. S. (2014). Three-dimensional reach trajectories as a probe of real-time decision-making between multiple competing targets. *Frontiers in Neuroscience*, *8*, 215.

Gallivan, J. P., Chapman, C. S., Gale, D. J., Flanagan, J. R., & Culham, J. C. (2019). Selective modulation of early visual cortical activity by movement intention. *Cerebral Cortex*, *29*(11), 4662–4678.

Gallivan, J. P., Chapman, C. S., Wolpert, D. M., & Flanagan, J. R. (2018). Decision-making in sensorimotor control. *Nature Reviews Neuroscience*, *19*(9), 519–534.

Gallivan, J. P., Chapman, C. S., Wood, D. K., Milne, J. L., Ansari, D., Culham, J. C., & Goodale, M. A. (2011). One to four, and nothing more: Nonconscious parallel individuation of objects during action planning. *Psychological Science*, *22*(6), 803–811.

Gallivan, J. P., Logan, L., Wolpert, D. M., & Flanagan, J. R. (2016). Parallel specification of competing sensorimotor control policies for alternative action options. *Nature Neuroscience*, *19*(2), 320.

Gallivan, J. P., McLean, D. A., Valyear, K. F., Pettypiece, C. E., & Culham, J. C. (2011). Decoding action intentions from preparatory brain activity in human parieto-frontal networks. *Journal of Neuroscience*, *31*(26), 9599–9610.

Gallivan, J. P., Stewart, B. M., Baugh, L. A., Wolpert, D. M., & Flanagan, J. R. (2017). Rapid automatic motor encoding of competing reach options. *Cell Reports*, *18*(7), 1619–1626.

Gellman, R., & Carl, J. (1991). Motion processing for saccadic eye movements in humans. *Experimental Brain Research*, *84*(3), 660–667.

Ghez, C., Favilla, M., Ghilardi, M., Gordon, J., Bermejo, R., & Pullman, S. (1997). Discrete and continuous planning of hand movements and isometric force trajectories. *Experimental Brain Research*, *115*, 217–233.

Ghez, C., Gordon, J., Ghilardi, M., Christakos, C., & Cooper, S. (1990). Roles of proprioceptive input in the programming of arm trajectories, In *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Psychology Press.

Glimcher, P. W. (2010). *Foundations of neuroeconomic analysis*. Oxford University Press.

Godfrey-Smith, P. (2016). *Other minds: The octopus, the sea, and the deep origins of consciousness*. William Collins.

Gold, J. I., & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, *404*(6776), 390–394.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Goldshtein, A., Handel, M., Eitan, O., Bonstein, A., Shaler, T., Collet, S., Greif, S., Medellín, R. A., Emek, Y., Korman, A., Et al. (2020). Reinforcement learning enables resource partitioning in foraging bats. *Current Biology*, *30*(20), 4096–4102.

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

Grattan, L. E., & Glimcher, P. W. (2014). Absence of spatial tuning in the orbitofrontal cortex. *PLoS One*, *9*(11), e112750.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Gupta, D., DePasquale, B., Kopec, C. D., & Brody, C. D. (2023). Trial-history biases in evidence accumulation can give rise to apparent lapses. *bioRxiv*, 2023–01.

Haarnoja, T., Moran, B., Lever, G., Huang, S. H., Tirumala, D., Wulfmeier, M., Humplik, J., Tunyasuvunakool, S., Siegel, N. Y., Hafner, R., Et al. (2023). Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *arXiv preprint arXiv:2304.13653*.

Hafner, D., Lillicrap, T. P., Ba, J., & Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hagura, N., Haggard, P., & Diedrichsen, J. (2017). Perceptual decisions are biased by the cost to act. *eLife*, *6*, e18422.

Haith, A. M., Huberdeau, D. M., & Krakauer, J. W. (2015). Hedging your bets: Intermediate movements as optimal behavior in the context of an incomplete decision. *PLoS Computational Biology*, *11*(3), e1004171.

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience*, *9*(5), 682–689.

Hanks, T. D., & Summerfield, C. (2017). Perceptual decision making in rodents, monkeys, and humans. *Neuron*, *93*(1), 15–31.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *Journal of Neuroscience*, *35*(6), 2476–2484.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*(7), 933–939.

Heess, N., Tb, D., Sriram, S., Lemmon, J., Merel, J., Wayne, G., Tassa, Y., Erez, T., Wang, Z., Eslami, S., Et al. (2017). Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150.

Hikosaka, O., Nakamura, K., & Nakahara, H. (2006). Basal ganglia orient eyes to reward. *Journal of Neurophysiology*, *95*(2), 567–584.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics*, *81*(7), 2288–2303.

Horowitz, A. C. (2002). *The behaviors of theories of mind, and a case study of dogs at play*. University of California, San Diego.

Horwitz, G. D., & Newsome, W. T. (1999). Separate signals for target selection and movement specification in the superior colliculus. *Science*, *284*(5417), 1158–1161.

Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, *436*(7047), 71–77.

Huang, S., Dossa, R. F. J., Ye, C., Braga, J., Chakraborty, D., Mehta, K., & Araújo, J. G. (2022). Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *The Journal of Machine Learning Research*, *23*(1), 12585–12602.

Hudson, T. E., Maloney, L. T., & Landy, M. S. (2007). Movement planning with probabilistic target information. *Journal of Neurophysiology*, *98*(5), 3034–3046.

Huk, A. C., & Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *Journal of Neuroscience*, *25*(45), 10420–10436.

Hutchinson, J. M., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: Can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour*, *75*(4), 1331–1349.

Ikeda, T., & Hikosaka, O. (2003). Reward-dependent gain and bias of visual responses in primate superior colliculus. *Neuron*, *39*(4), 693–700.

Jacob, G., Pramod, R., Katti, H., & Arun, S. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, *12*(1), 1872.

James, W. (1960). *The principles of psychology* (Vol. 1). Henry Holt and Company.

Johanson, M. B., Hughes, E., Timbers, F., & Leibo, J. Z. (2022). Emergent bartering behaviour in multi-agent reinforcement learning. *arXiv preprint arXiv:2205.06760*.

Johnson, H., Van Beers, R. J., & Haggard, P. (2002). Action and awareness in pointing tasks. *Experimental Brain Research*, *146*(4), 451–459.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633.

Kacelnik, A. (1984). Central place foraging in starlings (Sturnus vulgaris). I. Patch residence time. *The Journal of Animal Ecology*, 283–299.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kane, G. A., James, M. H., Shenhav, A., Daw, N. D., Cohen, J. D., & Aston-Jones, G. (2021). Rat anterior cingulate cortex continuously signals decision variables in a patch foraging task. *bioRxiv*.

Kane, G. A., Vazey, E. M., Wilson, R. C., Shenhav, A., Daw, N. D., Aston-Jones, G., & Cohen, J. D. (2017). Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *17*(6), 1073–1083.

Kang, Y. H., Petzschner, F. H., Wolpert, D. M., & Shadlen, M. N. (2017). Piercing of consciousness as a threshold-crossing operation. *Current Biology*, *27*(15), 2285–2295.

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254.

Katsumata, H., & Russell, D. M. (2012). Prospective versus predictive control in timing of hitting a falling ball. *Experimental Brain Research*, *216*(4), 499–514.

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*(7611), 285–288.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2015). Vacillation, indecision and hesitation in moment-by-moment decoding of monkey motor cortex. *eLife*, *4*, e04677.

Kauvar, I., Doyle, C., Zhou, L., & Haber, N. (2023). Curious replay for model-based adaptation, In *International Conference on Machine Learning*.

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132.

Keller, E. L., & McPeek, R. M. (2002). Neural discharge in the superior colliculus during target search paradigms. *Annals of the New York Academy of Sciences*, *956*(1), 130–142.

Kennerley, S. W., & Wallis, J. D. (2009). Encoding of reward and space during a working memory task in the orbitofrontal cortex and anterior cingulate sulcus. *Journal of Neurophysiology*.

Kepecs, A., & Mainen, Z. F. (2014). A computational framework for the study of confidence across species. *The Cognitive Neuroscience of Metacognition*, 115–145.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Kiani, R., Churchland, A. K., & Shadlen, M. N. (2013). Integration of direction cues is invariant to the temporal gap between them. *Journal of Neuroscience*, *33*(42), 16483–16489.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342.

Kiani, R., Cueva, C. J., Reppas, J. B., & Newsome, W. T. (2014). Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. *Current Biology*, *24*(13), 1542–1547.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017). Deep neural networks in computational neuroscience. *bioRxiv*, 133504.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Kilpatrick, Z. P., Davidson, J. D., & El Hady, A. (2021). Uncertainty drives deviations in normative foraging decision strategies. *Journal of the Royal Society Interface*, *18*(180), 20210337.

Kim, B., & Basso, M. A. (2008). Saccade target selection in the superior colliculus: A signal detection theory approach. *Journal of Neuroscience*, *28*(12), 2991–3007.

Kim, J.-N., & Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, *2*(2), 176–185.

Kim, S., Hwang, J., & Lee, D. (2008). Prefrontal coding of temporally discounted values during intertemporal choice. *Neuron*, *59*(1), 161–172.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Et al. (2017). Over-

coming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526.

Klaes, C., Westendorff, S., Chakrabarti, S., & Gail, A. (2011). Choosing goals, not rules: Deciding among rule-based action plans. *Neuron*, *70*(3), 536–548.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?

Klein-Flügge, M. C., & Bestmann, S. (2012). Time-dependent changes in human corticospinal excitability reveal value-based competition for action during decision processing. *Journal of Neuroscience*, *32*(24), 8373–8382.

Knights, E., Bultitude, J., & Rossit, S. (2015). Prism adaptation effects are not limited to dorsal visual processing: Evidence from pro-pointing and anti-pointing, In *British association for cognitive neuroscience*.

Kolling, N., & Akam, T. (2017). (Reinforcement?) Learning to forage optimally. *Current Opinion in Neurobiology*, *46*, 162–169.

Kolling, N., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2012). Neural mechanisms of foraging. *Science*, *336*(6077), 95–98.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857.

Krakauer, J. W., Hadjiosif, A. M., Xu, J., Wong, A. L., & Haith, A. M. (2019). Motor learning. *Comprehensive Physiology*, *9*(2), 613–663.

Krebs, J. R., Ryan, J. C., & Charnov, E. L. (1974). Hunting by expectation or optimal foraging? A study of patch use by chickadees. *Animal Behaviour*, *22*, 953–IN3.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Krotkov, E., Hackett, D., Jackel, L., Perschbacher, M., Pippine, J., Strauss, J., Pratt, G., & Orlowski, C. (2018). The darpa robotics challenge finals: Results and perspectives. *The DARPA robotics challenge finals: Humanoid robots to the rescue*, 1–26.

Krueger, P. M., van Vugt, M. K., Simen, P., Nystrom, L., Holmes, P., & Cohen, J. D. (2017). Evidence accumulation detected in bold signal using slow perceptual decision making. *Journal of Neuroscience Methods*, *281*, 21–32.

Lau, B., & Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, *58*(3), 451–463.

Leibo, J. Z., Hughes, E., Lanctot, M., & Graepel, T. (2019). Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*.

Lepora, N. F., & Pezzulo, G. (2015). Embodied choice: How action influences perceptual decision making. *PLoS Computational Biology*, *11*(4), e1004110.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038.

Liang, Y., Machado, M. C., Talvitie, E., & Bowling, M. (2015). State of the art control of atari games using shallow reinforcement learning. *arXiv preprint arXiv:1512.01563*.

Licata, A. M., Kaufman, M. T., Raposo, D., Ryan, M. B., Sheppard, J. P., & Churchland, A. K. (2017). Posterior parietal cortex guides visual decisions in rats. *Journal of Neuroscience*, *37*(19), 4954–4966.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, *7*(1), 13276.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lin, L. J. (1991). Self-improvement based on reinforcement learning, planning and teaching, In *International Conference on Machine Learning*.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031.

Link, S. W. (1975). The relative judgment theory of two choice response time. *Journal of Mathematical Psychology*, *12*(1), 114–135.

Lorenz, K., & Tinbergen, N. (1938). Taxis und instinkthandlung in der eirollbewegung der graugans. *Zeitschrift für Tierpsychologie*.

Lottem, E., Banerjee, D., Vertechi, P., Sarra, D., oude Lohuis, M. N., & Mainen, Z. F. (2018). Activation of serotonin neurons promotes active persistence in a probabilistic foraging task. *Nature Communications*, *9*(1), 1–12.

Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward value-based gain control: Divisive normalization in parietal cortex. *Journal of Neuroscience*, *31*(29), 10627–10639.

Ludwig, C. J., & Gilchrist, I. D. (2003). Target similarity affects saccade curvature away from irrelevant onsets. *Experimental Brain Research*, *152*, 60–69.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

Malavazi, F. B., Guyonneau, R., Fasquel, J.-B., Lagrange, S., & Mercier, F. (2018). Lidar-only based navigation algorithm for an autonomous agricultural robot. *Computers and Electronics in Agriculture*, *154*, 71–79.

Manohar, S. G., Chong, T. T.-J., Apps, M. A., Batla, A., Stamelou, M., Jarman, P. R., Bhatia, K. P., & Husain, M. (2015). Reward pays the cost of noise reduction in motor and cognitive control. *Current Biology*, *25*(13), 1707–1716.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 94.

Matsumoto, K., Suzuki, W., & Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, *301*(5630), 229–232.

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press.

Mazyn, L. I., Savelsbergh, G. J., Montagne, G., & Lenoir, M. (2007). Planning and on-line control of catching as a function of perceptual-motor constraints. *Acta Psychologica*, *126*(1), 59–78.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*(4), 287.

McDonnell, M. D., & Ward, L. M. (2011). The benefits of noise in neural systems: Bridging theory and experiment. *Nature Reviews Neuroscience*, *12*(7), 415–425.

McNamee, D. C., Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2021). Flexible modulation of sequence generation in the entorhinal–hippocampal system. *Nature Neuroscience*, *24*(6), 851–862.

McPeek, R. M., Han, J. H., & Keller, E. L. (2003). Competition between saccade goals in the superior colliculus produces saccade curvature. *Journal of Neurophysiology*, *89*(5), 2577–2590.

Megaw, E. (1974). Possible modification to a rapid on-going programmed manual response. *Brain Research*, *71*(2-3), 425–441.

Meyer, D. E., Irwin, D. E., Osman, A. M., & Kounois, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. *Psychological Review*, *95*(2), 183.

Michaels, J. A., Schaffelhofer, S., Agudelo-Toro, A., & Scherberger, H. (2020). A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *Proceedings of the National Academy of Sciences*, *117*(50), 32124–32135.

Michalski, J., Green, A. M., & Cisek, P. (2020). Reaching decisions during ongoing movements. *Journal of Neurophysiology*, *123*(3), 1090–1102.

Milani, S., Juliani, A., Momennejad, I., Georgescu, R., Rzepecki, J., Shaw, A., Costello, G., Fang, F., Devlin, S., & Hofmann, K. (2023). Navigates like me: Understanding how people evaluate human-like AI in video games, In *Conference on Human Factors in Computing Systems*.

Miller, G. A., Eugene, G., & Pribram, K. H. (1960). *Plans and the structure of behaviour*. Henry Holt and Company.

Miller, K. J., Eckstein, M., Botvinick, M. M., & Kurth-Nelson, Z. (2023). Cognitive model discovery via disentangled RNNs. *bioRxiv*, 2023–06.

Miller, M. L., Ringelman, K. M., Eadie, J. M., & Schank, J. C. (2017). Time to fly: A comparison of marginal value theorem approximations in an agent-based model of foraging waterfowl. *Ecological Modelling*, *351*, 77–86.

Milne, J. L., Chapman, C. S., Gallivan, J. P., Wood, D. K., Culham, J. C., & Goodale, M. A. (2013). Connecting the dots: Object connectedness deceives perception but not movement planning. *Psychological Science*, *24*(8), 1456–1465.

Minsky, M. (1988). *Society of mind*. Simon and Schuster.

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

Moher, J., & Song, J.-H. (2014). Perceptual decision processes flexibly adapt to avoid change-of-mind motor costs. *Journal of Vision*, *14*(8), 1–1.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, *377*(6551), 725–728.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

Moore, J., & Newell, A. (1974). How can MERLIN understand? Hillsdale, NJ: Erlbaum Assoc.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Morel, P., Ulbrich, P., & Gail, A. (2017). What makes a reach movement effortful? Physical effort discounting supports common minimization principles in decision making and motor control. *PLoS Biology*, *15*(6), e2001323.

Morimoto, J. (2019). Foraging decisions as multi-armed bandit problems: Applying reinforcement learning algorithms to foraging data. *Journal of Theoretical Biology*, *467*, 48–56.

Munoz, D. P., & Wurtz, R. H. (1995). Saccade-related activity in monkey superior colliculus. I. Characteristics of burst and buildup cells. *Journal of Neurophysiology*, *73*(6), 2313–2333.

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife*, *4*, e11946.

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, *22*(10), 1677–1686.

Mysore, S. P., & Knudsen, E. I. (2011). The role of a midbrain network in competitive stimulus selection. *Current Opinion in Neurobiology*, *21*(4), 653–660.

Nagel, T. (1980). What is it like to be a bat?, In *The language and thought series*. Harvard University Press.

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., & Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, *14*(7), 2152–2176.

Newsome, W. T., & Paré, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience*, *8*(6), 2201–2211.

Nicholson, D. A., & Prinz, A. A. (2021). Deep neural network models of object recognition exhibit human-like limitations when performing visual search tasks. *bioRxiv*, 2020–10.

Nijhawan, R. (2002). Neural delays, visual motion and the flash-lag effect. *Trends in Cognitive Sciences*, *6*(9), 387–393.

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior*.

Nonacs, P. (2001). State dependent behavior and the marginal value theorem. *Behavioral Ecology*, *12*(1), 71–83.

Noorani, I., & Carpenter, R. (2016). The LATER model of reaction time and decision. *Neuroscience & Biobehavioral Reviews*, *64*, 229–251.

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729–1735.

O'Connell, R. G., Shadlen, M. N., Wong-Lin, K., & Kelly, S. P. (2018). Bridging neural and computational viewpoints on perceptual decision-making. *Trends in Neurosciences*, *41*(11), 838–852.

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.

Open Ended Learning Team, Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., Et al. (2021). Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973.

Pacheco-Cobos, L., Winterhalder, B., Cuatianquiz-Lima, C., Rosetti, M. F., Hudson, R., & Ross, C. T. (2019). Nahua mushroom gatherers use area-restricted search strategies that conform to marginal value theorem predictions. *Proceedings of the National Academy of Sciences*, *116*(21), 10339–10347.

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: A good-based model. *Annual Review of Neuroscience*, *34*, 333–359.

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223–226.

Padoa-Schioppa, C., & Conen, K. E. (2017). Orbitofrontal cortex: A neural circuit for economic decisions. *Neuron*, *96*(4), 736–754.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 1–1.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, *113*, 54–71.

Pastor-Bernier, A., & Cisek, P. (2011). Neural correlates of biased competition in premotor cortex. *Journal of Neuroscience*, *31*(19), 7083–7088.

Pastor-Bernier, A., Tremblay, E., & Cisek, P. (2012). Dorsal premotor cortex is involved in switching motor plans. *Frontiers in Neuroengineering*, *5*, 5.

Patterson, A., Neumann, S., White, M., & White, A. (2023). Empirical design in reinforcement learning. *arXiv preprint arXiv:2304.01315*.

Pearce, T. M., & Moran, D. W. (2012). Strategy-dependent encoding of planned arm movements in the dorsal premotor cortex. *Science*, *337*(6097), 984–988.

Pearson, J. M., Hayden, B. Y., Raghavachari, S., & Platt, M. L. (2009). Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. *Current Biology*, *19*(18), 1532–1537.

Pearson, J. M., Watson, K. K., & Platt, M. L. (2014). Decision making: The neuroethological turn. *Neuron*, *82*(5), 950–965.

Peixoto, D., Kiani, R., Chandrasekaran, C., Ryu, S. I., Shenoy, K. V., & Newsome, W. T. (2018). Population dynamics of choice representation in dorsal premotor and primary motor cortex. *bioRxiv*, 283960.

Peixoto, D., Verhein, J. R., Kiani, R., Kao, J. C., Nuyujukian, P., Chandrasekaran, C., Brown, J., Fong, S., Ryu, S. I., Shenoy, K. V., & Newsome, W. T. (2021). Decoding and perturbing decision states in real time. *Nature*, *591*(7851), 604–609.

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

Pesquita, A., Chapman, C. S., & Enns, J. T. (2016). Humans are sensitive to attention control when predicting others' actions. *Proceedings of the National Academy of Sciences*, *113*(31), 8669–8674.

Peters, M. A. (2022). Confidence in decision-making, In *Oxford research encyclopedia of neuroscience*.

Pezzulo, G., & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences*, *20*(6), 414–424.

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience*, *27*(37), 9984–9988.

Platanios, E. A., Saparov, A., & Mitchell, T. (2020). Jelly bean world: A testbed for never-ending learning. *arXiv preprint arXiv:2002.06306*.

Platt, M. L., & Glimcher, P. W. (1997). Responses of intraparietal neurons to saccadic targets and visual distractors. *Journal of Neurophysiology*, *78*(3), 1574–1589.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238.

Pleasants, J. M. (1989). Optimal foraging by nectarivores: A test of the marginal-value theorem. *The American Naturalist*, *134*(1), 51–71.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864.

Port, N. L., & Wurtz, R. H. (2003). Sequential activity of simultaneously recorded neurons in the superior colliculus during curved saccades. *Journal of Neurophysiology*, *90*(3), 1887–1903.

Precup, D., & Sutton, R. S. (1997). Multi-time models for temporally abstract planning. *Advances in Neural Information Processing Systems*, *10*.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*, e223.

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, *20*(2), 262–270.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266.

Rideaux, R., & Welchman, A. E. (2020). But still it moves: Static image statistics underlie how we see motion. *Journal of Neuroscience*, *40*(12), 2538–2552.

Ringstrom, T. J. (2022). Reward is not necessary: How to create a compositional self-preserving agent for life-long learning. *arXiv preprint arXiv:2211.10851*.

Rizzolatti, G., Fogassi, L., & Gallese, V. (1997). Parietal cortex: From sight to action. *Current Opinion in Neurobiology*, *7*(4), 562–567.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, *22*(21), 9475–9489.

Rorie, A. E., & Newsome, W. T. (2005). A general mechanism for decision-making in the human brain? *Trends in Cognitive Sciences*, *9*(2), 41–43.

Rosenbaum, D. A., & Kornblum, S. (1982). A priming method for investigating the selection of motor responses. *Acta Psychologica*, *51*(3), 223–243.

Rushworth, M., Walton, M. E., Kennerley, S. W., & Bannerman, D. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*(9), 410–417.

Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach. *University of California, Berkeley*.

Russo, A. A., Bittner, S. R., Perkins, S. M., Seely, J. S., London, B. M., Lara, A. H., Miri, A., Marshall, N. J., Kohn, A., Jessell, T. M., Et al. (2018). Motor cortex embeds muscle-like commands in an untangled population response. *Neuron*, *97*(4), 953–966.

Salzman, C. D., Murasugi, C. M., Britten, K. H., & Newsome, W. T. (1992). Microstimulation in visual area MT: Effects on direction discrimination performance. *Journal of Neuroscience*, *12*(6), 2331–2355.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337–1340.

Samvelyan, M., Khan, A., Dennis, M., Jiang, M., Parker-Holder, J., Foerster, J., Raileanu, R., & Rocktäschel, T. (2023). MAESTRO: Open-ended environment design for multi-agent reinforcement learning. *arXiv preprint arXiv:2303.03376*.

Sarlegna, F. R., & Mutha, P. K. (2015). The influence of visual target information on the online control of movements. *Vision Research*, *110*, 144–154.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Schaffelhofer, S., Sartori, M., Scherberger, H., & Farina, D. (2014). Musculoskeletal representation of a large repertoire of hand grasping actions in primates. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *23*(2), 210–220.

Scherbaum, S., Dshemuchadse, M., Fischer, R., & Goschke, T. (2010). How decisions evolve: The temporal dynamics of action selection. *Cognition*, *115*(3), 407–416.

Schneider, D. M., & Mooney, R. (2018). How movement modulates hearing. *Annual Review of Neuroscience*, *41*, 553–572.

Schoemaker, P. J. (1991). The quest for optimality: A positive heuristic of science? *Behavioral and Brain Sciences*, *14*(2), 205–215.

Scholte, H. S., Losch, M. M., Ramakrishnan, K., de Haan, E. H., & Bohte, S. M. (2018). Visual pathways from the perspective of cost functions and multi-task deep neural networks. *Cortex*, *98*, 249–261.

Schouten, J., & Bekker, J. (1967). Reaction time and accuracy. *Acta Psychologica*, *27*, 143–153.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Schwartz, A. (1993). A reinforcement learning method for maximizing undiscounted rewards, In *International Conference on Machine Learning*.

Scialfa, C. T. (2002). The role of sensory factors in cognitive aging research. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *56*(3), 153.

Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience*, *5*(7), 532–545.

Selen, L. P., Shadlen, M. N., & Wolpert, D. M. (2012). Deliberation in the motor system: Reflex gains track evolving evidence leading to a decision. *Journal of Neuroscience*, *32*(7), 2276–2286.

Serences, J. T. (2008). Value-based modulations in human visual cortex. *Neuron*, *60*(6), 1169–1181.

Seth, A., Hicks, J. L., Uchida, T. K., Habib, A., Dembia, C. L., Dunne, J. J., Ong, C. F., DeMers, M. S., Rajagopal, A., Millard, M., Et al. (2018). OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLoS Computational Biology*, *14*(7), e1006223.

Shadlen, M. N., Hanks, T. D., Churchland, A. K., Kiani, R., & Yang, T. (2006). The speed and accuracy of a simple perceptual decision: A mathematical primer. *Bayesian brain: Probabilistic approaches to neural coding*, 209–37.

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791–806.

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: Seeing and deciding. *Proceedings of the National Academy of Sciences*, *93*(2), 628–633.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939.

Shadmehr, R., Huang, H. J., & Ahmed, A. A. (2016). A representation of effort in decision-making and motor control. *Current Biology*, *26*(14), 1929–1934.

Sharp, R., & Whiting, H. (1975). Information-processing and eye movement behaviour in a ball catching skill. *Journal of Human Movement Studies*.

Shenhav, A., Straccia, M. A., Botvinick, M. M., & Cohen, J. D. (2016). Dorsal anterior cingulate and ventromedial prefrontal cortex have inverse roles in both foraging and economic choice. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(6), 1127–1139.

Shuler, M. G., & Bear, M. F. (2006). Reward timing in the primary visual cortex. *Science*, *311*(5767), 1606–1609.

Shuvaev, S., Starosta, S., Kvitsiani, D., Kepecs, A., & Koulakov, A. (2020). R-learning in actor-critic model offers a biologically relevant mechanism for sequential decision-making. *Advances in Neural Information Processing Systems*.

Siegel, M., Buschman, T. J., & Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science*, *348*(6241), 1352–1355.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535.

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J., Pitchford, J. W., James, A., Ahmed, M. Z., Brierley, A. S., Hindell, M. A., Et al. (2008). Scaling laws of marine predator search behaviour. *Nature*, *451*(7182), 1098–1102.

Singh, S. P. (1992). Reinforcement learning with a hierarchy of abstract models, In *Proceedings of the National Conference on Artificial Intelligence*.

Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life*, *11*(1-2), 13–29.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, *32*(2), 135–168.

Soechting, J. F., Juveli, J. Z., & Rao, H. M. (2009). Models for the extrapolation of target motion for manual interception. *Journal of Neurophysiology*, *102*(3), 1491–1502.

Sohn, J.-W., & Lee, D. (2007). Order-dependent modulation of directional signals in the supplementary and presupplementary motor areas. *Journal of Neuroscience*, *27*(50), 13655–13666.

Song, H. F., Yang, G. R., & Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, *6*, e21492.

Song, J.-H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, *13*(8), 360–366.

Sontag, E. D., & Siegelmann, H. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences*, *50*, 132–150.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*(5), 207–211.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398.

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Computational Biology*, *16*(10), e1008215.

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, *48*(12), 1391–1408.

Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing*, *17*(3), 279–305.

Stănişor, L., van der Togt, C., Pennartz, C. M., & Roelfsema, P. R. (2013). A unified selection signal for attention and reward in primary visual cortex. *Proceedings of the National Academy of Sciences*, *110*(22), 9136–9141.

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., Et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, *372*(6539), eabf4588.

Stephens, D. W., & Krebs, J. R. (2019). *Foraging theory*. Princeton University Press.

Stewart, B. M., Gallivan, J. P., Baugh, L. A., & Flanagan, J. R. (2014). Motor, not visual, encoding of potential reach targets. *Current Biology*, *24*(19), R953–R954.

Stine, G. M., Zylberberg, A., Ditterich, J., & Shadlen, M. N. (2020). Differentiating between integration and non-integration strategies in perceptual decision making. *eLife*, *9*, e55365.

Stone, S. A. (2023). Eye and body tracking in the lab, in the wild, and in the clinic.

Striemer, C. L., Chapman, C. S., & Goodale, M. A. (2009). "Real-time" obstacle avoidance in the absence of primary visual cortex. *Proceedings of the National Academy of Sciences*, *106*(37), 15996–16001.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*(5), 363–375.

Sullivan, N., Hutcherson, C., Harris, A., & Rangel, A. (2015). Dietary self-control is related to the speed with which attributes of healthfulness and tastiness are processed. *Psychological Science*, *26*(2), 122–134.

Summerfield, C. (2022). *Natural general intelligence: How understanding the brain can help us build AI*. Oxford University Press.

Summerfield, C., & Parpart, P. (2022). Normative principles for decision-making in natural environments. *Annual Review of Psychology*, *73*, 53–77.

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, *18*(7), 1025–1033.

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

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., & Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction, In *International Conference on Autonomous Agents and Multiagent Systems*.

Tang, W., & Bennett, D. A. (2010). Agent-based modeling of animal movement: A review. *Geography Compass*, *4*(7), 682–700.

Tello-Ramos, M. C., Hurly, T. A., & Healy, S. D. (2015). Traplining in hummingbirds: Flying short-distance sequences among several locations. *Behavioral Ecology*, *26*(3), 812–819.

Thevarajah, D., Mikulić, A., & Dorris, M. C. (2009). Role of the superior colliculus in choosing mixed-strategy saccades. *Journal of Neuroscience*, *29*(7), 1998–2008.

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: Theory and experimental support. *Journal of Neurophysiology*, *108*(11), 2912–2930.

Thura, D., & Cisek, P. (2014). Deliberation and commitment in the premotor and primary motor cortex during dynamic decision making. *Neuron*, *81*(6), 1401–1416.

Thura, D., & Cisek, P. (2016). On the difference between evidence accumulator models and the urgency gating model. *Journal of Neurophysiology*, *115*(1), 622–623.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*(9), 907–915.

Todorov, E., Erez, T., & Tassa, Y. (2012). MuJoCo: A physics engine for model-based control, In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, *5*(11), 1226–1235.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, *12*(8), 291–297.

Tunyasuvunakool, S., Muldal, A., Doron, Y., Liu, S., Bohez, S., Merel, J., Erez, T., Lillicrap, T. P., Heess, N., & Tassa, Y. (2020). dm_control: Software and tasks for continuous control. *Software Impacts*, *6*, 100022.

Turing, A. M. (1950). Mind. *Mind*, *59*(236), 433–460.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Tyldesley, D., & Whiting, H. (1975). Operational timing. *Journal of Human Movement Studies*.

Urai, A. E., De Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *eLife*, *8*, e46331.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, *5*, e12192.

van der Wel, R. P., Eder, J. R., Mitchel, A. D., Walsh, M. M., & Rosenbaum, D. A. (2009). Trajectories emerging from discrete versus continuous processing models in phonological competitor tasks: A commentary on Spivey, Grosjean, and Knoblich. *Journal of Experimental Psychology: Human Perception and Performance*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vaziri-Pashkam, M., Cormiea, S., & Nakayama, K. (2017). Predicting actions from subtle preparatory movements. *Cognition*, *168*, 65–75.

Verneau, M., van der Kamp, J., de Looze, M. P., & Savelsbergh, G. J. (2016). Age effects on voluntary and automatic adjustments in anti-pointing tasks. *Experimental Brain Research*, *234*(2), 419–428.

Vertechi, P., Lottem, E., Sarra, D., Godinho, B., Treves, I., Quendera, T., oude Lohuis, M. N., & Mainen, Z. F. (2020). Inference-based decisions in a hidden state foraging task: Differential contributions of prefrontal cortical areas. *Neuron*, *106*(1), 166–176.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., & Stanley, H. E. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, *381*(6581), 413–415.

Von Helmholtz, H. (2013). *Treatise on physiological optics* (Vol. 3). Courier Corporation.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Waskom, M. L., Asfour, J., & Kiani, R. (2018). Perceptual insensitivity to higher-order statistical moments of coherent random dot motion. *Journal of Vision*, *18*(6), 9–9.

Welsh, T. N., & Elliott, D. (2004). Movement trajectories in the presence of a distracting stimulus: Evidence for a response activation model of selective reaching. *The Quarterly Journal of Experimental Psychology Section A*, *57*(6), 1031–1057.

Welsh, T. N., Elliott, D., & Weeks, D. J. (1999). Hand deviations toward distractors evidence for response competition: Evidence for response competition. *Experimental Brain Research*, *127*, 207–212.

Weng, T.-W., Dvijotham, K. D., Uesato, J., Xiao, K., Gowal, S., Stanforth, R., & Kohli, P. (2019). Toward evaluating robustness of deep reinforcement learning with continuous control, In *International Conference on Learning Representations*.

White, A. (2015). Developing a predictive approach to knowledge.

Wispinski, N. J. (2017). Modelling movement as an ongoing decision.

Wispinski, N. J., Butcher, A., Mathewson, K. W., Chapman, C. S., Botvinick, M. M., & Pilarski, P. M. (2022). Adaptive patch foraging in deep reinforcement learning agents. *Transactions on Machine Learning Research*.

Wispinski, N. J., Gallivan, J. P., & Chapman, C. S. (2020). Models, movements, and minds: Bridging the gap between decision making and action. *Annals of the New York Academy of Sciences*, *1464*(1), 30–51.

Wispinski, N. J., Truong, G., Handy, T. C., & Chapman, C. S. (2017). Reaching reveals that best-versus-rest processing contributes to biased decision making. *Acta Psychologica*, *176*, 32–38.

Wolfe, J. M., Cain, M. S., & Alaoui-Soce, A. (2018). Hybrid value foraging: How the value of targets shapes human foraging behavior. *Attention, Perception, & Psychophysics*, *80*(3), 609–621.

Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences*, *5*(11), 487–494.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

Wong, A. L., Goldsmith, J., Forrence, A. D., Haith, A. M., & Krakauer, J. W. (2017). Reaction times can reflect habits rather than computations. *eLife*, *6*, e28075.

Wong, A. L., & Haith, A. M. (2017). Motor planning flexibly optimizes performance under uncertainty about task goals. *Nature Communications*, *8*(1), 14624.

Wong, A. L., Haith, A. M., & Krakauer, J. W. (2015). Motor planning. *The Neuroscientist*, *21*(4), 385–398.

Wood, D. K., Gallivan, J. P., Chapman, C. S., Milne, J. L., Culham, J. C., & Goodale, M. A. (2011). Visual salience dominates early visuomotor competition in reaching behavior. *Journal of Vision*, *11*(10), 16–16.

Wood, D. K., Gu, C., Corneil, B. D., Gribble, P. L., & Goodale, M. A. (2015). Transient visual responses reset the phase of low-frequency oscillations in the skeletomotor periphery. *European Journal of Neuroscience*, *42*(3), 1919–1932.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, *12*, 579–601.

Woodgate, J. L., Makinson, J. C., Lim, K. S., Reynolds, A. M., & Chittka, L. (2017). Continuous radar tracking illustrates the development of multi-destination routes of bumblebees. *Scientific Reports*, *7*(1), 1–15.

Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2009). Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences*, *106*(40), 17199–17204.

Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2010). Economic choices can be made using only stimulus values. *Proceedings of the National Academy of Sciences*, *107*(34), 15005–15010.

Wurman, P. R., Barrett, S., Kawamoto, K., MacGlashan, J., Subramanian, K., Walsh, T. J., Capobianco, R., Devlic, A., Eckert, F., Fuchs, F., Et al. (2022). Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, *602*(7896), 223–228.

Yates, J. L., Katz, L. N., Levi, A. J., Pillow, J. W., & Huk, A. C. (2020). A simple linear readout of MT supports motion direction-discrimination performance. *Journal of Neurophysiology*.

Yates, J. L., Park, I. M., Katz, L. N., Pillow, J. W., & Huk, A. C. (2017). Functional dissection of signal and noise in MT and LIP during decision-making. *Nature Neuroscience*, *20*(9), 1285–1292.

Yoo, S. B. M., Tu, J. C., & Hayden, B. Y. (2021). Multicentric tracking of multiple agents by anterior cingulate cortex during pursuit and evasion. *Nature Communications*, *12*(1), 1985.

Yu, L., Simig, D., Flaherty, C., Aghajanyan, A., Zettlemoyer, L., & Lewis, M. (2023). MEGABYTE: Predicting million-byte sequences with multiscale transformers. *arXiv preprint arXiv:2305.07185*.

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., Et al. (2023). Catalyzing next-generation artificial intelligence through NeuroAI. *Nature Communications*, *14*(1), 1597.

Zago, M., Bosco, G., Maffei, V., Iosa, M., Ivanenko, Y. P., & Lacquaniti, F. (2004). Internal models of target motion: Expected dynamics overrides measured kinematics in timing manual interceptions. *Journal of Neurophysiology*, *91*(4), 1620–1634.

Zago, M., McIntyre, J., Senot, P., & Lacquaniti, F. (2008). Internal models and prediction of visual gravitational motion. *Vision Research*, *48*(14), 1532–1538.

Zago, M., McIntyre, J., Senot, P., & Lacquaniti, F. (2009). Visuo-motor coordination and internal models for object interception. *Experimental Brain Research*, *192*(4), 571–604.

Zgonnikov, A., Aleni, A., Piiroinen, P. T., O'Hora, D., & di Bernardo, M. (2017). Decision landscapes: Visualizing mouse-tracking data. *Royal Society Open Science*, *4*(11), 170482.

Zhang, Y., & Ross, K. W. (2021). On-policy deep reinforcement learning for the average-reward criterion. *arXiv preprint arXiv:2106.07329*.

Zhou, W., Chen, X., & Enderle, J. (2009). An updated time-optimal 3rd-order linear saccadic eye plant model. *International Journal of Neural Systems*, *19*(05), 309–330.

Zhu, J., & Thagard, P. (2002). Emotion and action. *Philosophical Psychology*, *15*(1), 19–36.

Zylberberg, A., Lorteije, J. A., Ouellette, B. G., De Zeeuw, C. I., Sigman, M., & Roelfsema, P. (2017). Serial, parallel and hierarchical decision making in primates. *eLife*, *6*, e17331.

Appendix A: Details for Chapter 3

A.1 Unit selectivity

We determined unit selectivity for CNN units by performing a linear regression of mean unit activity against coherence for each motion direction independently (i.e., one regression for left motion, and one for right motion). Significant linear regressions at a level of $p < 0.05$ were interpreted as units that were selective for that direction of motion. LSTM selectivity was performed instead on the difference between activity at stimulus onset and activity after the first full step of dot motion. Analysis was performed for all 10 trained agents in the 180 Hz saccade task.

Several CNN units were significantly selective for only leftward ($4.2\% \pm 0.7\%$), only rightward ($7.5\% \pm 1.1\%$), or both directions of motion ($63.3\% \pm 2.1\%$), while some units were not selective for either direction ($25.0\% \pm 1.4\%$). Several LSTM units were significantly selective for only leftward ($3.6\% \pm 0.5\%$), only rightward ($3.3\% \pm 0.5\%$), or both directions of motion ($66.0\% \pm 1.2\%$), while some units were not selective for either direction ($27.1\% \pm 1.3\%$).

A.2 Ablations

A set of agents (N = 9) were trained in line with the 60 Hz saccadic agents described in Fig 3.2. Here the only difference was in the training environment—agents were trained only on noiseless dots (i.e., only coherences of 100%). Theory predicts that noisy environments are necessary for accumulation mechanisms to emerge. On average, we find this to be the case; average agent performance at the end of training was not sig-

nificantly different from the one-sample accumulation threshold (80.3%) identified by the evidence accumulation model ($t(8) = 0.10, p = 0.54$). However, a small proportion of individual agents trained only on noiseless dots did develop accumulator-like internal dynamics (2 of 9 agents), albeit less clearly than agents trained with environmental noise. This is likely because poor initial representations of motion lead to noisy motion energy representations early on in training. Depending on network initialization, the learning of fine-tuned motion detection CNN kernels may be slower for some agents—enough for a rough accumulation-like mechanism to emerge within the downstream LSTM layer. Further work is needed on internal neural noise and the learning speeds of different layers to better understand this phenomenon.

A second set of agents (N = 9) were trained in line with the 60 Hz saccadic agents described in Fig 3.2. Here the only difference was in agent architecture—agents had a fully-connected layer in place of a recurrent LSTM layer with the same number of units. Note that because the recurrent agents also have parameters associated with their recurrent connections, the recurrent agents have more parameters than non-recurrent agents with a fully-connected layer in place of a LSTM layer. Average non-recurrent agent performance at the end of training was significantly higher than the one-sample accumulation threshold identified by the evidence accumulation model, $t(8) = 41.11, p = 6.75e\text{-}11$. In addition, non-recurrent agents learned the random dot motion task faster than recurrent agents, and displayed similar choice and response time sensitivity as their recurrent counterparts. While these agents did display dynamics that reflected momentary motion evidence in their CNN units, LSTM layer activation patterns strongly supported a non-primate-like extrema detection mechanism instead of a primate-like evidence accumulation mechanism. Recall that using an extrema detection mechanism, a decision maker waits until an individual sample is large enough to trigger a response (Stine et al., 2020). Under this mechanism, agents base their decisions on a single time step of evidence, rather than the accumulated history of evidence in time.

Overall, these two ablations support the idea that environmental noise and recurrent connections are necessary ingredients for primate-like decision making to reliably emerge via reinforcement learning.

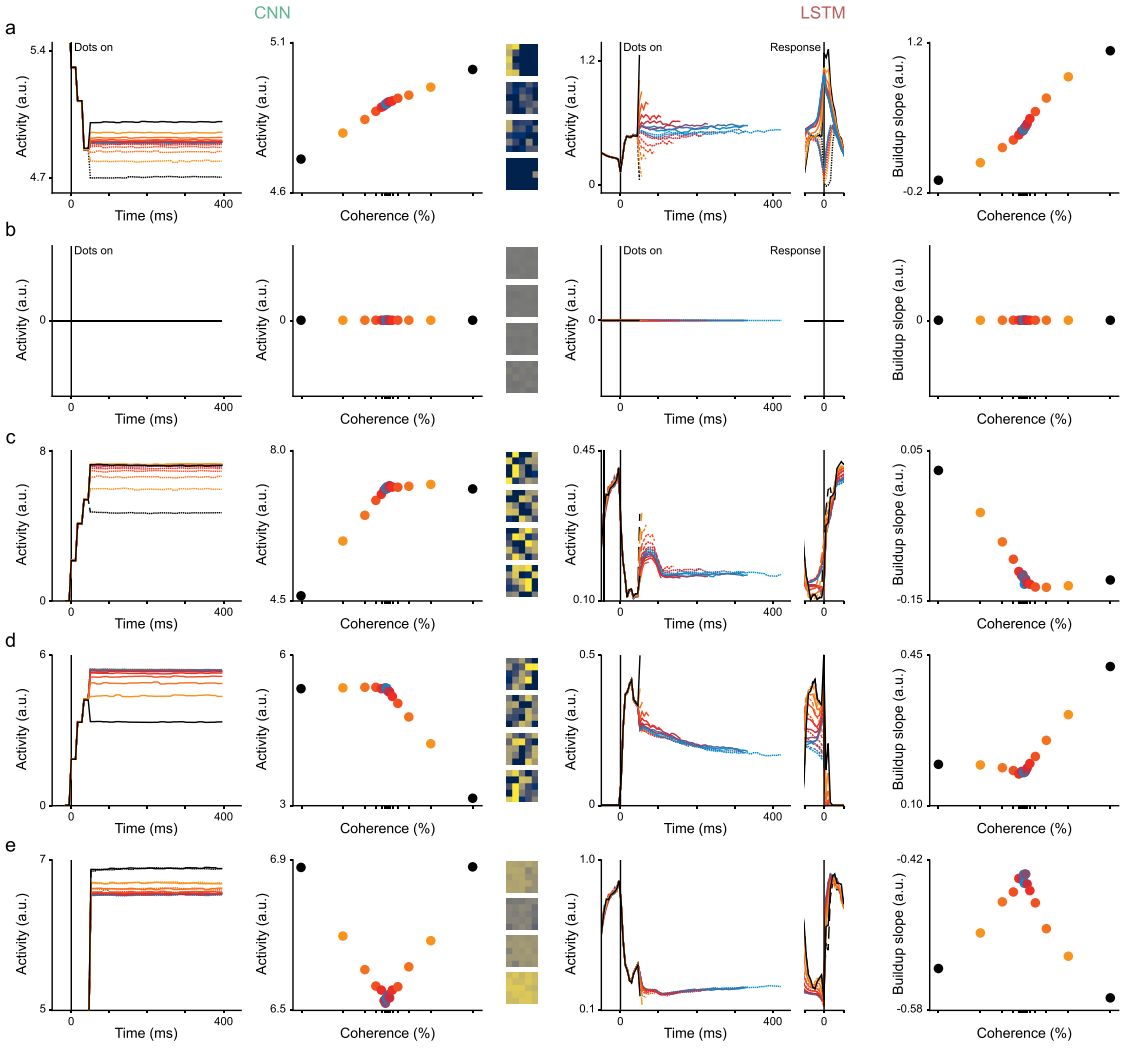

Figure A.1: Example CNN and LSTM unit selectivity profiles from a representative 180 Hz saccade agent. From left to right, the CNN unit activity over time, the CNN unit activity at a single point in time used for unit selectivity analyses, the 5 x 5 x 4 CNN kernel for this unit, the LSTM unit activity over time relative to stimulus onset, the LSTM unit activity over time relative to response, and the LSTM unit activity slope used for unit selectivity analyses. (a) Units shown in Fig 3.3. (b) Units with no selectivity. (c) Units selective only for leftward motion. (d) Units selective only for rightward motion. (e) Units selective for both motion directions in the same way. These units were excluded from microstimulation experiments because there was no hypothesis regarding how perturbing activity would impact the decision process.

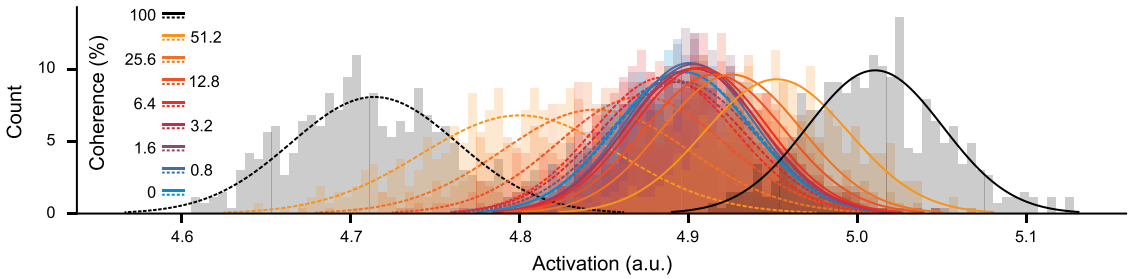

Figure A.2: Distribution of activity for the example CNN unit in Fig 3.3 from a single time step on independent evaluation trials. Gaussian distributions were fit to activity within each direction-coherence condition. Distributions of CNN unit activity were normally distributed within each direction-coherence condition (KS tests; ps > 0.05), in line with the hand-crafted evidence accumulation model (see Methods), standard evidence accumulation models (Gold & Shadlen, 2007), and recordings from primate area MT (Britten et al., 1992).

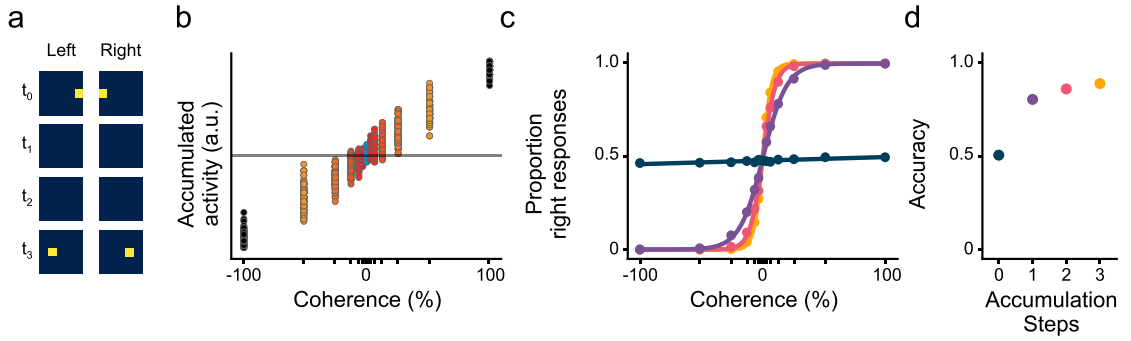

Figure A.3: Evidence accumulation model and results. (a) Two 5 x 5 x 4 convolutional kernels were hand-constructed with *a priori* knowledge of the dot motion stimulus—one for left motion and one for right motion at a speed of 3 pixels with 3 interleaved frames. The dot motion stimulus was convolved with each of these two kernels and summed across space. The difference between these two values was taken as a proxy for net motion energy. (b) This proxy for net motion energy was accumulated for n steps, and a left/right decision was made based on a signal detection threshold. Each dot represents the value for 3 steps of accumulated net motion energy on an independent simulated trial. (c) Model accuracy for 0, 1, 2, and 3 steps of accumulation. Accuracy increases with more steps of accumulation. See d for color legend. (d) Accuracy averaged across coherence conditions. The hand-constructed evidence accumulation model achieved 50% accuracy for 0 steps of accumulation, 80.3% accuracy for 1 step of accumulation, 86.2% accuracy for 2 steps of accumulation, and 89.1% accuracy for 3 steps of accumulation.

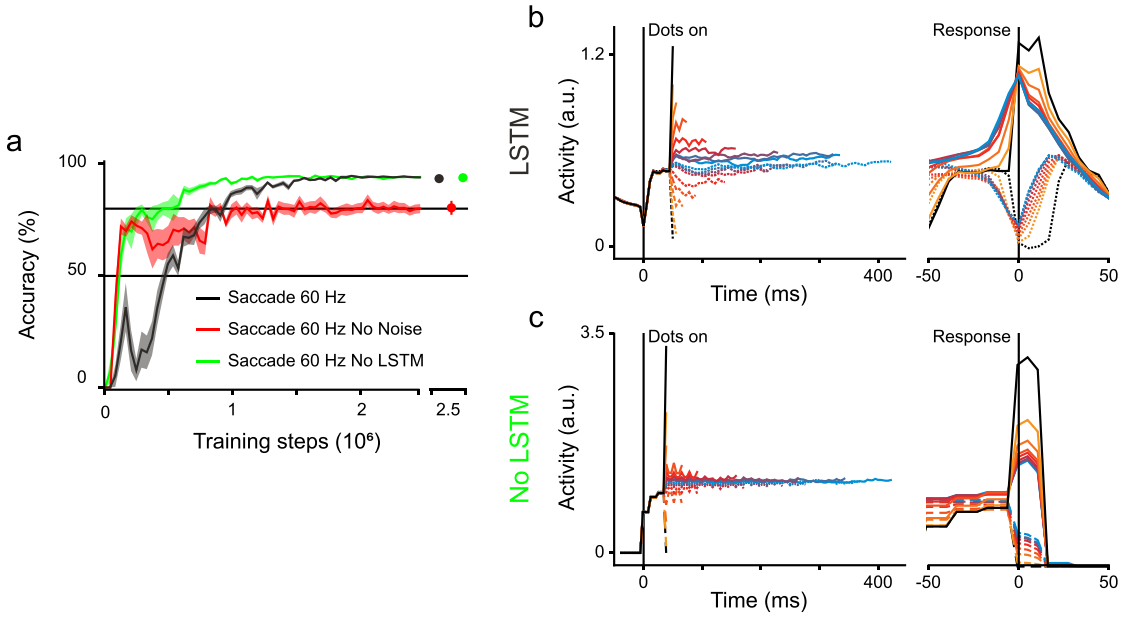

Figure A.4: Ablation results. Agents trained without environmental noise do not on average exceed the one-sample accuracy threshold determined from the evidence accumulation model. Agents trained without recurrence (i.e., no LSTM) display primate-like behaviour, but *not* primate-like mechanisms. (a) Accuracy during and after training. Agents trained without environmental noise (red), do not at any point exceed the one-sample 80.3% accuracy threshold. In contrast, agents trained without recurrence (green) exceed this threshold even earlier than recurrent agents (black). (b) Single-unit LSTM dynamics from a representative 180 Hz agent aligned to stimulus onset (left) and response (right). Reproduced from Fig 3.3c. Note the gradual accumulation of activity proportional to coherence before a response (right). (c) Single-unit dynamics from a representative 180 Hz agent aligned to stimulus onset (left) and response (right). Unit is from the fully-connected layer which replaced the LSTM layer for these agents. Since behaviour (the first key property of primate-like decision making) was similar between agents, we trained 180 Hz agents with and without recurrence to interrogate dynamics (the second key property of primate-like decision making). Note the response-aligned dynamics for this non-recurrent agent suggest that decisions are only based on a single time step before a response. In contrast, the response-aligned dynamics in b accumulate gradually before a response, except for the easiest decisions (black lines; 100%), which are also made based on a single time step of motion.

	Task/Network	
Hyperparameter	Saccade	Arm
No. of Conv kernels	64	64
No. of LSTM units	128	128
No. of environments	16	16
Steps per rollout	256	256
No. minibatches	2	8
Total time steps	250000	2,500,000
Simulation rate (Hz)	60 & 180	60
Steps before motion onset	(2, 11]	(0, 1]
Max time per trial (s)	3 & 2	3
Adam learning rate	3e-4 $\rightarrow$ 0.0	3e-4 $\rightarrow$ 0.0
Discount rate (γ)	0.99	0.99
PPO clip (ϵ)	0.2 $\rightarrow$ 0.0	0.2 $\rightarrow$ 0.0
PPO critic coef. (β_V)	0.25	0.25
PPO entropy coef. (β_{ent})	0	0
GAE parameter (λ)	0.95	0.95
Global norm grad clip	0.5	0.5
Move forward reward coef.	N/A	0.005

Table A.1: Hyperparameters for artificial agents responding via simulated saccades or reaches. Proximal policy optimization (PPO) algorithm hyperparameters for recurrent agents are described in Schulman et al. (2017).

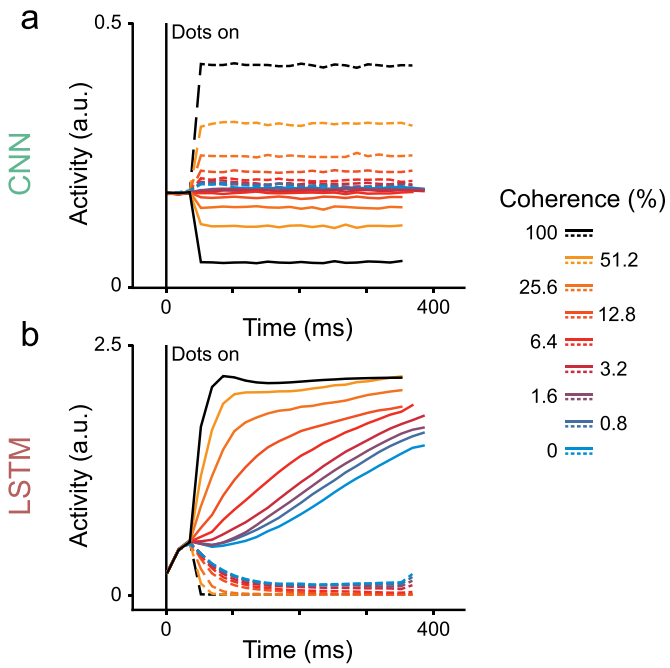

Figure A.5: Internal dynamics from a representative 60 Hz reaching agent. Solid lines indicate rightward motion conditions, and dashed lines indicate leftward motion conditions. For 0% coherence conditions (blue), traces are separated by when the agent ultimately chose the left or right target. Traces are averages of each condition for the agent. (b) Dynamics from an example LSTM unit.

Appendix B: Details for Chapter 4

B.1 Statistical reporting

B.1.1 Environment adaptation

We computed each agent's mean final score in each patch distance evaluation environment (e.g., single representative agent in Fig. 4.2b) to analyze whether agent score varied systematically with patch distance. Agents achieved a higher score on episodes where patches were closer together (Linear mixed effects regression with random agent intercept: $b = -5.82 \pm 0.21$, $p = 3.96 \times 10^{-166}$; Fig. 4.2b). Data consisted of 12 mean final scores (one for each agent) $\times$ 4 patch distances.

Trained agents adapted their patch leaving times to the environment, leaving patches later when travel distance is higher (Linear mixed effects regression with random agent intercept: $b = 9.60 \pm 0.87$, $p = 4.03 \times 10^{-28}$; Fig. 4.3a). Data consisted of 12 mean patch leaving times (one for each agent) $\times$ 4 patch distances.

B.1.2 Optimality

We take each agent's mean patch leaving time relative to the MVT across patch distance environments, and test if these 12 values are significantly different from zero (Fig. 4.3e) using a one-sample t-test. Comparing the difference between average observed and optimal patch leaving times, agents tend to overstay in patches relative to the optimal solution (One-sample t-test: $t(11) = 5.60$, $p = 1.60 \times 10^{-4}$; mean = 15.8 steps above MVT optimal, standard error = 2.7; Fig. 4.3e). Bonferroni-corrected one-sample t-tests show that agents significantly overstayed relative to the

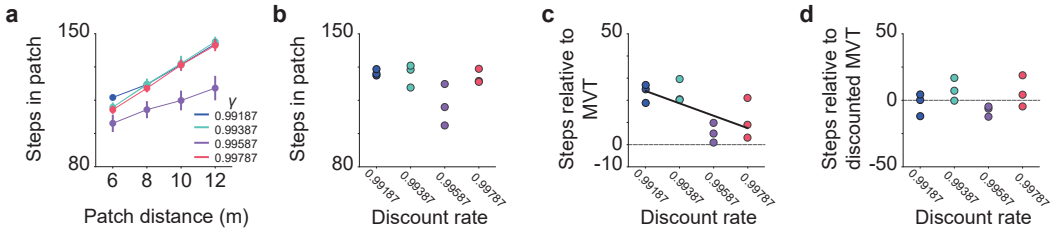

Figure B.1: Patch leaving times. (a) Agents grouped by discount rate. (b) Agents grouped by discount rate and collapsed by patch distance environment. (c) Mean difference between the observed and MVT patch leaving time collapsed by patch distance environment (Linear regression: $b = -2784.01 \pm 992.19$, $p = 0.019$). (d) Mean difference between the observed and discounted MVT patch leaving time collapsed by patch distance environment. All vertical lines denote standard errors.

MVT solution in all evaluation environments (ps < 0.0015), except for 12 m ($p = 0.047$).

We take the mean patch leaving time relative to the MVT across patch distance environments, and group these values by agent discount rate, leaving 3 agents per discount rate. Using linear regression, we test whether the difference between observed and MVT optimal patch leaving times decreases with discount rate. As expected, agents trained with higher temporal discounting rates tend to behave closer to MVT optimal (Linear regression: $b = -2784.01 \pm 992.19$, $p = 0.019$; Pearson correlation: $r(10) = -0.66$, $p = 0.019$; Figure B.1c). Data consisted of 3 agent's mean difference between observed and MVT patch leaving times (averaged across 4 environments) $\times$ 4 discount rates.

We take each agent's mean patch leaving time relative to the discounted MVT solution across patch distance environments, and test if these 12 values are significantly different from zero (Fig. 4.3h) using a one-sample t-test. Comparing the difference between average observed and discounted MVT optimal patch leaving times, agents were not significantly different from the optimal solution (One-sample t-test: $t(11) = 0.34$, $p = 0.74$; mean = 0.9 steps above discounted MVT optimal, standard error = 2.8; Fig. 4.3h). Bonferroni-corrected one-sample t-tests show that agents significantly overstayed in the 6 and 8 m evaluation environments (ps < 0.0067), understayed in

the 12 m ($p = 0.0023$), and were not significantly different from optimal in the 10 m environment ($p = 0.30$).

B.1.3 Dynamics and patch leaving time variability

Trained dynamics agents were evaluated on 30 episodes of each evaluation patch distance (i.e., 6, 8, 10, and 12 m). For the individual trained agent presented in Fig. 4.4, evaluation data were processed into 3790 unique patch encounters (mean = 31.6 unique patch encounters per evaluation episode). While single LSTM units (of 256) were analyzed, a principal components analysis (PCA) was also conducted to visualize patterns that accounted for larger amounts of variability in this layer. Exploratory analysis found that PC2 captured several key effects described in Hayden et al. (2011) while also accounting for 14% of LSTM layer variability (the second most in the network). Other PCs are included in Fig. B.2 for completeness.

For single-trial dynamics data aligned to patch entry and patch exit (Fig. 4.4c and g), the slope of LSTM unit or PC activity change across consecutive steps was regressed against patch leaving time quartile for 40 in-patch steps and 10 pre- or post-patch steps. Significance threshold for linear regressions at each dynamics time step was Bonferroni-corrected for 50 steps (i.e., p = 0.001) independently for patch entry- and patch exit-aligned data for both the example unit (Fig. 4.4c) and for the example PC (Fig. 4.4g).

We find in several LSTM units a significant relationship between the slope of rising activity and patch leaving time quartile for several steps after patch entry, but before patch exit (blue shaded bars; Fig. 4.4c). In the example unit shown, there are 19 consecutive steps where there is a significantly higher slope for encounters that had shorter patch leaving times (ps < 0.001; e.g., step 20 linear regression: $b = -0.0174 \pm 0.002$, $p = 2.36 \times 10^{-14}$; step 20 Pearson correlation: $r(3774) = -0.12$, $p = 2.36 \times 10^{-14}$; Fig. 4.4e). Data for each linear regression consisted of ∼944 activity slopes × 4 patch leaving time quartiles (Earliest, Early, Late, Latest). All results

were equivalent when using Pearson correlations instead of linear regressions.

Exit step activity was significantly different between patch leaving time quartiles for the example unit (One-way ANOVA: $F(3, 3775) = 7.15$, $p = 8.76 \times 10^{-5}$). Bonferroni-corrected pairwise follow-up tests show no significant differences other than the mean activity of the Latest patch leaving time quartile from all other conditions ($ps < 0.0012$). For the selected PC, analysis similarly showed exit step activity was significantly different between patch leaving time quartiles (One-way ANOVA: $F(3, 3775) = 320.79$, $p = 1.58 \times 10^{-185}$). Follow-up tests show significant differences between all conditions ($ps < 1.13 \times 10^{-11}$) other than between the Late and Latest patch leaving time quartiles ($p = 0.99$).

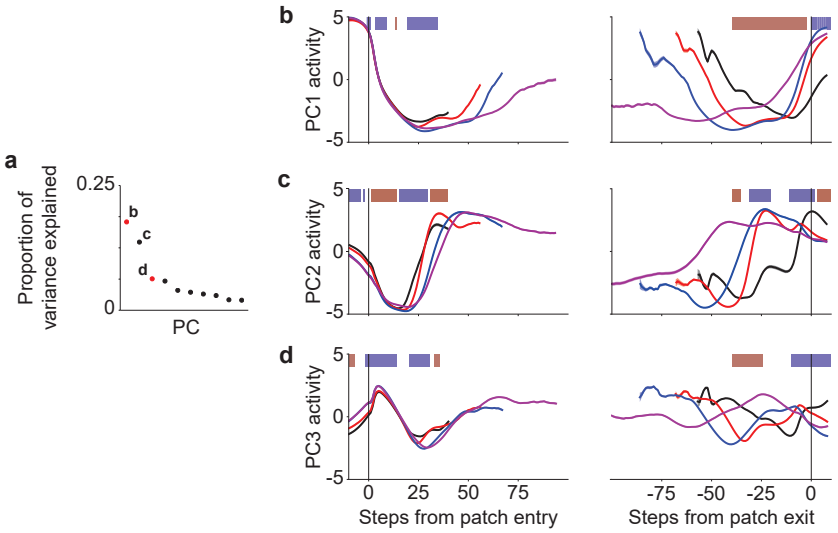

Figure B.2: Major principal components from the LSTM layer of a representative trained agent. (a) Proportion of LSTM layer variance explained by each of the first 10 PCs. PC2 (c) is highlighted in the main text. (b, c, d) Activation of the principal components accounting for the most amount of variance in the agent's LSTM layer dynamics. Average activation is aligned to patch entry (left), and patch exit (right). Shaded blue and red bars indicate steps where there is a significant slope-patch leaving time quartile relationship (negative in blue, positive in red). Traces in each quartile are plotted until median patch leaving time. Shaded regions along activity traces denote standard errors.

B.1.4 Dynamics and environment adaptation

We took the difference between activity at patch exit and activity at patch entry for every patch encounter, and separated these data by patch distance. We find a positive relationship between activity range and patch distance environment for the example unit (Linear regression: $b = 0.053 \pm 0.003$, $p = 5.88 \times 10^{-56}$; Pearson correlation: $r(3774) = 0.25$, $p = 5.59 \times 10^{-56}$; Fig. 4.4h), and for the selected PC (Linear regression: $b = 0.20 \pm 0.020$, $p = 1.70 \times 10^{-22}$; Pearson correlation: $r(3774) = 0.16$, $p = 1.70 \times 10^{-22}$; Fig. 4.4h). Data consisted of ∼944 differences in exit and entry activity × 4 patch distance evaluation environments (6, 8, 10, and 12 m).

Decreasing the slope with which a decision variable accumulates toward a threshold tends to prolong patch leaving times. For both the example unit and PC, we took the average slope of activity across all steps for which there was a significant predictive relationship between slope and patch leaving time quartile after patch entry (e.g., steps 5 to 23 after patch entry for the example unit; Fig. 4.4c). We find no significant relationship between the average slope of activity and patch distance environment in the example unit (Linear regression: $b = -0.00026 \pm 0.00020$, $p = 0.18$; Pearson correlation: $r(3785) = -0.022$, $p = 0.18$; Fig. 4.4i), nor for the selected PC (Linear regression: $b = -0.0034 \pm 0.002$, $p = 0.13$; Pearson correlation: $r(3784) = -0.025$, $p = 0.13$; Fig. 4.4i). Data consisted of ∼944 average activity slopes × 4 patch distance evaluation environments.

B.2 Accounting for discounting in the marginal value theorem

We accounted for temporal discounting rate by simulating individual stay and leave decisions at many patch leaving steps. Agents could either stay for an additional step of reward before leaving a patch, or immediately leave the patch, where the subsequent 5000 steps were simulated as alternating between a fixed number of steps

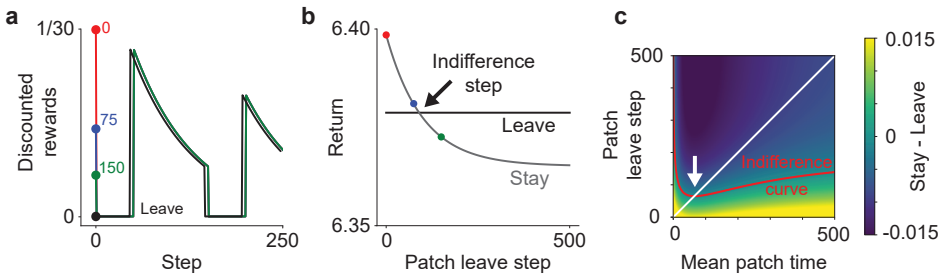

Figure B.3: Accounting for temporal discounting in the marginal value theorem. (a) Simulated discounted rewards for different artificial policies. After the first step, these simulated agents alternate between a fixed number of steps in a patch, and a fixed number of steps traveling between patches. (b) Computed discounted return for artificial policies in (a). (c) Difference in discounted returns between stay and leave policies at different fixed patch times (assuming a discount rate and a fixed travel time). White arrow indicates discounted MVT estimate for optimal mean patch time.

in a patch and a fixed number of steps traveling between patches. For example, on step 150 within a patch, and given a future fixed travel time of 50 steps and a future fixed patch time of 100 steps, is it more beneficial to leave immediately (Figure B.3a; "leave" in black) or stay in the patch for an additional step of reward (Figure B.3a; 150 in green)? By computing the discounted return for immediately leaving, and computing the discounted returns for remaining in the patch for one additional step of reward at every level of patch depletion, we generate a curve where the value of leaving can be compared to the value of staying at individual patch depletion points (Figure B.3b).

In Figure B.3b, we show the simulated indifference step when there are 100 fixed steps in a patch and 50 fixed steps traveling between patches. An indifference step can be estimated for every choice of subsequent fixed patch and travel steps (e.g., as in Figure B.3a), which gives rise to an indifference curve over these parameters (Figure B.3c). Using the observed mean travel steps from trained agents, we can instead sweep over only fixed steps in patch in order to evaluate an agent's choice of mean patch time. This assumes that the agent's mean travel time is constant and independent of mean patch time.

Where the stay/leave indifference curve matched the fixed patch steps (i.e., the unity line in Figure B.3c) provided an approximation of a single mean patch time that maximizes the discounted return (given a discount rate and observed mean travel steps).

B.3 Training details

We generally follow similar training procedures as for the models described in Cultural General Intelligence Team et al. (2022). Each agent was trained on an internal cluster for roughly 13 days, and used approximately 40 GiB RAM, 8 CPU, and 8 GPUs.