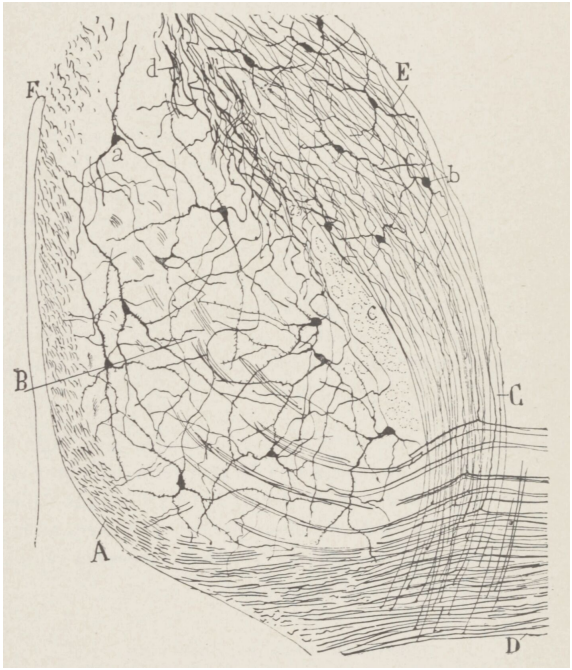
Neuronal and computational principles of individual long-term learning trajectories



Samuel Cian Liebana Garcia

Merton College

A thesis submitted for the degree of
Doctor of Philosophy

Supervised by
Professors Armin Lak, Andrew Saxe and Rafal Bogacz

Department of Physiology, Anatomy and Genetics
University of Oxford

Hilary Term, 2025

Dedicated to my grandfather, Abuelo Ernesto

Declaration

I declare that this thesis is entirely my own work and, except where stated, describes my own research.

Samuel Cian Liebana Garcia
Merton College

Acknowledgements

I generated the last novel panel for the figures used in this thesis on the 6th of January 2025. In Spain, the 6th of January marks the day of the Epiphany, popularly known as the 'Dia de Reyes', a Christian feast day remembering the visit of the Three Magi or Three Wise Men to Jesus after his birth. Instead of Santa Claus, it is the Three Wise Men who bring gifts to children in Spain during Christmas. Inevitably, the generation of that beautiful panel which I had struggled so much to obtain reminded me of the gifts that my parents used to prepare for my brother and I on this day as children. This time, however, I realised that the gift was indeed coming from the Three Wise Men who have guided me for the past 4 years. It has been a pleasure to work with you all.

Armin, without your incessant support, availability, kindness and motivation it would have been impossible for me to enter the world of experimental neuroscience which seemed so foreign at the beginning. I laugh at how naïve I was when I first started in your lab and admire your patience while listening through all the nonsensical ideas that I used to throw at you. I still remember the day when you said "if there's no reward, there's no paper!" after having listened through probably more than an hour of rambles about how we should try modelling dopamine without making reference to rewards. They were wise words, and I am very proud of the work we have achieved together. I admire many things about your work ethic, but especially your ability to take accurate 'executive decisions' when there has been enough thinking and we just have to get things done. We have gone through a long road together, and I am extremely grateful for your constant close supervision which has always led me down the right paths.

Andrew, how do you do it? I think there wasn't a single meeting between us after which I didn't come up with the solution for a problem I had been stuck on for days, if not weeks. Your carefully selected words, combined with your deep wisdom of everything to do with learning (and beyond!) is not only admirable but I dare say *magical*. At least that's how I felt when I used to wake up the day after a meeting with you and suddenly have your words assembling the solution to the problem I desperately needed to solve to get the model running or the maths correct. It has been a privilege to work alongside you, and I use this term because that's what it felt like the entire time. You place immense trust in your students and never impose your ideas on them, only providing suggestions which they can leave or take, allowing them to grow into successful independent researchers. I really admire your approach, and acknowledge that it must be stressful at times, especially when your student likes to stick around saddle points for a while before progressing with their work!

Rafal, your wealth of experience in academia has been invaluable. You have always known how to identify priorities and navigate difficult circumstances when it wasn't clear how to act. Moreover, your ability to take your students' complex, disorganised thoughts and help turn them into an elegant simple principle is astounding. Now, whenever I approach a new modelling problem I always have your wise words ringing in my head: "Make it simpler!". I am also very grateful for your eternally positive attitude, attending your lab meetings always brought a smile to my face - and this wasn't just because I was seeing mathematical neuroscience in Oxford!

The work in this thesis would also not have been possible without all the members of the Lak, Saxe and Bogacz labs with whom I always had stimulating discussions and many of whom directly contributed to the experiments. I would like to highlight a few characters whose friendship made a big difference to me. Stefano Sarao Mannelli, for inspiring me with your mathematical prowess,

Devon Jarvis, for your kindness and all you have taught me, Gabriel Ocaña Santero, for being my go-to Spanish-speaking biochemist in the office, Marko Tvrdic, for always encouraging me to stand up for myself, Matthias Fritsche, for impressing me with your rigour and principles, Sandra Tan, for your constant support and encouragement, and Chiara Toschi, for your warmth and friendly nature. I have also enjoyed a stimulating friend circle outside of the lab with my close friends from college who always supported me in difficult times and rejoiced with me in good ones. Thank you Cesare Vagge, Hiroto Takahashi, Kabir Bakshi, Jovan Jovanovic, Pavle Kravljanac, Vikaran Khanna, Emilio Chuvieco, Minghao Li and Uziel Gonzalez.

I would now like to thank my family, who have supported me all my life to allow me to reach this point where I am submitting the document to become a doctor. To my dear parents, Raquel and Ernesto, thank you for providing me with the best childhood one could desire. To my father, thank you for stimulating my scientific curiosity from the start. I still remember when you used to bring me those Petri dishes from the lab so that I could grow my own bacteria. To my mother, thank you for always believing in me and pushing me to reach my potential. To my dear brother, Diego, thank you for teaching me how to love. To all my uncles and aunts for their counsel and guidance throughout my life. And lastly, to my grandparents, thank you for your immense wisdom that I will never forget.

I cannot end without mentioning you, Ishu. As your name signifies, you have been my gift from God. Thank you for always being by my side.

This document was created using the LaTeX typesetting system.
Figure on the title page is Cajal's drawing of the mouse substantia nigra, from Ramon y Cajal S. Histologie du systeme nerveux de l'homme et des vertebres. Vol. 2. Maloine, Paris 1911.

Abstract

Striatal dopamine plays fundamental roles in learning to fine-tune decisions. However, in learning from naïve to expert individuals often exhibit diverse learning trajectories, defying understanding of its underlying dopaminergic mechanisms. Here we longitudinally measured and manipulated dorsal striatal dopamine signals in mice learning a decision task from naïve to expert. Mice learning trajectories transitioned through sequences of strategies, showing substantial individual diversity. Remarkably, the transitions were systematic; each mouse's early strategy determined its strategy weeks later. Dopamine signals matched individual trajectories, encoding teaching signals that update only a subset of the associations used for decision making. Optogenetic manipulations demonstrated the exclusive learning effects of these dopamine signals. A deep neural network using heterogenous teaching signals, each updating a subset of network weights, captured our results. Analysing the model's fixed points explained learning diversity and systematicity. Altogether, this work provides insights into the biological and mathematical principles underlying individual long-term learning trajectories.

Statement of contributions

This thesis contains work produced in collaboration with members of Armin Lak's lab at the University of Oxford, as well as my supervisors Prof Andrew Saxe and Prof Rafal Bogacz. The experiments were conceived and designed in collaboration with Armin Lak and Peter Zatka-Haas. In particular, I took the lead in the optogenetics experiments. I also trained and recorded neural signals from several mice, adding to those trained by Chiara Toschi, Aeron Laffere, Louisa Schilling, Peter Zatka-Haas and Armin Lak. Mouse surgeries were performed by Chiara Toschi, Jessica Moretti, and Peter Zatka-Haas. I led the analysis of the behavioural and neural data, with early inputs from Aeron Laffere and Peter Zatka-Haas, and contributions to specific analyses from Matthias Fritsche, Jacek Podlaski, all supervised by Armin Lak. I designed, implemented and simulated the computational model following advice from Andrew Saxe, Rafal Bogacz, and Armin Lak. Mathematical analysis of the model was performed with the guidance of Andrew Saxe. Lastly, parts of the main Introduction, and of each chapter's introduction, results and discussion have been adapted from text written with my supervisors Armin Lak, Andrew Saxe and Rafal Bogacz for the papers deriving from the work contained in this thesis.

List of publications

The following is a list of articles that were published as a result of the research carried out for this thesis.

Articles

- Liebana, S., Fritsche, M., Lak, A. Dopaminergic computations for perceptual decisions. Current Opinion in Behavioral Sciences 60, 101458 (2024).
- Fritsche, M., Majumdar, A., Strickland, L., Liebana Garcia, S., Bogacz, R., Lak, A. Temporal regularities shape perceptual decisions and striatal dopamine signals. Nat Commun 15, 7093 (2024).
- Lee, S., Liebana, S., Clopath, C., Dabney, W. Lifelong Reinforcement Learning via Neuromodulation. arXiv:2408.08446 (2024).
- Liebana Garcia, S., Laffere, A., Toschi, C., Schilling, L., Podlaski, J., Fritsche, M., Zatka-Haas, P., Li, Y., Bogacz, R., Saxe, A., Lak, A. Striatal dopamine reflects individual long-term learning trajectories. bioRxiv, 12.14.571653 (2023).

Table of contents

Table of contents xv

List of figures xvii

List of tables xix

Glossary xx

1 Introduction 1
- 1.1 Learning ... 3
- 1.2 Long-term learning ... 6
- 1.3 Learning in the brain ... 9
- 1.4 From a model to a theory ... 11
- 1.5 Outline of the thesis ... 15

2 Long-term learning trajectories 16
- 2.1 Introduction ... 16
- 2.2 Methods ... 17
  - 2.2.1 Mice ... 17
  - 2.2.2 Behavioural task ... 18
  - 2.2.3 Video monitoring ... 19
  - 2.2.4 Behavioural data analyses ... 19
  - 2.2.5 Statistics ... 21
- 2.3 Results ... 21
- 2.4 Discussion ... 29

3 Neuronal teaching signals 31
- 3.1 Introduction ... 33
- 3.2 Methods ... 34
  - 3.2.1 Mice ... 34
  - 3.2.2 Surgical procedures ... 34
  - 3.2.3 Imaging dopamine release ... 34
  - 3.2.4 Optogenetic manipulation of dopamine signals ... 35
  - 3.2.5 Histology and fibre track quantifications ... 36
  - 3.2.6 Mice trained to higher accuracy ... 36
  - 3.2.7 Reward size manipulation experiment ... 36
  - 3.2.8 Neural data analyses ... 36
- 3.3 Results ... 39
  - 3.3.1 Dopamine recordings ... 39
  - 3.3.2 Optogenetic manipulations ... 50
- 3.4 Discussion ... 54

4 Deep neural network model 57
- 4.1 Introduction ... 58
- 4.2 Methods ... 60
  - 4.2.1 Architecture ... 60
  - 4.2.2 Learning rule ... 61
  - 4.2.3 Model-derived dopamine signals ... 63
  - 4.2.4 Simulation of the optogenetic experiment ... 64
  - 4.2.5 Average dynamics ... 64
  - 4.2.6 Fitting procedure ... 65
  - 4.2.7 Saddle Points ... 66
  - 4.2.8 Generalisation to larger networks ... 67

4.3 Results ... 67
4.3.1 Deep 'tutor-executor' RL ... 67
4.3.2 Fixed point analysis ... 80
4.4 Discussion ... 83
4.4.1 Depth and heterogenous RPEs ... 83
4.4.2 A mathematical theory ... 84

5 Conclusion and future work **88**
5.1 Contributions to the study of behaviour ... 88
5.2 Implications for research on the neural bases of learning ... 90
5.3 Future perspectives for deep learning ... 92
5.4 Summary ... 94

Appendix A The Response Hypothesis **95**
A.1 Introduction ... 95
A.2 Theoretical framework: rewards as *responses* ... 98
A.3 Neural implementation: corticostriatal and striatonigrostriatal circuits ... 101
A.4 Experimental validation and predictions ... 103
A.5 Simulations ... 105
A.6 Conclusion ... 105

Appendix B Alternative Models **106**
B.1 Introduction ... 106
B.2 Flexible Automated Toolbox for Trial-by-Trial Model Fitting ... 107
B.3 Minimal Deep Linear Recurrent Neural Network Model ... 108
B.3.1 Descriptive Fits of DLRNN Parameters ... 109
B.3.2 Goal-directed RL and Habitual Learning Rules for the DLRNN ... 110
B.4 Discussion and Conclusion ... 112

Bibliography **113**

List of figures

1.1 Externalising the brain 2
1.2 Experimental set-up for studying long-term learning trajectories 8
1.3 Schematic of anatomy relevant for dopamine signals in perceptual decision making .. 10
1.4 Schematic of a deep neural network 12

2.1 Details of the behavioural task 18
2.2 Schematic of behavioural metrics 19
2.3 Colouring trajectories by slope asymmetry 20
2.4 Pupil keypoints 20
2.5 Lick keypoints 20
2.6 Learning a visual decision task from naïve to expert 21
2.7 Mice develop diverse strategies through learning 22
2.8 Behavioural metrics evolve over learning 23
2.9 Mice form diverse yet systematic learning trajectories from naïve to expert. 24
2.10 Quantification of diverse psychometric and chronometric phenomena 25
2.11 Non-learner mice and model assessing origin of initial bias 26
2.12 Eye movement and lick analyses 28

3.1 Dopamine encodes Temporal Difference (TD) error 32
3.2 Dorsolateral striatum (DLS) dopamine (DA) signals during task learning 39
3.3 DLS DA signals develop over learning encoding stimulus-choice associations 41
3.4 DLS DA stimulus responses are absent in initial days and zero-contrast trials 42
3.5 DLS DA in matched accuracy days 44
3.6 DLS DA signals during learning and in highly expert mice; split by stimulus type, response times and trial outcome 45
3.7 DLS and dorsomedial striatum (DMS) DA signals over learning 46
3.8 Further comparison of DLS and DMS DA signals and reward size manipulation experiment 47
3.9 DLS DA signals reflect learning trajectories from naïve to expert. 48
3.10 Further quantification of DLS DA responses to task events during learning 49
3.11 DLS DA key characteristics and causal predictions 50
3.12 DLS DA is required to learn 51
3.13 DLS DA updates behaviour distinct from classic RPE 52
3.14 Relative evolution of DMS and DLS DA signals and effect of DMS DA optogenetic stimulation 55

4.1 The deep 'tutor-executor' RL model and its mapping to anatomy 68
4.2 The tutor-executor model captures behavioural learning trajectories 69
4.3 The tutor-executor model learns over similar timescales and displays similar patterns of diversity 70
4.4 The tutor-executor model captures dopamine signals over learning 71
4.5 Further quantification of model-derived dopamine signals 72
4.6 Shallow networks do not capture learning trajectories 73
4.7 A model without partial RPEs does not capture optogenetic results 74
4.8 Exploring model constraints. 75
4.9 Effect of fixed non-uniform input representations 76
4.10 Effect of stimulus-pathway cross-connections 77
4.11 Generalisation to larger networks 78
4.12 Tutor-executor learning rule causes transfer to striatum 79
4.13 Saddle points of the model explain learning trajectories and dopamine signals 80
4.14 Non-learner simulations 81

4.15 A deep model with single-loss gradient descent captures behavioural trajectories with a teaching signal different from DLS DA 82

A.1 Schematic of TD-RPE Hypothesis 97
A.2 Schematic of brain anatomy relevant to the response hypothesis 102
A.3 Integrator-derivative model of corticostriatonigrostriatal connectivity 103
A.4 Simplified block diagram of integrator-derivative model 103
A.5 Neural network model of brain anatomy relevant to the response hypothesis 104

B.1 Deep Linear Recurrent Neural Network (DLRNN) Architecture 108
B.2 Descriptive fits of DLRNN parameters 110
B.3 RL-DLRNN model architecture 110
B.4 Fits of RL-DLRNN model 111

List of tables

Glossary

AI Artificial intelligence

BIC Bayesian information criterion

DA Dopamine

DLS Dorsolateral striatum

DMS Dorsomedial striatum

LLM Large language model

ML Machine learning

PC Predictive coding

RL Reinforcement learning

RPE Reward prediction error

RT Response time

SGD Stochastic gradient descent

SNc Substantia nigra pars compacta

TD Temporal difference

VTA Ventral tegmental area

Chapter 1
Introduction

Chapter contents

1.1	Learning	3
1.2	Long-term learning	6
1.3	Learning in the brain	9
1.4	From a model to a theory	11
1.5	Outline of the thesis	15

We are living in the age of learning

— *c.f. "We are living in the age of the machine"* by Paul B. Haviland*, 291 (magazine)*

Just as tools have served to enhance and extend the functionality of our limbs, humans and other animals have in the past found ways to extend the capabilities of their brain through its *externalisation* (Figure 1.1). Possibly one of the first examples of such externalisation is that of communication. Synapses in the brain enable different neurons and brain regions to communicate with each other and give rise to complex concerted function that is much more capable than each on their own. In a similar fashion, many organisms throughout the animal kingdom have developed communication techniques ranging in complexity [1], efficiency [2] and beauty [3]. In humans, language arose as a means to enable powerful collaboration between individuals, attested by the popular phrase "two heads are better than one". Language is continuously developing and enables ever-more effective communication on all fronts from speed, to accuracy, to intensity. The denser and wider the extent of our connectivity, the more elaborate the resulting patterns of activity will become and the greater the feats that humanity will be able to achieve.[1]

Emerging slightly slower was our externalisation of memory. One of the brain's most impressive abilities is that of being able to store and organise vast amounts of information in its patterns of neural connectivity with remarkable reliability. Nonetheless, each of our brains only has a limited capacity to store information, and one single brain cannot come even close to encompassing the amount of knowledge present in the world, although humans don't like to hear it [4]. As such, we came to realise that we should probably find ways to augment our brain's memory capabilities using tools from our environment. Although there are surely earlier examples of externalised memory, painting and writing (and more recently filming) are probably some of the first examples of sophisticated forms of fixing patterns of neural activity into our environment, with the intention of later retrieval. The development of these techniques not only greatly enhanced our memory, but also improved the communication techniques that we described above. Externalised memory could be transported long distances and stored for long timescales much more accurately than methods relying on personal encounters such as oral transmission. This took us a step closer to the dense and wide communication networks desirable for rapid progress.

We next started to externalise computation, that is, the processing of information. In addition to routing and storing information, the brain is constantly performing an immense number of

[1] Note that this is not too different from what distinguishes the cortex from other brain structures. The cortex being, in turn, one of the brain regions that most distinguishes us from other members of the animal kingdom.

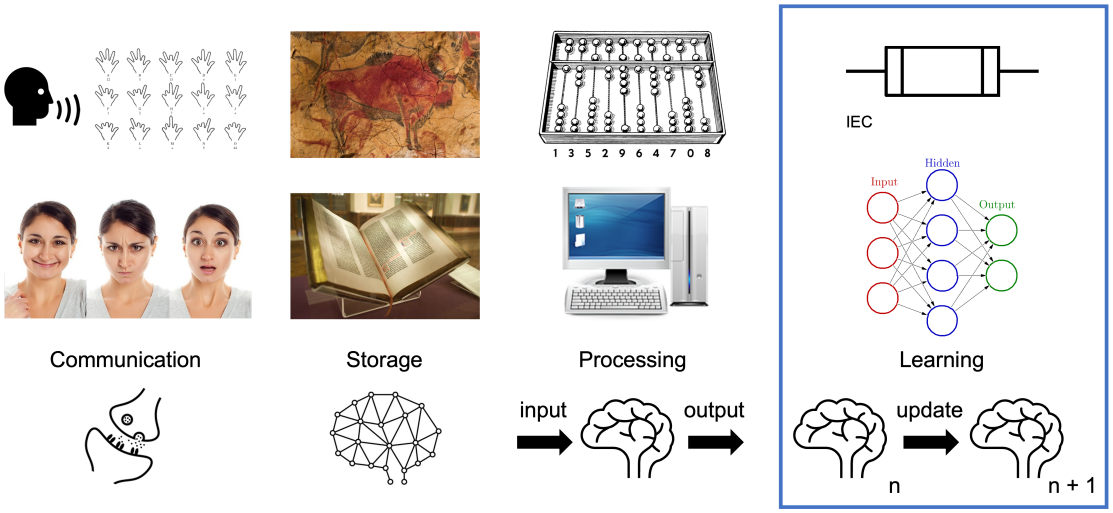


Figure 1.1: Externalising the brain. From left to right, progressive stages of brain externalisation over time. Hand signs are Gravettian hand stencils [5]. Image of woman displaying facial expressions of happiness, anger, and surprise cropped from a photo series by AlexSutula / shutterstock, reproduced from [6]. Bison Magdalenian polychrome by Museo de Altamira y D. Rodriguez / National Museum and Research Centre of Altamira. Image of Gutenberg bible by NYC Wanderer (Kevin Eng) / Flickr. Abacus by Pearson Scott Foresman / Wikimedia Commons. Computer icon by Everaldo Coelho / Crystal Project / KDE store. Fuse IEC icon by Quicksilver (a.k.a Hydrargyrum) / Wikimedia Commons. Deep neural network by Glosser.ca / Wikimedia Commons. Synapse icon by Lea Lortal / The Noun Project. Brain network icon designed by Freepik. Brain sketch and initial voice icon from Microsoft Powerpoint stock.

computations on the formidable amounts of sensory input that our senses can capture. However, at least in most individuals, the brain's ability to perform precise computations on restricted subsets of incoming information appears quite limited and fallible. Perhaps because it is not a situation that the brain has had to confront often in its evolutionary past. Regardless of the reason, it became clear to us that we could not rely on our brains for the implementation of precise logical input-output maps and began to design devices that could help us with the process. Amongst the many *machines* we designed, the abacus is an early example of those that enhanced the important ability of performing arithmetic accurately. The real computing revolution, however, came with the arrival of the modern-day *computer*. Its simple, intuitive interface and the flexibility of its computations (thanks to the simultaneous development of programming languages) brought the capacity to perform incredibly complex computations with remarkable accuracy to anybody who could engage with it. This gave rise to the field of computer science, which made remarkable breakthroughs in uncovering the principles of computation from notions of algorithmic complexity [7] to the information-theoretic characterisation of processing channels [8]. Although, due to physical restrictions, computers may be limited in the amount of information they can process compared to our brains, they have proved to be much more reliable at performing complex logical operations and are used across disciplines to both refine and extend humans' processing power. This marked the beginning of the digital age that we have been living in, which similarly revolutionised our memory and communication abilities in parallel.

Now, possibly for the first time in our history, we seem to have a grip on the problem of externalising learning. Despite the flexibility of computer programs, there always had to be a human behind the curtains writing and updating computer code for our increasing computing needs. With the advent of deep learning, this has radically changed. Computer programs can now essentially write themselves and implement any input-output map (within capacity) thanks to the

universal function approximator capabilities of deep learning systems [9]. All a practitioner has to do is provide the system with details of the task to complete and the artificial learning rules of deep learning systems tweak the weights of the connections between their neurons to learn the operation that has to be performed [10]. This framework has given rise to the current deep learning revolution, which once again is revolutionising multiple fields and disciplines, as well as vastly improving our previous externalisation of the brain's computing, memorisation/organisation and communication abilities. The recent development of large language models (LLMs) has demonstrated how deep learning has enabled processes as complex as language to be implemented in computers [11, 12]. These LLMs can convey vast amounts of information, and the results of computations performed on it, in human-parsable form. In addition, it can also write computer code in our programming languages to implement arbitrarily complex functions on previous computers and software. An interesting question that came to mind while writing this is: should we continue to call agents with such learning capabilities *machines*? 'Machine' may be more well-suited to describe computers of the computing age (notice how we say machine *learning* systems), which implemented precise yet fixed operations. Hence I wonder what word will eventually settle in common parlance to describe the more dynamic agents with which we are dealing now.

Our next steps in externalisation are difficult to predict, but the brain has many more capabilities that we may want to reproduce and enhance using our environment. To name a few, I think many would agree that large language models' thinking capabilities remain limited if they exist at all [13, 14], and similarly salient brain processes such as emotions and feelings have yet to be recreated in external agents. Another brain ability which we are currently struggling to externalise is that of generating complex, accurate and yet flexible concerted movements in our limbs. This is an open problem in the field of robotics [15], which may find inspiration for connecting the powerful language models of the present with their robotic effectors by investigating how the brain's cortical structures communicate with subcortical regions. Altogether, the future seems bright with opportunities for intelligence researchers. However, as Paul Haviland rightly remarks in his 291 article, what gives us control over 'machines' is that we have not endowed them with the capacity to think and act independently of our volition. This begs the question: should we really aim for this or will our wings melt and livers be eaten?

1.1 Learning

Defining 'learning' can be tricky, as with all attempts to explain our ultimately fuzzy use of language to describe the world [16]. It is generally conceptualised as a change in computation or behaviour that improves performance on a particular metric. In neuroscience, we think of it as changes in the strengths of synapses between neurons that enable improvements in the performance of a task. However, learning can also occur independently of its immediate relevance for the agent [17]. Being the broad concept that it is, learning is studied across many disciplines. Here we introduce concepts from three of these: Experimental Psychology, Neuroscience and Machine Learning, that this thesis will attempt to unify through its behavioural, neural and computational/theoretical investigations.

The characterisation of learning in the psychological sciences has set the scaffold for studying learning in many other disciplines. Through their detailed observation of behaviour, psychologists have designed sophisticated psychometric tests and measures that can quantify subtle changes in behaviour and identify core processes involved in decision making [18, 19]. Given its prominence, experimental psychology has often studied learning in infants and children. One of the seminal works in this regard is that of Jean Piaget [20], who identified distinct stages in the cognitive

development of infants, each of increasing conceptual complexity and generality, associated with typical age ranges. The presence of learning stages has also been reported in individuals acquiring and organising semantic knowledge [21, 22, 23], a process that is also observed in the learning dynamics of deep neural networks [24]. Stages in the development of numerical reasoning have also been reported [25, 26], though some argue that certain numerical abilities may be innate [27, 28]. A related phenomenon observed on shorter timescales is that tasks often display sigmoidal learning curves which start with an initial period of stasis followed by an abrupt jump from untrained to expert levels of performance [29, 30]. These jumps are often referred to as 'insight' or 'aha' moments [31, 32]. Superficially, this seems to reflect a discrete underlying learning mechanism. In fact, a common explanation is that agents test multiple solution strategies until suddenly finding the right one [33, 34, 35, 36]. However, such learning curves are also displayed by deep neural networks trained with continuous learning rules [37, 38, 39], which may provide a more mechanistic model of the phenomenon.

Edward Thorndike's work on animal learning provided an effective bridge from psychology to neuroscience [40]. He was one of the first to use animals for psychological research, introducing the use of 'puzzle boxes' or behavioural assays to study learning in animals. This enabled the study of behaviour in carefully constructed environments, as well as direct control over reinforcement schedules to assess their effect on learning. His work led to important ideas such as the law of effect [41], a precursor to Skinner's behaviourism [42], both of which claim that behaviours followed by satisfying outcomes ('rewards') will become more likely, and those followed by discomfort ('punishments') less likely. These ideas served as inspiration for the field of reinforcement learning (RL) [43] in Computer Science, which designed algorithms based on these principles that can learn complex tasks [44, 45] and reproduce several aspects of animal learning [46, 47, 48, 49, 50]. Following Thorndike's transition to animals, neuroscience started to combine these behavioural assays with neural recordings to investigate the neural bases of learning. An influential discovery made using this approach was the encoding of reward prediction error (RPE), a key teaching signal in RL algorithms [51, 43], in dopamine neurons of the midbrain [52]. Remarkably, this proved how independent research in Computer Science developed without the intention to explain brain function could be related back to neuroscience and psychology. Since then, other elements of RL algorithms have been related to patterns of neuromodulator release across the brain [53, 54, 55]. Despite their success, it has been highlighted that the class of *model-free* RL algorithms used in this line of research cannot account for all aspects of animal behaviour [56, 57] and neural activity [58, 59, 60]. They fail particularly in tasks where animals build a model of the statistical structure in the environment and use it to inform decisions [61, 62, 63], instead of relying purely on associations between cues and outcomes. This has led to the distinction between *model-free* and *model-based* RL algorithms [64, 65, 43], and proposals of how the brain may be balancing the use of both in different regions [66, 67, 62, 68, 69, 70]. Important advances have been made in our understanding of how the brain builds such models of the environment, including its encoding of space [71, 72], learning of task schemas [73, 63], and consolidation of memories [74, 75, 76].

In recent years, neuroscientific research on learning has seen a proliferation of experimental data and accompanying models to explain it. Though very insightful, such models are typically designed to capture a subset of the phenomena present in the data, and often make contradictory predictions about what would be observed in different settings. Inspired by the success of the unification of theories in physics [77], theoretical neuroscientists have proposed overarching frameworks that encompass multiple previous results and resolve conflicting views. Using the language of (Bayesian)

statistics, a prominent proposal is that humans and animals learn by decreasing the *uncertainty* with which they predict the future [78, 79]. This notion forms the basis of multiple models used to understand learning [80, 81, 82, 83]. One such model is predictive coding (PC) [80, 84, 85], which proposes that sensory cortices learn by comparing bottom-up sensory information with top-down probabilistic predictions of the input and updating proportional to the prediction error. This enables higher-order sensory cortices to gradually improve their probabilistic models and thus decrease their prediction errors and uncertainty regarding the future. Signatures of such prediction errors and their underlying probabilistic computations have been observed both in behaviour [86, 87, 88, 89] and neural activity [90, 91, 92, 93]. Indeed, a number of predictions from the framework have been verified experimentally, such as the progressive suppression of neural activity with repetition of stimuli [94, 95, 96], the gradual alignment of spontaneous and stimulus-evoked neural activity [97] or the tendency of perceived/remembered stimuli to be biased towards the mean of previous observations (i.e. 'contraction' bias) [98, 99]. Contrary to RL, predictive coding suggests that learning also occurs in settings without explicit reinforcement, where agents learn to passively predict the statistics of the inputs. Such learning is termed 'statistical' learning, and has been observed in both humans and animals [100, 97, 101, 102]. In an attempt to unify predictive coding with models of action selection such as RL, the PC model has been extended by the seminal work of Karl Friston on active inference [82]. This is derived from his even broader free-energy principle [78], that has been proposed as a unifying theory extending principles of predictive coding to the entire brain. Active inference and the free-energy principle both have supporting experimental evidence [103, 104, 105], yet some debate the usefulness of such grand unifying theories to uncover the complexities of the brain [106].

Despite the success of the above probabilistic models at describing neuroscientific phenomena, the dominant framework in machine learning and artificial intelligence is that of connectionist deep neural networks [107]. The origins of connectionist thought lie once again in psychology, with early pioneers William James [108] and Edward Thorndike [40] proposing that learning occurs through the strengthening of connections between neural units. In 1943, Warren McCulloch and Walter Pitts designed the first formal model of an artificial neuron [109], and showed that networks of such neurons could implement any boolean function if wired together with appropriate connection weights. Finding the right weight setting for each task, however, remained an open problem until Donald Hebb proposed a learning rule for synapses between biological neurons [110]. This rule was based on his famous principle 'neurons that fire together, wire together', known as Hebbian learning. Inspired by Hebb's idea, Frank Rosenblatt adapted Hebbian learning to a supervised setting (i.e., teaching the network to produce a particular output for each input by minimising a 'loss' function) and designed the first trainable *shallow* (single-layer) artificial neural network that he termed the 'perceptron' [111]. Rumelhart, Hinton and McClelland subsequently discovered a method to train *deep* (multilayer) neural networks [10, 112] giving rise to the booming field of deep learning as we know it today.

Perhaps one of the principal advantages of deep neural networks over the aforementioned probabilistic models is that they are easier to implement and train with modern technology [113, 114]. In addition, they are universal function approximators [115, 9], meaning that they can solve any task expressible as a function from inputs to outputs — this makes them extremely versatile. A drawback of their flexibility is that they can often solve the same task in multiple ways, all achieving optimal performance (i.e. minimal 'loss') for the input-output examples presented during training, but differing in their ability to generalise to unseen inputs (i.e. 'generalisation error') [116]. As

such, details of the learning rule are very important in determining what solution will ultimately be learned by the network. The No Free Lunch theorem [117] demonstrates that no learning rule can be universally better than any other when performance is averaged across all possible tasks, and hence each learning rule must be tailored to the structure of the particular task being learned. This tailoring is commonly referred to as the 'inductive bias' of the learning rule [118]. Sometimes such inductive biases are implemented *implicitly* in the learning rule, such as the implicit regularisation effect of gradient descent [119]. In other instances, the inductive biases are introduced *explicitly* to constrain the space of solutions to have specific desirable properties [120, 121]. This is the case of, for example, convolutional neural networks [122, 123] where the weights of certain connections in the first (input) layer are forced to be equal throughout training so that they compute a translation invariant 'convolution' operation. The practice of constraining the solution space to enforce specific symmetries is known as geometric deep learning [124] and has characterised a broad set of inductive biases for a range of applications [125, 126, 127].

Like neural networks the brain is also very versatile, yet it possesses a remarkable ability to learn solutions that generalise well. Hence, a natural question is what inductive biases best describe learning in the brain. To investigate this, one needs to analyse the learning dynamics of both behaviour and neural recordings in biological agents from a theoretical perspective. Machine learning theory has developed useful analysis tools to study the learning dynamics of neural networks [128, 37, 129]. One of the first analytical descriptions of learning in neural networks [130] used a framework known as the 'teacher-student' set-up, where a 'student' network learns from input-output examples generated by a 'teacher' network. In the online setting where student weights are updated for every example, the authors derived differential equations quantifying the degree of teacher-student overlap throughout learning. Since then, theoretical work has made progress by removing non-linearities [37], limiting connectivity [131], using mean-field techniques [129] or studying learning in the limit of infinite network width [132]. Among many other insights, this work has identified learning regimes such as the 'rich' and 'lazy' regime [133], present also in human learning [134, 135], and has also highlighted the importance of fixed points in governing learning trajectories [24, 136, 137]. Through the work in this thesis, we show that concepts from the machine learning (ML) theory literature can be readily applied to study biological learning, and can provide deep insights into the brain's learning mechanisms and their neural implementation.

1.2 Long-term learning

Learning can occur over multiple timescales. An expert cook can learn how to make a new dish in as little as a single observation of its preparation. Similarly, an expert pianist can sight-read a piece without having ever played it before. On the other hand, acquiring expert cooking and piano playing skills in the first place can take several months if not years of learning. Such long-term learning is omnipresent in our lives, from when we learn how to read and write as children, to when we learn the core principles of human knowledge as adolescents, all the way to when we learn how to exercise our professions in adulthood. Long-term learning is not limited to the period of initial acquisition; individuals continue to refine their skills and knowledge well into 'expertise'. Hence, it is a process that underlies our entire lives and can influence important decisions such as what degree to study, career to pursue or hobbies to practice. At first glance, long-term learning may seem like such a complex and vaguely-defined process that its scientific study is impossible. With the multitude of factors that influence what and how each individual learns, from their innate talents to their economic condition, it appears like a difficult problem to gain traction on. Although a full

characterisation of the mechanisms at work may be difficult to attain, in this thesis we demonstrate that by isolating core processes in a controlled setting, we can get closer to understanding the principles governing long-term learning.

Apart from its long timescales, long-term learning displays several characteristics that distinguish it from more rapid forms of learning. First, it does not happen all at once. That is, as opposed to the often steady progress of short-term learning, long-term learning typically occurs in stages. Take the example of learning how to play the guitar. An individual may experience steady progress as they first learn to strum while pressing their fingers on the fretboard to play a chord. However, unless they are Paco de Lucia [138], they will experience a plateau in their progress before moving on to more complex playing techniques, such as plucking. This oscillation between quick and slow learning is characteristic of long-term learning trajectories, and gives rise to the frequently observed stages in learning [112, 20, 24]. It is interesting to note that many curricula also divide learning objectives into distinct chunks to be learnt in succession, as opposed to teaching everything at once [139]. Second, long-term learning trajectories are vastly diverse across individuals. This diversity is not only observed in the speed with which individuals transition through the aforementioned stages, but also in the diversity of strategies that individuals develop as they learn. It is delightful how you can give the same exact materials to a handful of children and watch them develop completely different strategies to solve the same task [140, 141, 142]. This diversity may seem random or too reliant on unobserved variables (such as the background of each individual or the nature of the task) to be able to understand. However, there is often a latent structure in both the diversity and progression of the learning trajectories that can be uncovered and studied. Lastly, not all individuals reach the same level of expertise over long-term learning, with some remaining 'stuck' at lower levels of performance than others. Thus, another one of long-term learning's distinctive phenomena is that of learning stagnation. It is not uncommon for individuals who start to learn a new skill to reach a point where they give up [143, 144, 145]. This rarely happens at the very start of learning, or during periods of quick progress, and instead often happens in the interval between having completed a learning stage and moving on to the next. In this thesis we discuss all three of these characteristics of long-term learning, and provide a formal theoretical explanation for them in Chapter 4.

Long-term learning has been difficult to study in laboratory settings precisely due to its long-term nature. It is difficult to find a task that is both complex enough to give rise to multiple strategies and be learned over long timescales, yet simple enough to be learned by a model organism that can be kept in a controlled setting for multiple days with little to no interruption. Consequently, most neuroscientific studies of learning have analysed the fine tuning of behaviour in trained animals adapting to changing experimental contingencies [146, 147, 148, 149, 150]. This has led to important discoveries concerning animals' responses to rewards [47, 151, 152, 153, 154], animals' ability to re-learn and adapt to changes in their environment [155, 156, 63], or the computations that animals employ to solve tasks [157, 158, 159]. Recent developments in task design for the widely used mouse model organism have yielded sophisticated tasks that mice can learn over tens of days without human intervention or complex training protocols [160, 161, 159, 63]. Further, the tasks are complex enough to give rise to appreciably diverse behaviour across mice, such that we can investigate its potential sources. Experimental methods are thus starting to enable the study of long-term learning trajectories in laboratory settings, providing powerful tools to isolate their core processes and reveal the principles that underlie them.

To study long-term learning in the lab, it is most effective to motivate animals with rewards.

Hence, the type of learning studied is necessarily reward-based. Admittedly not all instances of long-term learning may be driven by external reinforcement, however, it does not seem far-fetched to posit that much of what motivates animal learning can be conceptualised as a reward [162, 43, 163]. An added benefit of using reward-based tasks is that the relevant teaching signal can be precisely controlled over learning, reducing the influence of internal motivational signals which may be difficult to observe and vary widely across individuals. To further reduce confounds when studying learning in the lab, one can control the environment so that animals' perception is known and accounted for at all times. To do this perceptual decision-making tasks are frequently used, where animals have to make decisions guided by precisely-defined perceptual stimuli that often vary across trials [164, 165, 166]. Perceptual decision making can then be combined with reward-based learning by rewarding correct decisions and not rewarding/punishing incorrect decisions based on the stimuli presented in each trial. Past studies of such perceptual learning have made important discoveries concerning its underlying neural implementation, though its interactions with reward-based mechanisms are less well known, as described in the following Section 1.3. Taken together, we believe that reward-based perceptual learning provides a suitable experimental setting for the study of long-term learning in the lab.

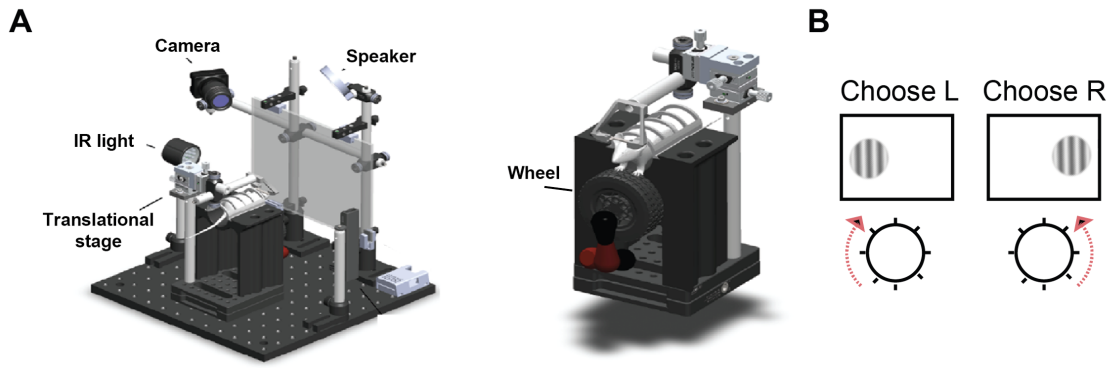


Figure 1.2: Experimental set-up for studying long-term learning trajectories. (A), Experimental rig used to train mice on the visual perceptual decision-making task. (B), Schematic of task trials with stimulus on the left and right. Wheels indicate the correct movement on each trial.

In this thesis, we study the long-term learning trajectories of mice learning to perform a visual perceptual decision-making task over multiple days (Figure 1.2). We adapted an existing task [160] such that we could collect useful data from the mice's entire learning trajectories from naïve to expert, shifting away from its previous focus on expert behaviour. One of the major changes we made was that of keeping the task unchanged throughout learning. We presented the full task from naïve to expert, with trials spanning all difficulty levels and the same reward contingencies as used in expert days. In this way, we removed the confound of experimentally imposed changes in behaviour as caused by common shaping protocols. Instead mice were allowed to form their learning trajectories independently, better isolating their intrinsic learning mechanisms. This approach also enabled a more accurate comparison of behaviour over days. We analysed the long-term learning trajectories that emerged from this experiment, and found that they display the characteristic features of diversity and systematic transitions through stages described above. Mice developed distinct strategies while learning the task; some displayed early biases to the left/right while others remained more balanced in the initial days of learning. Remarkably, this early bias (or lack of it) determined the strategies that each mouse developed later in learning. Mice persevered with each strategy before transitioning to the next, displaying the distinctive learning plateaus of long-term

learning trajectories. Further, several mice did not fully learn the task despite sufficient training. As such, our experiment was able to recreate principal characteristics of long-term learning and provided behavioural data from multiple individuals learning a complex task under controlled conditions that was amenable for the analysis of individual diversity (Chapter 2).

We next asked what could be the neural mechanisms responsible for the observed learning trajectories. The stereotypical nature of our perceptual decision-making task was well-suited for repeated measures of neural activity throughout learning. Hence, motivated by its likely involvement based on the literature, we collected dopaminergic data from the striatum as mice learned the task from naïve to expert. The results of the analysis are briefly described below and outlined in full in Chapter 3.

1.3 Learning in the brain

Perceptual and reward-based learning have traditionally been presented as two distinct processes engaging separate brain mechanisms. The former is mainly associated with sensory and association cortical circuits involved in perception and evidence integration [165], whereas the latter is concerned with frontal cortex and subcortical regions, such as striatum and midbrain dopamine neurons [167, 168, 48]. Owing to this separation, most studies investigating dopamine signals have been performed in so-called value-based (i.e. reward history-guided) decision tasks, where either no sensory stimuli are presented and animals choose actions based on reward history (e.g. two-arm bandit tasks) or the stimuli are readily distinguishable and animals track their 'value' based on association with past rewards [169, 48]. The RL framework has been immensely successful at explaining both behavioural and neural results from these experiments [162, 43, 170]. Phasic dopamine responses closely resemble the RPE term of RL models — the moment-by-moment difference between predicted and realised reward — which can be used as a teaching signal for learning to improve subsequent decisions [52, 171, 172, 173, 174, 175, 176].

Notably however, dopamine neurons are also active during perceptual decisions. In perceptual decision-making experiments, reward history alone does not provide an accurate estimate of upcoming outcome value. Instead, accurate decisions (and predictions about their outcomes) additionally require the evaluation of available sensory evidence, for example, discriminating between different sensory cues or detecting the presence/absence of a cue, which can be difficult due to weak sensory signals and perceptual noise. Furthermore, stimuli can be parametrically manipulated to enable precise control over sensory variables and investigate their effect on decision making. Studies across species and sensory modalities have shown that dopamine neurons, alongside several other nodes across the basal ganglia, respond to sensory stimuli and decision outcomes during perceptual decision making [177, 178, 49, 179, 180]. Given this marked involvement of dopamine in perceptual decisions, and its established role as a teaching signal in reward-based learning, we recorded dopamine signals from the dorsal striatum of mice as they learned a perceptual decision-making task from naïve to expert to investigate its role in long-term learning.

Generally, it has been observed that dopamine signals increase with the level of sensory evidence during stimulus processing and conversely decrease with sensory evidence during outcomes. For instance, dopamine-specific imaging of midbrain dopamine neurons in mice performing a visual contrast detection task revealed responses that were scaled by the contrast of the stimuli, being largest for high contrast stimuli — a relationship that reversed during reward delivery [49]. Further, monitoring dopamine release in different striatal regions in mice has confirmed a positive scaling of dopamine release with the strength of stimuli in visual, olfactory, and auditory decision tasks

[181, 182, 183, 184, 50]. Across many of these studies, the dopamine responses to stimuli encoded the accuracy of the pending choice: they were lower in error trials compared to correct trials after controlling for stimulus strength. These observations suggest that dopamine encodes the subjective sensory experience and perceptual decision uncertainty rather than the physical strength of sensory stimuli. Together, there is abundant evidence that dopamine signals are modulated by the strength of perceptual evidence during stimulus and feedback processing.

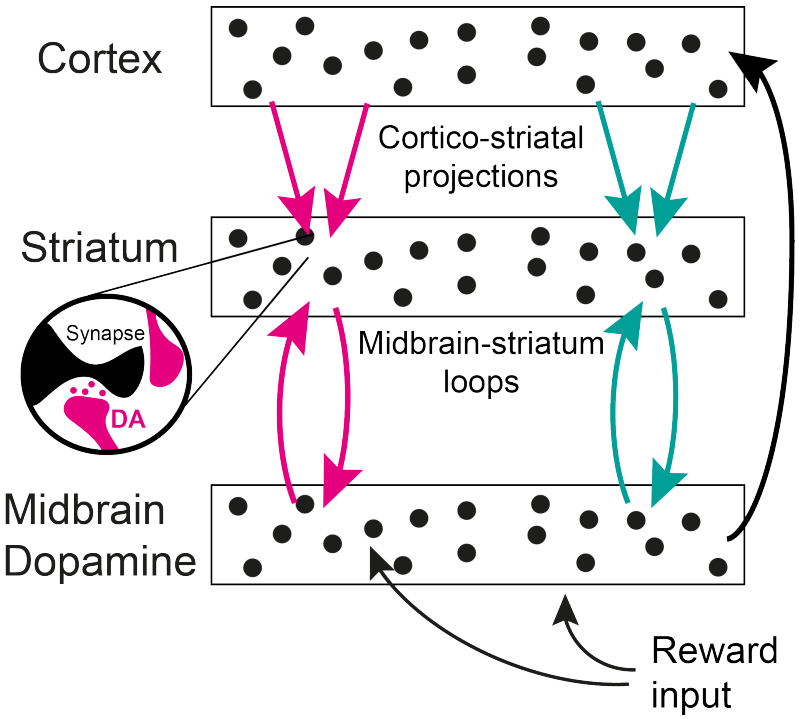


Figure 1.3: Schematic of anatomy relevant for dopamine signals in perceptual decision making. Sensory input represented in cortical regions is transmitted to the striatum via the cortico-striatal projections. The striatum then projects to downstream basal ganglia output nuclei and other motor regions to implement actions (not shown). Reciprocal projections between the striatum and midbrain dopamine neurons form functional closed midbrain-striatum loops (e.g. striatonigrostriatal connectivity) which, together with reward input, govern dopamine release in the striatum. Dopamine in turn modulates the strength of cortico-striatal projections. Midbrain dopamine neurons also project to the cortex, and likely play a reinforcing role there as well. Pink and aqua colours represent segregated channels for different cortical representations, e.g. pink for visual and aqua for auditory information.

The striatum receives dense dopaminergic innervation from midbrain dopamine neurons living in the substantia nigra pars compacta (SNc) and ventral tegmental area (VTA). Indeed, the connections between the SNc and striatum have been found to be reciprocal, with connections from neurons in the dorsal striatum in turn forming the majority of the input to SNc dopamine neurons [185], thus constituting the *striatonigrostriatal loops* of the midbrain (Figure 1.3). The striatum also receives diverse innervation from multiple cortical regions [186] through *cortico-striatal* projections, which have been found to play crucial roles in perceptual decision making [187]. Hence, this cortico-striatal and midbrain-striatum connectivity may be responsible for the dopamine signals during perceptual decisions described above. Perceptual stimuli are first represented in cortical regions deriving from sensory input, these representations are then transmitted to the striatum via cortico-striatal projections. Striatal activity is subsequently relayed to dopamine neurons in the midbrain which project back to striatum releasing dopamine around cortico-striatal synapses (Figure 1.3). Dopamine release promotes plasticity in these cortico-striatal synapses [188], which

could modify the strength of associations between cortical representations and actions implemented by the striatum and its downstream projections to basal ganglia output nuclei. This mechanism could thus underlie dopamine's well-documented reinforcing role on behaviour [52, 189, 190, 191]. An important feature of cortico-striatal connectivity is that not all cortical regions project to the same subregions of striatum. In fact, multiple past studies have demonstrated that different subregions of striatum are involved in the processing of different sensory input such as dorsomedial striatum for visual [182] and the tail of striatum for auditory input [183]. As such, there exist segregated channels for different cortical representations to reach the striatum that could similarly underlie the heterogenous tuning of dopamine neurons observed in recent studies [192]. The function of such heterogenous dopamine responses remains an open question, and challenges previous accounts of dopamine as a single global RPE signal broadcasted all across the brain.

The ubiquitous nature of dopamine responses during perceptual decision-making begs the question of what role they play in this process. Are they important for the current decision process or subsequent learning? Dopamine responses to stimuli are present before the behavioural manifestation of a perceptual decision, for example, a saccade [193]. However, consistent with value-based studies [194], these signals do not appear to have substantial and consistent effects on impending decisions. Optogenetic manipulation of dopamine signals during stimulus processing has provided differing results: manipulating midbrain dopamine neurons did not influence impending visual decisions [49], but manipulating dopamine axons projecting to posterior striatum influenced auditory decisions [195]. With regards to its effect on subsequent learning, recent studies in expert animals have suggested that dopamine signals play a significant role in learning to fine tune perceptual decisions [49]. Despite such discoveries, our understanding of the role of dopamine in long-term learning of perceptual decisions has seen little progress, as most studies have focused on well-trained animals refining their decisions over short timescales. Long-term learning presents several challenges for an individual, and understanding how the brain tackles these could yield important insights into the brain's learning mechanisms. For example, one of the first learning problems for a naïve individual is to discover which stimuli (states) are relevant for the task. Here we provide evidence for the progressive discovery of states in mice learning a perceptual task from naïve to expert, and we find that striatal dopamine signals reflect the evolving state representations, showing key characteristics of a teaching signal sculpting individuals' learning trajectories. We further demonstrated their role in learning through optogenetic manipulations that provided causal evidence for our claims.

Standard RL models commonly used to explain (well-trained) animal behaviour cannot account for our results because they assume that the relevant states have already been discovered, such that reward-driven updating of state–action associations can take place from the outset. However, as described briefly in the next Section 1.4 and outlined fully in Chapter 4, a deep RL model could capture both the learning trajectories and dopamine signals of mice learning from naïve to expert. Our model incorporates flexible state representations to capture the progressive state discovery and proposes that striatal dopamine reinforces associations between states and actions only if the states have been discovered. Taken together, examining dopamine signals can hence provide valuable insights into both the behavioural and neuronal bases of learning from naïve to expert.

1.4 From a model to a theory

Deep learning methods are revolutionising the fields of ML and artificial intelligence (AI) with their ability to learn almost any input-output map we can create training data for. Machine learning

had been going through periods popularly referred to as the 'AI winters' for several years, where little progress was being made in using the rapidly-advancing computing methods to reproduce human intelligence in machines. The first such winter happened in the 1960s-1970s, when early optimism in symbolic AI [196] resulted in frustration as systems failed to scale, and hardware limitations hindered performance. This led to a general disillusionment and the withdrawal of funding from many research programmes focused on AI. Interestingly, the second winter came in the 1980s-1990s, when the discovery of neural networks had already started to show promise [197], but the systems were unable to handle complex, real-world problems. The culprit of this second hiatus was mainly the lack of a suitable learning algorithm to train neural networks with more than one layer after the input layer (i.e., deep networks). This second winter began to thaw when, in 1986, Geoffrey Hinton, David Rumelhart and colleagues designed the backpropagation algorithm [10]. 'Backprop' enabled gradient-descent methods to be implemented in deep neural networks with many, arbitrarily wide layers without constraints on their connectivity and with flexibility in the choice of neuron activation functions. The combination of this learning algorithm with advances in computer hardware (such as the optimisation of GPUs) and the advent of 'big data' collected from the internet/digital devices gave rise to the deep learning revolution we are experiencing today.

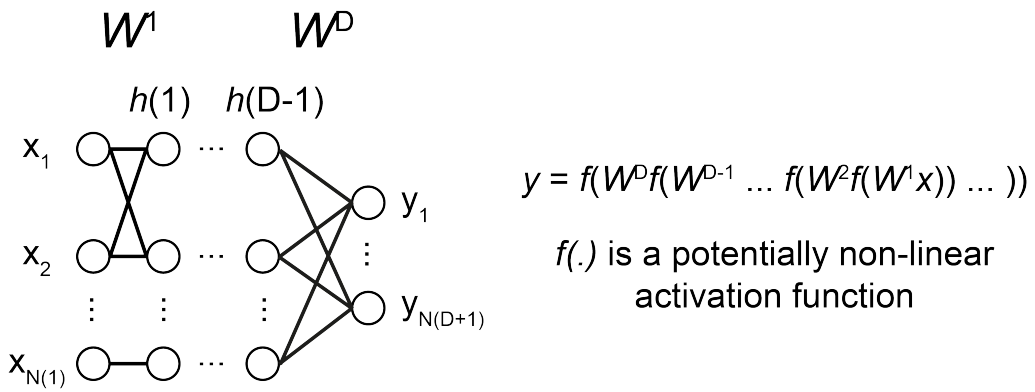


Figure 1.4: Schematic of a deep neural network. A deep neural network that processes an input $x \in \mathbb{R}^{N(1)}$ through layers $h(1)$ to $h(\text{D-1})$ with D sets of tunable weights W^1 to W^{D} producing the ultimate output $y \in \mathbb{R}^{N(D+1)}$. Neurons have a potentially non-linear activation function $f(.)$, applied element-wise.

In addition to being a remarkable tool for industry, deep learning methods are also enabling important contributions across multiple scientific disciplines. Indeed, the Nobel Prize in Chemistry from this year has been awarded to Demis Hassabis, the CEO and co-founder of DeepMind. The team at DeepMind designed the deep-learning based model AlphaFold [198] which has had a tremendous impact on chemistry research due to its ability to predict the structure of proteins with unprecedented accuracy. Given the uncanny similarity between the architecture of deep learning systems and the brain, they seem likely to be a promising tool for the Neuroscience community to study the brain [199]. However, they have proven to be a double-edged sword for neuroscientists, as their reliance on scale and complex representations to solve real-world problems has made them nearly as mysterious and difficult to comprehend as the brain [200]. The field of deep learning theory has emerged to tackle this problem, aiming to uncover the core principles that enable deep learning systems to achieve the feats that they do. Deep learning theorists have a significant advantage compared to neuroscientists, which is that they have access to details from the precise implementation of the systems they study. As such, powerful analysis methods have emerged from deep learning theory, ranging from the characterisation of learning dynamics [37, 201], to the study

of implicit and explicit biases [202, 203], to the effects of symmetries and other structural motifs on the operation of deep neural networks [204, 205]. Hence, although deep neural networks themselves may be black boxes like the brain, they have given rise to analysis methods from deep learning theory that could help shed light on the computational principles at work in the brain.

The distributed representations that evolve during the training of deep neural networks seem ideally suited to capture the progressive discovery of strategies throughout long-term learning. Previous work analysing the learning dynamics of deep neural networks has demonstrated that their 'depth' indeed gives rise to the characteristic learning plateaus and stage-like transitions of long-term learning trajectories that we described in Section 1.2 [37, 24]. As such, we decided to explore the capabilities of deep learning as a framework to study the long-term learning trajectories of mice performing the perceptual decision-making task described in the previous sections. This was an unconventional choice, as the standard modelling framework used to study reward-based tasks such as ours, especially if recording dopamine signals, is RL with predetermined states and potential variations on its standard learning algorithm [43, 52, 206, 49, 207]. Keeping this in mind, we opted for a combined approach using a *deep RL* system as the model to capture mouse learning trajectories. Deep RL systems have enjoyed great success in tackling complex problems in the past, such as learning how to play multiple board games at super-human level [208, 209]. One of their principal strengths is that they can take advantage of the rich representational structure expressible in the multiple layers of deep networks, while not requiring explicit supervised feedback of their outputs to learn them. Instead, deep RL systems learn from a unidimensional reward signal that indicates how good or bad their current policy (i.e. distribution over actions for each state) is [210]. They use this signal to update parameters and refine their policy to maximise total reward consumption in the future. Despite their achievements, they are known for having several drawbacks such as their sample complexity and tendency to exploit strategies developed early in learning unless explicit exploration bonuses are provided. They are hence not used as commonly as deep learning systems trained with supervised learning methods. This also reflects in the extent of theoretical work that has been performed to explain their underlying principles, which has progressed much less than for supervised learning. Taken together, the decision of employing a deep RL system seemed adequate due to the nature of the task and the observations we had made regarding mouse learning trajectories. However, it could have led us down a rabbit hole if we ended up with a model that either learned nothing like the mice or was too complex to understand with current theoretical tools.

To hedge our bets, we decided to strip away as much complexity from the deep RL model while maintaining its core features of evolving representations and reward-based learning that we deemed important to capture mouse data. We thus reduced the number of hidden layers to one, removed non-linear neural activation functions, reduced the dimensionality of the visual input to a simple one-hot encoding and modelled only choice data and not reaction times or wheel movements. This left us with the simple deep RL network described in Chapter 4. Although this seems like an overly simplistic model, it provided us with an interpretable system that possessed the core features of depth, reinforcement learning and gradient-descent based updates. The choice of using a simple network had several practical advantages too, enabling quick training of multiple agents and storage of their weights all throughout learning. To capture the striatal dopamine signals, we also designed a novel learning algorithm employing heterogenous RPEs to update the network's weights. Once again, the reduced dimensionality of our network made it much easier to both design and explain the mechanism of this unusual learning rule, termed the 'tutor-executor'. When

trained on the task in exactly the same way as the mice, our deep RL 'tutor-executor' model was able to capture both the diversity and systematicity of mouse learning trajectories. Further, its heterogenous RPEs mimicked the evolving dopamine signals over learning, and thus provided a formal characterisation of their function in long-term learning. Surprisingly, the model also learnt in a similar number of trials as the mice. To verify the role of dopamine signals proposed by the model, we performed both optogenetic inhibition and stimulation experiments which yielded results in line with predictions from the same manipulations in the model. Taken together, our model demonstrated that the deep RL framework is well-suited to capture long-term learning trajectories of a perceptual decision-making task for mice.

Thanks to multiple important discoveries from computational neuroscience in the recent past, the field has begun to advance rapidly. There now exist many modelling frameworks to capture brain processes, ranging in level of abstraction from spiking to rate models; in the degree of in-built structure from the more probabilistic and model-based [78, 211, 212, 213], to the frequentist and model-free [214, 215]; and in the flexibility of neural representations from deep learning to symbolic AI systems. Hence, a vast toolkit is available to the computational neuroscientist when analysing a new dataset, and it is often, if not always, the case that the data collected is not sufficient to uniquely constrain the 'correct' modelling framework. In fact, it is more likely that either multiple frameworks are able to capture the phenomena of interest, or that different frameworks can be used to study different phenomena present in the data. As such, now more than ever, it seems important to identify the principles underlying the phenomena rather than banking on the performance of each model to select the right modelling framework. To achieve this, it is not sufficient to show that a particular model or modelling framework is able to capture the data being analysed, or that it captures it better than others. It will never be possible for a computational neuroscientist to test all the modelling frameworks and identify those that work and those that don't. Instead, one has to go a step further and understand *why* a particular model may be capturing the data, and why others may not be able to. Obtaining in this way a more qualitative understanding of the mechanisms at work can enable a more holistic perspective and facilitate model selection (or creation!). In addition, it can better guide experimental investigation into the precise implementation of the phenomena in the brain. Empirical data can then serve to more effectively distinguish between models; instead of relying on *how well* a particular model fits the data (as all models can be refined, fit better, extended etc. and further, all models are wrong [216]) one can rather ask whether the brain data possesses the qualitative signatures predicted by the underlying principles of the modelling framework being tested.

In the work presented in this thesis, we analysed the principles governing the deep RL 'tutor-executor' model by performing a theoretical analysis of its learning dynamics. We employed the approach from [37] of deriving closed-form expressions for the average dynamics of the network's weights over learning and analysed their fixed points (i.e., network configurations where the dynamics go to 0). The fixed points of a dynamical system can often reveal important structure in the system's evolution over time, and can be stable, marginally stable or unstable. The fixed-point analysis of our model's average dynamics revealed a hierarchy of saddle points that governed the learning trajectories. Saddle points have both stable and unstable manifolds and in their vicinity, learning temporarily slows down along the stable manifold until the unstable manifold is discovered and learning speeds up again. Thus, the saddle point structure of our model was able to explain why long-term learning trajectories of both mice and model simulations displayed their characteristic learning plateaus and stage-like transitions through strategies. Further, the saddle

point structure spanned the diversity of strategies observed across mice, and the heteroclinic orbits connecting saddle points defined the same systematic transitions between early and late behaviour. Our fixed-point analysis could also be used to efficiently determine whether a model captured the mice's learning trajectories: if the model did not possess a similar saddle point structure, it would be unable to capture the patterns of diversity and systematicity that we observed across mice. This approach allowed us to rule out 'shallow' versions of the RL model, without multiple layers of tunable weights. Uncovering the principles governing the model's learning trajectories hence provided us with a framework to understand the patterns of diversity and systematicity in mice, indicating that they emerge from a (deep) learning algorithm with a hierarchy of saddle points in its dynamical landscape.

1.5 Outline of the thesis

In this thesis we address the questions outlined in the previous sections by examining learning trajectories and longitudinal dorsal striatal dopamine signals in tens of mice learning a visual decision-making task from naïve to expert. Chapter 2 shows how mice improved their performance by transitioning through sequences of strategies that became more stimulus-dependent over time. This sequence varied widely across individuals. However, the strategy transitions were remarkably systematic; each animal's future behaviour could be predicted weeks in advance. Chapter 3 describes how dorsal striatal dopamine signals reflected each individual's learning trajectory, showing strikingly similar patterns of diverse yet systematic transitions. These signals reflected prediction errors suitable to shape the intermediate strategies employed from naïve to expert. Optogenetic experiments provided causal evidence for the specific learning effects of these dopamine signals, which were distinct from the effect of reward on learning. The first part of Chapter 4 demonstrates that the learning trajectories, the dopamine signals and their optogenetic effects cannot be explained by standard 'shallow' RL models with predetermined state representations. Instead, our results suggest that mice refine their state representations throughout learning, a process well-captured by gradient-based RL in a *deep* neural network. Further, we demonstrate that networks require heterogenous teaching signals in order to capture the dopamine signals and their optogenetic effects. Lastly, the second part of Chapter 4 explains why our deep RL model reproduces the mice's diverse yet systematic stage-like transitions between behavioural strategies from naïve to expert, showing that they emerge from a hierarchy of saddle points. We conclude by demonstrating how the saddle point structure of a deep learning model provides a general framework for understanding long-term learning within and across individuals.

Chapter 2
Long-term learning trajectories

Chapter contents

2.1 Introduction 16
2.2 Methods 17
2.2.1 Mice 17
2.2.2 Behavioural task 18
2.2.3 Video monitoring 19
2.2.4 Behavioural data analyses 19
2.2.5 Statistics 21
2.3 Results 21
2.4 Discussion 29

Decades of experimental neuroscience research have discovered many of the neural mechanisms that enable animals to perform complex tasks. However, the mechanisms governing the learning of these tasks in the first place remain mostly unknown. Similar to the deep learning revolution, which has enabled computers to make the jump from being "computing" devices that execute pre-written code to being able to adapt their computation to changing circumstances, neuroscience is starting to indulge in learning too. The brain's remarkable ability to learn is one of its distinguishing features, and understanding it better could lead to important consequences in fields ranging from medicine, to education, to technology. Long-term learning from birth to adulthood and all the way to senescence is particularly relevant, as it defines multiple aspects of our lives from our careers to our hobbies and skills. However, its slow nature, high complexity and substantial individual diversity have made it difficult to study in the past. Here we present our attempt at simplifying the long-term learning problem by bringing it into a controlled environment with reduced dimensionality and faster timescales. We studied the learning trajectories of mice as they learn to perform a visual decision-making task over multiple days. Remarkably, we found that despite substantial individual diversity the learning trajectories are highly systematic: behaviour late in learning could be predicted several weeks in advance. We further show that mice exhibited distinct stages in their learning trajectories, corresponding to the diverse behavioural strategies that they transitioned through from naïve to expert. Mice persisted with these strategies for several days before progressing towards the next, giving rise to stage-like transitions reminiscent of the learning plateaus often observed in humans [20] and deep learning models [37].

2.1 Introduction

Past neuroscientific studies have often operationalised learning as the fine tuning of an already learned task over tens of trials in highly trained expert animals. This has enabled the study of adaptation to manipulations of the task structure [146, 147, 148, 149, 150, 155, 156, 63], analysis of the solution strategies employed by animals [157, 158, 159, 217], as well as the characterisation of neural circuits involved in solving the task [165, 167, 168, 48]. However, expert animals tend

to exploit their knowledge of the task, leading to only small changes in behaviour and negligible individual diversity. Long-term learning, in contrast, involves learning a new task from scratch, without any knowledge of the task structure, relevant environmental states or reward contingencies. Thus, its higher complexity and longer timescales often lead to substantial diversity in strategy transitions across individuals through learning.

A rigorous, controlled study of long-term learning has been difficult to achieve in the past. This is mainly due to the large number of animals and long training periods that are required to afford sufficient individual diversity and reduce noise in neural recordings, while keeping the experiment tractable. Here we overcome these obstacles through the use of a highly-stereotyped visual decision-making task that mice could learn over 20-30 days. The task was simple enough that mice could learn it over a reasonable period of time, yet complex enough that mice could develop different strategies. Crucially, we kept the task unchanged across training days, such that animals experienced the same task structure and contingencies from naïve to expert. This differs from the training protocols commonly used for this task, which shape animals by gradually increasing the task difficulty [218]. The static nature of our task enabled mice to form their own learning trajectories without experimentally-imposed changes, facilitating individual diversity and ensuring that changes in behaviour are a consequence of the animals' internal learning mechanisms. Further, it enabled an accurate, standardised comparison of day-by-day behavioural measures and neural recordings.

Despite substantial individual diversity across mice, the learning trajectories were highly systematic. We defined multiple behavioural metrics to analyse the mice's learning trajectories, such as their biases and psychometric performance over learning. Using these, we discovered that mice transitioned through sequences of distinct strategies over learning, differing considerably across mice. Further, when comparing early behaviour to late behaviour, we found a systematic relation between each mouse's early bias and their performance later in learning. Mice persisted with each strategy before progressing towards the next, giving rise to stage-like transitions in their behaviour over learning reminiscent of 'sudden-insight' events reported by previous studies [219, 220]. Some mice failed to learn the task despite similar training times, remaining 'stuck' at earlier strategies without transitioning further. Due to their key role in defining late behaviour, we also investigated the source of the mice's early biases. A simple modelling approach indicated that they emerge mainly from the early history of successful and failed decisions, with a potential influence of initial (pre-task) biases but little influence of day-by-day variation in experimental conditions. Thus, the factors that contribute to the diversity of mouse learning trajectories seem to derive from a combination of nature and nurture, yet we are able to define them more precisely thanks to our controlled setting.

2.2 Methods

2.2.1 Mice

The behavioural learning data was collected from 40 (30+10) male wild-type C57/BL6J mice, with their age ranging between 9 to 30 weeks. The first 30 mice were wild-type mice that were trained on the learning experiment from naïve to expert (i.e., learner mice). The next 10 mice were also trained on the task, but did not reach expertise despite more than 4300 trials of training (i.e., non-learner mice). All experiments were conducted according to the UK Animals Scientific Procedures Act (1986) under appropriate project and personal licenses.

2.2.2 Behavioural task

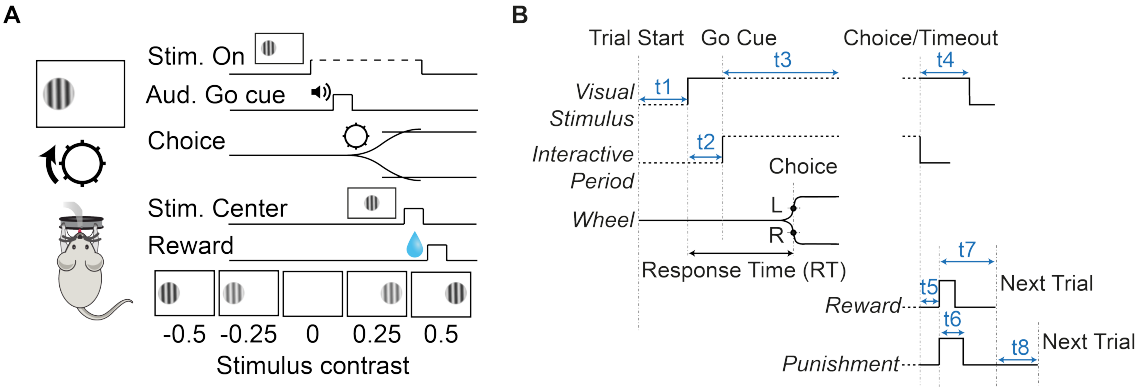


Figure 2.1: Details of the behavioural task. (A), The visual decision task for head-fixed mice (n=30). (B), The temporal structure of the task within a trial. Words in regular font indicate trial events, words in italics are labels for the traces in the timing diagram, and the solid vs dashed style of the traces indicate fixed vs variable time periods respectively. The duration of each labelled time period is provided in the paragraph below.

We trained mice in a complete psychometric visual decision-making task from day 1 until expertise. The task was further developed from an established task for head-fixed mice [160]. Following surgery recovery, mice were first habituated to the experimenter for 2-3 days, followed by 2-3 days of habituation to the experimental rig and head-fixation. In each day of the experiment, mice were head-fixed with their body and hind-paws resting on a stable platform with a covering, and their forepaws resting on a steering wheel that could be rotated left and right (Figure 2.1A). Each trial began after the wheel was held still for a short quiescence period (t1=0.7-0.8s) (Figure 2.1B). A sinusoidal grating stimulus of varying contrast (0%, 25% and 50%) was presented on either the left or right side of a screen 10cm in front of the mouse, followed by an auditory go cue t2=0.2s after visual stimulus onset. The go cue indicated the start of the interactive period, during which wheel movements were coupled with movement of the visual stimulus on the screen. It followed stimulus onset to encourage a fixation period and enable neural recordings to distinguish between responses to the visual stimulus and auditory cue. The mouse was required to indicate the position of the stimulus within a limited response time (t3=30s) by steering the wheel in the correct direction to move the stimulus to the centre of the screen, causing a water reward (3ul drop) to be delivered via a spout positioned close to the mouth. While the stimulus was held in the centre of the screen for t4=1s, a variable 'feedback delay' (t5=0.1-0.3s) separated the time of choice completion from reward delivery. Subsequently, the next trial started following an 'inter-trial interval' (t7=2-3s). When an incorrect choice was made, the mouse was presented with a t6=0.5s white noise auditory stimulus feedback via speakers positioned near each ear and had a brief timeout period (t8=2s) before the next trial. The contrast of the grating stimulus changed randomly across trials (with equal probability for 25% and 50% contrast and left/right stimuli), making the trials easier or harder. 10% of trials did not have a stimulus (i.e., zero-contrast trials), here animals were rewarded randomly (50/50) regardless of their choice. When a mouse responded incorrectly to an 'easy' high-contrast stimulus (50% contrast), there was a 50% chance that the same stimulus was repeated in the next trials, until the mouse responded correctly. Mice were trained for approximately 30 minutes each day. The behavioural experiments were delivered by custom-made software written in MATLAB (MathWorks) which is freely available [221].

2.2.3 Video monitoring

The left eye was monitored with a camera (Teledyne Flir CM3-U3-13Y3M-CS) fitted with a zoom lens (Thorlabs MVL50M23) recording at 20 Hz. Front body movements were monitored with another camera (same model but different lens, Thorlabs MVL16M23) also recording at 20 Hz. Mice were illuminated with infrared light (850nm, BW BWIR48) for the recording of eye and front body movements. Moreover, the box was lit by dim visible light so that the mouse's pupil was moderately dilated.

2.2.4 Behavioural data analyses

Behavioural data pre-processing

The behavioural data was pre-processed by removing the following trials: trials with response times more than 2 standard deviations above the mean per session, repeat trials (trials repeated after incorrect high contrast trials, see Section 2.2.2) and trials where mice did not make a choice in less than 30s.

Behavioural metrics

The main behavioural metrics we used to analyse the mouse trajectories were accuracy, psychometric slope and bias. Accuracy was defined as the proportion of rewarded choices in all trials except for those without stimuli (i.e., zero-contrast trials), where choices were rewarded randomly. Psychometric curves were calculated per session, and plotted the proportion of rightward choices (i.e., P('Right')) for different stimulus positions (left, right) and contrast values (0, 0.25, 0.5). The value of psychometric slope we used in all our analyses comes from a simplified psychometric curve collapsing across contrast levels to give left stimulus, zero-contrast and right stimulus x-axis values. Left (right) slope was defined as the absolute difference in P('Right') for left (right) stimulus and zero-contrast trials (Figure 2.2). Bias was defined as the difference between the P('Right') on zero-contrast trials and 0.5, thus representing the imbalance of choices on zero-contrast trials in left and right directions.

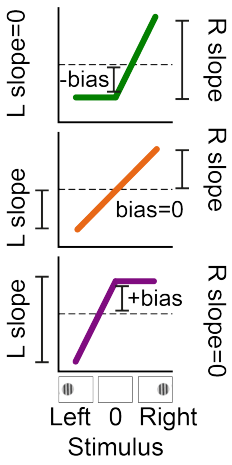


Figure 2.2: Schematic of behavioural metrics.

Learning trajectories

Each mouse was assigned a colour and cluster based on its learning trajectory. The rule for assigning colours used a weighted average of the difference between right and left per-session slopes over learning, where the weighting was equal to the sum of the left and right slopes on each session (Figure 2.3). The resulting average slope asymmetry metric was used to determine a colour for each mouse on a spectrum ranging from purple (for negative values) to orange (for values around 0) to green (for positive values).

The trajectories of the behavioural metrics were smoothed for better visualisation, highlighting their slow variation over learning. To do this, we used scikit-learn's Gaussian process regression package [222] to fit a Gaussian process with an RBF kernel (with tunable scaling and length-scale) to the session-by-session metrics. The predict method of the fit Gaussian process class could then be used to estimate the smoothed value of the metric at different time points over learning.

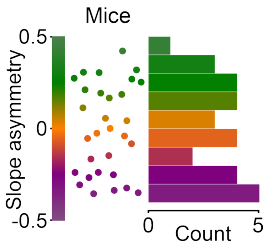


Figure 2.3: Colouring trajectories by slope asymmetry. Colour scheme (left), scatter (middle), and histogram (right) of average slope asymmetry per mouse.

A cluster label was assigned to each mouse to obtain cluster averages that highlighted the main trends in the diversity across learning trajectories. A dynamic time warping clustering algorithm was used to obtain these clusters [223]. This algorithm first looks for the time warping that best clusters the trajectories by shape. The cluster centres are then computed as the barycentres with respect to the time warped mouse trajectories, yielding cluster centres that are similar in shape to the individual trajectories, thus solving the problem of averaging across mice that learn at different speeds. This clustering was applied to the smoothed right vs. left slope trajectories in Figure 2.9E and the resulting cluster labels were used to compute the cluster averages in all other behavioural and neural plots throughout the paper. The same colouring, smoothing, and clustering methods were applied to the model simulations to obtain plots similar to those produced for the experimental data.

Pupil analysis

We used DeepLabCut [224] to track several points on the mice's left pupil throughout each task trial. We selected 4 points in the top, bottom, left and right portions of each mouse's pupil and recorded the x and y coordinates of each point over time (Figure 2.4). For our pupil motion analysis, we defined the average x and y coordinate of these 4 points as the position of the pupil and investigated its horizontal (Δx) and vertical (Δy) motion. For the pupil dilation analysis, we defined the pupil diameter as the mean of the Euclidean distances between the top & bottom and left & right points. The average x and y coordinates, as well as the pupil diameter measure, were z-scored and smoothed using a low-pass filter per session to enable more accurate comparisons across days. To reduce noise, we excluded sessions where the standard deviation of any of the non-zscored x-coordinates, y-coordinates or pupil diameter were above their 75th percentile. The alignment to stimulus onset and time warping was subsequently performed as for the neural data (Section 3.2.8).

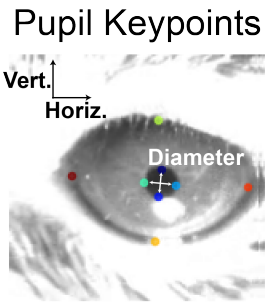


Figure 2.4: Pupil keypoints.

Lick analysis

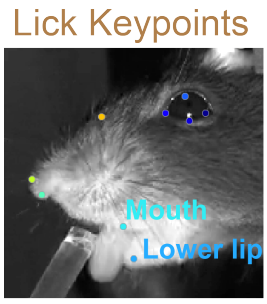


Figure 2.5: Lick keypoints.

We used FaceMap [225] to track the mouth and lower lip keypoints on videos of the mice's front body throughout each task trial (Figure 2.5). We recorded the x and y coordinate of each keypoint and defined a mouth-lip (Euclidean) distance measure that was used to detect licks. Licks were detected by applying the scipy.signal.find_peaks function to the mouth-lip distance with a minimum required prominence of 1 pixel. A lick rate measure was subsequently defined using a moving average with a window size of 4 (i.e., 0.2s), and was smoothed per session using a low-pass filter. The alignment to stimulus onset and time warping was subsequently performed as for the neural data (Section 3.2.8).

Wheel movement analysis

Wheel movements were recorded using a rotary encoder acting as the wheel's axel. The wheel position was saved in mm with a frequency of 200Hz, increasing for clockwise motion and decreasing for counterclockwise motion. The recordings were aligned to trial events and time warped as for the neural data (Section 3.2.8).

2.2.5 Statistics

The sample sizes are higher than sample sizes typically used in the field, due to the study's emphasis on individual variability. No statistical methods were used to determine sample size. Trial types (i.e., stimulus side and contrast level) were randomly determined by a computer program during behaviour. Details of all statistical analyses are provided in the figure legends. Analyses were performed using custom code written in Python 3 using standard analysis libraries: numpy, scipy, statsmodels.

2.3 Results

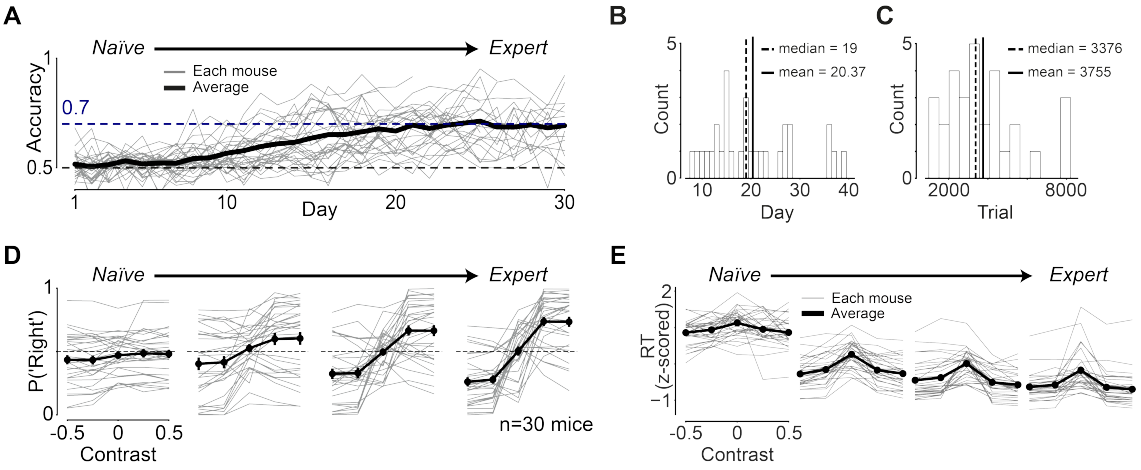


Figure 2.6: Learning a visual decision task from naïve to expert. (A), Accuracy over days per mouse (gray) and averaged across all mice (black). Dashed lines indicate chance level (black) and 70% accuracy level (blue). Number of days limited to 30. (B), Histogram of the number of days that mice required to reach 70% accuracy. (C), Histogram of the number of trials that mice required to reach 70% accuracy. (D), Psychometric curve over quartiles per mouse (gray) and averaged across all mice (black). Quartiles are defined per mouse by dividing days into 4 groups, with any remainder added to the last group. Negative (positive) contrast values indicate stimuli presented on the left (right) side of the screen. P('Right') indicates the probability of reporting 'right' side stimulus position. Error bars indicate +/- SEM across mice. (D), Chronometric curve over quartiles per mouse (gray) and averaged across all mice (black). RT (z-scored) indicates the mean z-scored response time (time from stimulus onset to choice completion) per contrast value. For each mouse, the z-score was computed by first finding the mean RT per contrast across trials in each day, these values were subsequently standardised across days and contrast values and finally averaged across days in each quartile. Error bars indicate +/- SEM across mice.

We trained 40 mice over multiple days with a single session (approximately 250 trials) each day. 30 of them learned the task reaching accuracies of at least 70% after many days of training (Figure 2.6A-C; median days = 19, median trial number = 3376), showing progressively steeper psychometric curves (Figure 2.6D), and faster choice response times (Figure 2.6E). Mice displayed a substantial spread in learning times, with some taking as little as 8 days to reach 70% and others close to 40 days (Figure 2.6B). Further, the mice's response time (RT) exhibited a considerably

different evolution over learning compared to the psychometric curves and accuracy. RTs experienced a sharp initial decrease followed by a prolonged gradual decrease (Figure 2.6E), whereas the accuracy and psychometric curves took longer to develop initially and then exhibited a sustained evolution until saturation (Figure 2.6D).

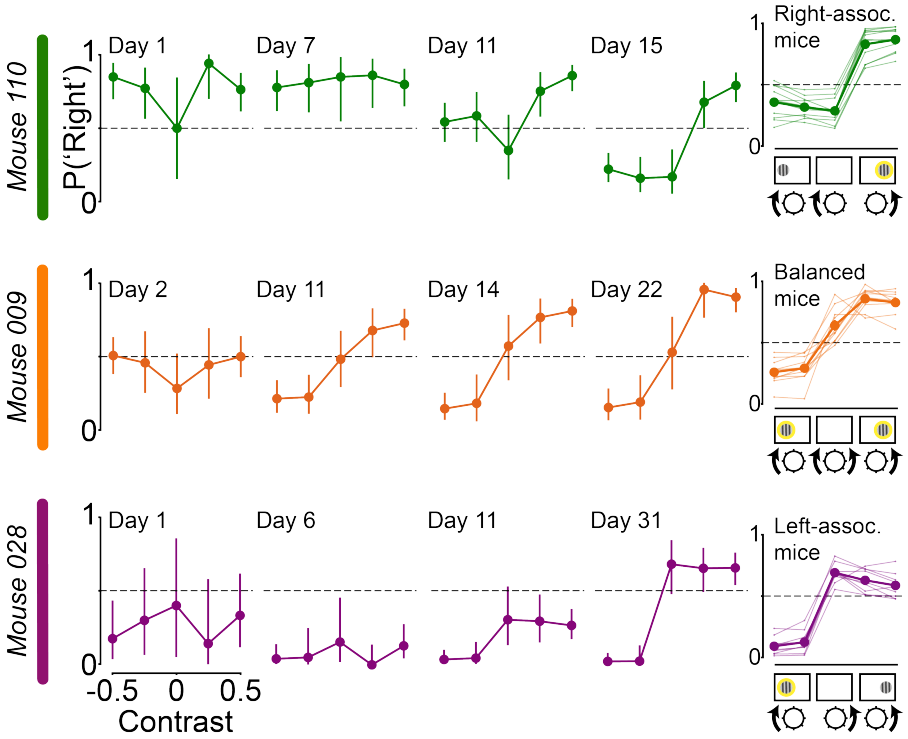


Figure 2.7: Mice develop diverse strategies through learning.
Left, psychometric curves from 3 example mice on 4 example days throughout learning. Error bars indicate the 95% confidence interval of a two-sided binomial test on P('Right'). See Figure 2.10F for corresponding chronometric curves. Right, per mouse (thin) and average expert psychometric curves clustered by trajectory type (thick): right-associating (green), balanced (orange) and left-associating (purple). See Figure 2.10G for corresponding chronometric curves. Cluster labels for each mouse were obtained from panel C, and colours were obtained from Figure 2.3.

Despite the apparent simplicity of the task, mice developed a diversity of strategies throughout learning. The general trends presented in Figure 2.6 reveal interesting patterns in the learning trajectories of the mice, such as the dynamics of accuracy and other psychometric measures. However, averages across mice hide important information about the diversity of behaviours that each mouse transitioned through before reaching expertise. Animals often find unexpected ways to solve the tasks that scientists give them, and this was the case for our experiment too. It turns out that the binary nature of the task allows it to be solved (i.e. achieve maximum accuracy) in three ways. The first is the way one would expect it to be solved, i.e. by choosing left when the left stimulus is presented, right when the right stimulus is presented and acting randomly on zero-contrast trials without stimuli. However, it can also be solved by using only one of the two stimuli to inform choices and making the other choice in its absence. In the case of only using the right stimulus, this would result in choosing right when the right stimulus is presented and choosing left when either the left stimulus or no stimulus is presented. Mice indeed found these three ways to solve the task and reached them by following different paths through behavioural space (Figure 2.7). The mice's learning trajectories and the sequence of behaviours that composed them are described in the subsequent paragraphs.

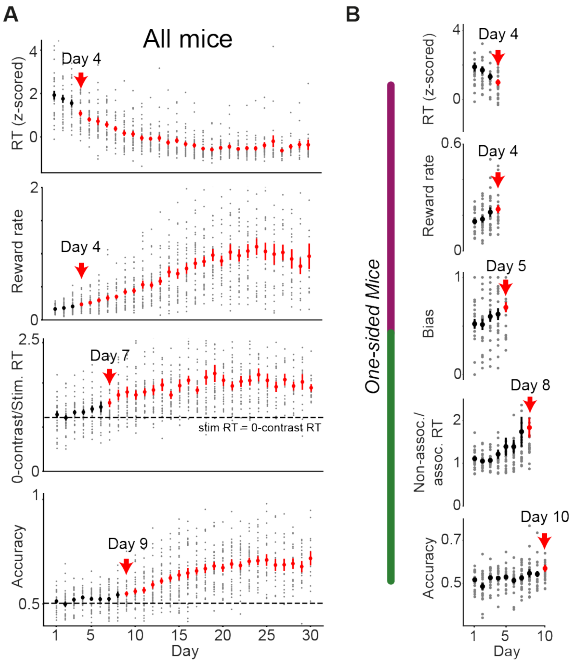


Figure 2.8: Behavioural metrics evolve over learning.
(A), Measures of general RTs (averaged across trials of all contrast levels and side), reward rate, the ratio of RTs on zero-contrast trials and trials with stimuli, and accuracy over the first 30 days of learning. First plot, z-scored general RTs over training days per mouse (gray) and averaged across all mice (black-/red). Red dots indicate data points significantly different from day 1 averages ($p<0.05$, estimated via two-sided t-test), and the red arrow indicates the first day this occurs. Z-scoring was performed by standardising each mouse's mean RT per day using data from all its days. Second plot, reward rate over training days, calculated as the choice accuracy divided by the mean (not z-scored) RT for each day. Third plot, the ratio of the average (non-zscored) RTs measured in zero-contrast trials to the average RTs in left and right stimulus trials. Fourth plot, accuracy over training days. For all measures, only days with more than 90 trials were considered in the analysis. (B), Similar analysis to panel A, visualised selectively for one-sided mice over the first days of learning until significance for each variable was reached. A measure of the average zero-contrast bias over days was also included. Further, the RT ratio was changed to plot the ratio of RTs on trials with non-associated stimuli (including zero-contrast) vs. associated stimuli.

In early days of learning, mice favoured left or right choices to varying degrees, exhibiting flat, but often biased, psychometric curves (Figure 2.7, first column). RTs averaged across all trial types decreased in these early days (Figure 2.8A-B), and the biases usually increased (Figure 2.8B). This decrease in RTs led to an early increase in reward rate, i.e., reward per unit time (Figure 2.8A-B). RTs then started to depend on visual stimuli, i.e. mice showed faster RTs for trials with visual stimuli compared to zero-contrast trials, resulting in chronometric curves that were no longer flat (Figure 2.8A). These changes in biases, RTs and reward rate occurred while choice accuracies were still at chance level, i.e. flat psychometric curves (Figure 2.7 and Figure 2.8A-B). Thus, early days were marked by strategies where the mice ignored the position of the visual stimulus for making choices, showing decreasing RTs and increasing biases with different directions across animals.

During later days of learning, mice's choices began to depend on the location of visual stimuli, resulting in psychometric curves with increasing slopes. Importantly, biases often reversed and slopes developed differently for left and right stimuli, with a vast diversity across mice (Figure 2.9A-D). To visualise this diversity, we coloured each mouse's learning trajectory based on the asymmetry of its psychometric slopes (Figure 2.3; Section 2.2.4). While the diversity of slopes formed a continuum across mice, to better visualise the main trends we clustered slope trajectories over learning (Figure 2.9E; Section 2.2.4). In some mice the slopes increased similarly on both sides, resulting in more balanced psychometric curves. This indicates that their strategy involved associating both left and right stimuli with their corresponding choice directions (Figure 2.7, middle row). However, in several other mice, the slope primarily increased on one side while the other side remained flat, forming a one-sided psychometric curve (Figure 2.7, top and bottom rows). These mice therefore solved the task by forming a single stimulus-choice association: they associated stimuli on one

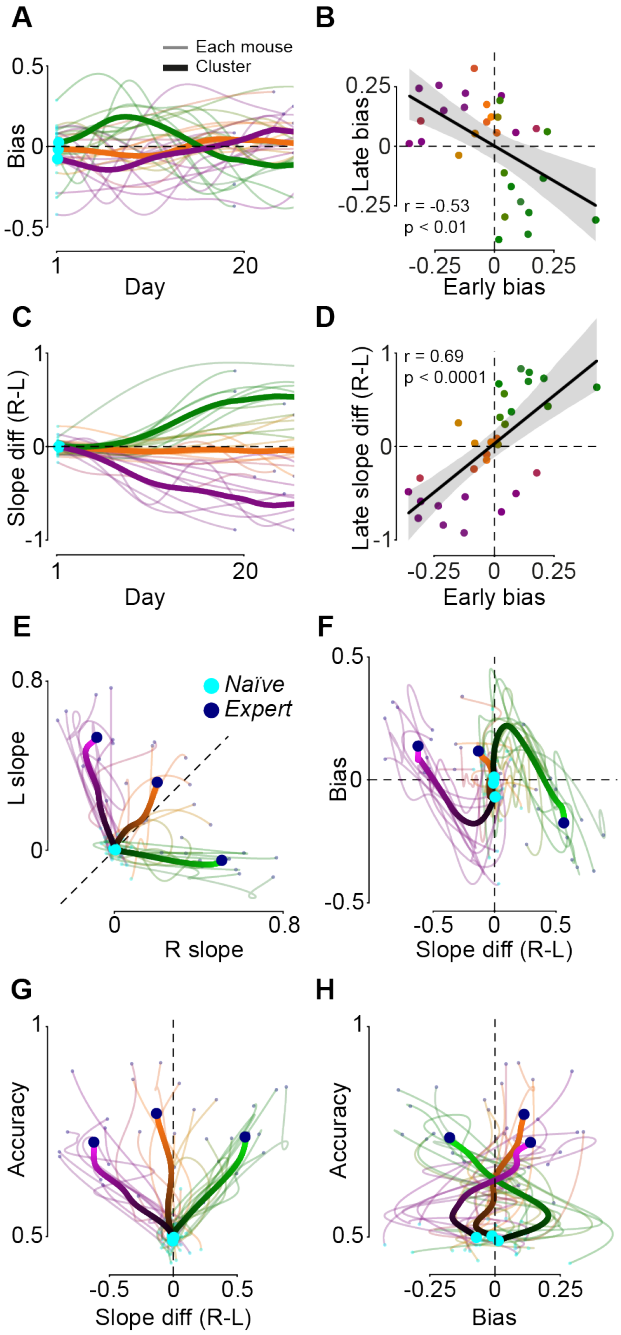


Figure 2.9: Mice form diverse yet systematic learning trajectories from naïve to expert.
(A), Bias over days per mouse (thin) and for the 3 clusters from panel E (thick). In right- and left-associating mice, biases increase before reversing ($p<0.05$; two-sided paired t-test first 2 days vs. days 5-6). Number of days limited to 25. (B), Regression of early bias (average across days 4-8) against late bias (average across final 5 days). Each point represents a mouse. p-value calculated from the exact distribution of r (stats.scipy.pearsonr). Shaded regions indicate 95% confidence interval across mice. (C), Difference between right and left psychometric slopes over days per mouse (thin) and for the 3 clusters from panel E (thick). Number of days limited to 25. (D), Regression of early bias against late slope difference. Each point represents a mouse. p-value from the exact distribution of r (stats.scipy.pearsonr). (E), Right vs. left slope across days per mouse (thin) and for 3 clusters (thick). The hue of the cluster lines changes from dark to light indicating progress through learning. Left-associating (purple), balanced (orange) and right-associating (green) clusters are obtained from dynamic time warping clustering applied to the mouse trajectories (Section 2.2.4). (F)-(H), In order, difference in right and left (R-L) slope vs. bias; R-L slope vs. accuracy and bias vs. accuracy across days per mouse (thin) and for the 3 clusters from panel E (thick).

side of the screen with their corresponding choice direction, and made the alternative choice in trials in which the associated stimulus was absent (Figure 2.7, rightmost column). Consistent with this strategy, choices in zero-contrast trials and trials with non-associated stimuli were indistinguishable, i.e., the psychometric curve was flat on one side (Figure 2.10A). These slope asymmetries often persisted as learning progressed (Figure 2.9E; Figure 2.10B-C). Importantly, despite strong asymmetries in slopes, one-sided left- and right-associating animals reached similar levels of choice accuracy compared to balanced animals (Figure 2.9G). This is because both the presence and absence of stimuli can inform correct choices in the task, as there is only one stimulus per trial. Overall, we observed that in later days mice transitioned to more stimulus-dependent strategies, while exhibiting diversity in their use of left and right stimuli to make choices.

Signatures of single stimulus use were also observed in measures of RT. A statistical test on

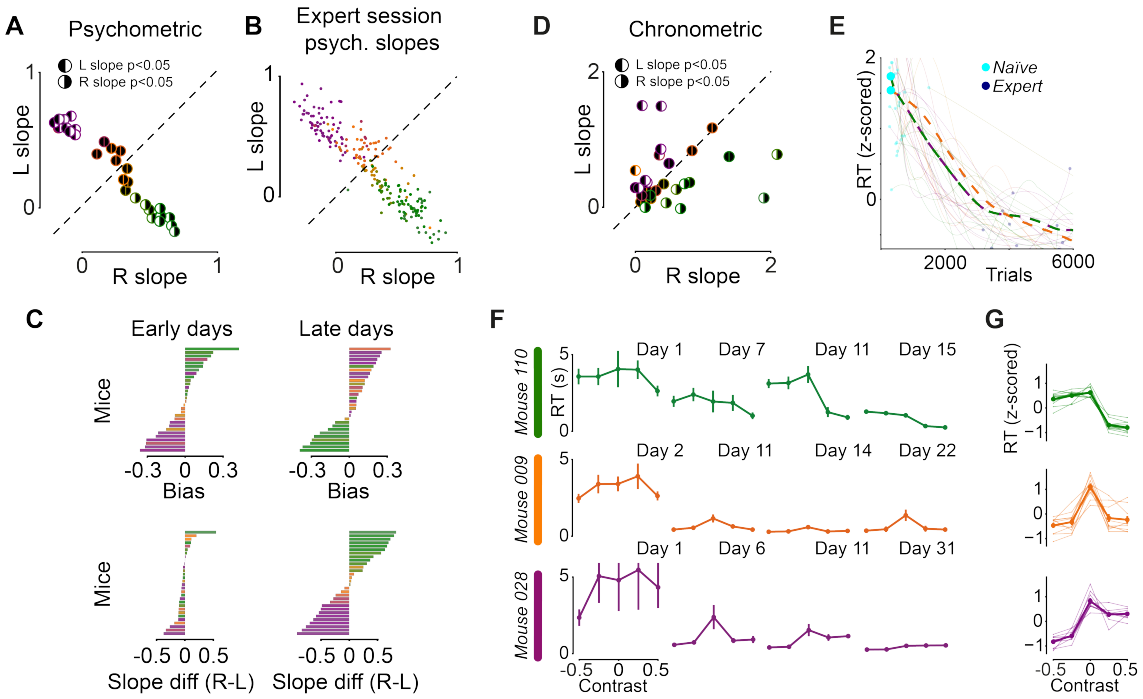


Figure 2.10: Quantification of diverse psychometric and chronometric phenomena. (A), Scatter of average right vs. left psychometric slopes in expert days for each mouse. Stroke colour indicates the average slope asymmetry of each mouse. The fill of the left/right half of each circle (black/white) indicates the statistical significance of the left/right slopes (i.e., whether the mean P('Right') for left/right stimulus trials is significantly different from that of zero-contrast trials, assessed with a Bernoulli statistic. Black means significant at 95% i.e., Bernoulli statistic >1.96 or <-1.96). (B), Scatter plot of expert day right vs. left psychometric slopes, coloured by the average slope asymmetry of the mouse trajectory they belong to. (C), Bar plots showing the average bias and slope difference in early (4-8) and late (final 5) days for each mouse, coloured by average psychometric slope asymmetry. (D), Scatter of average right vs. left chronometric slopes in expert days for each mouse. Chronometric slopes are defined as the difference between the median RT on zero-contrast trials and trials with right/left stimuli. (E), Z-scored RT over trials per mouse (thin) and for the 3 clusters from Figure 2.9E with left- and right-associating mice combined into a single group (thick dashed). Number of trials limited to 6000. (F), Chronometric curves for the same mice and days as in Figure 2.7. Error bars indicate +/- SEM across trials in each day. (G), Expert day chronometric curves per mouse (thin) and averaged per trajectory cluster (thick). Cluster labels for each mouse were obtained from Figure 2.9E.

the left and right slopes of chronometric curves showed that, in one-sided mice, the RTs for choices with the zero-contrast and non-associated stimulus were not significantly different (Figure 2.10D). Balanced mice instead had significantly lower RTs on trials with either of the left/right stimulus compared to zero-contrast trials. The evolution of general RTs through learning for one-sided and balanced mice also differed (Figure 2.10E). One-sided mice exhibited a very short initial plateau followed by a steady decrease to a subsequent lower plateau. Balanced mice instead showed a longer initial plateau followed by an uninterrupted exponential-like decrease to low values. The differences in chronometric curves across mice were also visible at the single mouse and single day level (Figure 2.10F). This day-by-day visualisation shows how RTs across mice initially decrease for all trial-types, followed by a further shortening of RTs as mice learn to associate visual stimuli with choices. The effect of single stimulus use is also seen clearly in the cluster-averaged expert chronometric curves (Figure 2.10G). We believe this effect is likely due to the one-sided mice's reliance on the auditory go cue as the indication of when to make a choice on trials with the zero-contrast and non-associated stimulus. Given that the go cue comes 0.2s after the onset of the

visual stimulus, and given that we defined RT as the difference between stimulus onset time and choice completion, this could explain the slower RTs on these trials.

Behavioural transitions throughout learning were systematic. The bias in early days strongly predicted the biases and psychometric curves in later stages of learning (Figure 2.9B and Figure 2.9D). Mice with early left bias developed a larger slope on the left side of their psychometric curve, whereas mice with early right bias developed a larger slope on the right side. To achieve high accuracy with asymmetric slopes, mice with more one-sided strategies reversed their early bias during learning (Figure 2.9A, F and H). Mice that did not learn despite performing enough trials (i.e. 4300, well above the median of learners), showed early choice biases but failed to develop psychometric slopes (Figure 2.11A-B).

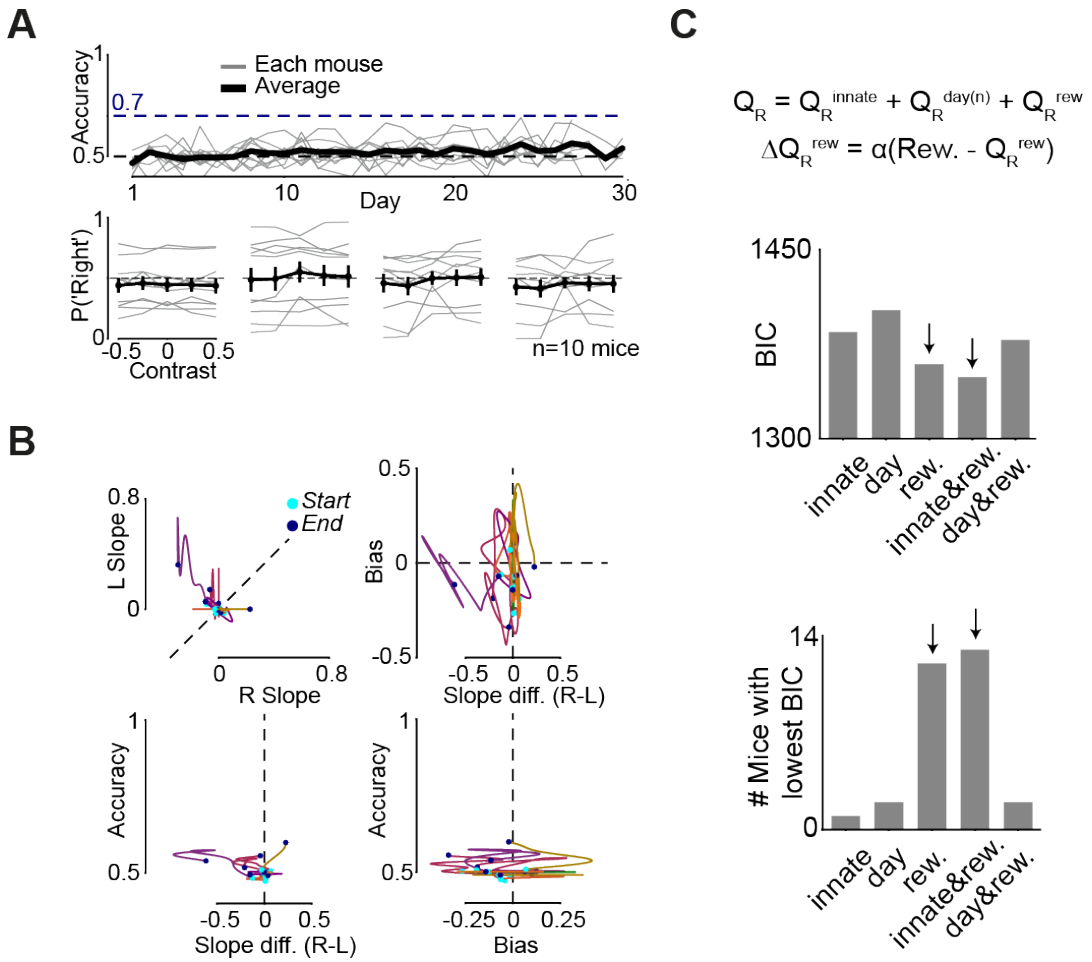


Figure 2.11: Non-learner mice and model assessing origin of initial bias.
(A), Top, accuracy over days per non-learner mouse (gray) and averaged across all non-learner mice (black). Dashed lines indicate chance level (black) and 70% accuracy level (blue). Number of days limited to 30. Bottom, psychometric curve over quartiles per mouse (gray) and averaged across all mice (black). In all panels, error bars indicate +/- SEM across mice. (B), In clockwise order from top left, right vs. left slope; difference in right and left (R-L) slope vs. bias; bias vs. accuracy and R-L slope vs. accuracy across days per non-learner mouse. Trajectory colours obtained from the average slope asymmetry as in Figure 2.3. (C), Results from a simple SARSA model assessing the effect of 'innate' biases, day-by-day tendencies, and reward-driven promotion of each choice. The model was fit to choice data from the first 5 days of training. Left, the BIC of each model in a nested model comparison selectively removing different combinations of model parameters. Arrows indicate the two model variations with the lowest BIC. Right, for each model variation, a count of the number of mice whose choice data had the lowest BIC for that variation compared to others. Arrows indicate the two model variations with the largest number of mice.

To investigate the origin of the diverse early biases across mice, we fit a simple SARSA model to mouse choices in the first 5 days of training (Figure 2.11C). The model had two Q-values for the left and right choices (Q_L and Q_R) independent of trial type (i.e., with a single state). Each Q-value was defined as the sum of an innate (fixed per mouse) choice tendency for each of the choices (Q^{innate}), a choice tendency that changes day-by-day (Q^{day}, one for each day), and a choice tendency that is shaped by reward history (Q^{rew}),

$$Q_{L/R} \;=\; Q^{innate}_{L/R} + \; Q^{day(n)}_{L/R} \;+\; Q^{rew}_{L/R}, \tag{2.1}$$

where n indicates the day (from 1 to 5) corresponding to each trial. The choice probabilities on each trial were determined by passing the resulting left and right Q-values through a softmax function with inverse temperature parameter β. On each trial, the corresponding choice was taken from the mice's choice history and Q^{rew} was updated using the regular SARSA update with two learning rates $\alpha^{+/-}$,

$$\Delta {Q_{ch}}^{rew} = \alpha^{+/-}(Rew. - {Q_{ch}}^{rew}), \tag{2.2}$$

where the Q-value corresponding to the choice (Q_{ch}) was updated using α^+ and a reward signal $Rew. = 1$ on trials when the mice were rewarded, and α^- with $Rew. = 0$ for when they were not.

We performed a nested model comparison by removing different combinations of the Q-value components (innate, day and rew) and fitting each model variation to each mouse's choice and reward history data. The values of Q^{innate}, $Q^{day(1-5)}$, β and $\alpha^{+/-}$ were fit per mouse for each version of the model by simulating the Q-value trajectories for each parameter setting, and minimising the resulting log-likelihood of the mouse's choices (based on the model-derived choice probabilities on each trial) with momentum-based projected gradient descent (optax.adam, run for 5000 iterations with learning rate 0.001 and parameters clipped to 0 on each update step if they go negative). Finally, the bayesian information criterion (BIC) for each model variation fit to each mouse's data was computed. The results showed that the best models were those with only the rew component, or both the rew and innate component (Figure 2.11C). This indicates that the early biases are likely due to a combination of unequal reward history for left and right choices caused by randomness in the task (e.g. the random sequence of stimuli or rewards on zero-contrast trials), together with a potential initial 'innate' bias for one of the choices (e.g. handedness).

We also investigated whether the early biases depend on imbalances in early sensory processing. However, we did not find evidence for this, as biases were present even prior to the animals' use of stimuli (Figure 2.8B). Further, although pupil and lick rate measures changed in response to stimuli and rewards (Figure 2.12A), these measures were similar for left and right stimuli in initial days (Figure 2.12B, left). In expert days, we did observe modulation of both pupil motion and diameter by stimulus use. The pupil of a one-sided mouse tracked the vertical and horizontal movement of an ipsilateral stimulus only if it was associated with a choice. Similarly, its pupil diameter was larger on trials with an associated than with a non-associated ipsilateral stimulus. On the other hand, the horizontal movement of a contralateral stimulus was tracked both when associated and when not associated with a choice. This could be because the eye ipsilateral to the associated stimulus searches for light even on trials with a contralateral stimulus. In balanced mice, the pupil tracked the vertical and horizontal movements of both left and right stimuli (Figure 2.12B, right).

Taken together, our behavioural results showed that naïve mice started by developing varying levels of bias shaped by their reward history, and this early bias determined which stimuli were associated with choices in later stages of learning.

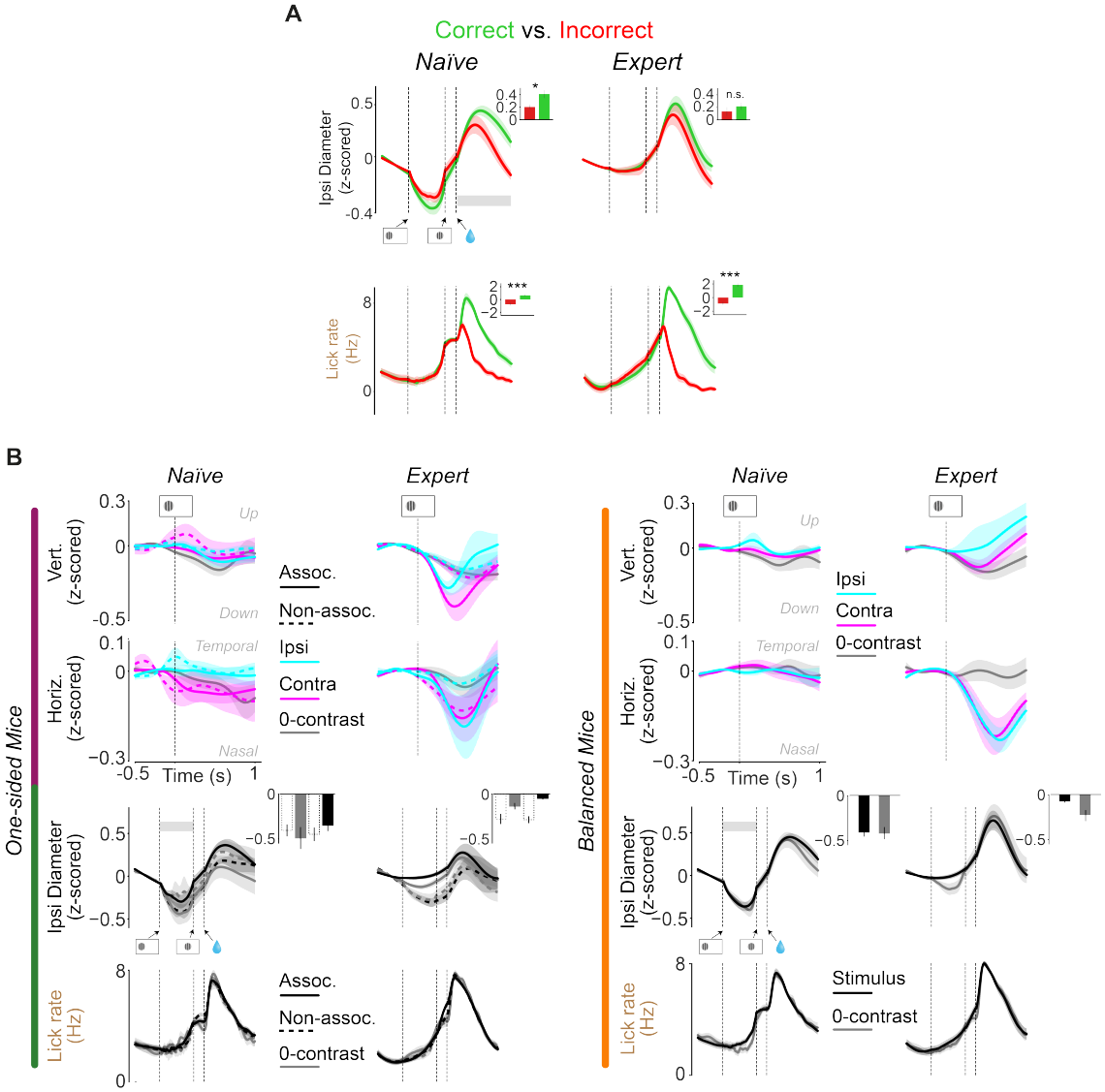


Figure 2.12: Eye movement and lick analyses. (A), Time warped measures of ipsilateral pupil diameter (top) and lick rate (bottom) on correct (green) vs incorrect (red) trials of balanced mice where the previous trial was rewarded. Naïve days are defined as the first 5 days of each mouse, and expert days have accuracy n.s. smaller than 0.7. Shaded bar shows the analysis window used for the bar plot insets. p-values calculated using two-sided t-test. (B), Top, average vertical and horizontal eye movements (pupil motion) aligned to stimulus onset in one-sided and balanced mice for correct trials with stimuli ipsilateral (cyan) and contralateral (fuchsia) to the pupil, as well as on zero-contrast rewarded trials (gray). In one-sided mice, solid lines indicate pupil motion in response to 'associated' stimuli and dashed lines to 'non-associated' stimuli. Bottom, time warped measures of ipsilateral pupil diameter (top) and lick rate (bottom) for correct trials with associated (black solid) and non-associated (black dashed) stimuli, as well as rewarded zero-contrast trials where the associated (gray solid) and non-associated (gray dashed) choice was made. Only trials that directly follow a rewarded trial are plotted. For one-sided mice, naïve days are defined as the first 3 days and for balanced mice as the first 5 days (like in panel A). Expert days have accuracy n.s. smaller than 0.7 for both groups. Shaded bar shows the analysis window used for the bar plot insets. n.s., *, ** and *** indicate $p>0.05$, $p<0.05$, $p<0.005$ and $p<0.0005$ respectively.

2.4 Discussion

In summary, our results show that in learning to make visual decisions from naïve to expert, mice exhibited diverse learning trajectories involving systematic transitions through behavioural strategies. Unlike conventional shaping methods that gradually change the task from easy to difficult, we maintained the full (i.e., relatively difficult) task all throughout the experiment. Given that this procedure might have slowed down learning, many of our mice were not trained enough to reach the level of asymptotic expertise often used in studies of decision making. Nevertheless, our results hold in animals with very high performance and graded levels of stimuli (Figure 3.6F). Our training procedure allowed mice to freely explore and self-define their trajectories, facilitating individual diversity. In fact, the substantial early side biases, and later one-sided strategies we observed could be due to the higher starting difficulty of our task. The effect of task difficulty in developing biases and one-sided strategies has been observed with other sensory modalities such as in whisker-based tasks [226]. Future studies can examine the effects of various shaping methods on learning trajectories.

The temporal resolution at which we observed systematic trajectories was on the order of hundreds of trials. However, there could also be behavioural strategy switches happening on faster time scales [227], which may be consistent with our observations. Investigating the effect of such switches on learning could reveal how they compound throughout long-term learning. Past studies have shown individual diversity in measures such as speed of learning across animals [228, 229]. Our results reveal that despite such individual diversity, learning trajectories can be highly systematic. Further, in our data the early variability in side bias emerges from uneven learning due to imbalances in the number of rewards after left and right early choices, and not day-by-day variation in animals' position. Consistently, analyses of eye movement and pupil size suggested that diversity in early side biases cannot be attributed to uneven detection of visual stimuli. However, these early side biases may be influenced by factors such as the mice's initial position and handedness. Beyond the effect of early side biases, different levels of confusion about the position of the stimuli in early days might also contribute to one-sided learning trajectories. As such, understanding the source of individual variability in these early days of learning requires further experiments.

A component of the mice's learning trajectory that this thesis does not explore is that of motor learning [230, 231]. The behavioural task was designed such that the movements required to turn the wheel are similar to the orienting movements that mice would make to bring the stimulus to the centre of their field of view. Regardless, we would expect there to be a period of time initially where the mice have to familiarise themselves with the dynamics of the wheel. Given that the RTs decreased sharply over the first quartile of learning (Figure 2.6E), we believe that this period is much shorter than the slower learning of associations between stimuli and choices. Nonetheless, we collected wheel movement data throughout learning whose analysis might reveal interesting motor learning dynamics in the task, and could be a fruitful subject of future studies.

Stage-like transitions through sequences of strategies are a characteristic feature of long-term learning. Recent studies have noted similar discrete transitions in mice [159] and humans [20, 24] learning new tasks from naïve. However, past work has struggled to analyse the diversity in these sequences across individuals, as doing so requires long-term learning data from a large number of individuals learning a sufficiently complex task that can afford the study of individual diversity. The visual decision making task for mice that we used yields a tractable level of diversity with sample sizes on the order of tens of mice. The diverse yet systematic transitions between early and late behaviour are indicative of a learning mechanism that establishes flows through the task

space such that early states influence late states. Further, the momentary persistence of strategies provide clues as to why some mice do not learn, as they get 'stuck' in exploiting a strategy that they discover early on. To investigate the brain mechanisms governing the mice's long-term learning trajectories, we recorded dopamine release signals across the dorsal striatum from naïve to expert. Chapter 3 describes our results, relating the dopamine signals we observed to the behavioural trajectories outlined in this chapter.

Chapter 3
Neuronal teaching signals

Chapter contents

3.1 Introduction 33
3.2 Methods 34
3.2.1 Mice 34
3.2.2 Surgical procedures 34
3.2.3 Imaging dopamine release 34
3.2.4 Optogenetic manipulation of dopamine signals 35
3.2.5 Histology and fibre track quantifications 36
3.2.6 Mice trained to higher accuracy 36
3.2.7 Reward size manipulation experiment 36
3.2.8 Neural data analyses 36
3.3 Results 39
3.3.1 Dopamine recordings 39
3.3.2 Optogenetic manipulations 50
3.4 Discussion 54

Learning is typically conceptualised as a system adapting towards an objective. As such, a learning system requires some notion of how far it is from that objective and how it should update to get closer. A signal that does this can be referred to as a 'teaching signal' and, in the brain, often appears in the form of a prediction error, i.e., the difference between a prediction formed by the brain and the ultimate outcome. Neurons encoding such prediction errors typically have the added ability of promoting synaptic plasticity, such that the teaching signal can update the connections in a network for better performance in the future. Discovering the teaching signals at work in the brain has been the subject of multiple studies in the past. Prediction errors and related signals at the service of learning have been reported in areas across the brain from the cortex [84], to the midbrain [52], to the cerebellum [232]. Finding these and understanding their mechanism contributes to uncovering the learning algorithms operating in the brain, which could lead to better autonomous systems as well as a better understanding of diverse learning abilities.

Here we study the dopaminergic system and its role in long-term learning. Dopamine (DA) neurons have been heavily implicated in learning from rewards [233, 190, 191, 189, 181] and seminal past studies have shown that they encode a teaching signal known as RPE [52]. This signal first emerged in the literature from a subfield of machine learning called RL, where it was used to train artificial agents to maximise reward consumption [43]. The RPE signals the difference between predicted and actual reward, and is thus positive when an agent receives more reward than expected and negative when it receives less. Such a signal can be used for learning by updating reward predictions proportional to its size (increasing them when positive and decreasing when negative), thus yielding more reliable predictions of reward that can be used to guide behaviour.

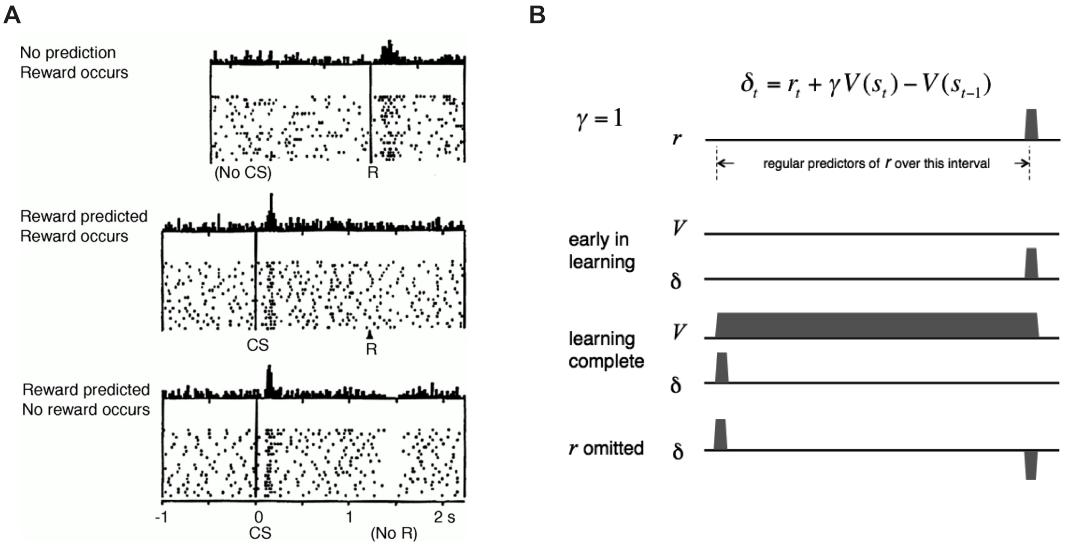


Figure 3.1: Dopamine encodes Temporal Difference (TD) error.
(A), Dopamine neuron firing rates for an unpredicted reward (top), for a predicted reward (middle), and for a predicted but omitted reward (bottom). Figure adapted from [52]. (B), Dynamics of value (V) and TD error (δ) in equivalent scenarios to those in panel A. Figure adapted from [43].

Dopamine has been found to encode a particular instantiation of RPE known as the temporal difference (TD) error [234, 51, 52]. This signal reflects both the difference between predicted and actual reward, and that between values (i.e., amount of reward predicted) of states around reward,

$$DA(t) \sim \delta_t = R_t + V(S_t) - V(S_{t-1}) \tag{3.1}$$

where $DA(t)$ denotes the level of dopamine at time t, δ_t denotes the TD error, $V(S_t)$ denotes the value of the state at time t and R_t denotes the reward. The TD-error (and similarly dopamine) thus moves backwards through time as the reward-predictive values of states preceding reward are learned (Figure 3.1A-B). In Pavlovian conditioning, where an agent learns to associate a cue with a reward, the TD-error early in learning would be largest at the time of reward, as the values of preceding states (i.e., cues) are yet to be learned. After learning, the TD-error at reward delivery would be small and has moved to the time of cue onset. If the cue has any predictors further back in time, the TD-error would continue to travel backwards. This backward-travelling dynamic of the TD-error has been observed in dopamine recordings [52, 235, 176].

Despite substantial research focused on dopamine the role of DA in long-term learning remains unknown, as past experimental studies have mainly recorded DA signals from expert animals learning to fine tune decisions. Instead, we recorded DA signals across the dorsal striatum of mice learning a visual decision-making task from naïve to expert over multiple days (Chapter 2). We found that DA signals mimic behavioural trajectories acting as a suitable teaching signal, yet are diverse across the striatum and thus differ from conventional formulations of RPE. To demonstrate the involvement of dorsolateral striatum (DLS) DA in learning, we performed an optogenetic experiment where the release of DLS DA was inhibited throughout learning. This severely impaired mouse learning, preventing them from surpassing chance-level accuracy despite training for as long as unmanipulated mice. Further, in line with our observations, we designed an optogenetic DLS DA stimulation experiment that verified the distinct effect on learning of the DLS DA signal from that of a classic RPE. We believe these results take us a step closer to understanding the brain's reward-based learning algorithms, while incorporating recent reports of heterogenous dopaminergic signalling across brain regions [236, 192, 237].

3.1 Introduction

Neuronal signals across a wide range of brain areas are important for learning [238, 48]. Many studies have shown that DA neurons and brain areas receiving DA signals, such as the striatum, are essential for learning to make decisions [239, 173, 240, 241, 242, 236, 243, 244, 245, 190, 246, 247, 248, 192]. These dopamine signals encode RPE, and have been shown to drive plasticity in cortical-striatal synapses [52, 188, 249, 171, 250, 174]. In particular, for learning to make perceptual decisions (i.e. decisions guided by incoming sensory information) dorsal regions of striatum, as well as the cortical and dopaminergic signals they receive, play critical roles [147]. However, the role of dorsal striatal DA signals in long-term learning to make perceptual decisions remains unknown. Crucially, if DA signals underlie long-term learning trajectories, these teaching signals should reflect and shape the diverse intermediate strategies employed by different animals throughout learning. Testing this hypothesis requires probing DA signals during learning of a decision task which admits different strategies, leading to sufficient yet tractable individual diversity. However, past studies have often recorded dopamine signals in expert animals, instead of recording throughout learning from naïve to expert.

Recent studies have shown considerable heterogeneity in the tuning of individual dopamine neurons during perceptual decision making [236, 243, 192, 182]. For instance, midbrain dopamine neurons of mice have been shown to encode a variety of sensory, motor, and cognitive variables during a visual evidence integration task [192]. Dopamine functional heterogeneity also appears to depend on dopaminergic axonal projection pathways. For instance, in a similar visual contrast detection task for mice, dopamine axons in the dorsomedial striatum (DMS) only responded to contralateral, not ipsilateral visual stimuli [182]. Instead, dopamine axons in the ventral striatum showed contrast-dependent responses to both contralateral and ipsilateral stimuli [182]. This projection-specific heterogeneity may vary throughout the perceptual decision-making process. For example, a study examining dopamine release in the ventral and posterior (tail) striatum during an auditory decision task found a similar representation of sensory evidence in both striatal regions, but projection-specific encoding of reward and stimulus expectations in pre-stimulus epochs [183]. Finally, compared to the relative consistency of studies examining dopamine signals in dorsal or ventral striatum, studies investigating dopamine signals in the tail of the striatum have argued for differing roles, including stimulus expectation, action prediction error, and stimulus novelty [251, 252, 253]. In our study, we also observe such heterogeneity of dopamine signals across the dorsal striatum. We find that DA release in the DMS only responds to contralateral visual stimuli and does not respond to reward. Whereas DLS DA release acts more similar to a classic RPE signal, responding to both contra- and ipsilateral stimuli and showing large responses to reward early in learning which decrease as mice become expert at the task.

Although the heterogeneity of dopamine responses is often presented as an argument against RPE encoding, we propose an anatomically plausible hypothesis which suggests that it may be compatible with the RL framework. Similar to a recent study [176, 254], our hypothesis attempts to explain dopaminergic heterogeneity with the concept of a 'partial' RPE, which reflects the difference between a reward prediction calculated using a subset of environmental features and a global reward signal. The hypothesis proposes that diverse cortico-striatal projections cause a segregated routing of information generating 'partial' reward predictions in the striatum which in turn generate 'partial' RPEs in midbrain dopamine neurons projecting back to the source cortico-striatal projections. Our DLS DA signals appeared to encode such 'partial' RPEs, and we thus hypothesised that its plasticity-inducing effects would be limited to the cortico-striatal synapses that are involved in the

formation of its associated prediction. In addition to verifying the brain's reliance on DLS DA for learning through an optogenetic inhibition experiment, we verified this hypothesis by stimulating DLS DA optogenetically and demonstrating its characteristic effect on behaviour. Although the concept of partial RPE can explain accounts of heterogenous DA signals tuned to different sensory, motor or cognitive variables, it cannot explain the striking lack of reward responses in DMS DA release. We are not aware of models in the literature that can account for such diversity, however, Appendix A provides a candidate framework which is in its early stages.

3.2 Methods

3.2.1 Mice

The neural learning data presented in this thesis was collected from 20 of the wild-type mice that were trained from naïve to expert on the task (Section 2.2.1). In addition, this section includes results from optogenetic studies involving 19 (5+7+7) male DAT-Cre mice, with their age ranging between 9 to 30 weeks. The first 5 were DAT-Cre mice where eOPN3 virus was expressed for the inhibition of DLS dopamine. The next 7 were DAT-Cre mice where ChrimsonR virus was expressed for the stimulation of DLS dopamine. The last 7 were DAT-Cre mice where ChrimsonR virus was expressed for the stimulation of DMS dopamine. Finally, it includes data from 6 wild-type mice that were trained with shaping, more contrast levels and more extensive training to higher accuracy levels.

3.2.2 Surgical procedures

Animals were anaesthetised with isoflurane and were kept on a feedback-controlled heating pad (Stoelting 53810). Hair overlying the skull was shaved and the skin and muscles over the central part of the skull were removed. The skull was thoroughly washed with sterile saline. A head plate was attached to the bone posterior to bregma using dental cement (Super-Bond C&B). After the head plate fixation, we made craniotomies over the target areas and injected 300nl of AAV9-hsyn-DA2m (for recording dopamine, titer 5×10^{12} genome copies/ml) into the right and/or left DLS (AP: +0.5mm from bregma; ML: +/-2.5mm from midline; DV: 2.8mm from dura) or DMS (AP: +1.25mm; ML: +/-1mm; DV: 2.8mm). For the optogenetic experiments, either 400nl of 4.1pAAV-hSyn1-SIO-eOPN3-mScarlet-WPRE (inhibition experiment, $\sim 4.67 \times 10^{12}$ genome copies/ml) or 150nl of 3.2pAAV-Syn-FLEX-rc[ChrimsonR-tdTomato] (stimulation experiment, $\sim 1.67 \times 10^{12}$ genome copies/ml) was injected into the right and left SNc (AP: -3mm; ML: +/-1.5mm; DV: -4.3mm) of DAT-Cre mice. The injections were performed slowly over 20 minutes using Nanoject II (Drummond). This was followed by implantation of the optical fibre over the DLS or DMS (core = 200 um, Neurophotometrics Ltd), which was secured to the head plate and skull using dental cement. Mice recovered for at least seven days following the surgery. We waited for an additional 8 weeks for opsin expression in mice with eOPN3 injections and 4-5 weeks in mice with ChrimsonR injections.

3.2.3 Imaging dopamine release

To measure DA release in the DLS, we employed fibre photometry. Photometry and behavioural data were collected simultaneously. We used chronically implanted optical fibres to deliver excitation light and collect emitted fluorescence (Neurophotometrics FP3002). We used multiple excitation

wavelengths (470 and 415nm), delivered on alternating frames (sampling rate of 40 Hz), serving as target and isosbestic control wavelengths, respectively. The isosbestic wavelength does not excite the fluorophore of the GRAB-DA dopamine sensor and was used as a control signal to correct for movement artefacts and other sources of noise correlated between the two measurements.

The recorded photometry signal was pre-processed following steps described previously [255]. We began by de-interleaving the recorded signal at 470nm and 415nm wavelengths. Both signals were then de-noised to remove short-pulse artefacts using a median filter with kernel size 5 (medfilt from scipy.signal). Subsequently, the signals were detrended with a zero-phase low-pass filter with a 10Hz cutoff frequency (2nd-order butterworth filtfilt from scipy.signal). Next, a photobleaching correction was applied to remove slow changes in the signal likely coming from fluorophore degradation due to light exposure throughout the recording session. To do this, we used a scipy filtfilt zero-phase high-pass filter with a cutoff frequency of 0.001Hz, thus removing signals varying with a timescale slower than 16 minutes. We then corrected for motion signals by fitting the 415nm isosbestic to the 470nm signal with a least squares polynomial fit of degree 1 (linregress, scipy.stats) and the resulting fitted signal was then subtracted from the 470nm signal. Finally, this quantity (ΔF) was normalised through division by the baseline fluorescence (F, defined as a low-pass filtering of the denoised 470nm signal with a cutoff frequency of 0.001Hz) to obtain ΔF/F which was subsequently z-scored per session to enable more accurate comparisons across days of recording.

High data quality was ensured by removing sessions with weak DA signals. We plotted the average relative amplitude of the raw 470nm and 415nm signal per session. If this ratio was smaller than 1 the session was discarded, since in such sessions most of the pre-processed ΔF/F fluctuations came from variation in the isosbestic signal instead of the informative 470nm channel. We also discarded sessions with maximum fluctuations smaller than one standard deviation (i.e., Z<1).

3.2.4 Optogenetic manipulation of dopamine signals

DLS DA inhibition experiment

Mice expressing the inhibitory opsin eOPN3 [256] were trained on the same visual decision-making task described above, following the same recovery and habituation protocol. Throughout each day of the experiment, laser pulses (each with duration 600ms, wavelength 532nm, and power $\sim$ 12mW measured at the tip of the patch cord) were delivered through the optical fibres in random intervals of 6-14s, independent of task trials. Mice were trained for approximately 30 minutes each day for a total of at least 4300 'good' trials (i.e., after preprocessing, see section below).

DLS DA stimulation experiment

After similar recovery and habituation protocols to those described above, we trained mice expressing the excitatory opsin ChrimsonR in DLS terminals. Upon reaching a stable accuracy above 70%, we began the optogenetic experiment. In alternating days, laser pulses (25ms on/25ms off, 635nm, $\sim$ 10mW measured at the tip of the patch cord) were delivered through the optical fibres over a period of 490ms at the outcome time of incorrect left (day n) or right (day n+1) choices (including non-rewarded zero-contrast trials). This alternation was performed for at least 8 days in each mouse (i.e., at least 4 days for each condition). As a control, a similar experiment was performed in expert animals with or without ChrimsonR, replacing laser stimulation with the delivery of a water reward. This experiment included a baseline day between manipulation days where mice performed the original task. This was done to avoid mice from becoming too biased towards one

side. The protocol was run for at least 12 days in each mouse (i.e. at least 3 days of each left/right condition). One mouse from the DLS dopamine stimulation cohort expressed the excitatory opsin pAAV-Ef1a-DIO-ChRmine-mScarlet (i.e., ChRmine). The surgery was performed similarly to the other mice used for optogenetics, injecting 200nl of the virus (titer $\sim 9 \times 10^{12}$ genome copies/ml) into SNc and then implanting optical fibres over DLS. For stimulation, we delivered laser pulses (25ms on/25ms off, 532nm, 0.25mW measured at the tip of the patch cord) over a period of 490ms.

DMS DA stimulation experiment

Following the recovery and habituation protocols defined above, we manipulated DMS dopamine levels from naïve while training mice expressing ChrimsonR with unilateral DMS fibres on the original task. In a random selection of 75% of trials with the stimulus contralateral to the DMS fibre, laser pulses (25ms on/25ms off, 635nm, 10mW measured at the tip of the patch cord) were provided over a period of 200ms through the optical fibres, locked to stimulus onset. This stimulation protocol was provided every day of the experiment for at least the first 5 days of training.

3.2.5 Histology and fibre track quantifications

Histology was performed after the experiment to confirm successful fibre positioning. Animals were deeply anaesthetised and perfused using 4% paraformaldehyde (PFA) and then decapitated. The brains were extracted, left in 4% PFA for 24h to post-fix in a refrigerator and then embedded in blocks of 1.5% agarose gel before collecting slices at 70 um thickness using a vibratome (Leica VT1000 S). Slices were then stained with DAPI for 15 min (1:1000 solution), mounted onto glass, coverslipped, and imaged using an epifluorescence microscope (Leica).

3.2.6 Mice trained to higher accuracy

In 6 mice, we gradually included stimuli with lower contrasts throughout training to have stimuli with contrasts of 0%, 6%, 12.5%, 50% and 100%, and trained them to higher performance ($\sim 85-90\%$) over a longer time period. We present both behavioural and DLS dopamine recordings from the days in which these mice performed with high accuracy.

3.2.7 Reward size manipulation experiment

In a subset of trained mice with dopamine recordings in DLS/DMS, we investigated the effect of reward size on choices and DA signals with our behavioural task (Section 2.2.2), similar to previous studies [49]. To do so, we provided twice as much water reward for correct choices after left or right stimuli (3ul vs. 1.5ul) in alternating days of testing. This alternation was performed for at least 6 days in each mouse (i.e. 3 days per condition).

3.2.8 Neural data analyses

Event alignment and time warping

DLS DA recordings were aligned to task events that caused significant DLS dopamine release, as determined by a linear deconvolution (see below). These events were visual stimulus onset, choice completion (i.e., correct trials: visual stimulus arriving in the centre of the screen, and incorrect trials: stimulus moving out of the screen) and trial outcome (reward/no reward). To do this, a

fixed time period around each event (-0.5s to +1s) was selected and a fixed number of elements for the resulting aligned neural trace were chosen (i.e., 100). The DA recording was then linearly interpolated to obtain a value for each of the desired time points in the chosen time period (using scipy.interp1d). The average value in the time period before the event (-0.5-0s) was used as a baseline and subtracted from the event-aligned traces.

Time warping was used to visualise DA signals in a single continuous trace including all trial events. This was achieved by warping the DA signals such that a fixed number of time points represented the time course between each event. In this way, varying time periods between events were accounted for by allowing different time intervals between data points. We chose to have 40 time points before stimulus onset, 30 between stimulus onset and choice completion, 12 between choice completion and outcome and 60 post-outcome. The recorded DA signal was then interpolated to obtain the fluorescence values at the corresponding time points. This time warping allowed us to compute and visualise average DA signals across trials with varying durations.

Normalisation across sessions

In addition to the pre-processing steps described above, we further corrected for session-by-session variation in fluorescence levels by normalising the DLS DA signals to the peak of the average DLS DA response to reward delivery in zero-contrast (i.e., no visual stimulus) trials. This was done for all analyses except those comparing DLS and DMS DA signals, due to lack of a suitable reference for normalising the DMS signal. To do this, we computed the average time warped DA trace from zero-contrast trials per session, took the mean of the 3 largest values in the post-reward response (peak), and then subtracted the mean fluorescence levels 3 time points before and after the time of reward delivery (baseline) to obtain a single number (peak – baseline) that was used to divide the entire DA recording from that session by. This normalization assumes that DA reward responses in zero-contrast trials should not vary session-by-session, as there are no cues that can predict the random reward delivery on these trials, meaning that the degree of 'surprise' (i.e., reward prediction error) should remain fairly constant throughout learning. This yielded normalised $\Delta F/F$ values that ranged from 0-1, expressing the signal as a proportion of the zero-contrast reward response in each session, thus resulting in more accurate averages across sessions and mice and more interpretable fluorescence values that could be compared with the model-derived DA signals.

Analysis windows

We defined average DA responses to 3 events (stimulus onset, choice completion and reward) which we use in several analyses throughout the paper. These were defined as the average DA signal in a specific time window after each event, relative to the signal before the event.

For the 'stimulus response', we defined an analysis window from +0.2s to +0.35s post-stimulus onset and took the peak response of a moving average (window size = 10) of the signal in that time window. We then subtracted a baseline fluorescence value (defined as the mean $\Delta F/F$ in a window -0.1s to +0.1s around stimulus onset) from the peak estimate to obtain the 'stimulus response' on each trial.

For the DA response to choice completion ('stim. centre' on correct trials), we used the time warped traces to define the analysis windows for the peak and baseline estimates. We did this because the times between stimulus, choice completion and reward delivery/absence are variable and can be short, so using time warping is an accurate method to obtain isolated estimates of DA

release values uniquely around this event. To do this, we used the entire 12 elements of the time warped trace between choice complete and reward delivery/absence and found the peak value of its moving average (window size = 6). From this, we subtracted the mean of the time warped DA signal one time point before and after the choice complete event.

For the DA response to reward delivery/absence we used the same procedure as for the 'stimulus response'. We only changed the analysis window from which we calculated the peak response, i.e., 0-1s after the reward time. The baseline was similarly defined using the average signal in a window -0.1s to +0.1s around the event.

For most neural analyses we defined a combined 'outcome response' which is the addition of DA responses to choice completion and reward delivery/absence. This was done because in rewarded trials, the 'stim. centre' cue is a perfect predictor of upcoming reward and thus rapidly acquires value, causing the DA responses to reward to become smaller. We did not want to include this learning process in our analyses as it does not involve decision making (i.e., is Pavlovian). Hence, by adding the responses for the two events, we obtained an outcome signal that did not change until events before choice completion became associated with reward.

Linear deconvolution

To find the events that caused significant variation in DMS and DLS DA release levels, we used a linear deconvolution algorithm. This algorithm works by regressing binary variables indicating the time period around events onto full trial DA signals. The design matrix (X) has a column per time point around the events of interest (regressors) and a row per time point of the DA signal, which was concatenated for all the trials used in the regression to form the dependent variable (y). The algorithm then finds the optimal scaling (β^*) of the regressors in X which produces a prediction of the concatenated DA signal ($\hat{y}^*$) that minimises the mean squared error $\sum_i (y_i - \hat{y}_i)^2$,

$$\beta^* = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \tag{3.2}$$

$$\hat{y}^* = X\beta^*. \tag{3.3}$$

The resulting elements in β^* compose the deconvolved signal for each event. The benefits of using linear deconvolution over an event-aligned average is that it isolates the effect of each event on the DA signal, removing the influence of other events occurring shortly before or after. This isolation is achieved as long as there is enough jitter between events across trials; if two events are separated by a fixed delay in all trials the regression will not find isolated signals due to correlated regressors.

To account for changing DA signals over learning in our linear deconvolution, we split all trials into 4 bins with increasing psychometric slope. We also performed separate regressions for rewarded and unrewarded trials. The events we considered were stimulus onset, choice start, choice completion and reward delivery/absence (Figure 3.2B). The explained variance for each event was computed by comparing the 5-fold cross-validated R^2 of a 'full' model with all the events against that of a model without the event being assessed.

3.3 Results

3.3.1 Dopamine recordings

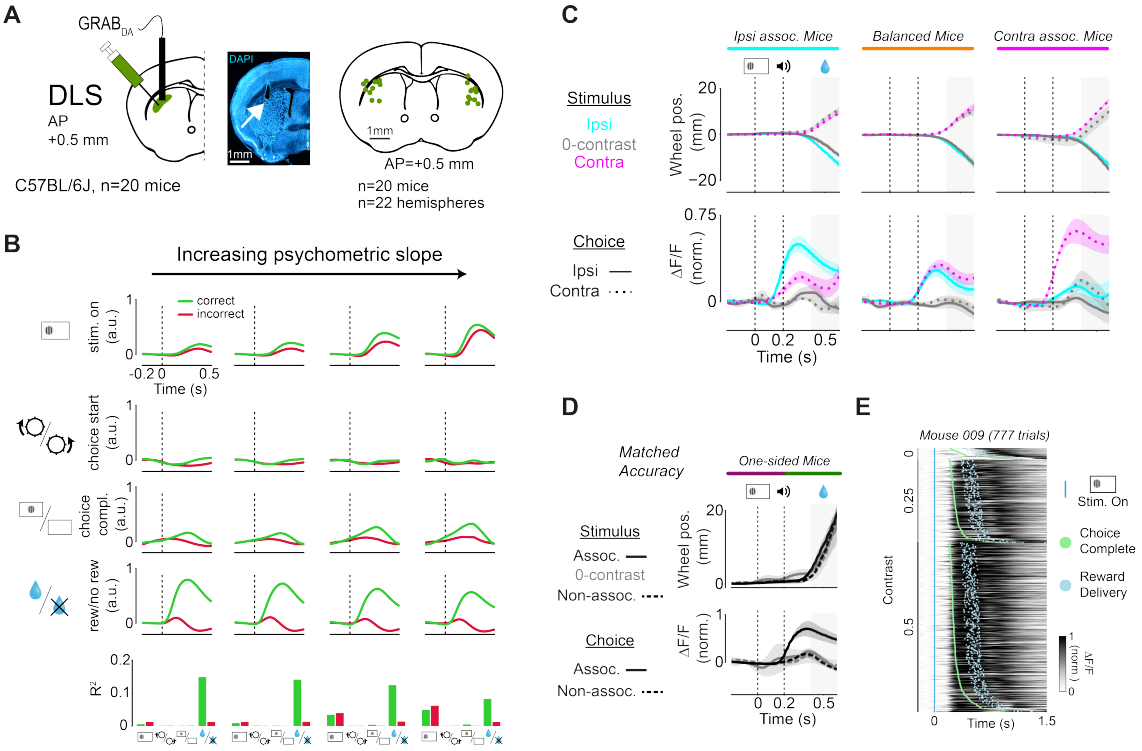


Figure 3.2: Dorsolateral striatum (DLS) dopamine (DA) signals during task learning. (A), Left, schematic of DLS DA fibre photometry recording experimental set-up. Middle, histology slice showing probe placement (white arrow). Right, the locations of DLS optic fibre implants estimated from post-mortem histological examinations overlaid on a brain slice schematic. (B), Top, regression coefficients from a linear deconvolution of DLS DA signals for different events in a trial: stimulus onset, choice start, choice completion, and outcome (reward/no reward). The resulting deconvolved signals were baselined using the average signal in the 0.2s before each event. Bottom, R^2 for each trial event used in the linear deconvolution. (C), Wheel position (positive for contralateral and negative for ipsilateral movements) and simultaneous DLS DA recordings aligned to stimulus onset in expert days (accuracy n.s. smaller than 0.7) of one-sided mice with fibre implants ipsilateral (left column) and contralateral (right column) to the associated stimulus, as well as balanced mice (middle column). Data is shown for correct trials with 0.4s<RT<0.6s and stimuli ipsilateral (cyan) or contralateral (fuchsia) to the recording fibre, as well as rewarded zero-contrast trials. Trials were further split by whether the choice made was ipsilateral (solid) or contralateral (dashed) to the fibre. The shaded region indicates the time period in which rewards were delivered. In all panels, error bars indicate +/- SEM across mice. When indicated, recorded DA levels were normalised to correct for non-task-relevant day-by-day variations in fluorescence. (D), Wheel position (mm, absolute value i.e. non-directional) and simultaneous DLS DA recordings aligned to stimulus onset in expert one-sided days with matched accuracy on trials with associated and non-associated stimuli (accuracy difference < 0.1). Data is shown for correct trials with 0.4s<RT<0.6s with associated (black solid) and non-associated (black dashed) stimuli, as well as rewarded zero-contrast trials where the associated (gray solid) or non-associated (gray dashed) choice was made. As in panel C, shaded region indicates the time period in which rewards were delivered. (E), Trial-wise stimulus-aligned DLS DA recordings from 4 example expert days of an example balanced mouse (only correct trials with reward delivery <1.5s after stimulus onset), grouped by the contrast of the stimulus presented on the trial (pooling left and right stimuli) and sorted by RT within each contrast group. Blue line indicates the time of stimulus onset, green and light blue dots indicate the time of choice completion and reward delivery respectively.

We recorded DA release in the DLS every day from day one until expert behaviour using fibre photometry and GRAB-DA [257] (Figure 3.2A). DLS DA release occurred in response to specific events within a trial and changed throughout learning. A linear deconvolution analysis revealed that visual stimulus onset, arrival of the stimulus to the centre of the screen in correct trials (i.e., 'stim. centre'), and water reward (or its absence in incorrect trials) were the main events modulating DA release (Figure 3.2B). Choice start modulated DA release much less compared to the other events. Consistently, wheel movements had negligible representation in DA responses (Figure 3.2C-E). Aligning wheel movement measures to the onset of the visual stimulus showed how DLS DA responses emerge before significant wheel movement and differ considerably across trial types despite similar wheel movements (Figure 3.2C-D). Further, the DLS DA response to the stimulus appeared reliably ~ 0.2s after stimulus onset across trials with variable RT (Figure 3.2E). This indicates that the DLS DA signals we recorded are better locked to the visual stimulus rather than to wheel movement. Though this does not refute previous accounts of dopamine's involvement in action initiation and execution, it suggests that visual stimulus onset and rewards are the principal events triggering DLS DA release in the task.

For our analyses, we defined a DA 'stimulus' and 'outcome' response to examine changes over learning (Section 3.2.8). The 'stimulus' response quantified DA release levels in a short time window after stimulus onset. The 'outcome' response was defined as the sum of DA responses to completion of choice (marked by the stimulus arriving to the centre of the screen in correct choices, or by the stimulus leaving the screen in incorrect choices) and water reward delivery (or its absence). We summed these responses because the final stimulus position determines whether water reward will be subsequently delivered. Overall, DLS DA exhibited a very similar progression through learning as the TD error (Figure 3.1). In early days, DLS DA release mostly occurred in response to rewarded outcomes but not visual stimuli. As learning progressed, DA responses to visual stimuli grew and DA responses to water rewards diminished (Figure 3.2B; Figure 3.3A-E). In incorrect trials, i.e. when reward was not delivered, DA signals transiently decreased (Figure 3.2B). This demonstrated that dopamine tracks the learning of cue-outcome associations even over the long timescales over which the mice learned the task.

The development of DA responses during learning reflected the diverse learning trajectories observed in behaviour. DA responses in individual mice showed strong correspondence to the development of their psychometric curves (Figure 3.3B). In mice developing more one-sided psychometric curves, DA stimulus responses emerged most strongly for stimuli presented on the associated side (top and bottom rows in Figure 3.3B-E). However, in mice developing more balanced psychometric curves, DA responses to stimuli presented on left and right sides were similar (middle rows in Figure 3.3B-E). The diverse behavioural trajectories were also reflected in the rewarded outcome DLS DA response; these DA responses were small after associated stimuli and large after non-associated stimuli (Figure 3.3B-E). From the perspective of the TD error hypothesis, this diversity in dopamine signals was surprising. All animals reached similar levels of accuracy and thus should have similar value estimates for the states preceding the outcome. Yet, only animals with balanced psychometric curves exhibited the pattern of DLS DA release expected from a conventional TD error encoding, with responses for both left and right stimuli.

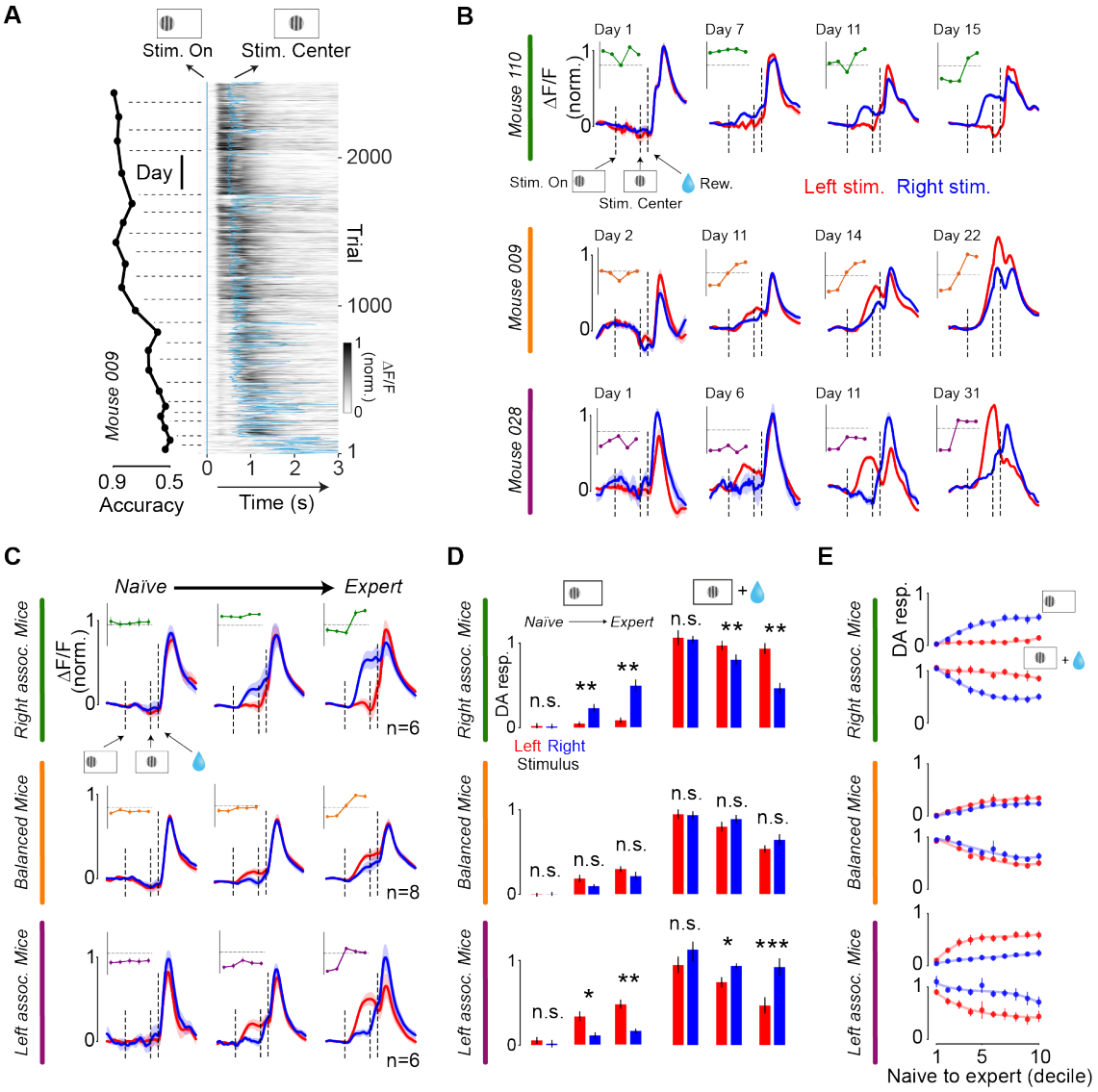


Figure 3.3: DLS DA signals develop over learning encoding stimulus-choice associations. (A), Accuracy over days and simultaneous trial-wise stimulus-aligned DLS DA recordings from an example mouse (only correct trials with reward delivery <5s after stimulus onset). Blue lines indicate the time of stimulus onset (left) and a 5-trial moving average of the time when the stimulus is brought to the centre (i.e., choice completion, right). When indicated, recorded DA levels were normalised to correct for non-task-relevant day-by-day variations in fluorescence. (B), Average time warped DLS DA signals in correct trials with stimulus on the left (red) and right (blue) for the same days and mice as in Figure 2.7A (psychometric curves shown as insets). Vertical dashed lines indicate stim. onset, stim. centre, and reward delivery time. Error bars indicate +/- SEM across trials. (C), Average time warped DLS DA signals in correct trials and average psychometric curves for the three clusters from Figure 2.9E in initial days (days 1-3, left), early days (days > 3 with accuracy n.s. greater than 0.5, middle), and expert days (accuracy n.s. smaller than 0.7, right). Error bars indicate +/- SEM across mice. (D), Quantification of DLS DA signals from panel C. Left, bar plots showing average DLS DA stimulus responses across mice in each cluster for correct trials in days 1-3, days >3 with accuracy n.s. greater than 0.5 and expert days (accuracy n.s. smaller than 0.7). p-values calculated using two-sided paired t-test. Right, same as left but for outcome responses. (E), Average stimulus and outcome DLS DA responses over deciles in correct trials for the three clusters from Figure 2.9E. Unless specified differently, error bars indicate +/- SEM across mice. Data points fit with a 3rd degree spline to visualise trends (scipy.interpolate.UnivariateSpline). p-values calculated using two-sided paired t-test. n.s., *, ** and *** indicate $p>0.05$, $p<0.05$, $p<0.005$ and $p<0.0005$ respectively.

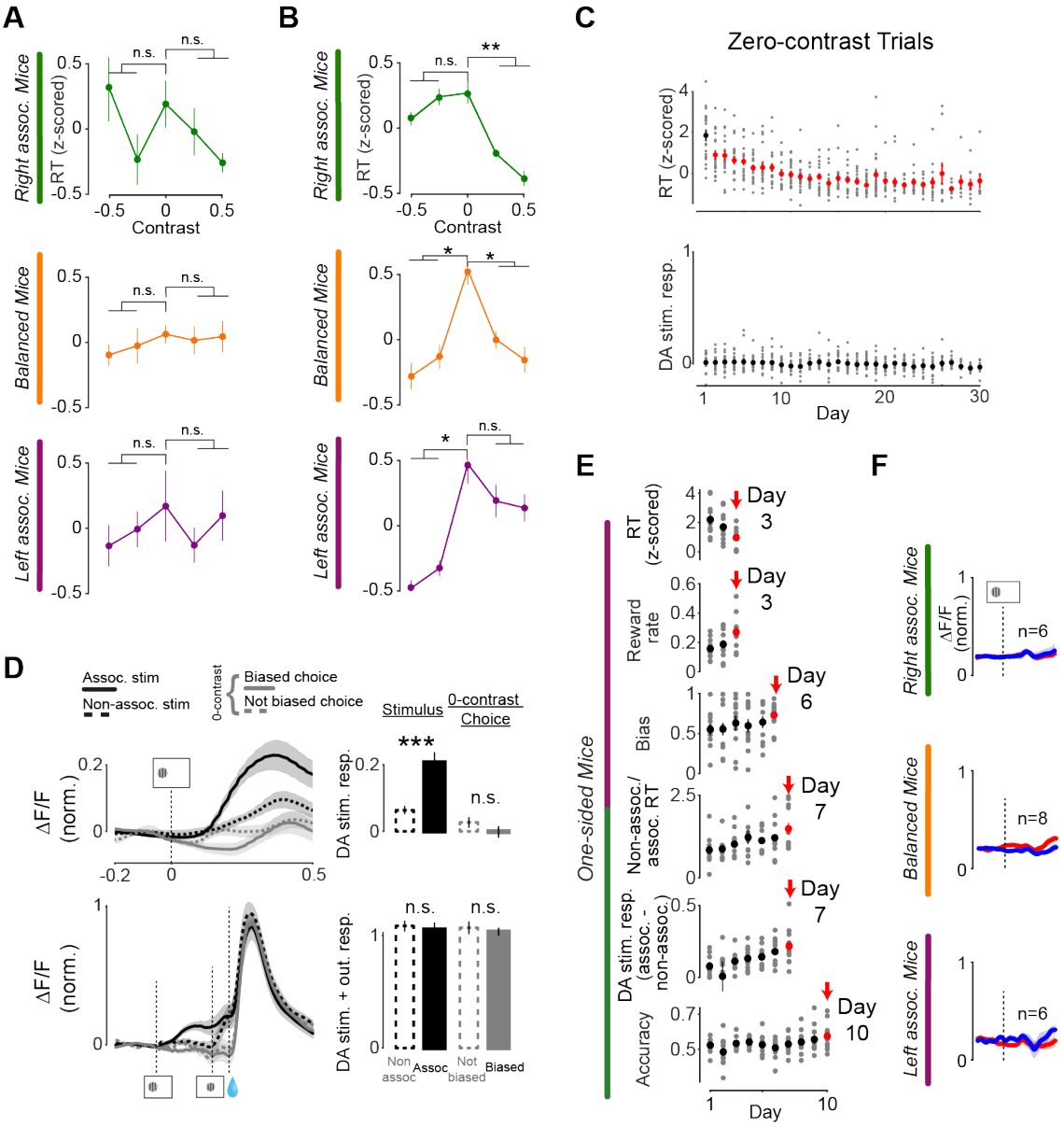


Figure 3.4: DLS DA stimulus responses are absent in initial days and zero-contrast trials. (A), Average chronometric curve across days 1-3 averaged across mice in each cluster. The z-score was computed per mouse by computing the mean RT per contrast across trials in each day, standardising these values across contrast values and days, and finally averaging across days. p-values calculated using two-sided paired t-test. (B), Average chronometric curve across days > 3 with accuracy n.s. greater than 0.5, averaged across mice in each cluster. p-values calculated using two-sided paired t-test. (C), Top, average z-scored RTs in zero-contrast trials over first 30 days of training per mouse (gray) and averaged across all mice (black/red). Red dots indicate data points significantly different from day 1 averages ($p<0.05$, estimated via two-sided t-test). Z-scoring was performed by standardising each mouse's mean zero-contrast RT per day using data from all its days. Bottom, average DLS DA stimulus responses in zero-contrast trials (average 0.2-0.35s after the stimulus would have been presented, i.e., 0.2 s before the auditory go cue) over days. (D), Top left, average stimulus-aligned DLS DA signals in correct trials from days with accuracy n.s. greater than 0.5 of one-sided animals. Solid and dashed gray lines show signals on zero-contrast trials (aligned to the time when the stimulus would have been presented, i.e., 0.2 s before the auditory go cue) for choices in (solid) and opposite (dashed) to the direction of each mouse's early bias. Bottom left, average time warped DLS DA signals. Top right, quantification of average DLS DA stimulus responses. Bottom right, the sum of average DLS DA responses to stimulus and outcome. p-values calculated using two-sided paired t-test. (E) Measures of general RTs, reward rate, zero-contrast bias, the ratio of RTs on trials without and with associated stimuli, the difference in the average DLS DA stimulus responses with and without associated stimuli and the accuracy over the initial days of learning for one-sided mice with DLS DA recordings. Data points are shown per mouse (gray) and averaged across all mice (black/red). Red dots indicate data points significantly different from day 1 averages ($p<0.05$, estimated via two-sided t-test). Data points for each measure are shown until significance was reached, indicated with a red arrow. (F), Stimulus-aligned DLS DA signals in correct trials of the first day of training. No stimulus responses were significantly greater than 0 ($p>0.05$; one sample two-sided t-test).

In early days, when both psychometric and chronometric curves were still flat, there were no DA responses to visual stimuli (Figure 3.3C-D, first column; Figure 3.4A). Subsequently, as RTs started to show signatures of mice using visual stimuli (i.e. non-flat chronometric curves), DA stimulus responses emerged while psychometric curves remained flat (Figure 3.3C-D, middle column; Figure 3.4B). These DA responses were strongest for stimuli with faster RTs, which appeared on the same side as the choice bias in one-sided mice (Figure 3.3C, middle column; Figure 3.4B). We asked whether these signals reflected associations that the animals were forming between stimuli and choices, or whether they simply reflected each animal's bias. To address this, we inspected trials without visual stimuli (zero-contrast trials). We did not observe any significant DA responses before outcome delivery in these trials, despite RTs decreasing over learning and choice biases in one-side mice (Figure 3.4C and D). Thus, DLS DA stimulus responses emerged only when animals started to form associations between visual stimuli and choices, the first signature of this appearing in RTs and not choice accuracy (Figure 3.4E). DA stimulus responses were not evident on the first day of the experiment (Figure 3.4F), indicating that they emerge over learning, rather than reflecting stimulus novelty [258, 242, 259]. Signatures of stimulus-choice associations being formed were also evident in DA responses at trial outcome (Figure 3.3E). Thus, during early days, DA responses emerged, reflecting the first signatures of learning to associate stimuli and choices.

In later days, as psychometric curves developed, DLS DA responses strongly reflected the growing psychometric slopes observed across animals (Figure 3.3B-E). In right-associating mice (i.e., those with a steep psychometric curve on the right and a flat curve on the left), DA responses to stimuli were evident in response to right but not left stimuli (Figure 3.3C, top row, right). This pattern was opposite in left-associating mice (Figure 3.3C, bottom row, right). In balanced mice (i.e., those associating both stimuli with their corresponding choices), both left and right stimuli elicited strong DA responses (Figure 3.3C, middle row, right). The relationship between psychometric slopes and DA responses to stimuli was maintained even after controlling for differences in the accuracy of choices after left and right stimuli, i.e., selecting days where the accuracies were matched but the slopes could differ (Figure 3.5A-D). Thus, in left- and right-associating mice, DA responses to associated stimuli were significantly larger than to non-associated stimuli, despite equally high choice accuracy (Figure 3.5A). This decoupling between accuracy and DA responses to stimuli could be seen throughout learning in one-sided mice (Figure 3.5C, bottom). It also held when matching wheel movement traces for associated and non-associated stimulus trials in addition to the accuracy Figure 3.2D.

DLS DA's encoding of stimulus-choice association held independent of stimulus laterality with respect to the recorded brain hemisphere (Figure 3.6A-B). Moreover, the encoding was largely invariant to RTs (Figure 3.6C-D) and the accuracy of pending choice (i.e., correct/incorrect, Figure 3.6E). These observations held in another cohort of animals (n=6) that we trained with more graded levels of stimulus contrast to high, asymptotic performance (accuracy 80-95%) (Figure 3.6F). DLS DA responded to the outcome of correct and incorrect trials, reflecting the difference between outcome value and the learned stimulus-choice association, i.e. the reward that animals learned to predict using visual stimuli. Hence, these DA responses decreased as stimulus-choice associations increased (Figure 3.6E, middle row; Figure 3.3E), instead of reflecting choice accuracy (Figure 3.5A-B). Taken together, in later days of learning, DLS DA signals developed substantial responses to visual stimuli, and a corresponding decrease in responses to trial outcome, only if those stimuli were associated with a choice. Thus, the DLS DA signals resemble a conventional RPE, but reflect the strategy each animal uses to perform the task.

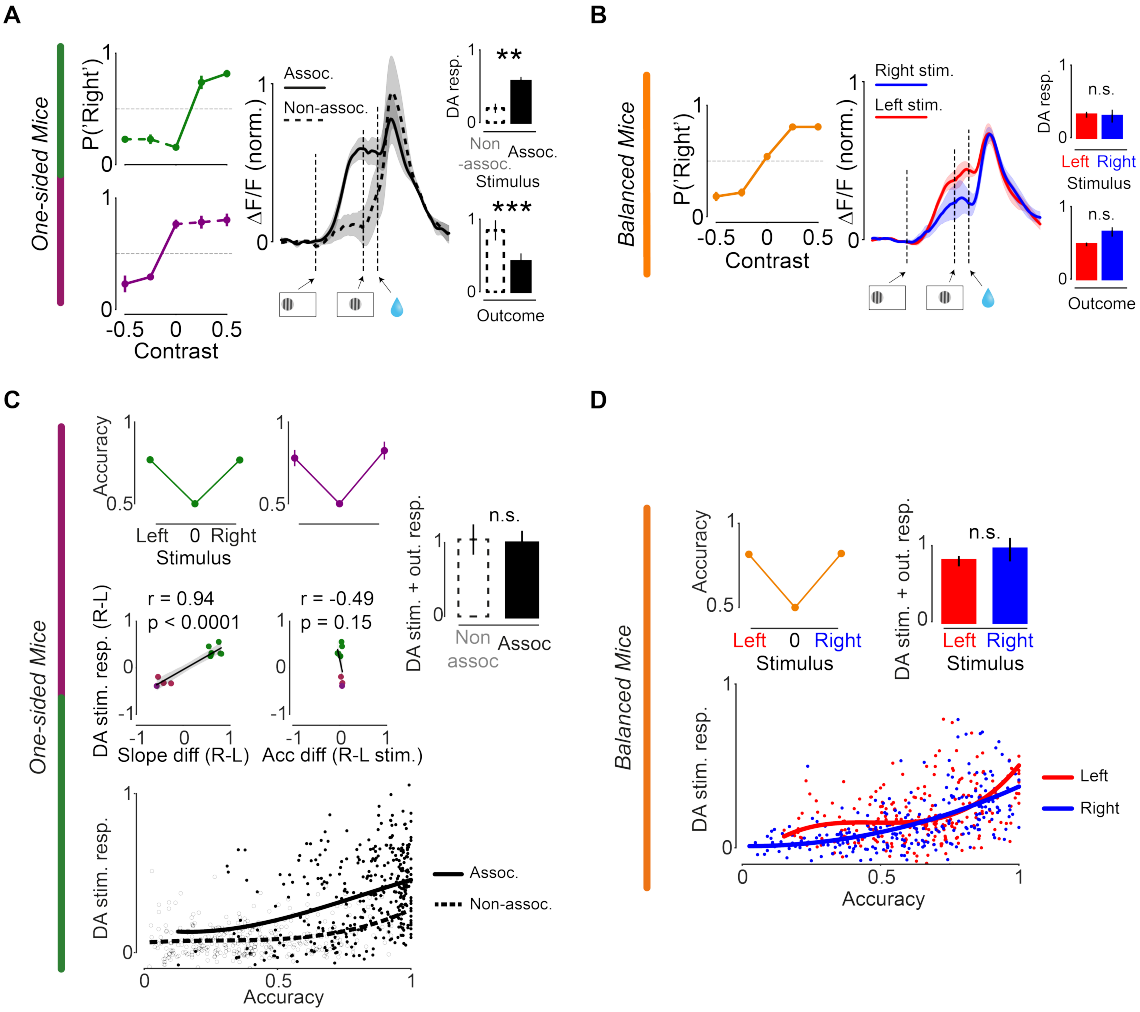


Figure 3.5: DLS DA in matched accuracy days. (A), Analysis of left- and right-associating mice DLS DA responses in expert 'matched accuracy' days (difference in left and right stimulus trial choice accuracy is <0.1). Left, average psychometric curves from matched accuracy days. Middle, average time warped DLS DA signals in correct trials. Right, DA responses to stimuli and outcome in correct trials. p-values calculated using two-sided paired t-test. (B), Same matched accuracy analysis as in panel A but applied to balanced mice. (C), Top, for one-sided mice in matched accuracy days, in clockwise order from top left, average accuracy across trial types; bar plots with sum of average DLS DA stimulus and outcome responses in correct trials; regression of difference in R-L psychometric slopes (left) and choice accuracy in R-L stimulus trials (right) against difference in DLS DA responses to R-L stimuli in correct trials. Each point in the regressions represents a matched-accuracy day, and the p-value is calculated from the exact distribution of r. Shaded region indicates 95% confidence interval. Bottom, for one-sided mice, scatter plot of average DLS DA stimulus responses against accuracy in associated (solid) and non-associated (dashed) stimulus trials. Each point represents a day. Data points fit with a 3rd degree spline to visualise trend (scipy.interpolate.UnivariateSpline). (D), Similar analysis to that in panel C reproduced for balanced mice. n.s., *, ** and *** indicate p>0.05, p<0.05, p<0.005 and p<0.0005 respectively.

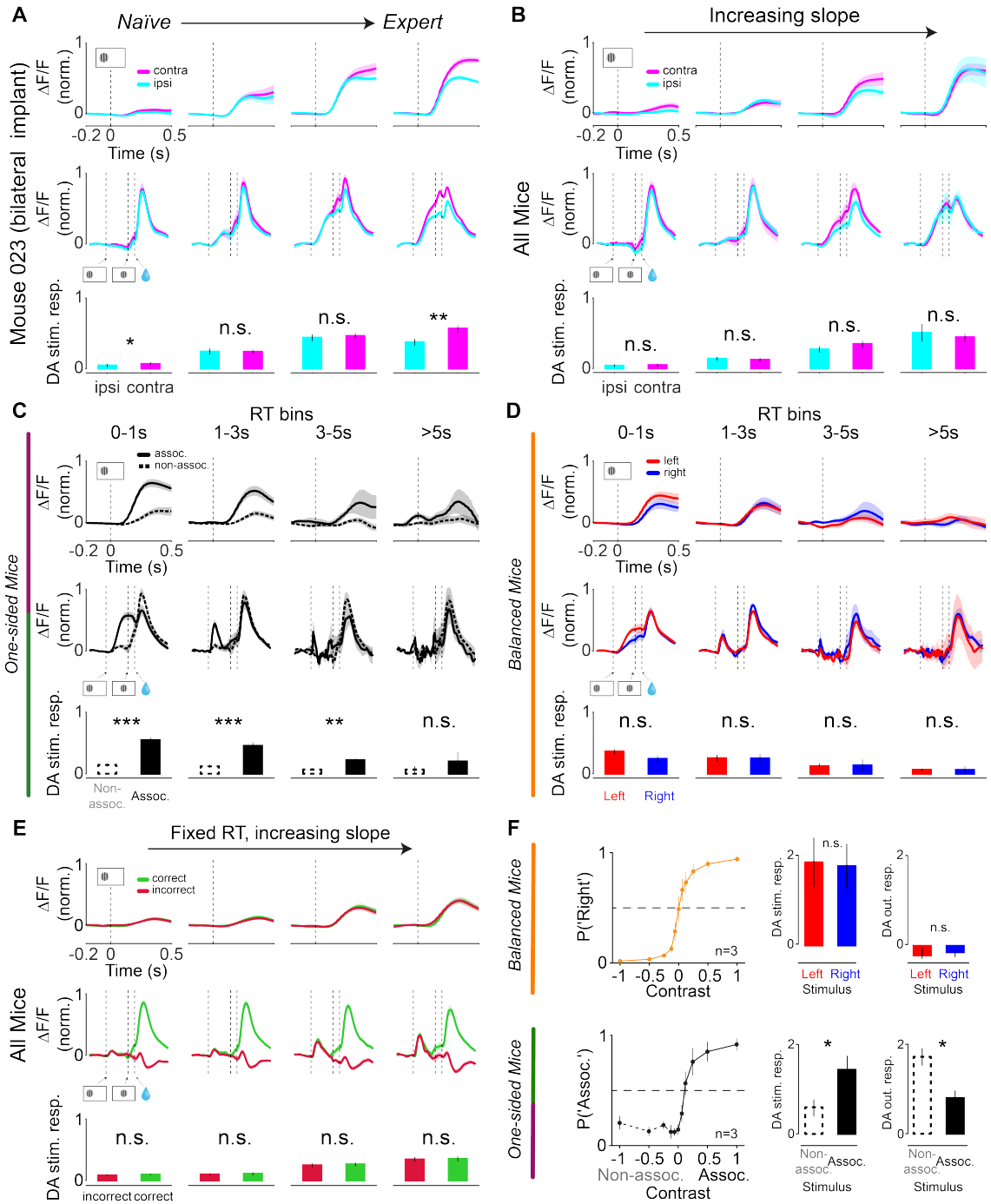


Figure 3.6: DLS DA signals during learning and in highly expert mice; split by stimulus type, response times and trial outcome.
(A), Average DLS DA stimulus-aligned signals (top), time warped signals (middle) and bar plots with average DLS DA stimulus responses (bottom) for rewarded trials from each quartile of an example mouse with bilateral implants, split by stimulus laterality with respect to the implanted fibre. Error bars indicate +/- SEM across days. p-value calculated from two-sided paired t-test. (B), Same as panel A for increasing psychometric slope values of all mice, also split by stimulus laterality with respect to the implanted fibre. In all panels, error bars indicate +/- SEM across mice, unless specified differently. (C), Average DLS DA stimulus-aligned signals (top), time warped signals (middle) and bar plots with average DLS DA stimulus responses (bottom) in rewarded trials for increasing RT values in expert days (accuracy > 0.7) of one-sided mice split by trials with the associated (solid) or non-associated stimulus (dashed). p-value calculated from two-sided paired t-test. (D), Same as panel C for balanced mice. (E), Average DLS DA stimulus-aligned signals (top), time warped signals (middle) and bar plots with average DLS DA stimulus responses (bottom) across all mice for increasing psychometric slope and a fixed RT range (1-3s) split by correct and incorrect trials. (F), Psychometric curves, average DLS DA stimulus responses and outcome responses in correct trials of highly expert (accuracy 80% – 95%) balanced (top) and one-sided (bottom) mice trained with shaping, more contrast levels and more extensive training. n.s., *, ** and *** indicate $p>0.05$, $p<0.05$, $p<0.005$ and $p<0.0005$ respectively.

The dopamine signals we observed are specific to DLS; we did not observe them in DA release recordings from the DMS under the same experimental conditions (Figure 3.7A and D). DMS dopamine release showed four main characteristics. First, DMS DA responses to reward were negligible from the start of learning, unlike DLS DA (Figure 3.7B, left column; Figure 3.8A-C). Second, DMS DA signals, unlike DLS DA, depended on the recorded brain hemisphere, showing particularly strong responses to stimuli contralateral to the recorded hemisphere and near zero responses to ipsilateral stimuli (Figure 3.7B, right column; Figure 3.8B, D and E). Third, these DMS DA contralateral stimulus responses were modulated by the animal's strategy; they were only present if the contralateral stimulus was associated with choice (Figure 3.7C; Figure 3.8E). Lastly, similar to DLS, DMS DA stimulus responses emerged over learning: they were absent in the initial days of the experiment (Figure 3.7B, left column; Figure 3.8F). We confirmed these observations using a linear deconvolution (Figure 3.7E). The negligible DMS DA responses to rewards in early days indicate that these signals do not encode the RPE that we observed in DLS DA, which showed strong responses to reward in initial days.

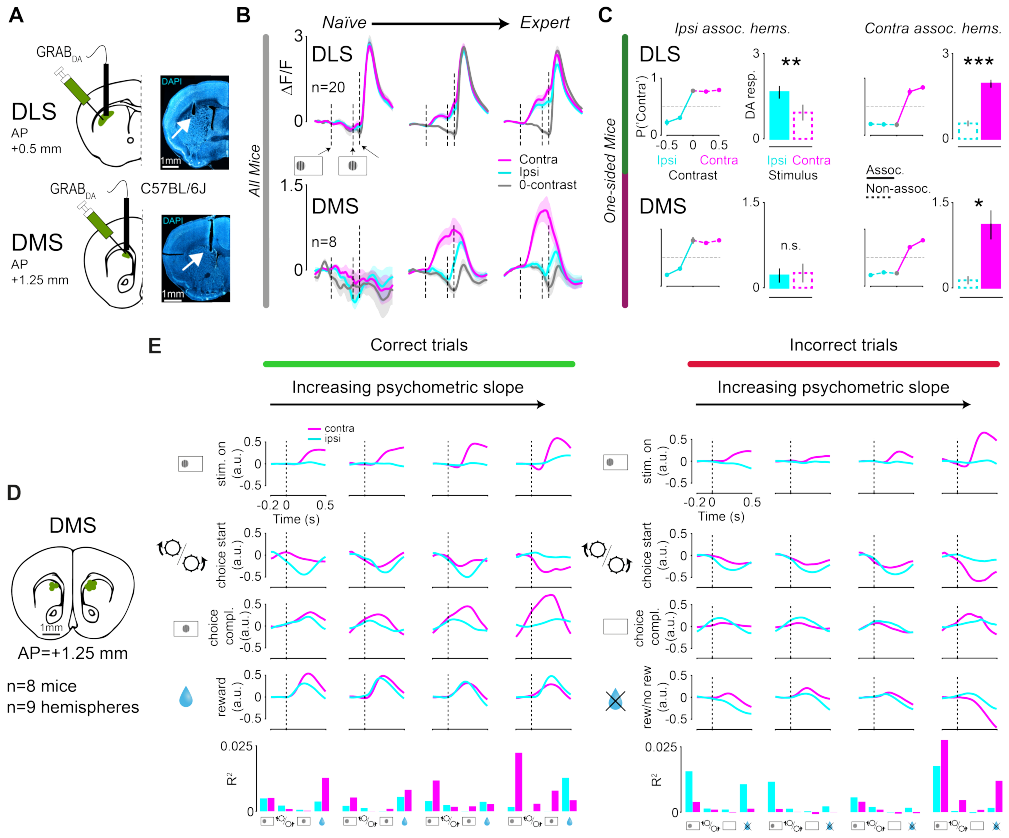


Figure 3.7: DLS and dorsomedial striatum (DMS) DA signals over learning.
(A), Fibre photometry set-up for recording DA release in dorsolateral (DLS) and dorsomedial striatum (DMS). (B), Average time warped DLS and DMS DA signals in correct trials with stimulus ipsilateral (cyan), contralateral (fuchsia) or 0-contrast trials (gray) in initial days (days 1-3, left), early days (days > 3 with accuracy n.s. greater than 0.5, middle), and expert days (accuracy n.s. smaller than 0.7, right). (C), Same matched accuracy analysis as in Figure 3.5A for one-sided mice, but now with recordings separated by recording site (DLS, DMS) and whether the associated stimulus is ipsilateral/contralateral to the recorded hemisphere. Error bars indicate +/- SEM across matched accuracy days. p-values calculated using two-sided paired t-test. (D), The locations of DMS optic fibre implants estimated from post-mortem histological examinations overlaid on a brain slice schematic. (E), Same linear deconvolution analysis as in Figure 3.2B, now applied to DMS DA signals across learning. For the DMS DA deconvolution analysis, trials were further subdivided into whether the stimulus was ipsilateral (cyan) or contralateral (fuchsia) to the fibre. n.s., *, ** and *** indicate p>0.05, p<0.05, p<0.005 and p<0.0005 respectively.

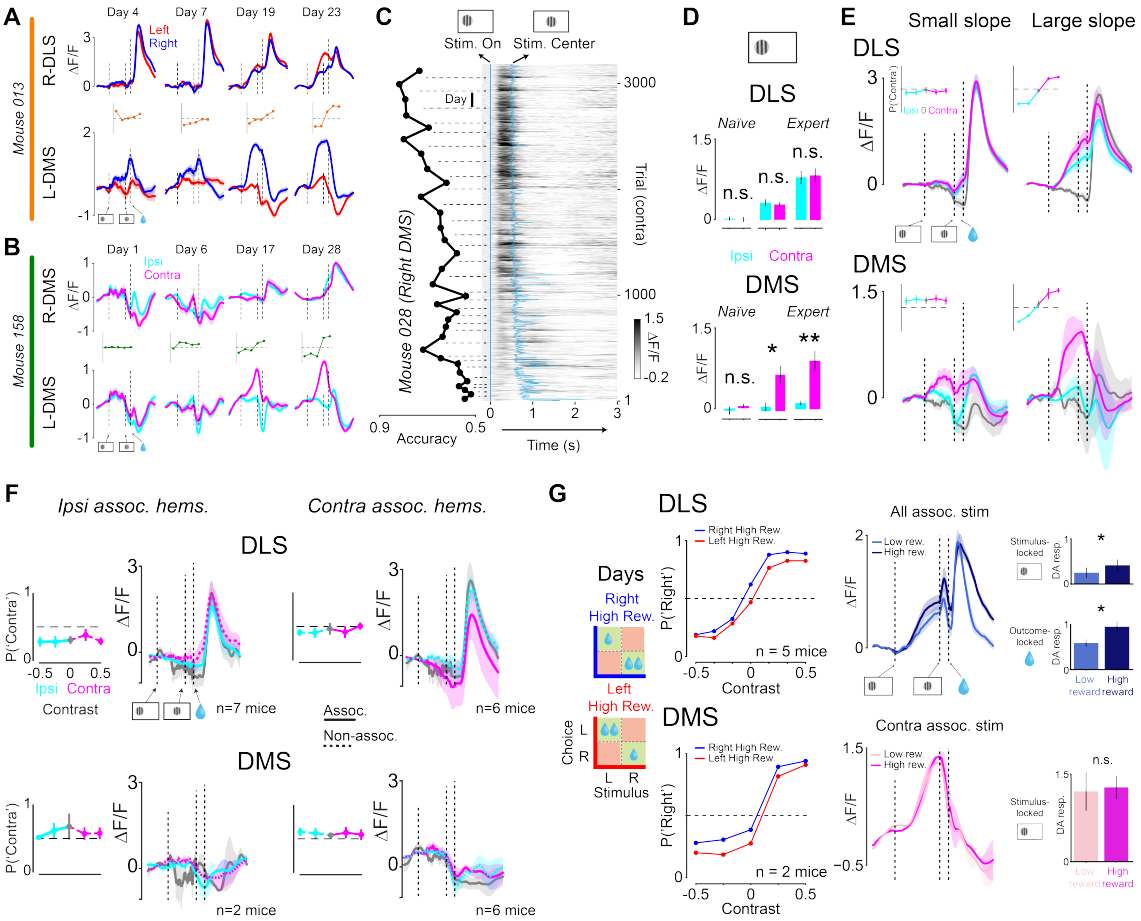


Figure 3.8: Further comparison of DLS and DMS DA signals and reward size manipulation experiment.
(A), Average time warped DLS (top) and DMS (bottom) DA signals in correct trials with stimulus on the left (red) and right (blue) for 4 example days of an example balanced mouse with bilateral DLS/DMS recordings. Average psychometric curves on each day shown as insets. Vertical dashed lines indicate stimulus onset, stimulus centre, and reward delivery time. Error bars indicate +/- SEM across trials. (B), Average time warped DMS DA signals in correct trials with stimulus ipsilateral (cyan) and contralateral (fuchsia) to the recording fibre for 4 example days of an example one-sided mouse with bilateral DMS recordings. (C), c.f. Figure 3.3A, accuracy over days and simultaneous trial-wise stimulus-aligned DMS DA recordings from an example mouse (only correct trials with stimulus contralateral to the recording fibre and reward delivery <5s after stimulus onset). Blue lines indicate the time of stimulus onset (left) and a 5-trial moving average of the time when the stimulus is brought to the centre (i.e., choice completion, right). (D), c.f. Figure 3.7B, bar plots showing average DLS and DMS DA stimulus responses across mice for correct trials in days 1-3, days >3 with accuracy n.s. greater than 0.5 and expert days (accuracy n.s. smaller than 0.7). In all panels, error bars indicate +/- SEM across mice, unless specified differently. p-values calculated using two-sided paired t-test. (E), Average time warped DLS and DMS DA signals on days with small (<0.05) and large (>0.2) psychometric slope in correct trials with stimulus ipsilateral (cyan) and contralateral (fuchsia) to the recording fibre, as well as zero-contrast trials (gray). (F), First day of training time warped DLS and DMS DA signals in correct trials with contralateral/ipsilateral stimulus, as well as zero-contrast trials. Average first-day psychometric curves shown as insets. No stimulus responses were significantly different from zero-contrast trial responses (p>0.05; one-way ANOVA). (G), In order, schematic, behavioural results and DLS (top) & DMS (bottom) DA results of the reward size manipulation experiment (Section 3.2.7). Left, schematic showing the alternating day structure of the experiment. Middle left, average psychometric curves on days with high reward for correct right stimulus (blue) and left stimulus (red) trials. Middle right, average time warped DLS and DMS DA signals in correct trials with the associated stimulus (only contralateral for DMS) separated by reward size. Right, bar plots quantifying stimulus and outcome DA responses. p-values calculated using two-sided paired t-test. Error bars indicate +/- SEM across mice. DMS stimulus responses for low and high reward are n.s. different both when two-sided t-test is performed over mice and over days. n.s., *, ** and *** indicate p>0.05, p<0.05, p<0.005 and p<0.0005 respectively.

The dynamics of DLS DA signals throughout learning showed striking similarities to the behavioural learning trajectories (Figure 3.9). Replacing left and right slope measures with average DLS DA responses to left and right stimuli in the plots from Figure 2.9 yielded neural trajectories that closely matched their behavioural counterparts. Similar to behaviour, DLS DA signals developed systematically. As such, from DA signals throughout the experiment, it was possible to infer the animals' past and future DA signals (Figure 3.9A; Figure 3.10A-B) and behavioural strategies (Figure 3.9B). Despite the similarity of the neural and behavioural trajectories, DA responses to stimuli developed across days reflecting the dynamics of the associations each animal formed between visual stimuli and choices (Figure 3.9C-E), and not directly psychometric slope. This is most clearly seen by comparing Figure 2.9C and Figure 3.9A where, in the former, the trajectories for one-sided mice evolve more sigmoidally than their neural counterparts. This agrees with our observation from Figure 3.4E that DLS DA stimulus responses grow before psychometric slopes develop, reflecting stimulus-choice associations whose first signature is seen in RTs and not choices. DA responses to rewards developed over learning mirroring the evolving stimulus responses (Figure 3.9F; Figure 3.10B-E), and hence also tracked the behavioural trajectories closely. These results demonstrate that dopamine signals during learning evolve reflecting each intermediate strategy employed by the mice, suggesting a crucial role for dopamine in shaping the learning trajectories.

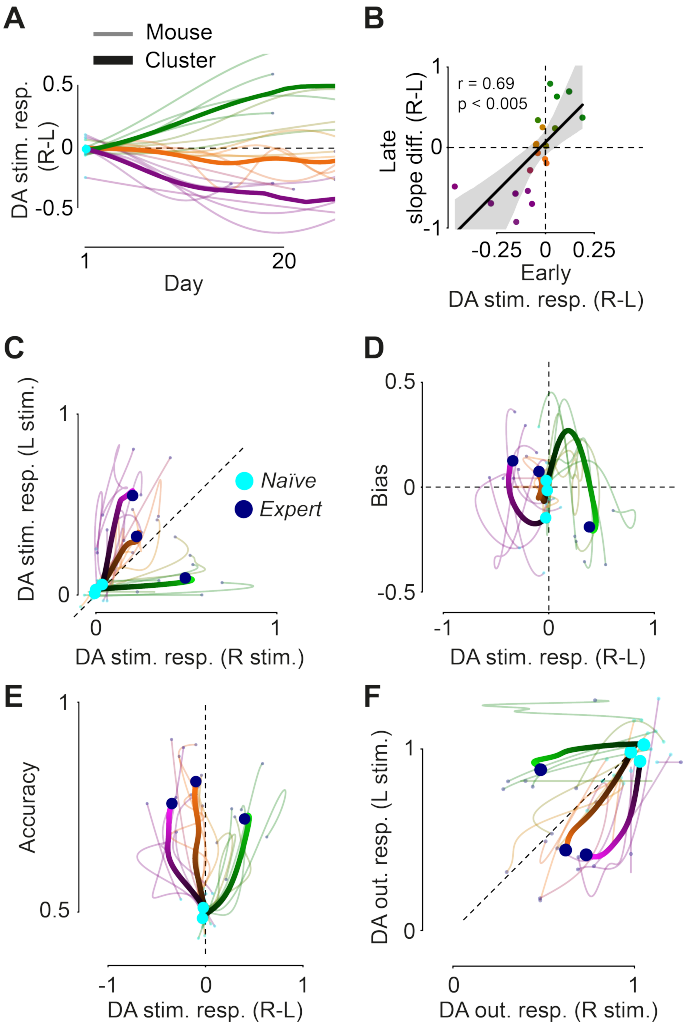


Figure 3.9: DLS DA signals reflect learning trajectories from naïve to expert.
(A), Difference in DLS DA responses to right and left stimuli (R-L) over days per mouse (thin) and for the 3 clusters from Figure 2.9E (thick). Number of days limited to 25. Neural trajectories here and in other panels plotted with colours and clusters obtained from behaviour. (B), Regression of early difference in DLS DA responses to R-L stimuli (average across days 4-8) against late R-L slope difference (average across final 5 days of training). Each point represents a mouse. p-value is calculated from the exact distribution of r. Shaded region indicates 95% confidence interval across mice. (C)-(F), In order, right vs. left DLS DA stimulus responses; difference in DLS DA responses to right and left stimuli (R-L) vs. bias; difference in DLS DA responses to right and left stimuli (R-L) vs. accuracy; and DLS DA rewarded outcome responses in right-stimulus trials vs. left-stimulus trials across days per mouse (thin) and for the 3 clusters from Figure 2.9E (thick). The hue of the cluster lines changes from dark to light indicating progress through learning.

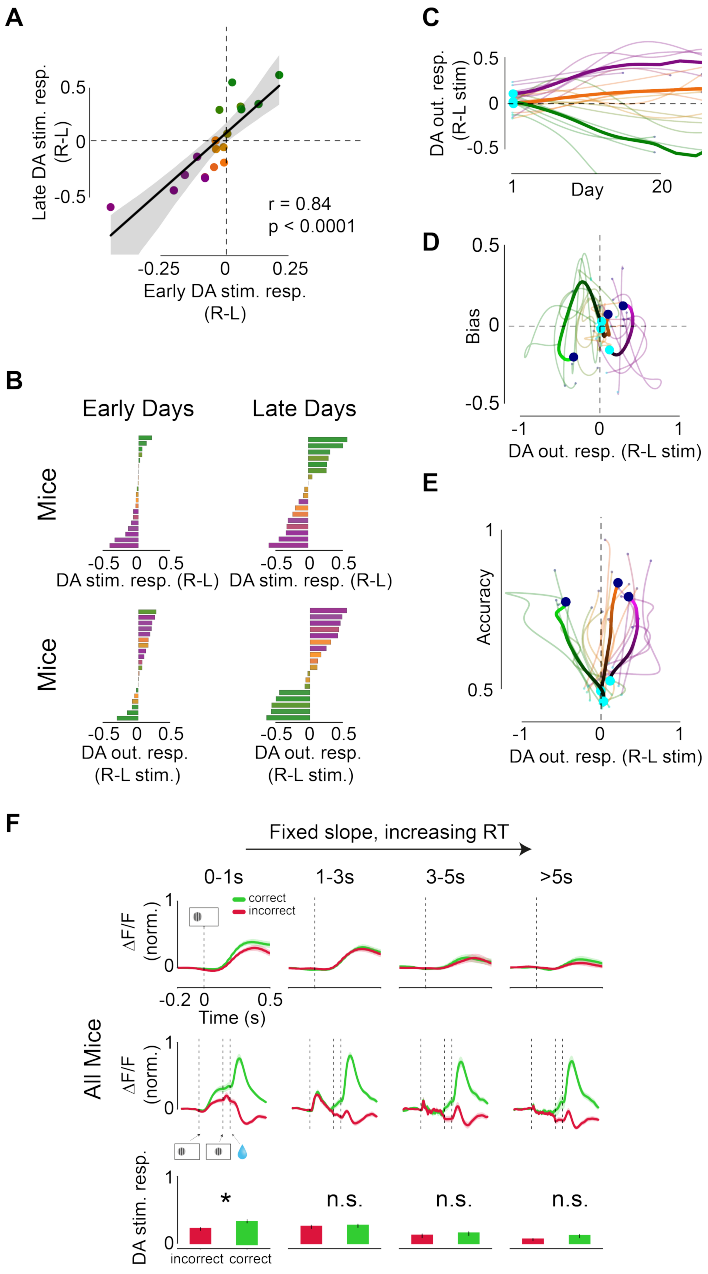


Figure 3.10: Further quantification of DLS DA responses to task events during learning. (A), Regression of early difference in DLS DA responses to R-L stimuli (average across days 4-8) against late difference in DLS DA responses (average across final 5 days of training). Each point represents a mouse. p-value calculated from the exact distribution of r. Shaded region indicates 95% confidence interval. (B), Average difference in DLS DA responses to right and left stimuli (R-L) and difference in DLS DA rewarded outcome responses after R-L stimuli in early days (4-8) and late days (final 5) for each mouse, c.f. Figure 2.10C for behavioural data. Colours come from average psychometric slope asymmetry (see Figure 2.9E). (C), Difference in DLS DA responses to rewarded outcomes after right and left stimulus (R-L) over days per mouse (thin) and for the 3 clusters from Figure 2.9E (thick). Number of days limited to 25, c.f. Figure 3.9A. (D), Difference in DLS DA responses to rewarded outcomes after R-L stimulus vs. bias, c.f. Figure 3.9D. (E), Difference in DLS DA responses to rewarded outcomes after R-L stimulus vs. accuracy, c.f. Figure 3.9E. (F), Average DLS DA stimulus-aligned signals (top), time warped signals (middle) and bar plots with average DLS DA stimulus responses (bottom) across all mice for increasing RT values and a fixed psychometric slope range (>0.325), split by correct and incorrect trials. p-values calculated using two-sided paired t-test. Error bars indicate +/- SEM across mice. n.s., and * indicates p>0.05 and p<0.05.

3.3.2 Optogenetic manipulations

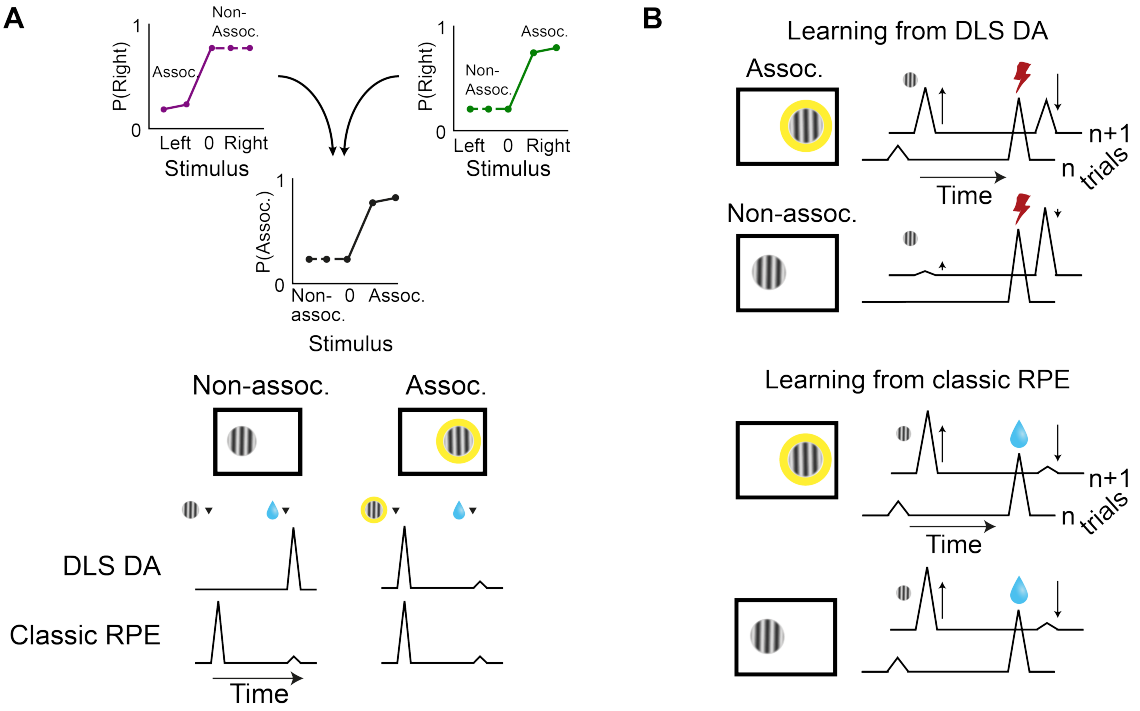


Figure 3.11: DLS DA key characteristics and causal predictions.
(A), Schematics describing the terminology and motivation for the optogenetic experiments. Top, schematic explaining the relabelling of the axes from left/right to non-associated/associated, such that all one-sided mice look like right-associating mice. Bottom, schematic showing the observed patterns of DLS DA release in expert one-sided animals for trials with non-associated and associated stimuli, alongside the signals expected from a classic RPE. (B), Schematic comparing the predicted learning effect of stimulating DLS DA with that expected from a classic RPE, on trials with the associated and non-associated stimulus.

The DLS DA responses we observed have two main characteristics. First, one-sided animals did not learn to associate one of the visual stimuli with a choice (Figure 3.11A), despite having DLS DA reward signals in initial days comparable to those in the trials with associated stimuli (Figure 3.3C). This suggests that these animals might not have a neuronal representation of the non-associated stimulus which can be trained using DLS DA reward signals. Second, in one-sided animals DLS DA did not respond during decisions made based on the absence of visual stimuli (e.g. trials with non-associated stimuli), even when these decisions were as accurate as those made based on the associated visual stimuli (Figure 3.11A and Figure 3.5A). These characteristics cannot be explained by a classic RPE estimated using all task-relevant cues (e.g. the visual stimulus or its absence) to predict the upcoming reward. Such RPEs would reflect the value (i.e., expected reward, proportional to accuracy) of each task-relevant cue, not only that of visual stimuli (Figure 3.11A). In contrast, the DLS DA responses reflect predictions and RPEs estimated only using a subset of the task-relevant cues, i.e., the visual stimuli. Based on these two characteristics, we predicted that DLS DA reward responses only update associations between choices and visual stimuli used for decisions (i.e., associated stimuli). This differs from the learning effect of classic RPEs that update associations between choices and all task-relevant cues (Figure 3.11B). To test this, we performed two optogenetic experiments: inhibiting DLS dopamine release throughout learning and stimulating DLS dopamine release in animals after learning.

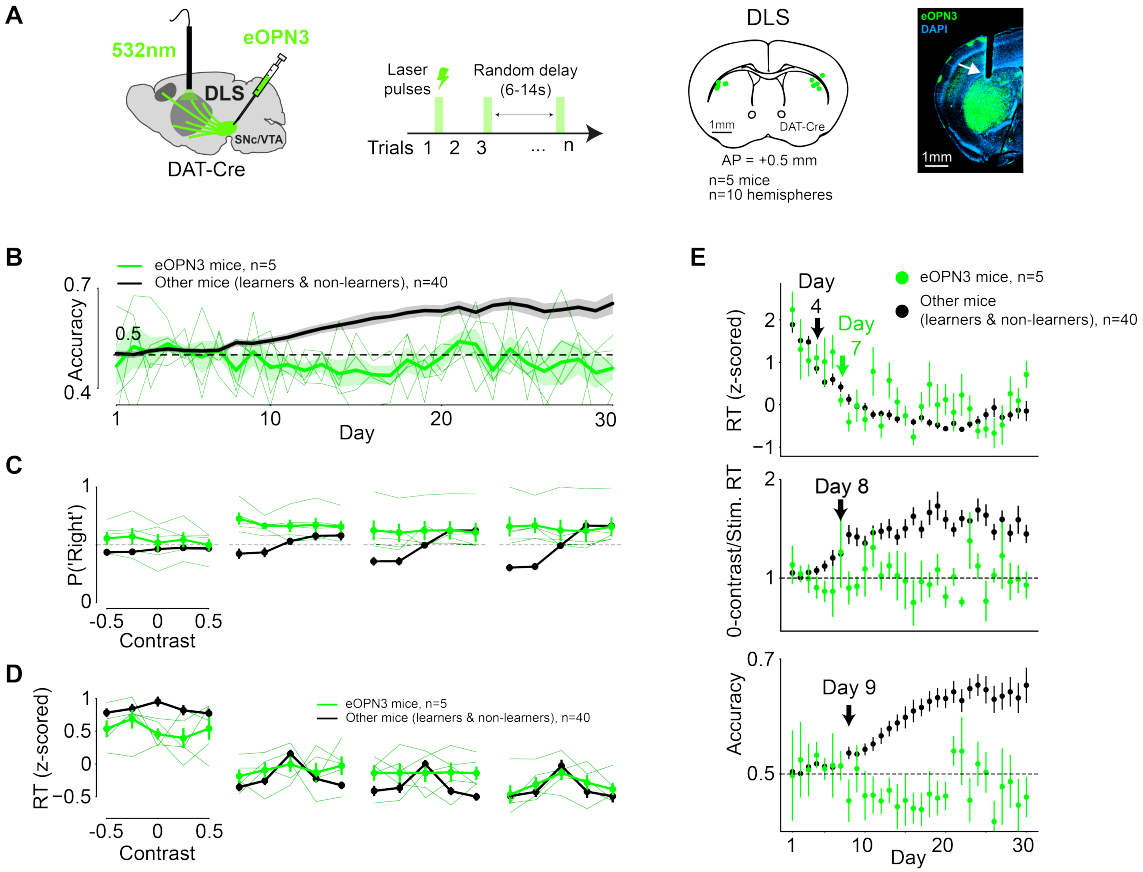


Figure 3.12: DLS DA is required to learn.
(A), Left, schematic for optogenetic experiment using the inhibitory opsin eOPN3. Laser pulses were delivered in random intervals of 6-14s, independent of task trials. Right, the locations of DLS optic fibre implants of eOPN3 mice estimated from post-mortem histological examinations, and example histology slice showing fibre tract (white arrow) with eOPN3 opsin expression. (B), Accuracy over days for individual eOPN3 mice (green, thin), averaged across eOPN3 mice (green, thick) and averaged across all the other mice that were trained on the task for at least 4300 trials (black). In all panels, error bars around averages indicate +/- SEM over mice. Number of days limited to 30. (C), Psychometric curve over quartiles. (D), Chronometric curve over quartiles. For each mouse, the z-score was computed by first finding the mean RT per contrast across trials in each day, standardizing these values across days and contrasts, and finally averaging across days in each quartile. (E), Measures of general RTs, the ratio of RTs on zero-contrast trials and trials with stimuli and accuracy over the first 30 days of training. First plot, z-scored RTs (across all trial types) over training days. Arrows indicate the first data points that were significantly different from day 1 averages ($p<0.05$, estimated via two-sided t-test). Z-scoring was performed by standardising each mouse's mean RT per day using data from all its days. Second plot, the ratio of the average (non-zscored) RT in zero-contrast trials to the average RT in left/right stimulus trials. Third plot, accuracy over training days. For all measures, only days with more than 50 trials were considered.

Longitudinal inhibition of DLS DA (Figure 3.12A) substantially impaired learning. Animals with inhibition showed significantly lower accuracy compared to all other mice that performed a similar number of trials (n=40, learners and non-learners, Figure 3.12B). The accuracy remained around 50% (chance level) and psychometric slopes did not develop, indicating that these mice did not learn to use the visual stimuli for decisions (Figure 3.12B-C). Moreover, compared to the other mice, RTs developed a dependence on visual stimuli much later in training (chronometric curves remained flat; Figure 3.12D-E). However, DLS DA inhibition largely spared other signatures of learning that did not rely on the stimuli, i.e., developing bias and an overall decrease in RT (Figure 3.12C-E). These results indicate that DLS DA is necessary for learning the task and forming stimulus-choice associations, but is not involved in learning other relevant associations.

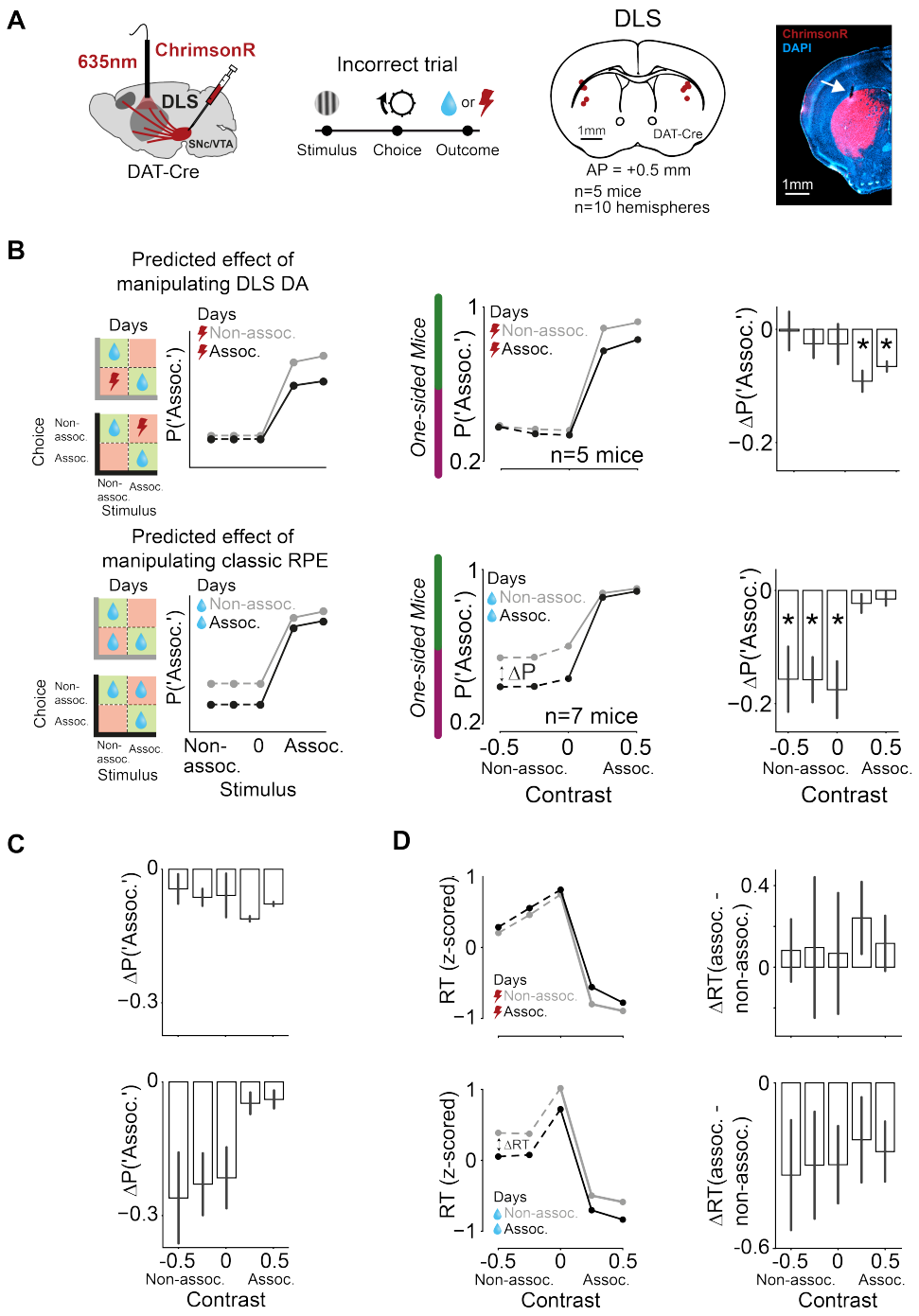


Figure 3.13: DLS DA updates behaviour distinct from classic RPE.
(A), Left, schematic for optogenetic experiment using the excitatory opsin ChrimsonR in DLS. Right, the locations of DLS optic fibre implants of ChrimsonR mice estimated from post-mortem histological examinations overlaid on a brain slice schematic, and example histology slice showing fibre tract (white arrow) with ChrimsonR opsin expression (Section 3.2.4). (B), Predicted and observed effects of ChrimsonR DLS optogenetic experiment and manipulating classic RPEs. Left, predicted results of stimulating DLS DA (top) and manipulating classic RPEs (i.e. delivery of water, bottom) at the outcome of error trials, switched in alternating days of experimentation. Right, average psychometric curves, and their difference, on days with DLS DA stimulation (top) and reward delivery (bottom) on error trials with non-associated (gray) and associated (black) stimuli. p-values calculated using two-sided one-sample t-test. * indicates $p<0.05$. (C), Average difference in psychometric curves with stimulation/water delivery in incorrect trials with the associated and non-associated stimuli (i.e., assoc. – non-assoc.) for mice in which both the ChrimsonR DLS stimulation and reward delivery experiments were performed (n=3). (D), Average chronometric curves (left) and a bar plot of their differences (right) for the mice and manipulations used in the ChrimsonR optogenetic (and water-delivery) experiment. Z-scored RTs defined as in panel B. Only trials with RT<12s were analysed for the stimulation experiment, and RT<5s were analysed for the water-delivery experiment.

Stimulation of DLS dopamine provided causal evidence for our hypothesised role of DLS DA in learning. We reasoned that stimulation of DLS DA signals at outcome time should update the association between stimuli used for decisions and choices, leading to a reduction of DLS DA reward signals over trials and the emergence of DLS DA responses to these stimuli (Figure 3.11B, top). However, stimulation of DLS DA signals at outcome time should not update the association between the stimuli that the animal does not use and choices, i.e. no learning effect in subsequent trials. Importantly, this is distinct from the learning effect of a classic RPE, e.g. triggered by extra water reward (Figure 3.11B, bottom). A classic RPE would update the association between all reward predictive cues and choices (not just used stimuli, but also their absence), having a learning effect on all trial types. These predictions arise from our hypothesis that the DLS of one-sided animals does not receive a representation of the non-associated stimulus that DLS DA release can train. Instead, one-sided animals relied on the absence of the associated stimulus to make correct choices on those trials, a signal whose value does not appear to be learned through DLS DA. This hypothesis was grounded in our recordings, which showed that DLS DA neither responds nor decreases its reward response after non-associated stimuli even late in learning (Figure 3.3C-E).

To test this, we first trained animals in the task until expert and selected mice that developed one-sided strategies. We then stimulated DLS DA axons at the outcome time of *incorrect* choices using 635nm laser pulses (Figure 3.13A). We performed this stimulation in alternating days for incorrect choices to the associated or non-associated stimuli (Figure 3.13B, left column). Using this protocol, we expected to see a change in the psychometric curve only on days with stimulation in associated stimulus trials. The expected change was an increase in the association of the associated stimulus with the incorrect choice, and thus a decrease in the proportion of correct choices for that trial type. This would manifest as a decrease in the slope of the associated (i.e. non-flat) side of the psychometric curve of one-sided mice (Figure 3.13B, left column, top). Importantly, we did not expect there to be a change in the psychometric curve for the other contrast levels, as such an effect would be driven by learning associations with cues other than the visual stimuli, and thus not governed by DLS DA. On the other hand, a similar protocol using the delivery of water reward instead of DLS DA stimulation should influence all associations used in the task, resulting in a general shift of the psychometric curve (Figure 3.13B, left column, bottom),

Indeed, the stimulation at the outcome of incorrect choices selectively influenced the associated side of the psychometric curve, decreasing the accuracy of choices in these trials (Figure 3.13B, upper row). We note that our prediction held despite there often being more stimulation events on days with stimulation for non-associated stimuli, due to choices often being less accurate on the flat side of the psychometric curve. In contrast, delivering water reward after incorrect choices in a similar experiment shifted the entire psychometric curve, decreasing the choice proportions particularly on the flat side (Figure 3.13B, lower row). The mice used in the DLS DA stimulation and water delivery experiment were not always overlapping, so we confirmed that these observations held within mice by measuring the changes in the psychometric curves selectively from mice for which both experiments were performed (Figure 3.13C). Similar effects were observed in the chronometric curves (Figure 3.13D), where DLS DA stimulation only influenced the RTs of choices made using the associated stimulus, whereas reward delivery shifted the entire chronometric curve. In addition, while DLS DA stimulation did not influence choices and RTs in zero-contrast trials, reward delivery significantly affected those trials, lending further support for our conclusion. Therefore, distinct from the effect of water reward, the learning effect of DLS DA relies on the animal using visual stimuli for decision making.

3.4 Discussion

We found that DLS DA reflected key characteristics of a teaching signal shaping individual trajectories. Two such characteristics were that it tracked both the evolving sequences of strategies as well as the diversity of learning trajectories across mice. We propose that these DLS DA signals encode 'partial' RPEs based on only a subset of all the task-relevant cues, i.e. the stimuli that animals use to make choices (and not their absence), providing a framework for explaining dopaminergic signals underlying individual diversity in learning observed here and previously [240]. This partial RPE encoding results in DLS DA reflecting the stimulus-choice associations corresponding to each strategy adopted throughout learning. This encoding persisted in highly trained mice (Figure 3.6F), and could also guide trial-by-trial learning during asymptotic expert behaviour, as our experiments with reward value manipulation showed (Figure 3.8G). Here, we alternated between days in which higher rewards were delivered for correct choices after right (left) stimuli and observed that both DLS DA stimulus and reward responses reflected the higher value of the stimulus.

Crucially, the DLS DA signals differ from a 'total' RPE that incorporates all task-relevant cues to compute reward predictions. The clearest demonstration of this was in our matched accuracy analysis (Figure 3.5A), which showed that DLS DA did not display pre-outcome responses in trials with non-associated stimuli even when choices on such trials were as accurate as with associated stimuli. The 'partial' stimulus-based RPE we observe in DLS DA signals could arise from the heterogeneous topography of cortico-striatal connectivity. DLS could be receiving inputs representing visual stimuli, but not the cues that inform choices which do not rely on visual stimuli. This partial information could be relayed to the dopamine neurons that project to DLS (employing the closed-loop striatonigrostriatal circuitry [260]), causing them to release dopamine only when the DLS's inputs are active. If so, this implies that there should be different DA signals across striatal regions (as demonstrated here and in previous studies [236, 243, 261, 262, 254]) depending on the information carried by their cortical inputs. For instance, our DLS DA signals were bilateral, in contrast to DMS DA signals which showed responses only to contralateral stimuli in our data, and in previous studies [182]. This difference could be due to DLS receiving more bilateral inputs from frontal association cortical areas compared to dorsomedial striatum [186]. The DLS DA signals we recorded are known to modulate the strength of local cortico-striatal synapses [188]. This could rapidly regulate the size of stimulus-evoked responses in striatal neurons, which form a functional closed loop by projecting back to midbrain DA neurons [263, 264, 260], causing the changes in DA stimulus and outcome responses we observed throughout learning.

Signatures of stimulus-choice association emerged in RTs before being evident in choice accuracy. DLS DA stimulus signals reflected this early signature, emerging precisely when mice showed signatures of using visual stimuli, i.e. when mice had a flat psychometric curve and choices for stimuli were becoming faster. Past studies using other behavioural tasks have shown DA responses aligned to reward-seeking movements or even spontaneous movements [265, 266]. Although the precise temporal evolution of DA signals might depend on the task and requires methods with better temporal resolution than imaging to study [242, 267], our results suggest that DLS DA signals were better locked to stimulus onset than choices. Nevertheless, these pre-outcome DA signals could contribute to reducing RTs by invigorating action [190, 49] (Figure 3.10F). The DLS and DMS DA signals also differed from previously reported DA novelty signals observed during initial exposure to stimuli [242, 258, 259, 237]. In our recordings, signals were absent in early days and grew over time, with DMS DA showing slightly faster growth over days than DLS

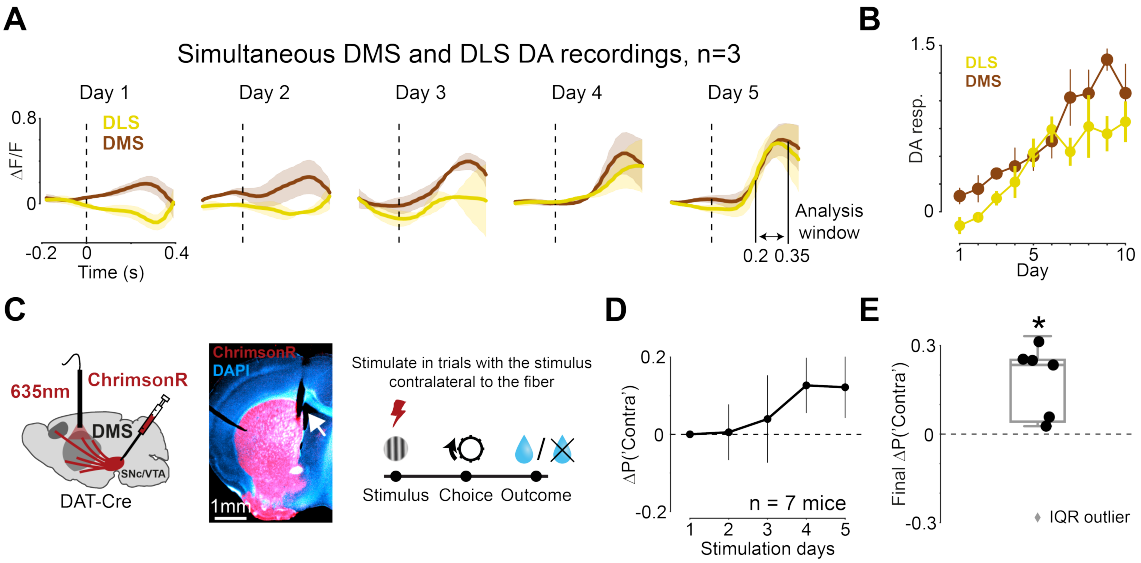


Figure 3.14: Relative evolution of DMS and DLS DA signals and effect of DMS DA optogenetic stimulation.
(A), Average stimulus-aligned DA signals for mice with simultaneous DLS and DMS recordings that associated the stimulus contralateral to the DMS fibre. Average is across correct trials with the stimulus contralateral (ipsilateral) to the DMS (DLS) fibre. (B), Average stimulus response over first 10 days of learning for the same trials and mice as in panel A (see analysis window). (C), Schematic explaining the optogenetic experiment using the excitatory opsin ChrimsonR in DMS, alongside example histology slice showing fibre tract (white arrow) and opsin expression (Section 3.2.4). (D), Change from day 1 in the probability of making the choice contralateral to the stimulation fibre, i.e. P('Contra'), over 5 stimulation days. (E), Box plot of day 5 P('Contra'). Each point is a mouse. p-value calculated using two-sided one-sample t-test against 0, excluding interquartile range (IQR) outlier.

DA (Figure 3.14A-B). The absence of DA novelty signals could be because the visual stimuli we used were not salient enough in the early stages of learning, before their task-relevance was discovered. Surprisingly, DMS DA did not respond to reward delivery (Figure 3.7B), and consistent with previous studies [182], did not reflect reward size in the reward manipulation experiment (Figure 3.8G). These characteristics suggest that DMS DA's role in learning differs from that of a RL RPE. Optogenetic stimulation of DMS DA signals at stimulus time led to contralateral biases developing over days (Figure 3.14C-E). However, the absence of DMS dopamine stimulus responses in early days suggests that DMS dopamine might not contribute to the early biases we observed in mice.

Optogenetic experiments showed that DLS dopamine is necessary for learning the task, and dissociated the effect of stimulus-based partial RPEs from total (i.e. classic) RPE on learning. These results, and the different DA signals we observed in DLS and DMS, are consistent with recent studies showing heterogeneity of dopamine signals [236, 243, 192, 182]. Our results build on these by showing that these heterogenous signals are at the service of learning. The optogenetic stimulation results also indicate that DLS DA only causes a learning effect if the stimulus is represented in the cortical inputs to the DLS. This can be seen from the lack of a behavioural effect of manipulating DLS DA on trials with the non-associated stimulus (Figure 3.13B), as well as from our recordings that show persistent DLS DA responses to rewards on trials with the non-associated stimulus which never form associations and consequent stimulus-locked responses (Figure 3.3C-E). Furthermore, in our inhibition experiment the lack of DLS DA release did not impair learning that did not depend on stimulus-choice associations, such as general decreases in RTs and the development of biases (Figure 3.12C-E).

In Chapter 4, we describe a model that accounts for the behavioural, neural and optogenetic data and provides a formal framework for defining the function of DLS DA in learning (Figure 4.1). The model has parameters analogous to the cortico-striatal synapses that associate stimulus inputs to the left and right choices, and these are updated via a partial RPE that mimics DLS DA. Analysing the model's learning trajectories revealed the characteristics of the learning algorithm with heterogenous RPEs that we propose. At surface level it can seem similar to a model trained with a single RPE (Figure 4.15), but in reality it gives rise to very different dynamics of the synaptic weights between neurons. One such dynamic is the transfer of knowledge from cortex to striatum over learning (Figure 4.12), whose precise mechanism we explain in more detail in Chapter 4. A similar phenomenon has also been observed experimentally [268, 269], where learning a new task relies on the cortex but executing a learned task does not. In our model, this dynamic causes the decision-making circuit to maintain the same behaviour but reduce the neural activity in cortical layers by compensating with increased strength of cortico-striatal synapses. Our model suggests that decreased cortical activity may make the cortico-striatal synapses less malleable and thus make the animal more prone to perseverative behaviour. However, lowering the cortical load could free the cortex for other activities or further learning, allowing new task-relevant stimuli to be discovered or multiple tasks to be performed simultaneously. This demonstrates the potential benefits of the learning algorithm we propose.

Chapter 4
Deep neural network model

Chapter contents

4.1 Introduction 58
4.2 Methods 60
4.2.1 Architecture 60
4.2.2 Learning rule 61
4.2.3 Model-derived dopamine signals 63
4.2.4 Simulation of the optogenetic experiment 64
4.2.5 Average dynamics 64
4.2.6 Fitting procedure 65
4.2.7 Saddle Points 66
4.2.8 Generalisation to larger networks 67
4.3 Results 67
4.3.1 Deep 'tutor-executor' RL 67
4.3.2 Fixed point analysis 80
4.4 Discussion 83
4.4.1 Depth and heterogenous RPEs 83
4.4.2 A mathematical theory 84

Deep learning is one of the major scientific breakthroughs of the century to-date. It has not only revolutionised the fields of ML and AI, but is now starting to have important applications in fields as diverse as healthcare, law enforcement and defence. Although deep learning researchers might claim that the methods emerged independently from neuroscience research, they share a non-coincidental similarity with the brain's architecture and neural physiology. The scepticism is reciprocal, with several neuroscientists claiming that they function nothing like the brain and thus there is little to learn from deep learning systems about brain function. An additional barrier to the use of deep learning systems as a tool to study the brain is their ever-increasing complexity, which has made their operating mechanisms difficult to understand even for the ML researchers who designed them.

A common trend in neuroscience studies involves presenting novel behavioural results, describing their neural correlates, and then tying both together with a computational model. More recently, studies have also started to demonstrate the causal role of the neural signals they record through optogenetic methods [270, 271]. However, few studies analyse their computational model to uncover the principles that allow it to capture the data [157, 272, 273, 83]. We believe this is a crucial step that greatly enhances the value of the model, as it passes from being something that works, to being a symbol for a more general principle that underlies brain function. There are often many models that can capture the data, as the limited data we collect in experiments is rarely able to fully constrain modelling choices. Further, the signals in the data may be influenced by unobserved variables that we do not (or cannot) record. As such, it is not surprising that models built solely to capture the data often do not provide deep insights into brain function. However, if further analysis

is performed to discover *why* the model captures the data, the exact instantiation of the model becomes much less relevant and the more widely-applicable mechanisms at work are highlighted. A benefit of employing this approach is that it is often able to unify observations from a range of different experiments and brain regions. In addition, it can guide future experiments more holistically, as the model's predictions no longer rely on a precise one-to-one mapping between its parameters and the brain.

In this chapter, we show that deep learning methods can indeed be useful to better understand the brain. We demonstrate that the deep learning framework of using multiple layers of neurons connected by tuneable weights to create a flexible input-output map can shed light on the brain's learning mechanisms. We even demonstrate that the seemingly unphysiological learning rule of gradient descent leads to learning trajectories that mimic those of mice surprisingly well. Although we are aware that deep learning methods have several known incompatibilities with brain physiology [274, 275, 276], we believe that the framework can help uncover the principles governing learning in the brain. We further show that it can guide future experimental research with more targeted questions about the physiological mechanisms at work. We conclude by showing that our deep RL model's learning trajectories are governed by saddle points and their connecting manifolds, providing a formal account for how a biological learning mechanism can steer and yield diverse yet systematic long-term learning trajectories. We hope our research demonstrates the importance of establishing a dialogue between theory and experiment, bringing the experiment-theory loop one step closer to the community of intelligence researchers.

4.1 Introduction

Many abilities are learned over long time periods spanning weeks or months. For instance, a novice can take many months to learn how to play tennis. Such long-term learning is distinguished by two key features. First, individuals typically progress through stages as they learn, adopting and mastering increasingly effective strategies. For example, a novice may start by mastering forehands before learning backhands. Second, individuals tend to follow diverse trajectories from naïve to expert, discovering different strategies at different times, just as tennis players often develop distinct playing styles. Decades of experimental and theoretical work have provided fundamental insights into the neuronal circuits and computations underlying learning [47, 146, 151, 147, 148, 155, 152, 153, 149, 158, 154, 159, 277]. However, the behavioural correlates, neuronal underpinning, and computational principles governing long-term learning trajectories and their individual diversity are not yet understood.

Past studies of learning in neuroscience have mainly focused on short-term learning, where animals have to fine tune their behaviour in an already learned task to adapt to changing contingencies over short timescales. The results from these studies have been broadly captured by standard RL models that (re-)learn the value of predetermined states through RPE updates [43, 206, 151]. Long-term learning, in contrast, involves discovering state representations that enable the adoption of increasingly effective strategies. This often leads to substantial diversity in strategy transitions across individuals through learning, which standard RL models might not be able to explain. Thus, it remains an open question whether the RL framework can account for long-term learning trajectories within and across individuals.

Deep neural networks are particularly suited to capture the process of discovering novel state representations throughout long-term learning. Their architecture with multiple layers of neurons connected by tuneable weights enables the formation of deep distributed representations with

rich structure that have been shown to share similarities with neural representations in the brain [272, 278, 279]. Deep reinforcement learning indeed exists as a subfield of ML, showing much promise as a method for autonomous systems to tackle complex problems where their outputs do not have to match a specified label, instead the objective being to maximise the attainment of a reward signal [208, 209, 215]. The deep RL framework seems thus well-suited to modelling the behaviour of mice learning the visual decision-making task outlined in Chapter 2. Here mice were not taught by explicitly showing them how to turn the wheel on each trial, instead they received a water reward for correct choices and a short punishment sound and timeout for incorrect choices. In addition, the depth of a deep RL network seems apt to capture the mice's progressive (and selective) discovery of stimuli observed in the behavioural and neural results from Chapter 2 and Chapter 3. We considered alternative models throughout our investigations Appendix B, which further helped to inform our decision.

For a deep RL network to provide useful insights into the mechanisms at work in the mice's brain, it has to be interpretable. Deep RL networks are often large, complex systems with astounding capabilities but little interpretability. Such a network would have captured our behavioural data, but is unlikely to have a similar structure to the mice's brain at the neuron and weight level due to the indeterminacy of large networks. Further, understanding the structure of the network's representations becomes more difficult at large scales. We also aimed to understand our bulk-level recordings of dopamine release through the lens of the model, and as such required a level of abstraction suited for this type of data. We thus opted for a very simple architecture that we believe reduced the network to the bare minimum required to capture the mouse learning trajectories. This enabled us to design a learning rule with teaching signals that mimicked our dopamine recordings, and also provided an interpretable framework to understand the evolving neural representations in the mice. The dopamine-based learning rule further allowed us to simulate the effects of optogenetic stimulation and reproduce the results of our experiment from Chapter 3 in silico. Lastly, the model's simplicity enabled mathematical analysis of its learning dynamics in closed-form.

It has been difficult to determine the degree of systematicity in what appears to be an inherently stochastic long-term learning process – does behaviour early in learning determine future behaviour? If so, why? We attempted to answer this question through our model. Our deep RL model reproduces the mice's diverse yet systematic stage-like transitions between behavioural strategies from naïve to expert. We find that, similar to mice, model simulations that developed an early left/right bias developed left/right-associating strategies later in learning, whereas simulations with more balanced early choices became balanced in their later stimulus-choice associations. However, without further analysis, the reason behind the systematicity of the model's learning trajectories remains unclear. One could argue that this is simply the result of running the dopamine-based RPE updates of our model, and thus that the principle at work is none other than the well-known RPE mechanism [52]; but what about it yields the characteristic structure of mouse and model learning trajectories across individuals? Finding this would explain what properties of the model determine the sequence of strategies that an individual transitions through before reaching expertise, thus explaining why different network architectures, learning rules and/or tasks may vary in the distribution of learning trajectories they generate.

To tackle this, we mathematically analysed the learning dynamics of our deep RL model. The simplicity of its learning rule and architecture allowed us to derive closed-form expressions for the average dynamics of the weights across layers. Using these expressions, we performed a fixed point analysis of the dynamics that revealed a hierarchy of saddle points corresponding to

the diverse strategies observed in mice and simulations throughout learning. Learning dynamics temporarily slow down in the vicinity of a saddle point, before speeding up again as they leave. This deceleration explains why we could detect the sequences of distinct strategies in the mice learning trajectories. The saddle point structure further explained the systematicity of learning trajectories with the heteroclinic orbits between saddle points, which traced paths through the parameter space that captured the distribution of mouse and simulation trajectories. Studying the fixed points of a model to understand its behaviour has proved to be a successful approach in the past, leading to important advances in our understanding of the mechanisms of context-dependent computation [157]. In this chapter, we outline how this approach provides a general framework that may help understand the patterns of diversity and systematicity in long-term learning trajectories across tasks, individuals and learning rules.

4.2 Methods

4.2.1 Architecture

The model is a 3-layer deep neural network with linear activation functions (Figure 4.1A). We denote the weight matrix connecting the input layer to the hidden layer W^1, and the matrix connecting the hidden layer to the output layer W^2, such that the function computed by the network is given by

$$y = W^2 W^1 x, \tag{4.1}$$

where x is the vector of inputs and y is the vector of outputs. The weights in both W^1 and W^2 are constrained to be nonnegative, and W^1 is constrained to be a diagonal matrix. These constraints are motivated later in Section 4.3. The network has two binary input neurons which encode the presence or absence of the left and right visual stimulus and one input neuron that has an activation of 1 for every trial, representing any non-stimulus cues that the mice may use to make choices on each trial, e.g. the auditory go cue. The inputs did not vary with time within a trial. Thus, the network receives three different input vectors depending on the trial type,

$$\text{Left stimulus: } \begin{bmatrix}1\\1\\0\end{bmatrix}, \text{ Zero-contrast: } \begin{bmatrix}1\\0\\0\end{bmatrix}, \text{ Right stimulus: } \begin{bmatrix}1\\0\\1\end{bmatrix}. \tag{4.2}$$

We opted not to model the different contrast levels of stimulus trials (25% and 50%) because in most mice we did not see a significant difference in accuracy between these two levels.

In every trial, the input vectors are multiplied by the weights in W^1 and W^2 to obtain the activation of the two output neurons which encode the learned value of left and right choices, Q_{L} and Q_{R}. These action values are used to determine choice through a softmax function with inverse temperature β that determines the choice probabilities,

$$\mathrm{P}(\text{`Right'}) = \frac{1}{1 + e^{-\beta(Q_{\mathrm{R}} - Q_{\mathrm{L}})}} \text{ and } \mathrm{P}(\text{`Left'}) = 1\text{–}\mathrm{P}(\text{`Right'}), \tag{4.3}$$

from which a choice is sampled on each trial.

For each simulation, the initialisation of W^1 was sampled from a Gaussian distribution centered on fit values of the initial constant input weight μ_k and stimulus input weights μ_{w_0} (identical for

left/right) such that,

$$W^1_{init} = \begin{bmatrix} k & 0 & 0 \\ 0 & w_0 & 0 \\ 0 & 0 & w_0 \end{bmatrix} + 10^{-4}\mathcal{N}(0,\ 1), \tag{4.4}$$

where $\mathcal{N}(\mu, \sigma)$ indicates a Gaussian distribution with mean μ and standard deviation σ. Here k was sampled from $\mathcal{N}(\mu_k = 1,\ 0.1)$ and w_0 was sampled from $\mathcal{N}(\mu_{w_0} = 0.05,\ 0.05)$ for each tutor-executor simulation, and from $\mathcal{N}(\mu_k = 1,\ 0.01)$ and $\mathcal{N}(\mu_{w_0} = 0.05,\ 0.001)$ for each single-loss simulation. The values of k and w_0 were re-sampled if not strictly positive, and all weights were reset to their absolute value and the off-diagonal terms to 10^{-5} after each update step. This resetting was also applied when computing the losses on each step. When investigating the effect of the nonnegativity and connectivity constraints in Figure 4.8, this resetting was either not applied or only the constant-pathway off-diagonal terms were reset to 10^{-5} for the fully-connected stimulus-pathway model.

Similarly, the initialisation of W^2 was sampled from a Gaussian distribution centered on fit values of the constant pathway weights μ_{c_0} (Figure 4.1A, aqua – all identical) and stimulus pathway weights μ_{s_0} (Figure 4.1A, pink – all identical),

$$W^2_{init} = \begin{bmatrix} c_0 & s_0 & s_0 \\ c_0 & s_0 & s_0 \end{bmatrix} + 10^{-4}\mathcal{N}(0,\ 1), \tag{4.5}$$

where c_0 was set to $\mu_{c_0} = 0$ and s_0 was sampled from $\mathcal{N}(\mu_{s_0} = 0.2,\ 0.05)$ for each tutor-executor simulation, and to $\mu_{c_0} = 0$ and sampled from $\mathcal{N}(\mu_{s_0} = 0.7,\ 0.01)$ for each single-loss simulation. These values were also re-sampled if not strictly positive. The W^2 weights were similarly reset to their absolute value when computing the loss and after each update step to enforce the nonnegativity constraint.

The softmax inverse temperature parameter β and learning rate α were also sampled from a Gaussian centered on a value fit to the experimental data. More specifically, β was sampled from $\mathcal{N}(\mu_\beta = 9,\ 1)$ and α from $\mathcal{N}(\mu_\alpha = 0.0026,\ 0.001)$ for each tutor-executor simulation and from $\mathcal{N}(\mu_\beta = 13,\ 0.01)$ and $\mathcal{N}(\mu_\alpha = 0.0015,\ 0.001)$ for each single-loss simulation. The fitting procedure is described in the corresponding subsection. Simulations were run for 10,000 trials and those that reached 70% accuracy in less than 8,500 trials (approximately highest number of trials required for mice to learn) were included in our analyses.

4.2.2 Learning rule

Tutor-executor gradient descent

We refer to the model presented in Figure 4.1 as the 'tutor-executor' model due to its learning rule, which uses different RPEs to train the weights in W^1 and W^2. The updates minimise three different losses through stochastic gradient descent (SGD): the 'cortical' loss for the weights in W^1, the 'stimulus cortico-striatal' loss for the weights in the stimulus pathway of W^2 (Figure 4.1A, pink), and the 'constant cortico-striatal' loss for the weights in the constant pathway of W^2 (Figure 4.1A, aqua). Each of these losses is a RPE^2 comparing predictions based on different subsets of inputs against trial outcome,

$$\text{'Cortical' loss: } \mathcal{L}^{tot} = \frac{1}{2}\delta^{tot\,2} = \frac{1}{2}\left(Rew - Q_{ch}\right)^2, \tag{4.6}$$

$$\text{'Stimulus cortico-striatal' loss: } \mathcal{L}^{stim} = \frac{1}{2}\delta^{stim\,2} = \frac{1}{2}\left(Rew - Q_{ch}^{stim}\right)^2, \tag{4.7}$$

$$\text{'Constant cortico-striatal' loss: } \mathcal{L}^{const} = \frac{1}{2}\delta^{const\,2} = \frac{1}{2}\left(Rew - Q_{ch}^{const}\right)^2, \tag{4.8}$$

where Rew is a binary variable indicating whether the trial was rewarded or not, and the subscript ch indicates the choice made on each trial (left/right). Here Q_{ch} is the 'total' Q-value that uses all inputs to form its predictions, while Q_{ch}^{stim} and Q_{ch}^{const} are 'partial' Q-values based on the stimulus and constant inputs in turn,

$$Q_{ch} = W_{ch,0}^2 W_{0,0}^1 + W_{ch,1}^2 W_{1,1}^1 VSL + W_{ch,2}^2 W_{2,2}^1 VSR, \tag{4.9}$$

$$Q_{ch}^{stim} = W_{ch,1}^2 W_{1,1}^1 VSL + W_{ch,2}^2 W_{2,2}^1 VSR, \tag{4.10}$$

$$Q_{ch}^{const} = W_{ch,0}^2 W_{0,0}^1 , \tag{4.11}$$

where VSL and VSR are the binary inputs indicating the presence or absence of the right and left stimulus respectively and ch used as a subscript for the weights is 0 for left and 1 for right choices.

Gradient descent on these losses yields updates which depend on the trial outcome and choice,

$$\Delta W^1 = -\alpha \frac{\partial \mathcal{L}^{tot}}{\partial W^1} = \alpha \begin{bmatrix} W_{ch,0}^2 & 0 & 0 \\ 0 & VSL\ W_{ch,1}^2 & 0 \\ 0 & 0 & VSR\ W_{ch,2}^2 \end{bmatrix} \delta^{tot}, \tag{4.12}$$

$$\Delta W^{2,stim} = -\alpha \frac{\partial \mathcal{L}^{stim}}{\partial W^2} = \begin{cases} \alpha \begin{bmatrix} 0 & VSL & W_{1,1}^1 & VSR & W_{2,2}^1 \\ 0 & 0 & & 0 & \end{bmatrix} \delta^{stim}, \text{ if } ch = L, \\ \alpha \begin{bmatrix} 0 & 0 & & 0 & \\ 0 & VSL & W_{1,1}^1 & VSR & W_{2,2}^1 \end{bmatrix} \delta^{stim}, \text{ if } ch = R, \end{cases} , \tag{4.13}$$

$$\Delta W^{2,const} = -\alpha \frac{\partial \mathcal{L}^{const}}{\partial W^2} = \begin{cases} \alpha \begin{bmatrix} W_{0,0}^1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \delta^{const}, \text{ if } ch = L, \\ \alpha \begin{bmatrix} 0 & 0 & 0 \\ W_{0,0}^1 & 0 & 0 \end{bmatrix} \delta^{const}, \text{ if } ch = R, \end{cases} , \tag{4.14}$$

where α is the learning rate and $\Delta W^2 = \Delta W^{2,stim} + \Delta W^{2,const}$. Notice how the updates for W^1 are proportional to the total RPE, δ^{tot}; the updates for the stimulus pathway in W^2 are proportional to the stimulus-based RPE, δ^{stim}; and lastly, the updates for W^2's constant pathway are proportional to the constant-based RPE, δ^{const}. Interestingly, the general tendency of the learning rule is to minimise the total 'cortical' loss, $\mathcal{L}^{tot}$, as the learning in W^1 tutors downstream learning in W^2 by determining the relative salience of the inputs and balancing updates in the executor pathways. Shallow versions of this learning rule set the W^1 updates to 0.

Single-loss gradient descent

This learning rule updates W^1 and W^2 through SGD to minimise a single total RPE^2. This corresponds to the conventional method of training deep RL networks, where all parameters are updated to minimise a single loss and thus share the same objective [280]. The loss we use for this learning rule is the same as the 'cortical' loss in the tutor-executor model,

$$\mathcal{L}^{tot} = \frac{1}{2}\delta^{tot\,2} = \frac{1}{2}\left(Rew - Q_{ch}\right)^2, \tag{4.15}$$

where Rew is a binary variable indicating the outcome of a particular trial and Q_{ch} is a reward prediction calculated based on all inputs to the network. The updates for W^1 and W^2 can be written as follows

$$\Delta W^1 = -\alpha \frac{\partial \mathcal{L}^{tot}}{\partial W^1} = \alpha \begin{bmatrix} W^2_{ch,0} & 0 & 0 \\ 0 & VSL\ W^2_{ch,1} & 0 \\ 0 & 0 & VSR\ W^2_{ch,2} \end{bmatrix} \delta^{tot}, \tag{4.16}$$

$$\Delta W^2 = -\alpha \frac{\partial \mathcal{L}^{tot}}{\partial W^2} = \begin{cases} \alpha \begin{bmatrix} W^1_{0,0} & VSL & W^1_{1,1} & VSR & W^1_{2,2} \\ 0 & 0 & 0 & 0 & \end{bmatrix} \delta^{tot}, \text{ if } ch = L, \\ \alpha \begin{bmatrix} 0 & 0 & 0 & & \\ W^1_{0,0} & VSL & W^1_{1,1} & VSR & W^1_{2,2} \end{bmatrix} \delta^{tot}, \text{ if } ch = R, \end{cases}, \tag{4.17}$$

where α is the learning rate and VSL and VSR are binary variables indicating the presence or absence of the right and left stimulus respectively. Here, all the updates are proportional to the same total RPE, δ^{tot}. Shallow versions of this learning rule set the W^1 updates to 0.

4.2.3 Model-derived dopamine signals

Our neural network model captures empirical dorsolateral striatal dopamine (DLS DA) signals through the weights in its stimulus pathway. Over learning, the model reproduces DLS DA outcome responses with a stimulus-based RPE,

$$\text{Outcome-time DLS DA} \equiv \delta^{stim} = Rew - Q^{stim}_{ch}. \tag{4.18}$$

This differs from conventional temporal difference reward prediction errors (TD-RPEs) commonly used in DA studies [52] in that it does not use the full reward prediction based on all inputs to define the RPE. Instead, it compares trial outcome with a prediction based only on the stimulus inputs. This was motivated by our matched accuracy analysis in Figure 3.5A, which showed that DLS DA at both stimulus and outcome time does not reflect reward predictions based on the constant input (i.e., the absence of a stimulus).

At stimulus time, in accordance with the TD-RPE hypothesis, the model captured DLS DA signals with a pre-choice reward prediction. Importantly, this prediction is also based only on the stimulus inputs,

$$\text{Stimulus-time DLS DA} \equiv p^{stim}_L Q^{stim}_L + p^{stim}_R Q^{stim}_R, \tag{4.19}$$

where p^{stim}_L and p^{stim}_R are the left and right choice probabilities determined by the stimulus-based Q-values passed through the choice function in Equation (4.3). Here, there is no need to subtract the value of the previous state as there is no cue before the stimulus that is predictive of reward.

To test whether the network could be trained with a learning signal analogous to our DLS DA recordings, the tutor-executor model uses a learning rule based on these 'partial' prediction errors to update its weights. Specifically, the updates of the W^2 stimulus pathway weights are proportional to the stimulus-based RPE, δ^{stim} (Equation (4.13)). Comparing the evolution of δ^{stim} with the DLS DA outcome response over learning shows a striking similarity (Figure 4.4F; Figure 4.5A and D-F), suggesting that a similar dopamine-based learning mechanism could be governing the learning process of the mice.

4.2.4 Simulation of the optogenetic experiment

We simulated the optogenetic experiment in the model to investigate the predicted effect on behaviour of the tutor-executor and single-loss gradient descent learning rules undergoing similar manipulations. For the tutor-executor learning rule, the effect of DLS DA stimulation was modelled as a 'reward' signal provided only to the 'stimulus cortico-striatal' loss $\mathcal{L}^{stim}$. However, for the single-loss learning rule the 'reward' was provided to the total loss $\mathcal{L}^{tot}$ used to update weights in W^2. As for mice, simulated DLS DA stimulation was provided to expert one-sided simulations for incorrect left or right choices on alternating days (i.e. 220 simulated trials) for 10 days (i.e. 5 of each condition). The 'reward' signal used to model the stimulation was x0.25 the size of the reward provided by water, and a learning rate boost of $x10^3$ on incorrect trials was used to reproduce the empirical results. This requirement was surprising, and indicated to us that the learning effect of exogenously triggered DA release is much larger than that released endogenously. Networks were initialised using weights from right-associating average dynamics at expertise.

We also simulated the incorrect trial water delivery control experiment. As in correct trials, the effect of water delivery on incorrect trials was simulated as a global reward signal provided to all the losses used for weight updates in both the tutor-executor and single-loss models. As for mice, incorrect trial water delivery was provided to expert one-sided simulations for incorrect left or right choices on alternating days, with a baseline day in between where simulations performed the original task. The alternation was performed for 20 days (i.e. 5 days of each left/right condition, with each day correspoding to 220 trials) for each simulation. A learning rate boost of x5 on incorrect trials with water delivery was used to capture the empirical results. This indicated that the unexpected reward on incorrect trials causes more learning than on correct trials. Networks were initialised using weights from right-associating average dynamics at expertise.

4.2.5 Average dynamics

Deriving the average dynamics of the model allowed us to obtain an analytical description of its learning mechanism. We took the continuous time limit of the average gradient descent updates for W^1 and W^2 from the tutor-executor and single-loss learning rules, averaging over trial type (left stim., right stim., or zero-contrast) and choice (left or right). This yielded a 9-dimensional system of coupled differential equations describing the evolution of each weight in the network. For sufficiently small learning rates α, this 'gradient flow' limit provides a good description of the average dynamics for both learning rules. The resulting differential equations were numerically integrated to obtain average weight trajectories over training time. To capture the three main types of learning trajectory (i.e., left-associating, balanced and right-associating), we initialised the integration with a network configuration yielding different degrees of initial choice bias (i.e., imbalanced connections from the const. input to Q_L and Q_R). We then overlayed the resulting trajectories (thick dashed lines in model figures) on those from trial-by-trial simulations.

Tutor-executor average dynamics

The average dynamics of the network weights for the tutor-executor learning rule are governed by three main differential equations; one for each of the cortical, stimulus cortico-striatal and constant cortico-striatal weight subsets (black, pink, and aqua in Figure 4.1A). These can be derived by taking the average over trial-types and choices of the gradient descent updates in

Equation (4.12)-Equation (4.14). Doing this for the cortical weights in W^1 we obtain

$$\begin{aligned}\frac{1}{\alpha}\frac{d}{dt}W^1 &= \langle\langle -\frac{\partial \mathcal{L}^{tot}}{\partial W^1}\rangle_{ch}\rangle_{trial} = -\langle\langle \delta^{tot}\frac{\partial \delta^{tot}}{\partial W^1}\rangle_{ch}\rangle_{trial} \\ &= -p_{VSL}\Big(p_{L,VSL}\left(1-Q_{L,VSL}\right)\frac{\partial\left(1-Q_{L,VSL}\right)}{\partial W^1} + p_{R,VSL}\left(0-Q_{R,VSL}\right)\frac{\partial\left(0-Q_{R,VSL}\right)}{\partial W^1}\Big) \\ &- p_0\Big(0.5p_{L,0}\left(1-Q_{L,0}\right)\frac{\partial\left(1-Q_{L,0}\right)}{\partial W^1} + 0.5p_{L,0}\left(0-Q_{L,0}\right)\frac{\partial\left(0-Q_{L,0}\right)}{\partial W^1} \\ &\quad + 0.5p_{R,0}\left(1-Q_{R,0}\right)\frac{\partial\left(1-Q_{R,0}\right)}{\partial W^1} + 0.5p_{R,0}\left(0-Q_{R,0}\right)\frac{\partial\left(0-Q_{R,0}\right)}{\partial W^1}\Big) \\ &- p_{VSR}\Big(p_{L,VSR}\left(0-Q_{L,VSR}\right)\frac{\partial\left(0-Q_{L,VSR}\right)}{\partial W^1} + p_{R,VSR}\left(1-Q_{R,VSR}\right)\frac{\partial\left(1-Q_{R,VSR}\right)}{\partial W^1}\Big)\end{aligned} \tag{4.20}$$

where t is a continuous time variable counting the number of trials; δ^{tot} represents the total RPE (Equation (4.15)); $p_{VSL} = p_{VSR} = 0.45$ and $p_0 = 0.1$ are the probabilities of there being a left vis. stim., right vis. stim., and zero-contrast trial; and lastly $p_{A,B}$ and $Q_{A,B}$ indicate the choice probabilities and total Q-values for choice A in a trial of type B. The choice probabilities are calculated using the sigmoidal choice rule in Equation (4.3). The partial derivatives of the Q-values can then be expanded by writing them in terms of the network weights (Equation (4.9)-Equation (4.11)) and differentiating w.r.t. the W^1 matrix.

The same procedure can be followed to find the differential equations for the weights in the stimulus and constant pathways of W^2, with gradient flow equations

$$\frac{1}{\alpha}\frac{d}{dt}W^{2,stim} = \langle\langle -\frac{\partial \mathcal{L}^{stim}}{\partial W^{2,stim}}\rangle_{ch}\rangle_{trial} = -\langle\langle \delta^{stim}\frac{\partial \delta^{stim}}{\partial W^{2,stim}}\rangle_{ch}\rangle_{trial}, \tag{4.21}$$

$$\frac{1}{\alpha}\frac{d}{dt}W^{2,const} = \langle\langle -\frac{\partial \mathcal{L}^{const}}{\partial W^{2,const}}\rangle_{ch}\rangle_{trial} = -\langle\langle \delta^{const}\frac{\partial \delta^{const}}{\partial W^{2,const}}\rangle_{ch}\rangle_{trial}. \tag{4.22}$$

Single-loss average dynamics

The single-loss average dynamics can be derived in a similar fashion, where now all the weights are minimising the same loss function (Equation (4.15)), yielding the following gradient flow equations

$$\frac{1}{\alpha}\frac{d}{dt}W^1 = \langle\langle -\frac{\partial \mathcal{L}^{tot}}{\partial W^1}\rangle_{ch}\rangle_{trial} = -\langle\langle \delta^{tot}\frac{\partial \delta^{tot}}{\partial W^1}\rangle_{ch}\rangle_{trial}, \tag{4.23}$$

$$\frac{1}{\alpha}\frac{d}{dt}W^2 = \langle\langle -\frac{\partial \mathcal{L}^{tot}}{\partial W^2}\rangle_{ch}\rangle_{trial} = -\langle\langle \delta^{tot}\frac{\partial \delta^{tot}}{\partial W^2}\rangle_{ch}\rangle_{trial}, \tag{4.24}$$

which can be expanded as exemplified with the tutor-executor dynamics.

4.2.6 Fitting procedure

To reproduce the learning trajectories in the data with our model, we fit the mean value of the network weights at initialisation (see Equation (4.4) and Equation (4.5)) and the β parameter of the choice function (Equation (4.3)). Specifically, we fit the initial mean W^1 weights for the constant μ_k and stimulus inputs μ_{w_0} (equal for left/right), the initial mean W^2 weights for the constant μ_{c_0} and stimulus pathways μ_{s_0} (all equal within each pathway) and the mean value of the inverse temperature μ_β.

To do so, we took advantage of our expressions for the average dynamics of both the tutor-executor and single-loss learning rules. We integrated the average dynamics of networks initialised

with different values of the mean initial weights, of the inverse temperature μ_β and with a small amount ε of left, right or no initial bias in the constant-pathway W^2 weights (i.e. $W^2_{1,0} = W^2_{0,0} + \varepsilon$, ε was also fit when non-zero) until 73% accuracy. We then compared the resulting behavioural and neural trajectories with the left-assoc., right-assoc. and balanced trajectory clusters from the data (from Figure 2.9E-H and Figure 3.3E).

The fitting procedure minimised the mean squared error between each trajectory measure for the three behavioural and neural clusters emerging from integrating the average dynamics and the three trajectory clusters from the data using momentum-based gradient descent on the parameters. The optimisation was performed using a chained optimiser from the JAX-based optax module with optax.clip(1), optax.adabelief(learning_rate=1e-3) and optax.keep_params_nonnegative(). We started the optimiser with 5 different initial parameter settings sampled randomly from a suitable range for each parameter. We then ran the optimisation for 1000 iterations, finally selecting the parameter combination with the lowest mean squared error. This method using the average dynamics proved faster than with simulations. Lastly, the learning rate α was fine-tuned by hand to obtain learning trajectories that were stable and learned in a similar number of trials as the mice.

The resulting parameter values were used as the means of Gaussians (with variances set manually to reproduce similar patterns of diversity as observed in mice) from which the parameters of each simulated network were sampled for the simulations shown in the figures across the paper (see Equation (4.4) and Equation (4.5)). Further, the resulting parameters (including ε) were also used as initialisation for the trajectories derived from integrating the average dynamics shown as dashed thick lines across the paper.

4.2.7 Saddle Points

Saddle points in the model's learning dynamics provide an explanation for the systematic transitions between behavioural strategies and dopamine release patterns observed in the mice. We used the average dynamics derived in Section 4.2.5 to demonstrate the existence of these saddle points. We first made informed guesses at fixed points by looking for network configurations where the average dynamics go to 0, and then investigated the dynamics around these points to verify their nature. These derivations can be found in the following Mathematica notebooks:

`https://www.wolframcloud.com/obj/samuelliebanagarcia/Published/tutor_executor_stationary_points.nb`
`https://www.wolframcloud.com/obj/samuelliebanagarcia/Published/single_loss_stationary_points.nb`

We also provide evidence for heteroclinic orbits connecting the saddle points, represented by the arrows in Figure 4.13A and Figure 4.15B. To do this, we used the string method [281] to find the minimal energy path between each pair of saddle points. This allowed us to distinguish between points that are directly connected by such paths (i.e., heteroclinic orbits), and points that are only connected through another one of the saddle points. The orbits discovered by the string method are shown in the Mathematica notebooks, and their schematic form is shown in Figure 4.13A and Figure 4.15B.

4.2.8 Generalisation to larger networks

To investigate whether the learning principles we identified in small networks generalise to larger, more conventional neural networks, we trained such networks on our behavioural task (Figure 4.11). First, we improved the faithfulness of the model to the mice's perceived environment by replacing the one-hot encoding of the stimulus inputs with a pixel-level representation of the screen presented to the mice. We used a downsampled version of the 1280x1024 gray screen with a Gabor stimulus presented at an x-offset of +/-423 pixels from the centre with a frequency of 0.01348 cycles/pixel and a Gaussian envelope with variance 98.69 pixels^2 in both the x- and y-directions. Second, we increased the number of neurons in each layer and the number of layers while maintaining the connectivity patterns of the smaller network (i.e., one-to-one input channels, fully-connected stimulus pathways or a completely fully-connected network). Third, we introduced a non-linear activation function (i.e. ReLU) to further improve the faithfulness of the model to brain networks. The output of the large networks remained as the same as the small network: two Q-values represented by single neurons, which were subsequently fed to a softmax function to determine the probability of each choice. A complete depiction of the large network architectures can be found in Figure 4.11A-C. The networks were trained with the single-loss gradient descent learning rule, i.e., minimising the total loss $\mathcal{L}^{tot}$, as implementing the tutor-executor learning rule in deeper multi-layer networks is outside the scope of our study. We kept the nonnegativity constraint as we found it worked best for the smaller network.

For each simulation, the large networks were initialised with a similar effective network to the smaller one-to-one network (set to the fit values of W^1 and W^2 from before). More specifically, the weights in the first and last layers were sampled from Gaussians with variances set to the fit values from the small network (negative weights clipped to 0), and weights in intermediate layers were initialised small. For fully-connected networks, the cross-connections were also initialised small (except where we explicitly investigated manipulating their magnitude). The softmax inverse temperature parameter β and learning rate α were also sampled from a Gaussian centered on the value fit to the experimental data for the smaller network. The fitting procedure is described in Section 4.2.6. Simulations were run for 10,000 trials and those that reached 70% accuracy in less than 8,500 trials (approximately highest number of trials required for mice to learn) were included in our analyses.

4.3 Results

4.3.1 Deep 'tutor-executor' RL

To understand the computational principles underlying the learning trajectories, dopamine signals and optogenetic effects, we designed several neural network models with different architectures and learning rules. We first outline a reduced deep network that captured the data. This reduced architecture allowed us to probe the role of specific components, through which we showed that the model requires depth, i.e. multiple layers of tunable weights, and heterogenous teaching signals to account for the data. We subsequently demonstrated that the requirement of these key features generalised to larger deep networks.

In essence, the deep network models all contain multiple layers of neurons that learn to predict the reward associated with each stimulus and choice. The reduced model has a simple architecture: an input layer, a hidden layer, and an output layer (Figure 4.1A). To capture our DLS DA signals, and their learning effects in the optogenetic experiments, the model is organised

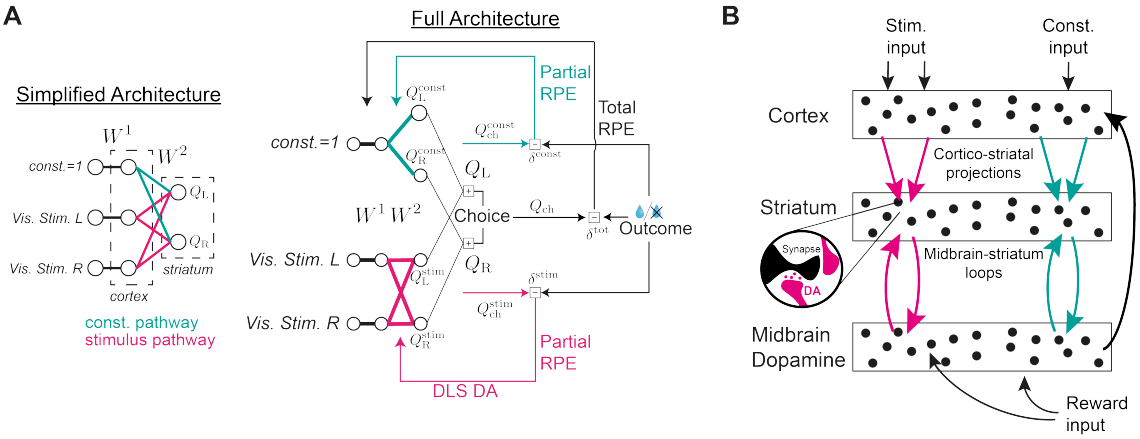


Figure 4.1: The deep 'tutor-executor' RL model and its mapping to anatomy.
(A), Simplified (left) and full (right) schematic of the reduced deep linear 'tutor-executor' network architecture and learning rule. (B) Diagram showing the proposed mapping between the model and brain anatomy.

into parallel pathways: the visual stimulus pathway (Figure 4.1A, pink) and the 'constant' pathway encompassing all other inputs that do not vary across trials, e.g. auditory go cue (Figure 4.1A, aqua). A matrix of nonnegative weights denoted by W^1 connects the input layer to the hidden layer, and similarly nonnegative weights in W^2 connect the hidden layer to the output layer. The weights in W^1 form one-to-one connections between input layer neurons and the hidden layer ('cortical') neurons, resembling the anatomical segregation of visual inputs in each brain hemisphere. The weights in W^2 are divided into two pathways, those connecting hidden layer neurons receiving visual inputs with the output layer (pink), and those connecting the hidden layer neuron receiving 'constant' input with the output layer (aqua). These represent the brain's segregated 'cortico-striatal' projections (Figure 4.1B). The outputs of the network are two action values, Q_{L} and Q_{R}, reflecting the learned value (i.e., reward prediction) of making each choice as a function of the inputs. These action values are compared with the reward obtained on each trial to compute an RPE that is used to update the weights (Section 4.2.2). Anatomically, this computation could be achieved through the relaying of the Q-values to midbrain dopamine neurons through closed-loop striatonigrostriatal projections [282], which are known to be functional [260].

On each trial, the model receives inputs, makes a choice, and learns from the resulting outcome. Two binary inputs 'Vis. stim. L' and 'Vis. stim. R' (represented as 0/1) indicate which of the stimuli is presented on a particular trial. The third input 'constant' is always 1 and reflects stimulus-independent input, capturing environmental features that do not change trial-by-trial. The model makes choices by comparing Q_{L} and Q_{R} using a softmax choice rule. The model then compares the outcome of the choice (reward/no reward) with its corresponding reward prediction to calculate RPEs used to update the weights W^1 and W^2 (Section 4.2.2). The model uses three different RPEs to update its weights. For W^1, the RPE is calculated using a 'total' reward prediction using all the inputs to the network (Q_{ch}). For W^2, the updates differ for the two pathways (Figure 4.1A, pink and aqua arrows respectively). The RPE in the 'stimulus' pathway is calculated using a partial reward prediction (Q_{ch}^{stim}) based on the stimulus inputs (Vis stim. L and R), whereas the RPE in the 'constant' pathway is calculated using a prediction based only on the constant input (Q_{ch}^{const}). The 'stimulus' pathway's reward prediction and RPE is the model's account of DLS DA stimulus and outcome responses respectively (Figure 4.1A-B, Section 4.2.3). This learning rule yields updates that minimise three different losses through gradient descent: the 'cortical' loss equal to the total RPE^2, and the two 'cortical-striatal' losses equal to the pathway-specific RPE^2's

(Section 4.2.2). We term this model the 'tutor-executor' network because the 'cortical' learning (W^1) tutors downstream 'cortico-striatal' learning (W^2) by determining the relative salience of the inputs and balancing updates in the executor pathways to minimise its 'total' loss.

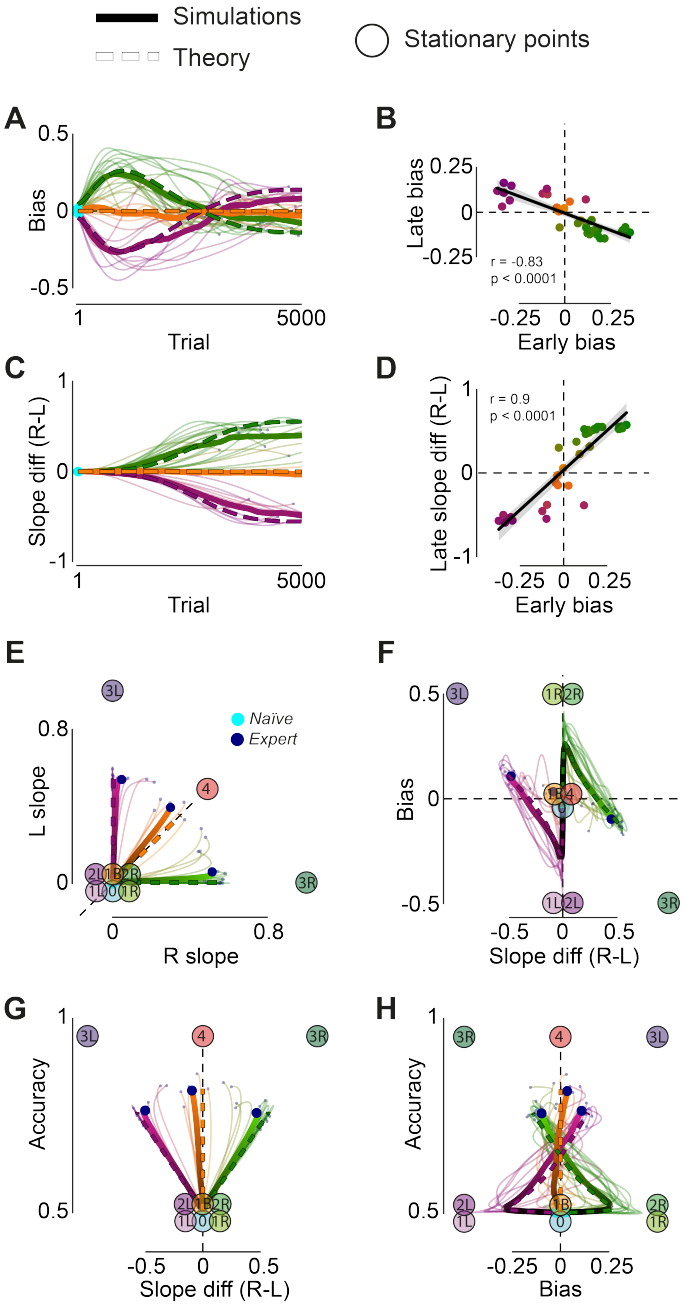


Figure 4.2: The tutor-executor model captures behavioural learning trajectories. (A), c.f. Figure 2.9A, Bias over trials per simulation (thin), for the 3 clusters from panel E (thick), and for the average dynamics (thick dashed). Here, and in panel C, the number of trials was limited to 5000. Thick dashed lines in all panels indicate trajectories derived from the average dynamics (see Section 4.2.5). (B), c.f. Figure 2.9B, Regression of early bias (average across trials 1000-2000) against late bias (average across final 1000 trials). Each point represents a simulation. p-value is calculated from the exact distribution of r. In all panels, shaded regions indicate 95% confidence interval across mice. (C), c.f. Figure 2.9C, Difference between right and left psychometric slopes over trials per simulation (thin) and for the 3 clusters from panel E (thick). (D), c.f. Figure 2.9D, Regression of early bias against late slope difference. Each point represents a simulation. p-value is calculated from the exact distribution of r. (E), c.f. Figure 2.9E, Right vs. left slope over trials per simulation (thin) and for 3 clusters (thick). Clusters and colours obtained using the same procedure as for the behavioural data in Figure 2.9E (Section 2.2.4). The clusters from this analysis are used in all other panels. Here and in panels F-H, numbered circles represent fixed points of the learning dynamics (see Figure 4.13) plotted using the average behaviour arising from their corresponding weight configurations. (F)-(H), c.f. Figure 2.9F-H, in order, difference in right and left (R-L) slope vs. bias, R-L slope vs. accuracy, and bias vs. accuracy over trials per simulation (thin) and for the 3 clusters from panel E (thick).

The model captured the diverse learning trajectories across mice. Similar to mice, the model started learning by developing varying degrees of left/right bias (Figure 4.2A-B cf. Figure 2.9A-B). Subsequently, the model's biases reversed as its psychometric slopes grew, exhibiting a similar diversity of left/right slope differences across trajectories to that seen in the mice (Figure 4.2C-D cf. Figure 2.9C-D). Further, the model's bias early in learning predicted bias and psychometric slopes later in learning (Figure 4.2B and D cf. Figure 2.9B and D; Figure 4.3A cf. Figure 2.10C).

As such, the model's entire learning trajectories resembled those of mice, capturing their diversity and systematicity (Figure 4.2E-H cf. Figure 2.9E-H), and exhibiting similar sigmoidal accuracy curves over comparable time scales (Figure 4.3B cf. Figure 2.6C; Figure 4.3C-E cf. Figure 2.6A, D and Figure 2.7). The model's simplicity allowed us to derive expressions for its average learning dynamics (Section 4.2.5), which showed close proximity to our behavioural data and model simulations (Figure 4.2A, C, and E-H, thick dashed lines). Any discrepancies between the theoretical average dynamics and cluster averages can be attributed to the clustering algorithm finding distorted barycentres due to the truncation of simulation trajectories after reaching 70% accuracy (Section 4.2.1).

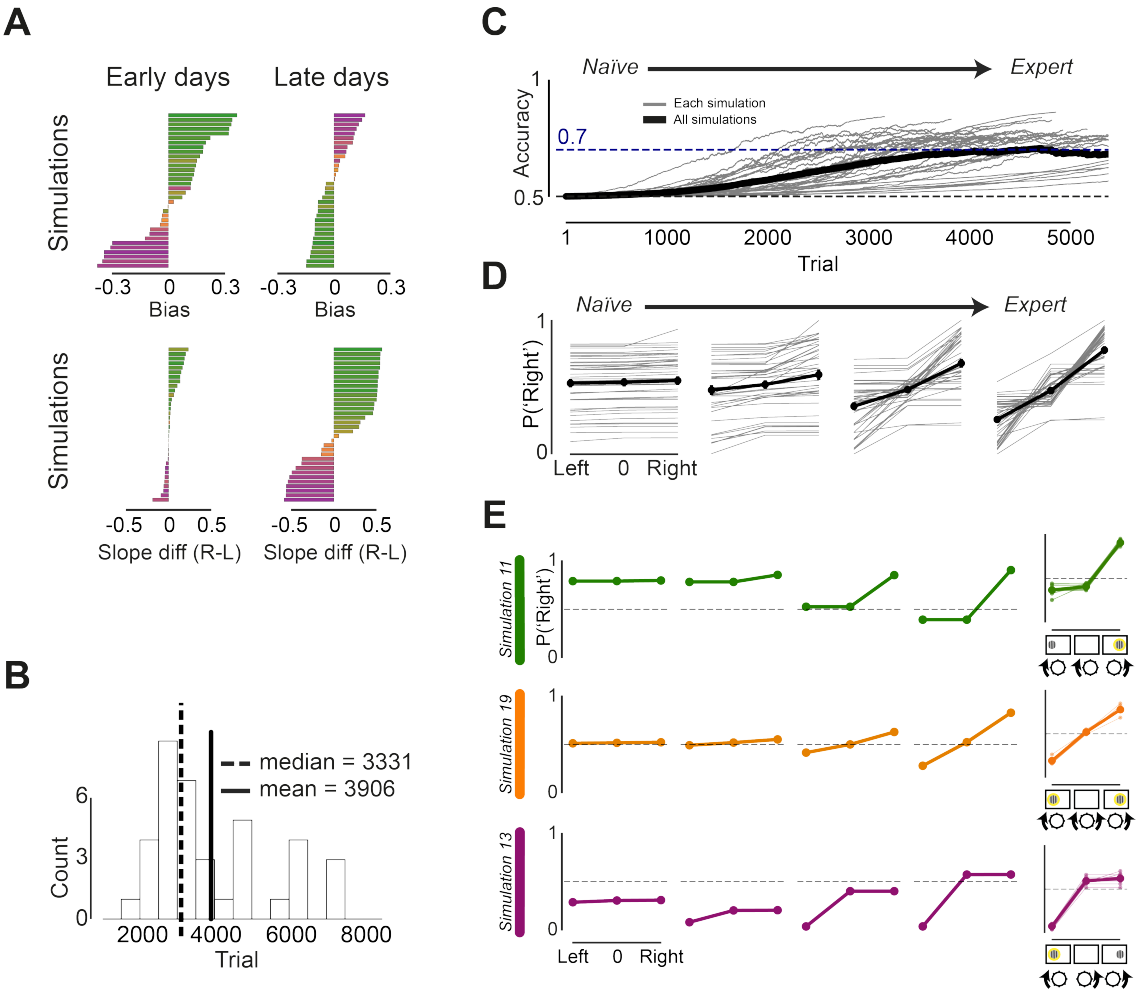


Figure 4.3: The tutor-executor model learns over similar timescales and displays similar patterns of diversity.
(A), c.f. Figure 2.10C, The model's average bias and slope difference early in training (average across trials 1000-2000) and late in training (average across final 1000 trials) for each simulation. Colours come from average psychometric slope asymmetry of each simulation (see Figure 2.3). (B), c.f. Figure 2.6C, Histogram of the number of trials that simulations required to reach 70% accuracy. (C), c.f. Figure 2.6A, Accuracy over trials per simulation (gray) and averaged across all simulations (black). Dashed lines indicate chance level (black) and 70% accuracy level (blue). Number of trials limited to the average simulation length. (D), c.f. Figure 2.6D, Psychometric curves over quartiles per simulation (gray) and averaged across all simulations (black). Error bars indicate +/- SEM across simulations. (E), c.f. Figure 2.7, Left, psychometric curves from 3 example simulations over quartiles throughout learning. Right, per simulation (thin) and average expert psychometric curves (acc.>70%) clustered by trajectory type (thick). Cluster labels for each simulation obtained from Figure 4.2E, colours obtained from average psychometric slope asymmetry. Error bars indicate +/- SEM across simulations.

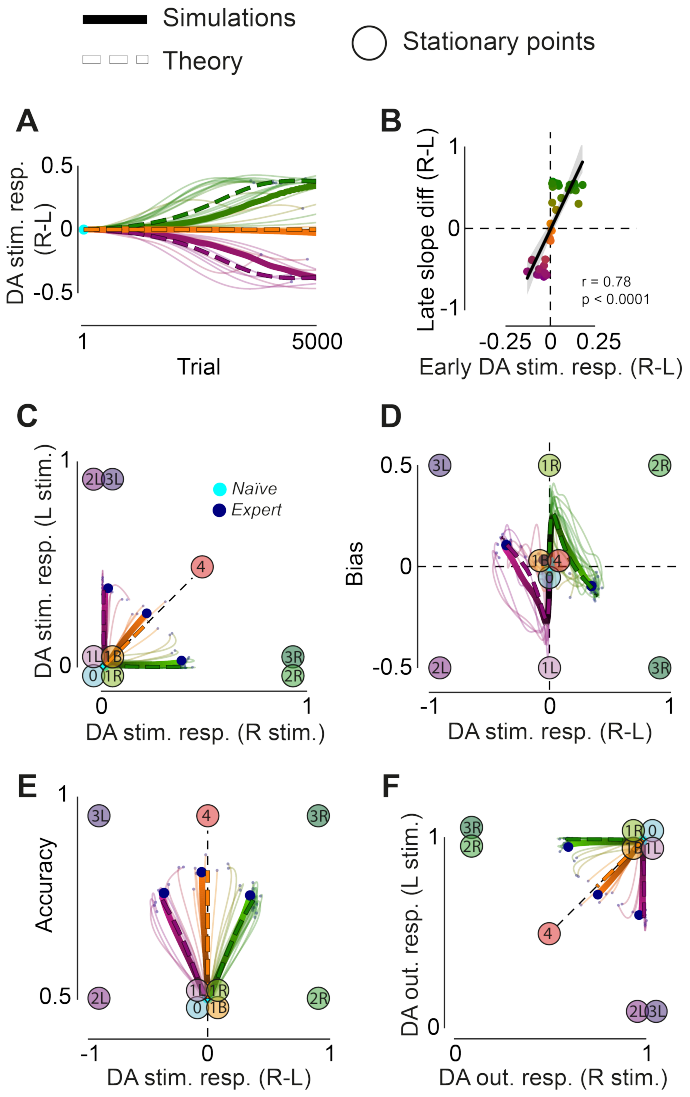


Figure 4.4: The tutor-executor model captures dopamine signals over learning.
(A), c.f. Figure 3.9A, Difference in DLS DA responses to right and left stimuli (R-L) over trials per simulation (thin) and for the 3 clusters from Figure 4.2E (thick). (B), c.f. Figure 3.9B, Regression of early difference in DLS DA responses to R-L stimuli against late slope difference. Each point represents a simulation. p-value is calculated from the exact distribution of r. (C)-(F), c.f. Figure 3.9C-F, in order, right vs. left DLS DA stimulus responses; difference in DLS DA responses to right and left stimuli (R-L) vs. bias; difference in DLS DA responses to right and left stimuli (R-L) vs. accuracy; and DLS DA rewarded outcome responses on right-stimulus vs. left-stimulus trials per simulation (thin) and for the 3 clusters from Figure 4.2E (thick). Fixed points are plotted using the average DLS DA responses arising from their weight configurations.

DLS DA responses over learning were also well captured by the model. We derived expressions for the trial-by-trial DA responses to stimuli and outcomes using reward predictions and RPEs from the 'stimulus' pathway (Section 4.2.3). Analogous to the empirical data, model-derived DA responses to stimuli grew over learning, reflecting the stimulus-choice associations that each animal formed (Figure 4.4A cf. Figure 3.9A). The partial prediction and RPE of the stimulus pathway best captured the DLS DA responses over learning, compared to the 'constant-based' and 'total' predictions and RPEs (Figure 4.5A cf. Figure 3.3E). The 'constant-based' RPE displayed negligible encoding of the diversity of learning trajectories, and the 'total' exhibited increasing stimulus responses and decreasing reward responses for the non-associated stimulus, which were not observed in the data. The model-derived DA signals early in training predicted both the model's slope difference (Figure 4.4B) and DA signals (Figure 4.5B-C) late in training, akin to DLS DA data (Figure 3.9B; Figure 3.10A-B). Thus, model-derived DA responses to stimuli across learning exhibited the diverse yet systematic progression of empirical DA signals (Figure 4.4C-E cf. Figure 3.9C-E). Finally, model-derived DA responses to outcome also showed strong similarity to our data, encoding the difference between the value of each trial's outcome and the reward predicted by the stimulus, i.e. RPEs in the 'stimulus' pathway (Figure 4.4F cf. Figure 3.9F; Figure 4.5D-F cf. Figure 3.10C-E).

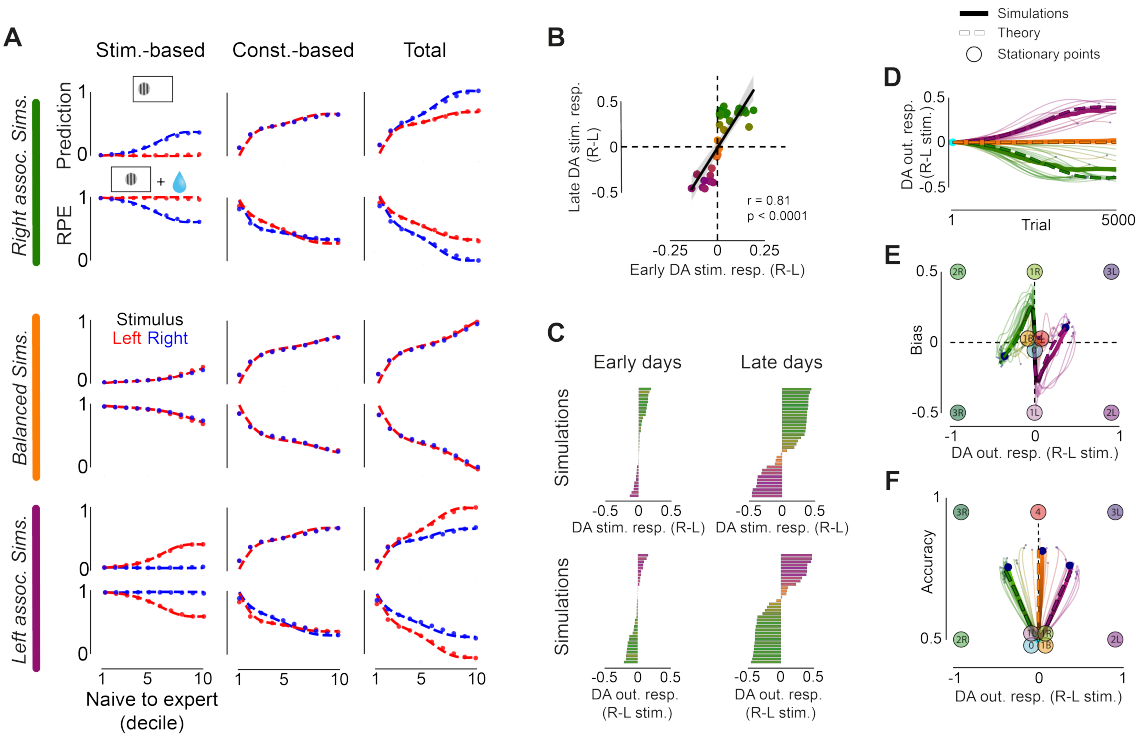


Figure 4.5: Further quantification of model-derived dopamine signals.
(A), c.f. Figure 3.3E, in order from left to right columns: average stimulus-based (i.e. model-derived DLS DA signals), constant-based and total predictions and RPEs over deciles in correct trials with stimulus on the left (red) and right (blue) for the three clusters from Figure 4.2E. Points show simulation averages and thick dashed lines show average dynamics. (B), c.f. Figure 3.10A, Regression of early difference in model-derived DLS DA responses to R-L stimuli against late difference in model-derived DLS DA stimulus responses. Each point represents a simulation. p-value is calculated from the exact distribution of r. Shaded region indicates 95% confidence interval. (C), c.f. Figure 3.10B, The average difference in model-derived DLS DA responses to right and left stimuli (R-L) and difference in model-derived DLS DA rewarded outcome responses after R-L stimuli early in training (average across trials 1000-2000) and late in training (average across final 1000 trials) for each simulation. Colours come from average psychometric slope asymmetry. (D), c.f. Figure 3.10C, Difference in model-derived DLS DA responses to rewarded outcomes after R-L stimuli over trials per simulation (thin) and for the 3 clusters from Figure 4.2E (thick). Number of trials limited to 5000. (E), c.f. Figure 3.10D, Difference in model-derived DLS DA responses to rewarded outcomes after R-L stimulus vs. bias. Fixed points here and in panel F plotted from average behaviour and model-derived DLS DA of their weight configurations. (F), c.f. Figure 3.10E, Difference in model-derived DLS DA responses to rewarded outcomes after R-L stimulus vs. accuracy.

Without a deep architecture, standard 'shallow' RL models do not capture mouse learning trajectories. We simulated learning in a shallow version of the tutor-executor model with fixed state representations for all inputs (Figure 4.6A). Behavioural and neural trajectories did not display the characteristic one-sided strategies of mice, and instead followed trajectories similar to those of balanced mice. The accuracy curve over learning of the shallow model also differed, lacking the initial plateau around 50% accuracy (Figure 4.6D). However, the trajectories of a deep RL model reproduced the individual diversity mice and the initial accuracy plateau (Figure 4.6B and D). The reason for the narrower diversity of the shallow model is that fixed state representations cannot implement one-sided strategies, where one of the two stimuli is ignored. Hence, the shallow model always learns to associate both stimuli with reward, displaying only transient imbalances if choices lead to more rewards on left/right stimulus trials (e.g., if biased). The deep model instead achieves long-lasting one-sided strategies, as observed in mice, by weighting the inputs with its W^1 weights such that weights for the non-associated stimulus remain near 0. The absence/presence of the initial

plateau in the accuracy curve is explained by each model's fixed point structure (Figure 4.6C). More details on the influence of fixed points on learning dynamics is provided in Section 4.3.2.

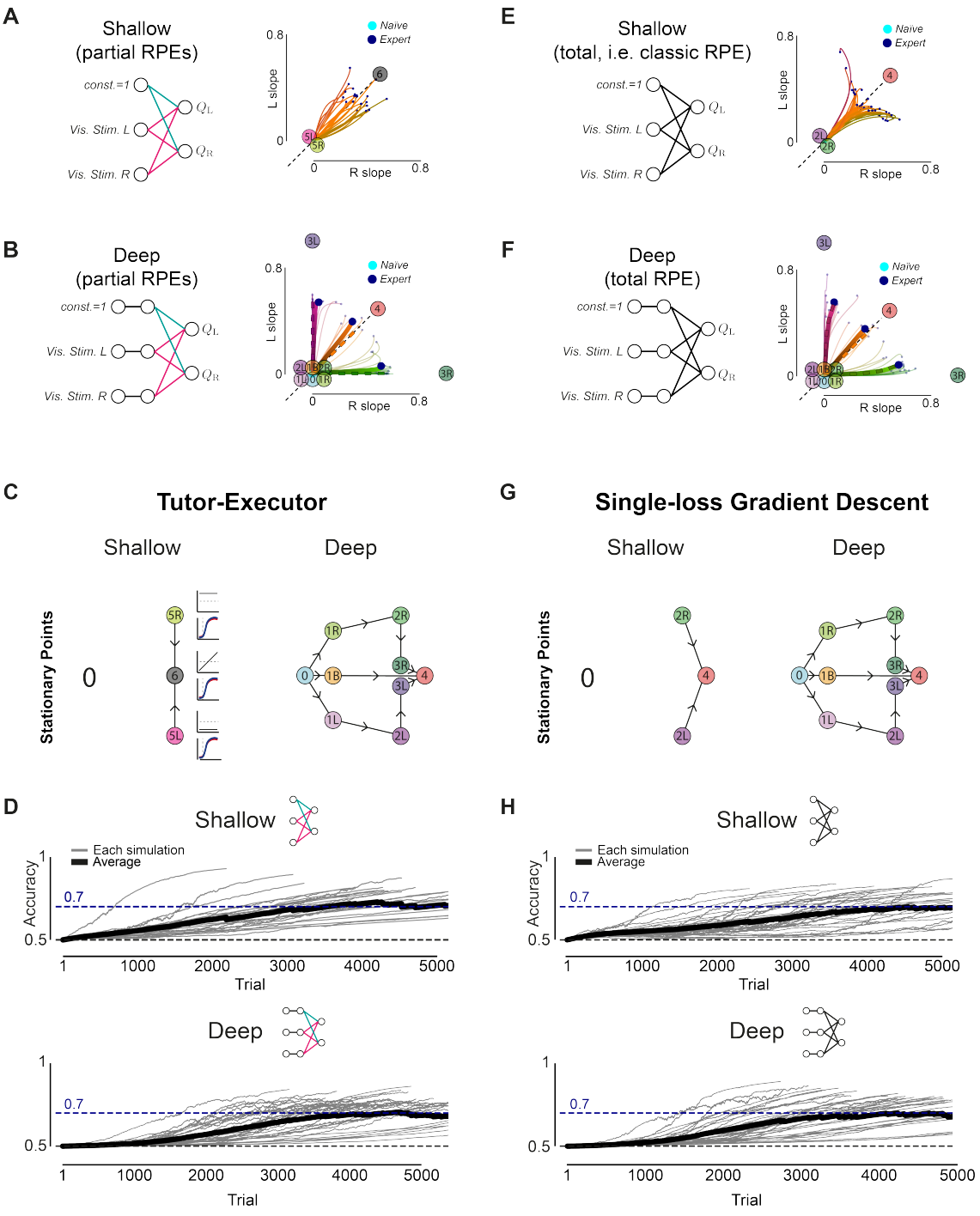


Figure 4.6: Shallow networks do not capture learning trajectories.
(A), Simplified architecture of a shallow version of the tutor-executor network (left), and right vs. left psychometric slopes over trials of its simulations (right). Trajectory colours obtained from the average slope asymmetry as in Figure 2.3. Numbered circles represent fixed points of the network learning dynamics (see panel C) plotted using the average behaviour arising from their corresponding weight configurations. (B), Same as A for a deep tutor-executor network. (C), Fixed points of a shallow (left) and deep (right) tutor-executor network. Note that 0 is not a fixed point of the shallow network, and is included for reference. (D), Comparison of the accuracy over trials for simulations from a shallow (top) and deep (bottom) tutor-executor network. Gray lines show accuracy curves for each simulation, and the black line is the average across simulations. (E)-(H), Same as (A)-(D) for the single-loss gradient descent network.

We simulated the optogenetic ChrimsonR stimulation experiment (Section 3.2.4) in silico using our neural network model (Section 4.2.4). This allowed us to test how different features of the model influence the behavioural effect of optogenetic manipulations (and the control manipulation using water reward). We tested two different learning rules in deep networks: the tutor-executor rule that included partial RPEs for the different pathways and the 'single-loss' gradient descent rule that only used a 'total' RPE (Section 4.2.2). Both learning rules in deep architectures captured the mice's behavioural learning trajectories (Figure 4.6A-D c.f. Figure 4.6E-H). This is because the tutor-executor rule effectively minimises the same 'total' loss of the single-loss rule, as its W^1 updates are based on the total RPE, δ_{tot}, and 'tutor' the 'executor' pathways in W^2 by determining the relative salience of the inputs, in turn governing the size of their corresponding W^2 updates. However, only the deep tutor-executor model could capture the differential effects on behaviour of DLS dopamine and water reward manipulations in the optogenetic experiment (Figure 4.7A-B c.f. Figure 3.13B). While both manipulations with the single-loss learning rule caused the entire psychometric curve to shift (Figure 4.7B), the tutor-executor learning rule correctly predicted a shift only on the associated side when stimulating DLS DA (Figure 4.7A). This lent further support to our hypothesis about the selective learning role of DLS DA as a partial RPE that modulates only stimulus-choice associations; showing how mice's behaviour is inconsistent with it acting as a classic total RPE.

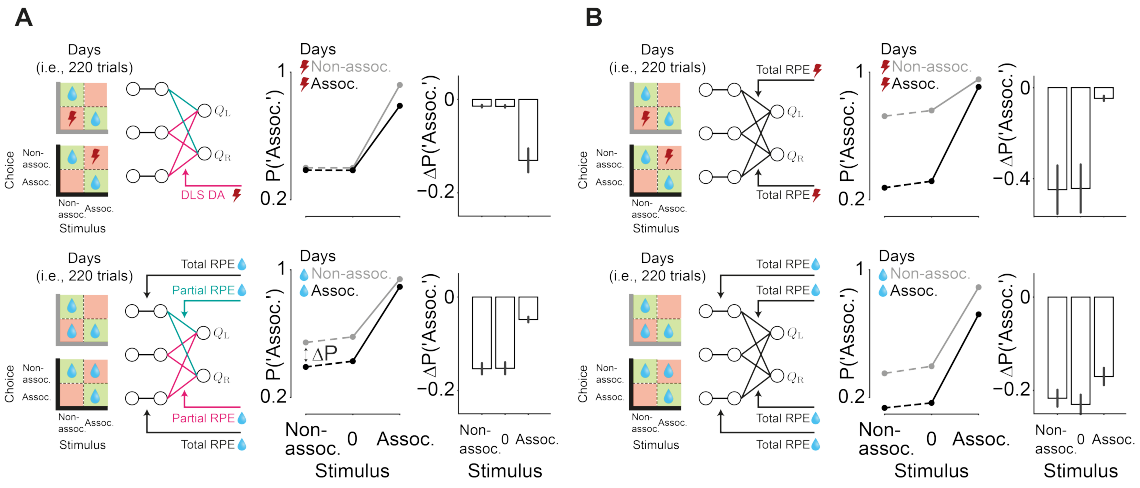


Figure 4.7: A model without partial RPEs does not capture optogenetic results. (A), Schematic and behavioural results of simulating the DLS ChrimsonR optogenetic experiment with the tutor-executor model. Left, schematic showing the alternating block structure and effect of DLS DA stimulation (top) and total (i.e. classic) RPE manipulation on the model's learning rule (bottom). Right, c.f. Figure 3.13B, average psychometric curves, and their difference, on blocks with DLS DA stimulation (top) and reward delivery (bottom) on error trials with non-associated (gray) and associated (black) stimuli. Error bars indicate +/- SEM across simulations. (B), Schematic and behavioural results of simulating the DLS ChrimsonR optogenetic experiment with the single-loss gradient descent model.

Having established that our deep tutor-executor network with nonnegative weights and W^1 one-to-one connectivity accounts for our data, we explored the role of the network's constraints. First, we explored the effects of network connectivity constraints while maintaining the nonnegativity constraint. A fully-connected network could not replicate the data (Figure 4.8C, left). The learning trajectories of this model were not able to reverse their bias and develop one-sided strategies. However, similar to our simpler one-to-one model (Figure 4.8A, left), a model in which the visual stimulus pathway was fully connected captured the behavioural trajectories (Figure 4.8B, left). Second, we relaxed the nonnegativity constraint on network weights, finding that while the one-to-

one model could still capture the trajectories (Figure 4.8A, right), a network with full connections in its stimulus pathway develops only balanced trajectories (Figure 4.8B, right). Similarly, a fully-connected network exhibited trajectories that balanced their slopes before developing one-sided strategies (Figure 4.8C, right). Exploring the effect of these constraints on models trained using a single total RPE gave similar results (Figure 4.8D-F). These results validated our choice of connectivity and nonnegativity constraints, which indeed agree with anatomical considerations of the segregation of left (right) visual (non-visual) inputs in sensory cortices, and abide by Dale's law (such that neurons don't change from excitatory to inhibitory).

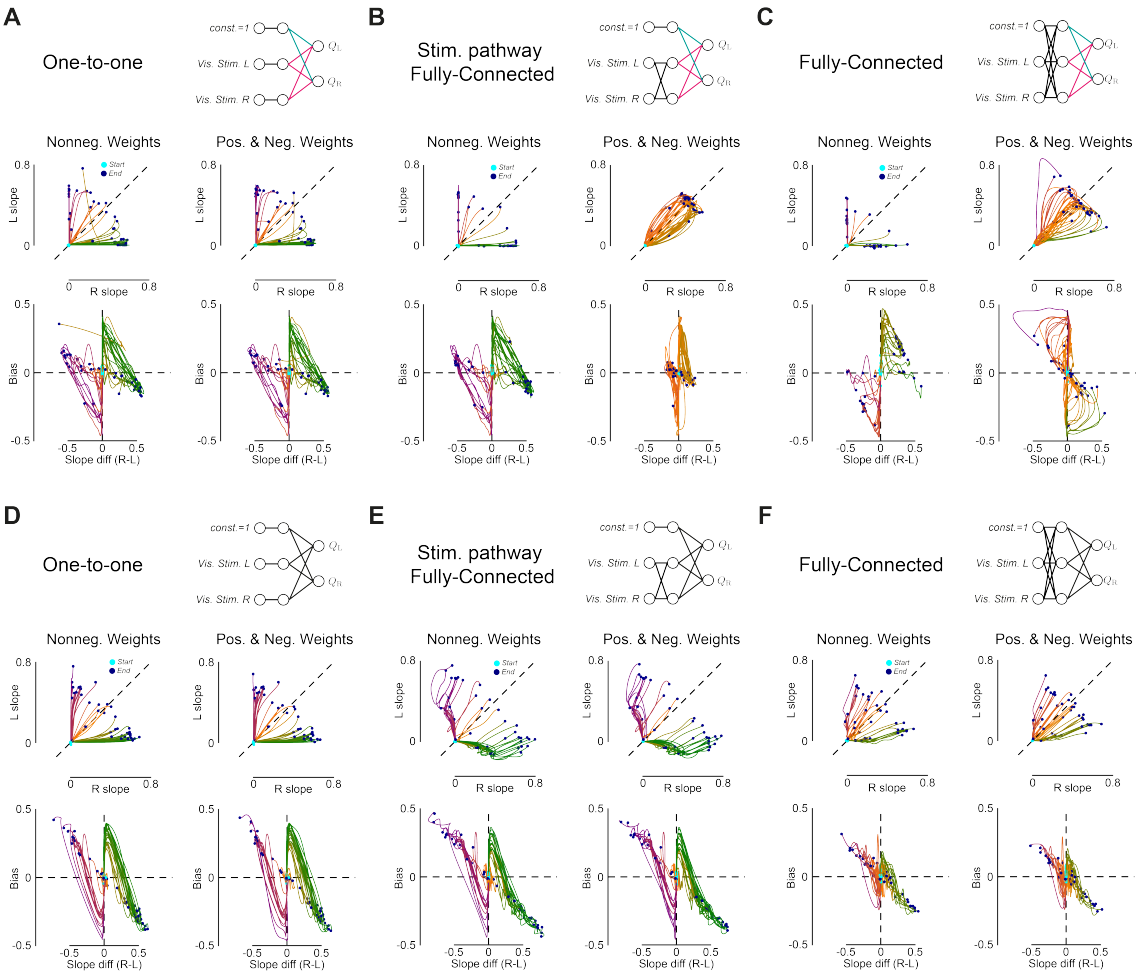


Figure 4.8: Exploring model constraints.
(A), Top, schematic of a tutor-executor network with one-to-one connections in W^1. Bottom, right vs. left slope (c.f. Figure 2.9E) and difference in right and left slope (R-L) vs. bias (c.f. Figure 2.9F) over trials per simulation for simulations with (left) and without (right) the nonnegativity constraint applied to the network weights. Trajectory colours obtained from the average slope asymmetry as in Figure 2.3. All simulations (n=40) are depicted regardless of whether they reach expertise (i.e., acc.>70%). Simulations were initialised as in Figure 4.2 (Section 4.2.1). (B), Same as panel A for a tutor-executor network with cross-connections between the stimulus inputs in W^1. (C), Same as panels A and B for a fully-connected tutor-executor network. (D)-(F), Same analysis as in panels (A)-(C) applied to a network trained with the single-loss gradient descent learning rule. Simulations were initialised as in Figure 4.15 (see Section 4.2.1).

We were often asked whether one-sided animals were blind in one eye, or in model terms, whether we really needed variable input representations (i.e., depth) or a fixed but different salience for the left and right stimulus was sufficient to capture the one-sided learning trajectories. To answer this question, we trained a version of the tutor-executor model where the weights in W^1 were

fixed and initialised to those of trained one-sided model simulations (Figure 4.9A-D). Simulations from this one-eyed model did not capture the learning trajectories, as they all developed the one-sided strategy determined by their W^1 setting (i.e., their 'good' eye) regardless of their early bias (Figure 4.9B-D). Hence, the simulations did not exhibit the dependence of the late slope imbalance on the early bias that we robustly observed in mice (Figure 4.9D c.f. Figure 2.9D).

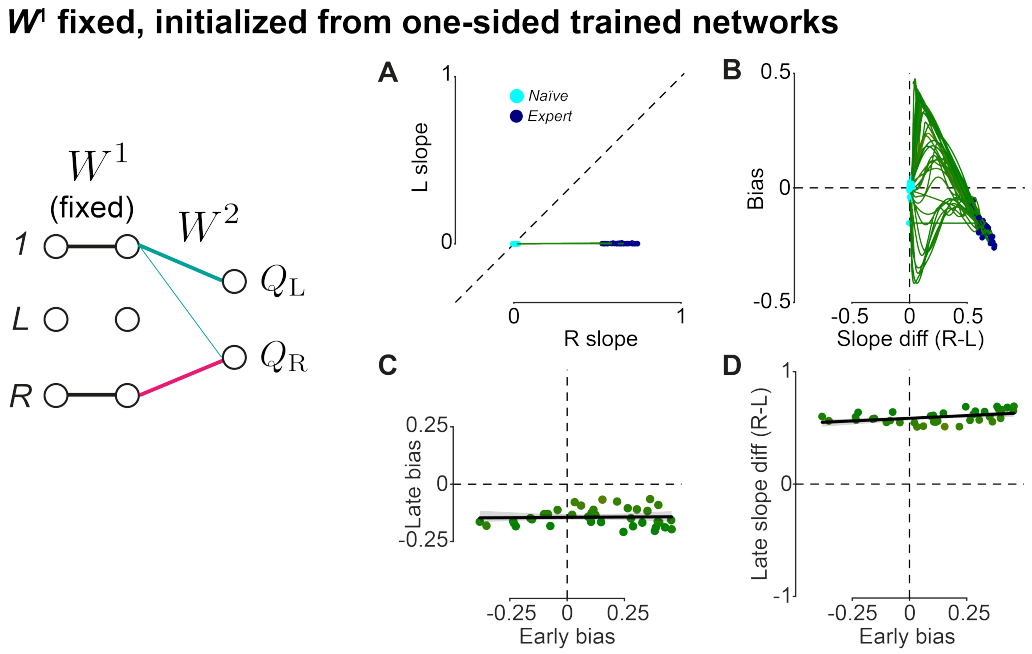


Figure 4.9: Effect of fixed non-uniform input representations.
(A)-(D), Effect of fixing W^1 to values from the average network configuration of right-associating simulations from Figure 4.2. Only simulations that reached expertise (i.e., acc.>70%) from n=40 simulations are shown. Simulations are from tutor-executor networks where the W^1 update is set to 0. Trajectory colours obtained from the average slope asymmetry as in Figure 2.3. (A), c.f. Figure 2.9E, right vs. left slope over trials per simulation. (B), c.f. Figure 2.9F, difference in right and left slope (R-L) vs. bias. (C), c.f. Figure 2.9B, Regression of early bias (average across trials 1000-2000) against late bias (average across final 1000 trials). Each point represents a simulation. In all panels, shaded regions indicate 95% confidence interval across simulations. (D), c.f. Figure 2.9D, Regression of early bias against late slope difference. Each point represents a simulation.

After our above investigation of the connectivity constraints, we were intrigued by the ability of the model with the fully-connected stimulus pathway to capture the behavioural trajectories (Figure 4.8B and E, left). In addition, in our behavioural data it was not uncommon to observe (slightly) V-shaped psychometric curves in one-sided animals (Figure 2.7), evidence for the misperception of one stimulus for the other. This is an effect that could be well-captured by cross-connections between the left and right stimulus inputs. We hence decided to further investigate the role of cross-connections in the stimulus pathway on learning (Figure 4.10). To do so, we parametrically manipulated the initialisation of the cross-connections in simulations all trained on the exact same trial sequence and found that larger connectivity between the L and R stimulus inputs led to greater imbalances in the psychometric slopes for left and right stimuli (Figure 4.10A). This is likely because, as the simulations transitioned from an early bias without stimulus use to discovering the stimuli, they associated both the 'associated' stimulus corresponding to the bias and the (previously) 'non-associated' stimulus with the same choice, due to the frequent misperception in the case of large cross-connectivity. Hence, the slope would grow one one side and become negative on the other, leading to V-shaped psychometric curves. Though potentially relevant, this cross-connection

effect cannot alone explain the one-sided trajectories, as it did not fully replicate the bias reversal (Figure 4.10C c.f. Figure 2.9B), nor the relation between early bias and late psychometric slopes (Figure 4.10D c.f. Figure 2.9D). Further, it did not capture the mice's DLS dopamine signals (Figure 4.10E c.f. Figure 3.9C). Its prediction of one-sided animals associating both stimuli with a choice predicted DLS DA stimulus responses for both stimuli, whereas we did not see responses to the 'non-associated' stimulus in mice (Figure 3.5A).

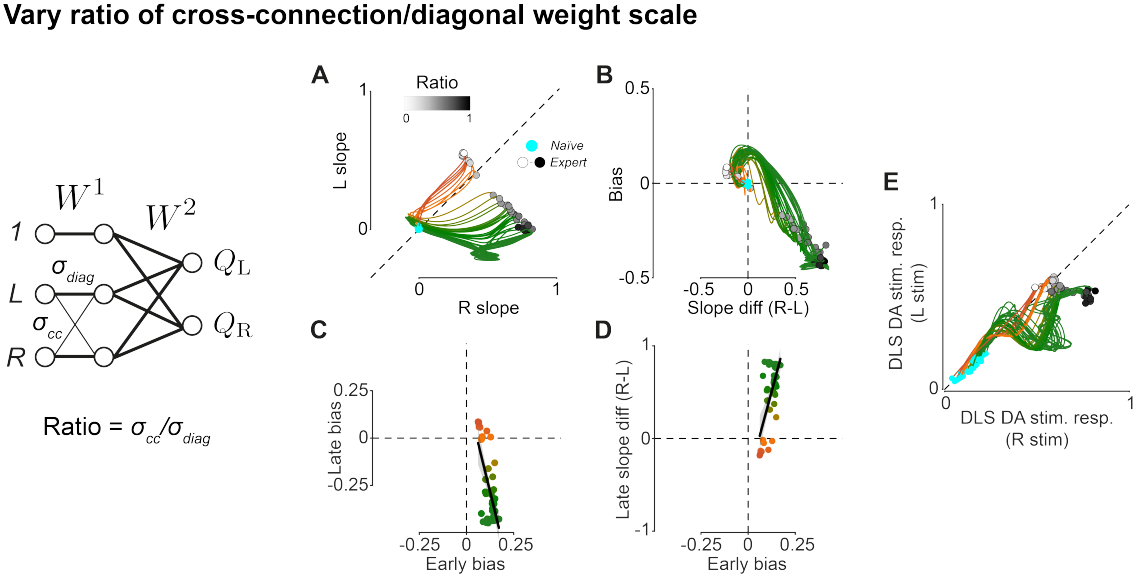


Figure 4.10: Effect of stimulus-pathway cross-connections.
Effect of varying the relative magnitude of W^1 stimulus-pathway cross-connection to diagonal weights at initialisation. Only simulations that reached expertise (i.e., acc.>70%) are shown. Simulations are from large single-loss gradient descent networks. (A)-(D), In order, left vs. right slope, difference in right and left slope (R-L) vs. bias, regression of early bias (average across trials 1000-2000) against late bias (average across final 1000 trials) and regression of early bias against late slope difference. All simulations used the same seed for the generated stimulus/reward sequence, softmax choice rule and initialisation weights, but with different variances for the Gaussians from which the cross-connection weights were sampled. (E), c.f. Figure 3.9C, right vs. left model-derived DLS DA stimulus responses over trials per simulation.

The neural network used for our main investigations had a reduced architecture that made it amenable to interpretation and analysis (Figure 4.1A). However, the neural circuitry involved in this task is surely much more complex than that of the reduced model presented thus far, making it crucial to verify whether the principles identified generalise to larger neural networks with more realistic input representations (Section 4.2.8). To do this, we provided a vectorised pixel-level version of the screen on each trial to a larger ReLU network with multiple input neurons, multiple layers and two output neurons representing left and right Q values as in the reduced model (Figure 4.11, left). The network was trained with the single-loss gradient descent learning rule to minimize the 'total' RPE^2 on each trial with weights constrained to be nonnegative. We chose the single-loss as opposed to the tutor-executor learning rule because the latter involved arbitrary design choices whose justification we believed to be outside the scope of this analysis. As for the reduced model, we also investigated the effect of different connectivity constraints on the learning dynamics (Figure 4.11A-C, left). Simulating the task and learning rule with these larger networks yielded overall similar learning trajectories to those of the reduced model (Figure 4.11A-C, right c.f. Figure 4.8). Only networks with segregated pathways for the stimulus and constant inputs generated learning trajectories similar to those of mice, with the fully-connected network leading to

more balanced simulations. Although the analysis of such large non-linear networks is difficult, the learning dynamics of their linear counterparts are better understood and known to produce similar trajectories under certain assumptions [37, 283]. It was thus possible to provide a high-level explanation of these results through the fixed point structure of their learning dynamics, as outlined in Section 4.3.2.

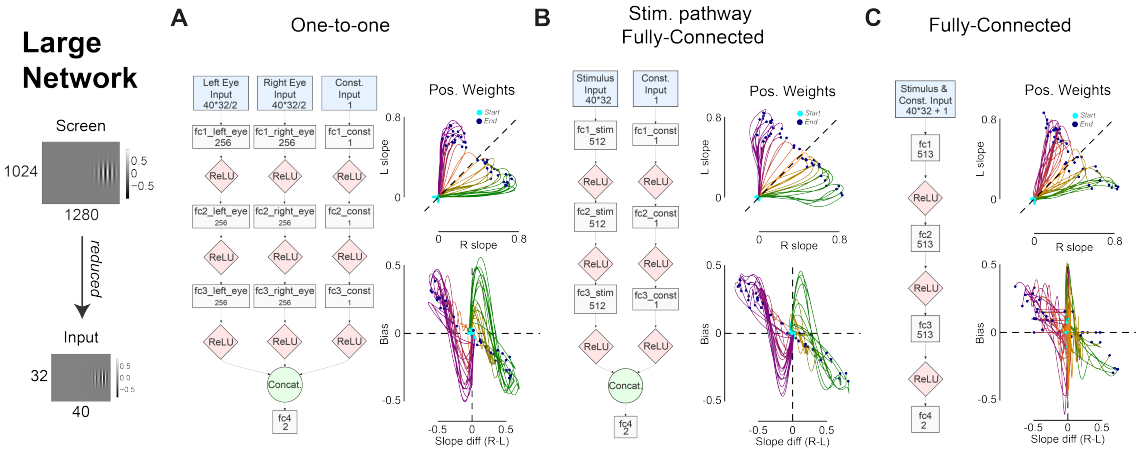


Figure 4.11: Generalisation to larger networks.
(A), Left, mermaid diagram of a scaled-up non-linear version of the one-to-one network with segregated fully-connected channels for each input which merge only in the final layer. Right, right vs. left slope and difference in right and left slope (R-L) vs. bias over trials per simulation trained with the single-loss rule and the nonnegativity constraint applied to the network weights. Trajectory colours obtained from the average slope asymmetry as in Figure 2.3. Only simulations that reached 70% accuracy in less than 8500 trials are depicted. Simulations were initialised such that the effective weights in each channel were as in Figure 4.15 (see Section 4.2.1). (B), Same as panel A for a scaled-up non-linear version of the network with cross-connections only between the stimulus inputs. (C), Same as panels A and B for a scaled-up version of the fully-connected network.

An intriguing feature of the tutor-executor learning dynamics is that, with extensive training, weight magnitudes transfer from W^1 ('cortical') to W^2 ('cortico-striatal') while maintaining the value of their product W^2W^1 (Figure 4.12). This resembles the 'transfer to striatum' reported in past studies, where cortex was found to be crucial for task performance only in early stages of learning [268, 269]. This feature comes from the tutor-executor's use of three different RPEs for updating, and particularly due to the hierarchy of information in their predictions. The total RPE δ_{tot} that updates W^1 combines information from both the stimulus and constant pathways to compute its prediction (Figure 4.1A), whereas the stimulus-based δ_{stim} and constant-based δ_{const} RPEs for W^2 only have access to predictions deriving from their respective inputs. This causes W^1 weights to 'learn' faster due to the quicker decrease of their teaching signal δ_{tot}, and also makes them responsible for 'tutoring' the evolution of the behaviourally relevant reward prediction (as choices were determined by the 'total' Q-values). Hence, the dynamics would often involve W^1 updates decreasing weights to compensate for excessively large W^2 updates due to their partial information. This effect was most prominent in the vicinity of fixed points, where a stable 'total' RPE had been reached and was maintained by W^1 decreasing its weights in response to increasing W^2 weights. Over time, this caused the weights in W^1 to 'transfer' to W^2 while maintaining the same effective network. In fact, the fixed points of the tutor-executor network existed only in the limit of infinitely small W^1 weights and infinitely large W^2 weights (Section 4.2.7). Though this may seem implausible biologically, the influence of noise being excessively boosted by large W^2 weights, as well as physiological limits on the strength of synapses, may cause this mechanism to

eventually halt. This transfer might explain past results involving DLS in perseverative behaviour [284], as the decrease in W^1 weights leads to lower flexibility in remapping input-choice associations due to W^2 updates being proportional to their magnitude (Section 4.2.2). On the other hand, it could be beneficial for multi-tasking and further learning due to the progressive decrease in reliance on the cortex over training, whose limited resources could then be employed for other purposes.

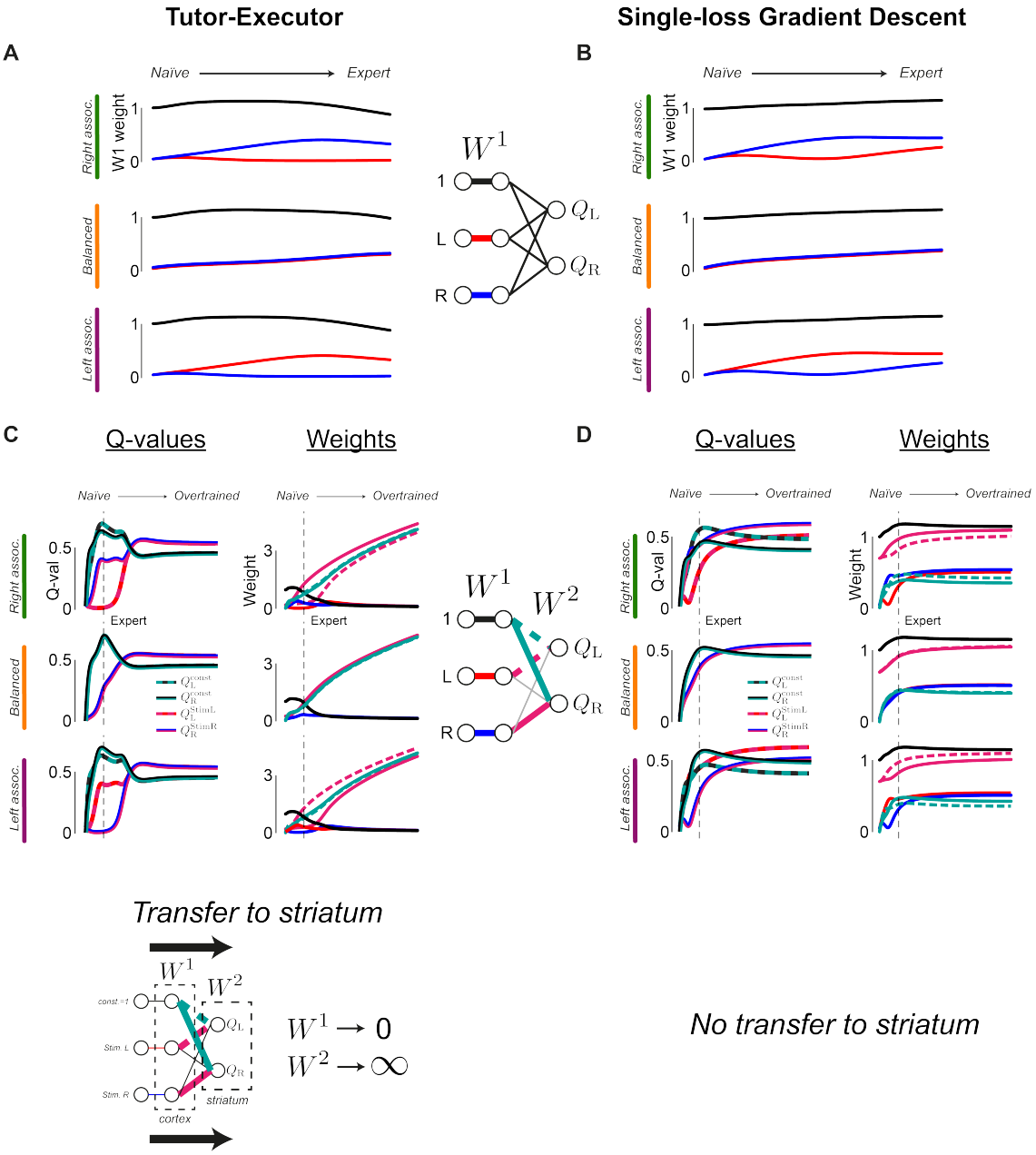


Figure 4.12: Tutor-executor learning rule causes transfer to striatum.
(A), W^1 weights from naïve to expert in examples of right- (top), balanced (middle) and left-associating (bottom) average dynamics of the tutor-executor network (same as dashed trajectories in Figure 4.2). (B), Same as A for the single-loss learning rule. (C), Left, stimulus- and constant-pathway Q-values of the tutor-executor average dynamics from naïve to overtrained. Overtrained: trained 8 times longer than the training used for naïve to expert. We do not plot the Q-values of the 'incorrect' associations (i.e., left stimulus with right choice and right stimulus with left choice) as these remain around 0. Right, W^1 and W^2 weights from naïve to overtrained. Again, we do not plot the weights that connect inputs with the wrong choices as these remain around 0. (D), Similar to C but for the single-loss learning rule.

4.3.2 Fixed point analysis

Analysis of the model's average dynamics revealed a hierarchy of saddle points that explained the behavioural and neuronal trajectories (Figure 4.13A). We found these saddle points by deriving the fixed points of the average dynamics (i.e., weight configurations where the average update across inputs and choices goes to zero, see Section 4.2.7). Saddle points have both stable and unstable manifolds, and in their vicinity learning momentarily slows down (Figure 4.13B-C). These points span the entire learning process, starting from a 'naïve' weight configuration (0) and converging on a final 'expert' global minimum (4), and reflect the diversity of trajectories across mice. The saddle points establish a systematic flow through the parameter space, whereby points approached early in learning influence those approached later in learning.

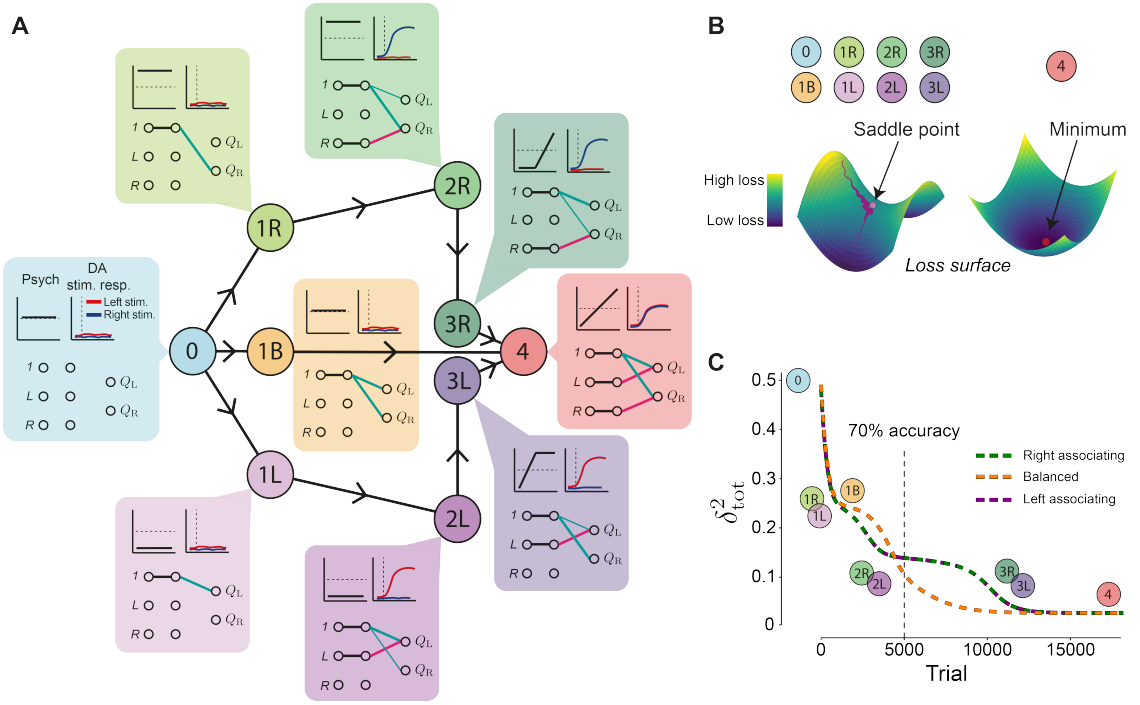


Figure 4.13: Saddle points of the model explain learning trajectories and dopamine signals. (A), Schematic of the fixed point structure including behavioural and neural predictions, as well as corresponding network weight configurations. The connecting lines with arrows represent the steepest heteroclinic orbits into/out of each fixed point. (B), Schematic of the loss surface around a saddle point and a minimum, along with the classification of each model fixed point. (C), Average dynamics of the total RPE^2 over learning for each of the clusters from Figure 4.2. Fixed points are plotted at approximate positions to depict their influence on the dynamics. A similar pattern of plateaus can be observed in the average RTs over learning for one-sided and balanced mice (Figure 2.10E).

Each fixed point has a characteristic behavioural and dopaminergic signature, similar to mouse behavioural strategies and DLS DA responses. Simulations start close to the first saddle point (0 in Figure 4.13A, corresponding to a network configuration with all weights set to zero) though with higher W^1 weights for the constant than the stimulus inputs, capturing its larger initial salience for the mice. All simulations are hence initialised similarly (Section 4.2.1), and what determines their eventual learning trajectory is the stochasticity of choices and rewards on zero-contrast trials. Simulations that, for example, develop an early right side bias learn in the direction of the 1R saddle point, developing a strong association between the 'constant' input and Q_R, evident in its corresponding weight configuration diagram (1R in Figure 4.13A). This happens when simulations receive more rewards for right than left choices, which can happen due to imbalances in either

the zero-contrast reward or choice history. These simulations then move preferentially towards the next saddle point, 2R, developing an association between the right stimulus and Q_R while maintaining a strong right bias. This happens as the simulation is still making mostly right choices (weight between 'constant' and Q_R is larger than between 'constant' and Q_L) but starts to learn the correlation between right stimuli and reward after right choices. As this correlation is learnt, the association between the 'constant' input and Q_R weakens through small negative 'total' RPEs that correct overpredictions of the reward on right stimulus trials. This encourages left choices in the absence of the right stimulus, such that simulations approaching 2R move towards 3R, where psychometric slopes emerge, and the bias starts to reverse. Here, the simulation maintains its association between the right stimulus and Q_R while developing a strong weight between 'constant' and Q_L, all without using the left stimulus. Thus, in the vicinity of this saddle point the simulation infers correct left choices from the absence of the right stimulus, showing psychometric slopes and DA signals similar to expert 'right-associating' mice (green cluster in Figure 2.9 and Figure 3.9). A mirror image of this trajectory is observed in simulations that develop an early left bias (follow 1L, 2L and 3L), whereas more balanced simulations (those with a negligible early bias) move from 1B towards 4, i.e. the global minimum. Hence, the saddle points of the deep tutor-executor network govern the learning trajectories (see saddle points visualised in Figure 4.2E-H and Figure 4.4C-F), explaining their diverse yet systematic transitions between strategies and corresponding DA signals.

The saddle point structure of a model substantially influences its learning dynamics. First, it governs the systematic transitions between strategies during learning, defining the relation between early and late behaviour. Further, the multiple layers of the deep model give rise to saddle points that are critical for explaining learning trajectories, in particular 3R/L, that capture mice's one-sided strategies. These saddle points emerged in our tutor-executor network as well as networks trained with standard gradient descent on a single loss (Figure 4.6G and Figure 4.15). In contrast, the saddle point structures of shallow models lack these one-sided saddle points, and succinctly demonstrate why such an architecture cannot capture the mice's learning trajectories (Figure 4.6B and G). The fixed point structure of fully-connected networks also lacked these saddle points, due to their ability to associate both left/right stimuli to the same W^2 weight. Second, saddle points give rise to learning plateaus, where learning is slower. This can be observed in the dynamics of the model's total RPE, which exhibits plateaus in the vicinity of the saddle points from balanced and one-sided trajectories (Figure 4.13C). Further, in the deep network, the 0 and 1 saddle points capture the initial plateaus in accuracy observed in mice (Figure 4.6D and H c.f. Figure 2.6A). These plateaus also explain why some mice did not learn despite sufficient training (Figure 2.11A-B). Similar to these mice, the trajectories of model simulations that did not learn stayed in close vicinity of 1L/R or 2L/R saddle points (Figure 4.14A and B).

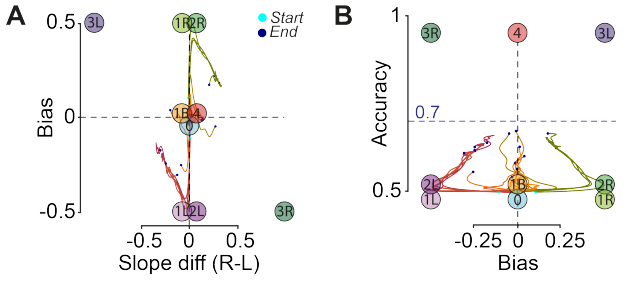


Figure 4.14: Non-learner simulations.
(A)-(B), c.f. Figure 2.11B, slope difference vs. bias and bias vs. accuracy over trials of tutor-executor simulations that did not reach accuracy>=70% in less than 8500 trials from a total of n=120.

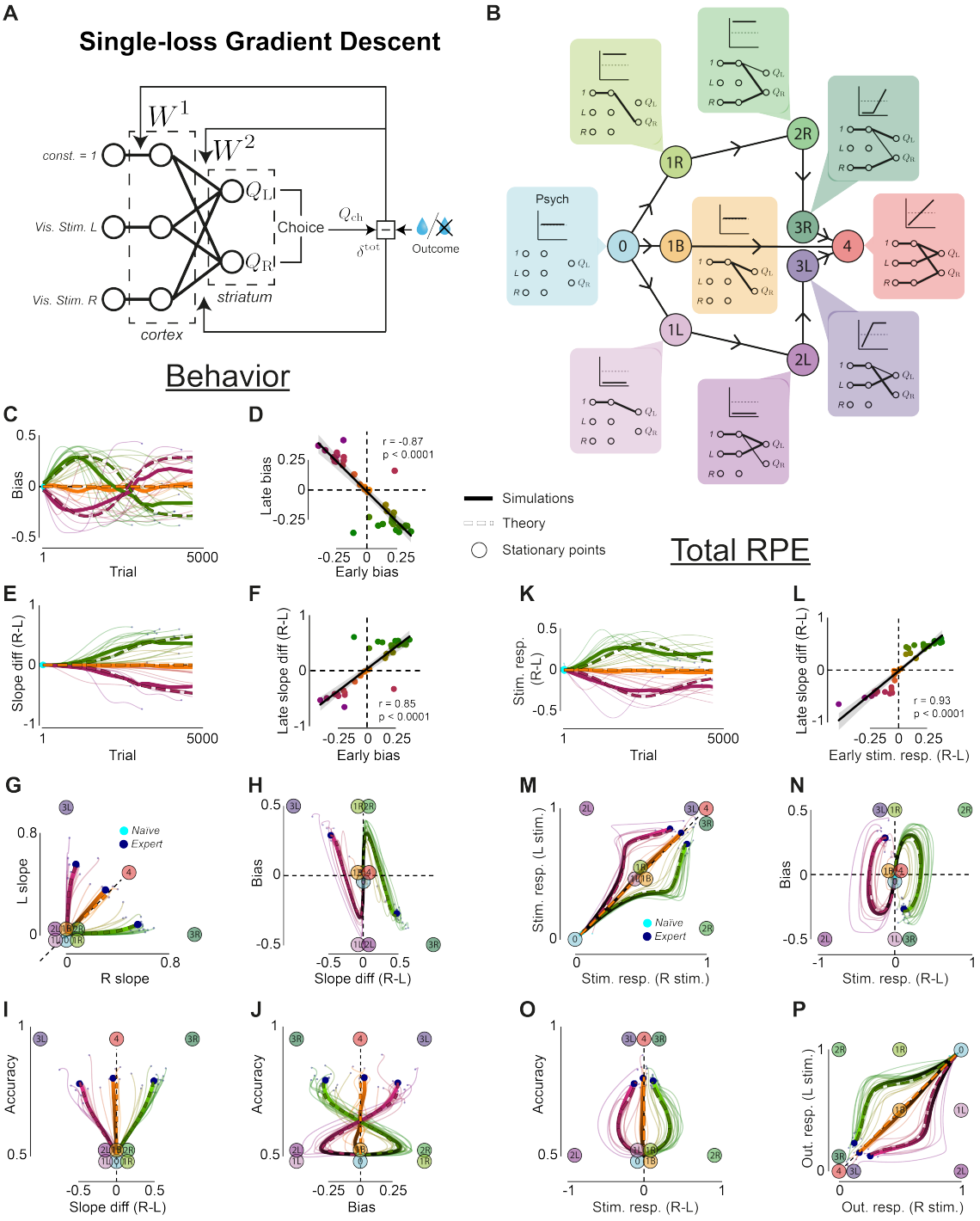


Figure 4.15: A deep model with single-loss gradient descent captures behavioural trajectories with a teaching signal different from DLS DA.
(A), Schematic of the deep linear 'single-loss' RL network architecture and learning rule (Section 4.2.2). (B), Schematic of the fixed point structure with behavioural predictions as well as corresponding network weight configurations. The connecting lines with arrows represent the steepest heteroclinic orbits into/out of each fixed point (Section 4.2.7). All the fixed points are saddle points except for 4, which is the global minimum. (C), c.f. Figure 2.9A, Bias over trials per simulation (thin), for the 3 clusters from panel G (thick), and for the average dynamics (thick dashed). Here, and in panels E and K, the number of trials is limited to 4667. Thick dashed lines in all panels indicate analytical trajectories derived from the average dynamics (Section 4.2.5). (D), c.f. Figure 2.9B, Regression of early bias (average across trials 1000-2000) against late bias (average across final 1000 trials). Each point represents a simulation. p-value calculated from the exact distribution of r. In all panels, shaded regions indicate 95% confidence interval across mice. (E), c.f. Figure 2.9C, Difference between right and left psychometric slopes over trials per simulation (thin) and for the 3 clusters from panel G (thick). (F), c.f. Figure 2.9D, Regression of early bias against late slope difference. Each point represents a simulation. p-value calculated from the exact distribution of r. (G), c.f. Figure 2.9E, Right vs. left slope over trials per simulation (thin) and for 3 clusters (thick). Clusters and colours obtained using the same procedure as for the behavioural data in Figure 2.9E. The clusters from this analysis are used in all other panels. Fixed points here and in panels H-J are plotted using the average behaviour arising from their weight configurations. (H)-(J), c.f. Figure 2.9F-H, in order, difference in right and left (R-L) slope vs. bias, R-L slope vs. accuracy and bias vs. accuracy over trials per simulation (thin) and for the 3 clusters from panel G (thick). (K), c.f. Figure 3.9A, Difference in total RPE signals after right and left stimuli (R-L) over trials per simulation (thin) and for the 3 clusters from panel G (thick). (L), c.f. Figure 3.9B, Regression of early stimulus-evoked total RPE (average across trials 1000-2000) against late slope difference (average across final 1000 trials). Each point represents a simulation. p-value calculated from the exact distribution of r. (M)-(P), c.f. Figure 3.9C-F, in order, right vs. left stim.-evoked total RPEs; difference in total RPE evoked by right and left stimuli (R-L) vs. bias; difference in total RPE evoked by right and left stimuli (R-L) vs. accuracy; and outcome-evoked total RPE signals after right stimulus vs. after left stimulus per simulation (thin) and for the 3 clusters from panel G. Fixed points are plotted using the average total RPEs arising from their weight configurations.

4.4 Discussion

We tied the behavioural and neural results together through a deep neural network model trained using heterogenous dopamine-like teaching signals, which reproduced the empirical distribution of mouse learning trajectories as well as dopamine signals and optogenetic results. We further demonstrated that the data cannot be explained by standard 'shallow' RL models with predetermined state representations. Instead, our results suggest that mice refine their state representations throughout learning, a process well-captured by gradient-based reinforcement learning in a deep neural network. Further, we demonstrated that networks require heterogenous teaching signals in order to capture the dopamine signals and their optogenetic effects. Lastly, we performed a fixed point analysis of the model's learning dynamics and discovered a hierarchy of saddle points that govern the diversity and systematicity of the learning trajectories.

4.4.1 Depth and heterogenous RPEs

Depth is a crucial element of our model, which enables the network to capture both the diversity and systematicity of the mice's behavioural and neural trajectories. The chaining together of weights in different layers through a product enables the reproduction of one-sided strategies where W^1 selectively gates away one of the stimulus inputs. One could argue that this effect could also be captured with non-uniform, fixed state representations as the model's input e.g., setting Vis. stim. L to 0 for a right-associating mouse. This would be equivalent to shutting off the input from the left stimulus, as if the mouse was blind in one eye. To explore this, we tried capturing the learning trajectories with a network with W^1 weights initialised and fixed to their average value in expert right-associating simulations, only updating W^2 over learning. However, we found that although such a learning algorithm indeed reproduces the strategy that right-associating mice developed at expertise, it does not capture the dependence of late behaviour on early behaviour (Figure 4.9). Hence, although such 'fixed' imbalances in perception of the two stimuli may be partially contributing to the development of one-sided strategies, they cannot explain the full learning trajectories of mice from naïve to expert. Further, the analysis of eye movements and pupil size from Chapter 2 suggest that such early imbalances in perception might not be present at all (Figure 2.12B). We thus opted for deep networks (training both W^1 and W^2) initialised with similar W^1 weights for the right and left stimulus to capture the mouse data. Importantly, the model was able to reproduce the diversity of learning trajectories despite this balanced initialisation. This indicates that initialisation effects may only play a minor role in determining mouse learning trajectories, these instead being mainly influenced by each mouse's individual choice and reward history.

The heterogeneity of teaching signals is an important element of the learning rule operating in the model. This heterogeneity gives rise to unique phenomena which would not be expected from a system using a single RPE for learning, such as the 'transfer-to-striatum' effect (Figure 4.12). Our behavioural and DLS dopamine results did not distinguish between learning rules that used heterogenous RPEs vs. a single total RPE. However, our optogenetic results distinguished these, showing that the deep model requires heterogenous teaching signals to account for the distinct behavioural effects of DLS DA manipulation (i.e. partial RPE) and reward size manipulation (i.e. total RPE). From the perspective of calculating RPEs based on a subset of inputs, our model is similar to a recent study [254]. Critically however, unlike our model, their model sums heterogenous RPEs to provide a 'globally-broadcasted' RPE (akin to our 'total' RPE) for updating all weights

into striatum. We instead update corticostriatal weights using the partial RPE from the prediction they contribute to. In fact, we propose a mapping from model to anatomy (Figure 4.1B) where diverse corticostriatal projections compute partial reward predictions in different regions of striatum which, in turn, project to different dopamine neurons in the SNc. Here the partial prediction is compared with a total reward signal to produce partial RPE encoding of dopamine activity. Finally, following recent evidence [260], we propose that the patterns of dopamine release across the striatum inherit the diversity of corticostriatal connectivity by means of closed striatonigrostriatal loops. Regarding the 'total' RPE signal that projects to W^1 in our model, we believe this could be computed in the dopamine neurons of the VTA rather than the SNc. The VTA receives projections from a wide range of brain regions, and is known to project to the cortical association areas that project to DLS and other striatal regions (e.g., frontal association areas, MOS, etc. [186]), making it a promising candidate. In fact, dopamine release patterns resembling total RPEs have been observed in the nucleus accumbens [285], a region that also receives dense innervation from the VTA. Other RL models, such as distributional RL, attribute DA heterogeneity to diverse sensitivities to positive and negative RPEs [286]. However, such a model does not capture our data because, despite choices with associated and non-associated stimuli having similar reward distributions when accuracies are matched, DLS DA signals differed substantially on these two trial types.

After probing the model's constraints, we believe that our choice of capturing the mice's learning trajectories and dopamine signals with a deep non-negative tutor-executor network with one-to-one connections in W^1 is well-founded. This not only proved to capture the learning trajectories and optogenetic manipulations better than other models, but also yielded a simpler, more interpretable network. Thus, this suggests that mice might not be employing flexible, non-linear, deep distributed representations for this task. However, we think that the cortex and cortico-striatal connectivity are involved (c.f. less flexible brain regions), given that inactivation of cortical regions projecting to striatum impairs mice's performance of this task [187]. Further, the anatomical segregation of left/right visual pathways in the visual cortices is consistent with the one-to-one mapping we use. Our investigation of deeper and wider networks demonstrated that the principles which govern the learning trajectories of our simple main model generalise to larger networks. Importantly, this provides further evidence that these principles may be operating in the brain, which is much larger and more complex than the simple network we use. The simple network, however, proved useful to design the tutor-executor learning rule, understand and visualise the changes in weights over learning, and derive the analytical expressions for the average learning dynamics which enabled its fixed point analysis.

4.4.2 A mathematical theory

Our experiments highlighted two hallmarks of long-term learning: its systematic stage-like progression through strategies and its marked individual diversity. Many other well-studied abilities such as semantic and numerical cognition are characterised by similar structured transitions through strategies [112, 24]. However, the source of individual differences in learning has been difficult to understand. Our model reveals that both hallmarks could arise from a single dynamical process approaching saddle points in the reward landscape of a deep neural network. As the learning dynamics approach a saddle point, small gradients in its vicinity cause learning to momentarily slow down until the saddle point's unstable manifold is discovered [287, 37]. The dynamics then speed up as gradients grow, following the steepest path to the next saddle point in the sequence where the process repeats itself. This alternation between slow and fast learning along sequences of saddle

points explains the systematic stage-like transitions during long-term learning, which manifest as distinct stages in animal learning trajectories [159]. These stage-like transitions are evident in the average dynamics of the model's total RPE (Figure 4.13C), which resembles mice response times over learning (Figure 2.10E). Meta-cognitive processes such as hypothesis testing [227, 288] or switching between exploitation and exploration [289, 290, 291, 292] may help agents spend less time in the vicinity of saddle points as they learn, thus shortening the periods of slower learning. We also predict that direct supervision from a teacher will be effective at driving dynamics away from saddle points, as they can point out relevant stimuli and encourage particular behaviours to push agents out of saddles or avoid them completely throughout learning.

The divergence in the sequences of saddle points demonstrate how choices made early in learning influence strategies developed later on. Thus, the multiplicity of systematic paths through the saddle point structure explains individual diversity. Consistently, the trajectories of mice that failed to learn remained in the vicinity of initial saddle points (1L/R or 2L/R). This saddle-point approach to explain the diversity and systematicity of learning trajectories can in principle be extended to more complex tasks with higher-dimensional inputs and actions. In fact, the learning dynamics of networks trained on such tasks are likely to have many more saddle points with more complex heteroclinic orbits and thus richer patterns of diversity. Finding the sequences of saddle points that govern learning in these settings is an open problem in ML theory. Nonetheless, progress is being made by identifying core principles in restricted settings [136, 131, 137, 283]. Flipping the problem, insights on the circuits and learning rules involved in learning different tasks can be obtained from data by finding the stages where individuals' learning slows down. Hence, collecting long-term learning data from a variety of tasks will be useful to constrain model design choices and identify the teaching signals and changes in the brain's connectivity that underlie learning in each task.

Simulating the task and learning rule with larger networks yielded overall similar learning trajectories to those of the simpler main model (Figure 4.11). Only networks with segregated pathways for the stimulus and constant inputs generated learning trajectories similar to those of mice, with the fully-connected network leading to more balanced simulations. This is because fully-connected networks do not have one-sided saddle points (equivalent to 3R/L) in their fixed point structure (this also applies to the reduced network), similar to shallow networks. The lack of these one-sided saddle points causes any imbalances in right and left slopes to be transient, and hence such models were unable to recreate the sustained single-stimulus use observed in mice, as well as their concomitant bias reversal. This distinct dynamic of the bias reversal was surprising for many behavioural researchers, and is explained by the model through the evolution of its constant pathway weights. These transition from the 1R/1L saddle points where they associate the constant input with either the right/left choice to compute a reward prediction of 0.5 as the network makes primarily right/left choices, to later the 3R/3L saddle points where they associate the constant input with the other choice. This flip happens as the stimulus on the same side of the bias is gradually associated with the biased choice (due to its more reliable prediction of reward), causing the association with the constant input to be weakened to avoid overpredicting the reward. This weakening causes the network to make less biased choices in the absence of the associated stimulus, until the bias eventually reverses as the constant input learns to predict the reward for the other choice in the absence of the associated stimulus. This all happens without the other stimulus ever being associated due to both its corresponding W^1 and W^2 weights being close to 0, and thus gradient descent updates also being 0. The phenomenon created by the 3R/L saddle points

has been previously described in the machine learning literature as unimodal bias in multimodal networks with late fusion [283]. Zhang et al. demonstrate that unimodal bias arises in arbitrarily large multimodal networks with late (i.e., "deep") fusion of the modalities, and that it is similarly governed by unimodal saddle points. They also show that the deeper the layer at which the fusion occurs, the longer the unimodal phase.

Now, why might the brain have such a learning mechanism that gives rise to saddle points which slow down learning at periodic intervals? At first glance, this seems like a terribly inefficient mechanism that regularly interrupts steady periods of performance improvement (when an individual can get 'stuck' and risk losing motivation) every time the brain arrives to a configuration in the vicinity of a saddle point of its learning mechanisms. Wouldn't it be much more efficient to learn via trajectories that reach the final objective with a stable rate of progress? One potential reason is that these stage-like transitions enable the 'chunking' of tasks. The saddle points we found in our deep RL model mostly correspond to network configurations implementing partial solutions of the task. Examples are the 1L/1B/1R and 3R/3L saddle points which respectively correspond to network configurations forming the most accurate reward predictions using only the const. input or the const. input and one of the two stimuli. As such, the saddle point structure reflects the partial strategies that individuals progressively master as they learn, persevering with each strategy for a while before transitioning to the next. A prediction from the tutor-executor learning rule is that during these periods of perseverance, representations that were in the cortex during the early phases of learning new associations would transfer to subcortical structures such as the striatum [268, 269]. In doing so, this might reduce the brain's cortical load, freeing it for further learning or multitasking. In this way, the cortex could focus on representing novel stimuli that may improve performance on the task, such as learning to pay attention to traffic lights as one cycles. As such, the brain might exploit or have evolved to operate with a learning mechanism involving these dynamics in order to circumvent the limited capacity of the cortex and enable the learning of ever-more complex strategies and representations. If this prediction holds true, another interesting consequence of the tutor-executor learning rule is that once learned associations are transferred to the subcortex, they become very difficult to change, potentially giving rise to what we commonly refer to as 'habits' [293, 294]. This arises due to the gradient descent updates of W^2 being proportional to the magnitude of W^1 weights. Hence, although such perseverative behaviour might make it difficult for animals to adapt to changing circumstances [295, 296], it may be crucial to reduce cortical load and enable the formation of rich representations that permit further learning and task mastery. This transfer from cortex to striatum resembles the complementary learning systems theory of memory consolidation from the hippocampus to the neocortex [297, 74, 298, 299], yet in the opposite direction. Both processes may be working in tandem to first consolidate useful memories into rich cortical representations during learning [159], and later transfer them to striatum such that previously engaged regions of cortex can be repurposed for new learning problems.

In addition to explaining the characteristics of mouse learning trajectories from naïve to expert, our fixed point analysis elucidates why some mice did not reach expertise. By definition, the learning dynamics of a network configuration exactly at a saddle point are 0. Consistently, gradients become increasingly smaller as you move towards a saddle point along its stable manifold. This, in turn, makes it increasingly difficult to escape a saddle point the closer the dynamics get to it. Hence, mice that did not reach 70% as quickly as 'learner' mice might have developed brain network configurations exceedingly close to a saddle point, causing them to never proceed further in learning. In fact, the trajectories of non-learner mice from Figure 2.11A-B indicate that such an

explanation might be quite plausible for many of the mice that remain with highly biased strategies and never use the visual stimuli. This property of saddle points might also explain why most of the one-sided mice don't ultimately converge at the global minimum 4. Although we observe early signs of 'bending in' towards 4 in the trajectories from Figure 2.9E-H, one-sided left- and right-associating mice do not seem to ultimately associate both stimuli with their corresponding choice and develop balanced strategies. One reason for this could be purely practical: several mice were not trained for much longer after reaching 70% accuracy, because our aim was to acquire data from a larger number of mice so that we could investigate individual diversity. Mice appeared to persevere with one-sided strategies for quite long and hence in the few sessions that we trained past 70%, the mice were not able to fully associate the other stimulus with its corresponding choice. This perseverance is consistent with the behaviour of a network in the vicinity of a saddle point, which would imply that highly one-sided mice may take many more days of training before they can escape the 3R/3L saddle points. This could be further exacerbated by the "transfer-to-striatum" effect of the tutor-executor learning rule, which might be causing the learned associations of one-sided mice to become less flexible and thus further slow down their approach towards 4.

Taken together, we have demonstrated that knowing the fixed point structure of a learning problem can provide deep insights into the learning trajectories that emerge when individuals tackle it. This fixed point structure is determined by the architecture, task, and learning rule of the agent tackling the problem, and can be discovered by finding the network configurations where learning dynamics go to 0. However, finding the fixed points of more complex networks can become difficult very quickly, and thus future work can investigate how to obtain these fixed point structures efficiently (or even automatically) for different classes of common network architectures. The identification of symmetries in the architecture with respect to the learning rule [300, 205, 204] appears to be one promising avenue for finding methods to generalise this approach to more complex networks and tasks.

Chapter 5
Conclusion and future work

Chapter contents

5.1 Contributions to the study of behaviour 88
5.2 Implications for research on the neural bases of learning 90
5.3 Future perspectives for deep learning 92
5.4 Summary 94

Long-term learning defines many aspects of our life, from our career to our much-loved hobbies. However, we know little about the brain mechanisms underlying such learning. In this work, we showed that mice learning a perceptual decision-making task from naïve to expert display diverse yet systematic transitions through behavioural strategies. Dopamine signals acted as teaching signals shaping the learning trajectories. These signals reflected the intermediate strategies mice adopted from naïve to expert, encoding the stimulus-choice associations determined by each individual's strategy. Consistently, optogenetic manipulation of these dopamine signals altered each individual's behaviour, in a manner distinct from the effect of water reward. We tied these behavioural and neural results together through a deep neural network model trained using heterogenous dopamine-like teaching signals, which reproduced the empirical distribution of mouse learning trajectories as well as dopamine signals and optogenetic results. Lastly, we demonstrated that learning trajectories were qualitatively governed by saddle points and their connecting manifolds, providing a formal account for how a biological learning mechanism can steer and yield diverse yet systematic long-term learning trajectories.

5.1 Contributions to the study of behaviour

Our behavioural results prove that it is indeed possible to study long-term learning trajectories in the lab. Despite the many limitations of the artificial environments used in laboratory settings, we were able to reproduce the characteristic features of long-term learning trajectories in mice. Albeit shorter and much simpler than the long-term learning problems faced by humans in the real-world, mice learning the perceptual decision-making task we designed displayed similar learning plateaus and stage-like transitions as humans when e.g. learning a new skill. The controlled laboratory conditions removed several confounds of the real-world such as genetic diversity, differences in living conditions, and differences in education and general past life experience, allowing us to focus on core brain mechanisms involved in long-term learning. Further, the head-fixed perceptual decision-making design provided a similar, highly-stereotyped perceptual experience to each individual over trials, which greatly enhanced the quality of our neural recordings. As such, any individual differences in learning trajectories that emerged could be largely attributed to differences in each individual's experience during the task, and how this interacts with their learning mechanisms. As with all experiments, we designed ours with a specific intention, i.e. that of finding the neural bases of long-term learning at the systems level. However, our design can be adapted to tackle

other questions such as the effect of genetic diversity, differences in curricula or details of sensory parameters (e.g. multisensory integration) on long-term learning.

Our use of a wheel coupled with the movement of the visual stimulus made the action required to make a choice quite similar to the naturalistic orienting movements of mice. As such, this considerably reduced the duration of the motor learning component of long-term learning in our experiment, which instead focused on its more cognitive aspects. However, this component is present in the learning trajectories of our mice and forms an important part of long-term learning in humans. To study this component more adequately, one could either modify the effector of the task or introduce multiple steps such that motor planning is required. This highlights another important characteristic of our task, which is that the actions required for the mice to implement their decisions are ballistic and of a stimulus-response nature. The RL framework is suited to modelling behaviour with these properties, as the actions are well-defined and can be considered as a single entity to be associated with stimuli. However, the learning mechanisms at work in tasks with more sophisticated motor requirements are likely more complex and worthy of future study. Similar to our extension of the definition of states from standard RL models achieved with the deep RL model, we may have to refine our theoretical framework for describing actions to be able to understand the principles underlying learning in such settings.

One interesting observation from our data is that the 'balanced' trajectory involves transitioning through less saddle points and thus less learning stages than the one-sided trajectories (Figure 4.13A). The effect of this can be seen in Figure 4.13C for the model and Figure 2.10E for the data. Indeed, we noticed that balanced mice tended to reach higher accuracies than one-sided mice after training for a similar number of trials (Figure 2.9G). We believe this could have implications for designing curricula that enable faster learning in both animals and machines. In the case of our task, avoiding the development of early biases through shaping methods or explicit supervision could enable mice to reach expertise in less trials. As such, using the stationary point structure of a dynamical model of a learning problem could be a powerful tool to find the optimal route through the task space for rapid and efficient learning. Admittedly, we are quite far from having a concrete understanding of the learning algorithms operating in human brains during long-term learning, and hence it may be difficult to obtain accurate stationary point structures from theoretical analyses. Further, rapid learning may not always be the objective of a learning agent, who could instead be seeking to develop a rich representational structure that enables better generalisation to other problems. However, in both cases, curriculum design may benefit from empirical methods on large learning datasets as a first step to identify the diversity of learning trajectories and strategies that they transition through and thereby obtain an estimate of their underlying stationary point structure.

Lastly, on a more speculative note, the principles of long-term learning that we have identified in this thesis may generalise beyond the individual and apply to the learning dynamics of collectives as well. Attested by our use of the phrases 'the age of' or 'the era of', it is clear that societies, nations and humanity as a whole also go through stages in their development. We also observe remarkable diversity in the approaches used by different collectives to solve similar problems, and it is not unheard of that societies and nations (hopefully not humanity!) 'fail' or remain stagnant for several years if not forever. As such, the analysis methods we have developed to study the individual learning trajectories of mice could be used to better understand the dynamics of larger collectives of individuals over long timescales. Of course, we are nowhere close to knowing the algorithms governing collective learning, but given its phenomenology, it is likely that they will also possess stationary point structures with diverse orbits through their landscape. As

with the design of curricula for individuals that could improve the achievement of their learning objectives, governments could use these properties of the dynamical system they govern to design and implement policies that lead it down the route most adequate for each problem. To uncover the principles underlying collective learning, it will have to be studied in the controlled setting of the lab, as this thesis has done for long-term learning in individuals. Efforts to do so have already started [301, 302, 303], relying on model organisms that can be studied in large numbers such as flies and ants which although simple, may display dynamics that are also present in more complex organisms.

5.2 Implications for research on the neural bases of learning

A distinctive feature of our neural results is the heterogeneity of dopamine signals across the striatum. Previous studies have found that the activity of dopamine neurons in the midbrain is tuned to different sensory, motor and cognitive variables such as speed, acceleration, accuracy of pending choice etc [192]. However, it remained unclear whether this diversity is present also in the patterns of downstream dopamine release, or whether the diverse signals are added together and broadcasted as a single teaching signal across the brain [254]. Our GrabDA recordings [257] of dopamine release across dorsal striatum revealed a remarkable diversity in dopaminergic responses to stimuli and rewards over learning. Through our model, we proposed a function for such heterogenous release patterns i.e., that they each reinforce synaptic connections selectively in the cortico-striatal pathway that projects to the dopamine neurons they derive from. We described this role as that of a 'partial' RPE, which compares a prediction formed using a subset of the stimuli relevant for a task (hence 'partial', e.g. RPEs based on visual vs. auditory stimuli) and in turn updates only a subset of the cortico-striatal associations used for decision making. We subsequently provide causal evidence for this hypothesis through our optogenetic manipulations of dopamine release in the dorsolateral striatum (DLS). Although the function we propose is consistent with our results, the data we collected cannot verify whether its anatomical implementation through cortico-striatal projections and striatonigrostriatal loops is indeed responsible for the patterns of dopamine release that we observe. To do so, future studies would need to additionally record from both the cortical and striatal neurons that project to a particular dopamine subpopulation and verify that their activity over learning evolves as our model predicts. This could answer the question of how much of the heterogeneity of dopamine responses can be explained by the source of their inputs.

One of the main signatures of dopaminergic heterogeneity in our dataset was the lack of DLS dopamine responses to the non-associated stimulus. We found it striking that even when mice were highly accurate at making choices on such trials, dopamine signals in the DLS never reflected any expectation of reward that could be driving those choices. This would not be expected from a dopamine signal acting as a classic RPE that compares a prediction computed based on all reward-predictive cues with a reward. This observation led us to the concept of 'partial' RPE described above, but also made us wonder why we observed DLS dopamine responses to the associated stimulus but not the non-associated stimulus. We thought this might be because, although it could, the cortical representation of the non-associated stimulus does not reach the areas of cortex that project to the dorsolateral striatum. Another explanation, that we thought less likely, is that the non-associated stimulus is represented in a different cortico-striatal pathway and thus does not affect dopamine signals in the DLS. In our model, we captured the effect by dividing the network into a 'stimulus' pathway and a 'constant' pathway. We posited that the partial 'stimulus-based' predictions formed by the stimulus pathway were used to compute the

partial RPEs that captured DLS dopamine release, and then captured the lack of a prediction on trials with non-associated stimuli by setting the W^1 weight corresponding to the non-associated stimulus to 0. As before, to formally verify this implementation we would have to record from the cortical regions that project to DLS and demonstrate that a representation of the non-associated stimulus is indeed lacking. This could be the subject of a future experimental study, which does not necessitate recording over learning in the experiments suggested above.

Consistent with the lack of a stimulus response, DLS dopamine reward responses on correct trials with the non-associated stimulus did not decrease over learning. Once again, we found this observation very surprising given the well-known cortico-striatal plasticity-inducing effects of dopamine release in the striatum. Further, this could not be explained solely by the concept of partial RPE, as such a teaching signal should always modulate the strength of stimulus-choice associations. However, the effect was consistent with the updates expected from a deep network learning by gradient-descent based updates. As nicely illustrated by the backpropagation algorithm, the size of a gradient-descent update in any particular layer is proportional to both the backpropagated error in the output and the activation of adjacent layers. As such, the gradient-descent updates in a network such as the one we used to capture the behavioural and neural data of one-sided mice would also predict the negligible updates of associations between the non-associated stimulus and downstream brain regions. This is because setting both the weight corresponding to the non-associated stimulus in W^1,and its downstream weights in W^2 to 0 would cause the updates of weights in both layers to also be of size 0, despite a non-zero error in the output. As such, if a similar mechanism is at work in the brain, the lack of associations between early visual cortices and the cortical regions that project to DLS, together with the lack of associations between the same cortical regions and the DLS would similarly lead to no 'learning' even in the presence of a large error signal such as the reward responses of DLS dopamine. To verify this hypothesis causally, we performed the optogenetic stimulation experiment outlined in Chapter 3. We showed that stimulating DLS dopamine release indeed did not lead to associations being formed between the non-associated stimulus and choices, but did lead to changes in behaviour when the stimulation was performed after associated stimuli. This optogenetic experiment also demonstrated the stimulus-specific effect of DLS dopamine, as replacing dopamine stimulation with water rewards led to behavioural effects on both trial types. This provided important causal evidence for our proposed function of heterogenous dopamine signals across the striatum.

Implementing the 'tutor-executor' learning rule with heterogenous RPEs based on our dopaminergic results led to the unexpected effect of the progressive transfer of weights from the 'cortico-cortico' connections of W^1 to the 'cortico-striatal' connections of W^2. Previous experimental studies have observed a similar early reliance on the cortex for learning which later becomes less relevant or even unnecessary for performing the task [268]. Although our model still requires there to be non-zero representations in W^1 to perform the task, one could posit that the decrease in cortical activations that the effect gives rise to in the model could translate to the brain as a reliance on a progressively smaller proportion of cortical neurons. In the model, the effect comes about because the 'partial' losses used to train the weights in W^2 only have access to a subset of the inputs required to predict the reward on each trial, whereas the 'total' cortical loss training W^1 has access to the full prediction based on all inputs. As a consequence, even when the 'total' loss is at its minimum (or a fixed point), the 'partial' losses continue to produce error signals that update the weights in W^2 and have to be compensated for in W^1 by a corresponding decrease. As such, this effect was particularly prominent in the vicinity of saddle points. In fact, the dynamics of the

'tutor-executor' model only went to 0 at saddle points defined by fixing the product of W^1 and W^2 and then taking the limit of infinite weights in W^2. This was not the case for the single-loss model which could not produce this transfer effect in the first place. Although our current implementation predicts that this process continues *ad infinitum*, had we included noise in our model there would likely come a point where the large weights in W^2 would boost the noise excessively such that performance would be compromised and the transfer would stop. Hence it is possible that, if a similar process occurs in the brain, the transfer would also stop when performance on the task starts to be compromised. This once again opens an interesting line of experimental investigation into the mechanisms governing the transfer from cortex to striatum in the brain, and what effect this has on behaviour.

Albeit insightful, it never stopped puzzling us why the brain would have a learning rule governed by saddle points. Having saddle points interspersed in its dynamical landscape seems like a very inefficient design for a learning algorithm, which slows down in the vicinity of each of them (and frequently gives up!). However, the transfer-to-striatum effect of the 'tutor-executor' network gave us some ideas as to why evolution may have given us this present. Since the transfer effect seems to be most prominent in the vicinity of these saddle points, a possible hypothesis is that they enable us to divide learning problems into bite-size 'chunks' that we learn in stages. In this way, we can use the limited capacity of our brains to progressively learn each chunk with an early reliance on the cortex, and then transfer the learnt representations to subcortical structures such as the striatum to free the cortex for further learning. In the model, transferring the weights to W^2 causes the associations to become much less flexible, as the size of the gradient-descent updates are proportional to the cortical activations which become progressively smaller. As such, the 'transferred' representations could be similar to what we commonly refer to as *habits*, which are known to die hard. Thus, unless something causes the cortical representations to resurface and remain high despite error signals promoting their decay, the subcortical associations will remain largely fixed once the transfer is complete. The little reliance on cortical representations to perform the task after transfer is reminiscent of the common experience of performing habits 'without thinking', and it could be precisely this effect that enables us to continue learning ever more complex strategies over our lives. Deepening our knowledge of this process could once again help in the design of learning curricula that take advantage of the transfer to teach useful 'chunks' that can be effortlessly combined to perform more complex tasks. As such, future research in this direction will surely bring important discoveries.

5.3 Future perspectives for deep learning

There is an appreciable reticence to incorporating knowledge about the brain into deep learning models. This seems to emerge from a structural barrier between more mathematical or technical and the biological sciences, let alone the humanities. This is surprising given how the basis for modern connectionism was likely put forward by a physiologist for the first time [304]. Back then it was known as the neurone doctrine, and its main adversary was reticularism as opposed to symbolic AI today. Maybe as a remnant from the computing age, the brain is seen as a slower, less precise version of machines whose incorporation into technology will only hinder its progress. We hope that the work in this thesis demonstrates otherwise. Evolution has equipped the brain with multiple solutions for the difficult problems it faces every day in the real world. Although often machines are not intended to fully reproduce human capabilities, instead to focus on a few core competences and both automatise and enhance them, there are many such capabilities that we are

yet to discover, and combining our *externalisations* with previous technology may yield even more powerful autonomous systems.

One of the core contributions of this thesis to deep learning is the demonstration of using fixed point structures to understand learning dynamics. Finding such structures for the deep learning systems used in industry today could significantly improve the protocols used to train them. Inter alia, it could help understand the effect of early learning (e.g. biases) on later performance. Deep learning theory has already started to tackle the problem of finding the fixed points of modern-day systems. However, their scale and complexity makes this a difficult problem even when they are discovered, as visualising their structure and understanding their effect on the dynamics remains a challenge. As such, analysing simple problems such as the perceptual decision-making task for mice from this thesis can provide a change in perspective and help research become unstuck. Despite its simplicity, the network we analysed revealed principles that generalised to larger networks. For instance, we found that saddle points typically arise for 'incomplete' solutions of the task using a subset of the input modalities (e.g. using only one of the two stimuli) if the network has late fusion of the modalities [283]. Our work also identified that the scaling symmetry of deep networks gives rise to additional saddle points that are not present in 'shallow' networks without this symmetry. By the principle of symmetric criticality [300], there may be other symmetries in deep networks that similarly give rise to fixed points. This approach may enable the categorisation of fixed points that derive from the same underlying symmetry, thus providing some structure to the multiple fixed points that exist in industry-standard networks. A benefit of using such holistic principles is that one can also distinguish between fixed points that emerge solely from the architecture, data, initialisation, learning rule or combinations of these. Furthering this line of research promises to improve our understanding of deep learning systems, and may continue to benefit from the study of animals learning complex tasks over long timescales.

It is often the case that more than one model can account for the observations in an experimental dataset. After analysing the mice's learning trajectories and dopamine signals, we found that both the 'tutor-executor' network and a version trained with a single loss were able to capture mouse behaviour over learning. Although the teaching signal of the single-loss model differed from our recordings of DLS dopamine, the learning trajectories of its weights were similar to those of the 'tutor-executor' network and thus it could also compute partial RPEs that mimicked DLS dopamine signals. Given that we only had correlative evidence for the function of DLS dopamine as a partial RPE, we could not fully distinguish between these two models. We thus designed the optogenetic inhibition and stimulation experiments from Chapter 3 to be able to distinguish between them by providing causal evidence for the tutor-executor's proposed role of DLS dopamine. We found this use of optogenetic methods to test conflicting hypotheses from different models very compelling, and believe that such causal approaches should be used more often in the future. These results gave much more credence to our tutor-executor model. Taken together, the tutor-executor model showed that a network trained with an unconventional learning rule based on updating subsets of network weights with different losses can reproduce both the learning trajectories and fixed point structure of its equivalent trained conventionally with a single loss; all while being more faithful to the implementation of learning mechanisms in the brain. In addition, it gives rise to interesting phenomena such as the transfer-to-striatum effect which may prove beneficial for continual learning. Extending the use of multiple losses for training larger networks on more complex tasks may thus be an interesting avenue for future deep learning research.

Finally, we believe that features such as heterogenous dopaminergic RPEs and their unique

consequences highlight that it is important to consider both cortical and subcortical structures when studying learning. The deep learning systems commonly used today are deep, wide and extremely flexible, and are mostly initialised with random weights and trained with gradient descent on a single loss to update all weights in the network. As such, they appear to be a good model of the flexible parts of the cortex, which can be quickly re-configured to solve multiple tasks over an individual's lifetime. Although the cortex is one of the brain's most powerful regions (being indeed the region that most distinguishes humans from other animals), the brain has many other regions that are continuously involved in learning and computing. It hence seems naïve to think that a model based primarily on the cortex, despite its immense flexibility, will be able to solve every problem as efficiently as the brain does. Integrating knowledge from the computational principles of other brain regions may hence lead to advances in addressing the problems that have been identified with modern deep learning systems, such as their catastrophic forgetting when trained on new tasks, their dependence on extremely large datasets to learn, and their tendency to overfit. The field of robotics also stands to gain from this approach, as the brain possess a remarkable ability to coordinate the movement of thousands of muscles which relies on multiple subcortical and brainstem structures before reaching the muscles and motor effectors. In conclusion, we believe that deep learning can benefit greatly from identifying the operating principles of the brain and may change radically in doing so, leading to more capable autonomous systems that externalise and enhance the plentiful virtues of the brain.

5.4 Summary

In summary, we have identified behavioural, dopaminergic, and computational principles underlying long-term learning trajectories. A saddle point view of long-term learning and its diversity across individuals can generalise beyond our task and could be a signature of learning in the 'deep' cortico-striatal circuit that we investigated. Building on our results, future studies can investigate the dynamics of behaviour and neural activity in other dopaminergic circuits or brain regions during long-term learning. Further, studies in deep learning theory can continue to build on the saddle point view to characterise learning in more complex architectures and problems. Lastly, from a more practical perspective, our results can be used to optimise curricula for long-term learning in animals and machines. Through our work, we hope to have demonstrated that the approach of (i) collecting a dataset with novel experimental results, (ii) finding a suitable model to capture the data, (iii) analysing the model to understand its underlying principles, and finally (iv) verifying the model's predictions through optogenetics (or other controlled manipulation methods) is a recipe for collaborations between experimental and theoretical neuroscience that will bear much fruit in the future.

Appendix A
The Response Hypothesis

Chapter contents

A.1 Introduction 95
A.2 Theoretical framework: rewards as *responses* 98
A.3 Neural implementation: corticostriatal and striatonigrostriatal circuits 101
A.4 Experimental validation and predictions 103
A.5 Simulations 105
A.6 Conclusion 105

As my supervisor Prof. Armin Lak told me when I described this idea to him, all dopamine researchers go through a period where they think dopamine is not encoding RPE. This appendix contains my thoughts during that period. I thought I might as well include them in my thesis as they probably won't see the light anywhere else. The hypothesis is unfinished, so take the contents with a pinch of salt, and do let me know your thoughts after reading!

A.1 Introduction

Reinforcement learning (RL), a computational approach to goal-directed learning from interaction, has been very successful in describing a multitude of biological behaviours [43]. This framework posits that all of behaviour emerges from the maximisation of an abstract numerical 'reward' signal [305], defined to be positive for anything 'good' for the agent and negative for anything 'bad'. RL was first developed as a framework for the field of Artificial Intelligence (AI), to design agents that can navigate complex environments by learning policies (functions mapping states to actions) that maximise an objective function defined in terms of accumulated reward.

One of the simplest and most elegant algorithms designed to learn such policies is *Temporal Difference (TD)* learning [306]. Here the agent learns to assign values to different states through a *value function*, while simultaneously learning the values of different actions in each state through a *Q-function*. The value function and Q-function are tightly linked, in that the actions that lead to the states with the highest values (V's) are those with the highest action values (Q's). A TD agent updates its value/Q-functions proportionally to a quantity known as the *TD error (δ)*. This is a retrospective quantity, defined as the difference between the discounted value of the current state (plus any reward) and the value of the previous state,

$$\delta_t = R_t + \gamma V(S_t) - V(S_{t-1}). \tag{A.1}$$

The TD error is used to update the value of the previous state,

$$V(S_{t-1}) \leftarrow V(S_{t-1}) + \alpha\delta_t, \tag{A.2}$$

such that if an agent is pleasantly surprised by a state transition (i.e., has a positive TD error δ) it will increase the value of the previous state, S_{t-1}, thus learning to approach that state in the future

(and the opposite if disappointed). These equations can also be written in terms of the Q-function [43]. A TD agent is considered to have learnt when its TD error is minimal, indicating that its value function has converged and thus it has learnt the values of all states as well as possible.

In neuroscience, a common experimental setting used to study learning is known as classical (or Pavlovian) conditioning, where an animal learns to pair a stimulus with a reward [307]. These experiments tend to have a trial structure, consisting of a single stimulus followed by a single reward, with trials presented repeatedly at regular intervals to enable the learning of the association. Inter-trial intervals are usually sufficiently long that one can consider the value of the state at the time of reward (or more generally, outcome) to be 0 due to accumulating effects of the temporal discount factor γ. In these settings, the TD error at outcome time t_{out} reduces to,

$$\delta_{t_{\text{out}}} = R_{t_{\text{out}}} - V(S_{t_{\text{out}}-1}) \tag{A.3}$$

and is typically referred to as the *reward prediction error (RPE)*. Here the learnt value of the state preceding the reward $V(S_{t_{\text{out}}-1})$ is typically referred to as the *predicted reward* or, more simply, the *prediction*. When modelling these experiments, the environment can be simplified to have only two states and two relevant times: the stimulus state/time and the outcome state/time (the latter having a default value of 0). There are hence only two TD errors computed on each trial, one at the time of stimulus and equal to the *prediction*, and another at the time of outcome and equal to the *RPE*.

In the late 90's, a seminal paper by Schultz, Dayan and Montague [52] revealed a signal in the brain that correlated very well with the above-described TD-RPE errors of classical conditioning experiments. This signal was the spiking activity of dopaminergic neurons (Figure A.1A), a type of neuron living in the midbrain which communicates through the release of the neurotransmitter *dopamine* (DA) [308]. DA neurons had been previously studied in relation to pleasure and addiction [309, 310]. This led to a series of studies investigating the effects of dopamine on plasticity, to see whether it could be implementing the learning effect that TD errors have in RL agents. Dopamine was indeed found to promote synaptic plasticity [311, 188, 312, 313], especially in the cortico-striatal circuitry (Figure A.1B), one of the major efferent projections of midbrain dopamine neurons and an area thought to be involved in the learned selection of actions (or the learning of state-action Q-values, in the language of RL) [314]. These observations have now been reproduced in multiple studies, and have consolidated what is known as the *TD-RPE hypothesis of dopamine function* [315].

Recently, there have been several publications in the literature reporting that the TD-RPE hypothesis may not completely account for all of DA activity and function [316, 266, 317]. These are just the latest examples in a long line of work scrutinising the TD-RPE hypothesis [318, 319, 320, 321, 322]. Novel experimental techniques have enabled the high-resolution recording of DA from different midbrain areas that selectively project (receive projections) to (from) other brain areas [323, 257]. This has revealed a region- and projection-specific diversity of DA that had been previously ignored [324, 236, 192, 254, 325]. Single-cell gene expression analyses have also identified that DA neurons have diverse molecular phenotypes [326, 327], which show differences in activity and function [328]. In addition, several studies have revealed an important relationship between DA and action [329, 330, 317, 266], such as its role in action initiation and movement vigour [320, 243, 331]. This relationship is further supported by the loss of motor abilities associated with Parkinson's Disease [332], a condition characterised by the progressive death of midbrain DA neurons [333]. Lastly, DA has been found to reflect reward-independent properties of stimuli such as salience [334], novelty [335], generalisation [336], and information content [337].

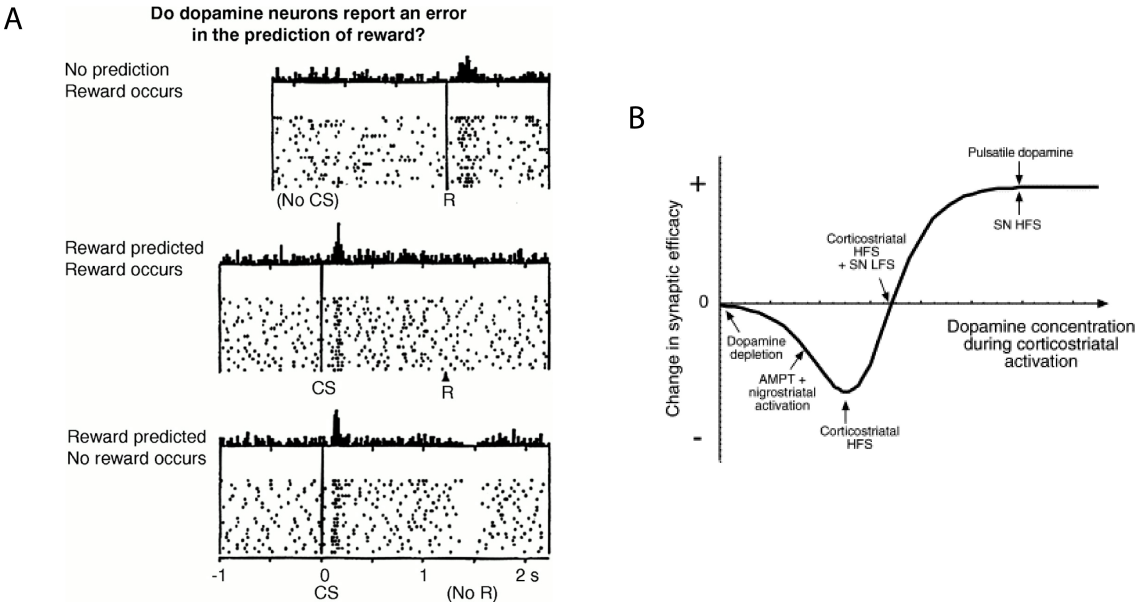


Figure A.1: Schematic of TD-RPE Hypothesis A. As the animal learns to associate the reward (R) with the conditioned stimulus (CS), the TD error shifts from outcome time to stimulus time. This dynamic can also be seen in the activity of the dopamine neurons. If the reward is omitted once the association has been learnt, a dip in the firing rate at outcome time is observed. This is similar to the negative TD error predicted by the TD-RPE hypothesis. Figure adapted from [52]. B. Function that relates dopamine release to changes in cortico-striatal synaptic efficacy. Figure adapted from [312].

Underlying this line of work are hence three major criticisms of the TD-RPE hypothesis:

1. the (original) TD-RPE hypothesis is incompatible with the diversity of DA activity, release and gene expression patterns; as it assumes a single prediction error represented by all DA neurons broadcast uniformly to all efferent projection targets.

2. the (original) TD-RPE hypothesis does not account for the motor effects of DA; instead focusing on its role in learning predictions of upcoming reward.

3. the (original) TD-RPE hypothesis does not account for reward-independent DA signals; due to its reliance on reward (positive or negative) as the only driver of DA.

In response, theoretical studies have used mainly three approaches to address these anomalies:

1. separate the TD-RPE signal into multiple prediction dimensions [254] and/or multiple reward dimensions [338] according to the DA neurons' afferent projections and molecular phenotype.

2. augment the TD-RPE hypothesis to include action-related prediction errors in addition to RPEs [339]

3. modify the TD-RPE hypothesis to include reward 'bonuses' for salience, novelty, information and generalisation [340, 341] or remove the hypothesis of reward prediction altogether [316]

However, to date, these approaches have addressed each criticism in isolation. This patchwork process has hampered the discovery of a unifying theoretical framework describing DA activity and function. We believe that such a framework exists, but at a higher level than the single functions proposed in the above studies. Instead, we propose that plasticity-promoting (reinforcing) and excitability-enhancing (motor) effects of DA release are indeed common across DA neuron

populations, but the role of each neuron cannot be understood without considering its inputs, outputs and genotype. To support and exemplify this idea, here we present an anatomically-justified role for the DA neurons living in a midbrain area known as the *substantia nigra pars compacta* (SNc) [263, 342, 343].

In this work, we propose that the main function of SNc DA neurons is to encode the difference between consecutive *responses*, defined as the strength of a reaction to an internal/external stimulus[1]. We also identify a candidate circuit mechanism for the implementation of this computation and provide evidence from the literature to validate our claim. In essence, our framework proposes that it is mainly strong *responses* which activate SNc DA neurons and promote both learning and action, if not matched in strength (or 'predicted') by an immediately-preceding *response*. Examples of *responses* could be swallowing water when thirsty or hugging an old friend (rewarding), but also spitting out food in disgust or avoiding a punch when attacked (aversive). These *responses* could have been learnt throughout an agent's lifetime or be innate and shaped through evolution to facilitate our survival. The framework, however, suggests that the behaviour of animals with a SNc DA system can be in large part described by the complex chaining of *responses* through time to approach a limited set of innate, pre-encoded *responses* useful for survival.

A.2 Theoretical framework: rewards as *responses*

Our proposed framework for describing SNc DA function boils down to a very simple re-definition of reward, enabling us to maintain the architecture and machinery of RL learning algorithms. We replace the reward signal with an outcome *response*, i.e., the strength of the reaction to an outcome state (e.g., swallow (action) the-water-in-my-mouth-when-thirsty (state)), and employ the same learning mechanisms as in RL to describe behaviour. For a TD agent, this reduces to simply replacing reward with *response* in the definition of the TD error for Q-values,

Original TD error for Q-values :

$$\delta_t = R_t + \gamma Q(S_t, A_t) - Q(S_{t-1}, A_{t-1}) \tag{A.4}$$

Proposed TD error reflected by SNc DA neurons :

$$\delta_t^{(r)} = |S_{\text{out}}| q(S_{\text{out}}, A_{\text{out}}) + \gamma |S_t| q(S_t, A_t) - |S_{t-1}| q(S_{t-1}, A_{t-1}) \tag{A.5}$$

where $|S_{\text{out}}| q(S_{\text{out}}, A_{\text{out}})$ is the *response* during the transition from time $t-1$ to t. One may note another small modification compared to the original expression, which is nothing but the extraction of the 'salience' of a state (denoted by $|S|$) from the definition of state, instead introduced as a multiplicative gain factor on the state's q-value (now a parameter directly proportional to the Q-value). We do this because,

1. it makes it easier to explain the effect of changing the magnitude of outcome *responses*, where earlier we would double the size of the reward signal, we can now double the salience of the outcome state $|S_{\text{out}}|$,

2. it removes the (unrealistic) requirement of the agent having to learn independent q-values for versions of the same state with different saliences,

[1]We use the word *response* to indicate that the actions require a certain degree of information processing to be triggered, thereby distinguishing them from reflexes which do not require complex cortical circuitry to activate and can hence be implemented in simpler neural architectures

3. it is directly mappable to the brain's corticostriatal circuitry, where the cortex represents salience-weighted states, and the q-value is reflected in the strength of the corticostriatal synapses, acting multiplicatively on the cortical representation to generate actions (as explained in Section A.3).

Our framework then uses the novel *response TD (R-TD) error*, $\delta_t^{(r)}$, as in TD-RL to update the q-value of the previous state-action pair,

$$q(S_{t-1}, A_{t-1}) \leftarrow q(S_{t-1}, A_{t-1}) + \alpha |S_{t-1}| \delta_t^{(r)}, \tag{A.6}$$

where the only difference is the multiplication of the update by the salience of the previous state, such that the update is proportional to the strength of the cortical representation. We call this learning algorithm *R-TD Learning*.

This re-definition of reward has many shared properties with the original concept of reward, but removes the reliance on states being 'good' or 'bad'[2] and instead shifts the emphasis onto the agent's *perception* and *interaction* with states. This means that the extent to which one learns to approach a state is no longer determined by a complex reward calculation, but rather by how readily an agent responds to that state, where this *response* can be either learnt or innate. As a consequence, both 'good' and 'bad' states that readily trigger actions can lead to equally positive R-TD errors, and hence increases in DA activity and approachive learning. In fact, the activation of DA neurons to noxious stimuli is something that has been observed in multiple past studies [344, 345, 346, 347].

But why would an agent learn to approach a 'bad' state only because it generates a strong *response*? We don't think that the reinforcing aspect of R-TD errors or DA activity triggered by 'bad' states is normative *per se*, but that it instead emerges as the by-product of a learning mechanism that uses the same signal both for reinforcing and immediately promoting (or *gating*) learnt behaviour, which has been observed to be the case for the DA system [348, 243, 330, 264]. The widespread neuromodulatory effects of the DA released during a *response* to a 'bad' state can temporarily heighten the activation of the circuits involved in action generation, and lead to faster and more vigorous actions to avoid the noxious state [349, 331]. However, if an agent has particularly large DA activity bursts (or R-TD errors) when responding to 'bad' states, or encounters such states more regularly than 'good' states, agents can overly reinforce and become addicted to 'bad' activities maladaptively, a phenomenon that is not uncommon in humans [350].

If it can pose such a large risk to survival, why would evolution have favoured the SNc DA system? We have identified four potential reasons,

1. the combination of reinforcing and gating functions in a single signal can be useful to reduce the amount that synaptic strengths have to be modulated to learn new behaviours, instead boosting the effective strength of existing synapses through temporary dynamical effects. Thus, if synaptic strength modulation is more energetically expensive than the temporary boosting of neural activity, the SNc DA system could have emerged from the constrained optimisation of survival with limited energy stores. **(minimum norm)**

2. the combination of reinforcing and gating functions also favours the activation of the most strongly-encoded *responses* when encountering both new and old states, since the gating

[2] the definition of 'good' and 'bad' is conditional on an underlying morality, which can be defined in many ways, this being one of the reasons why the reward hypothesis of RL can at times seem tautological when used to describe behaviour. When we use the words 'good' and 'bad' in this paper, we restrict their definition to states that immediately increase or decrease the chance of survival.

function boosts the effect of any differences in the strengths of active synapses. This will encourage the use of existing circuitry to learn new behaviours and thus again favour minimal learning, in this case with regards to the formation of new synapses or associations (c.f. strengths in point 1). Hence, the SNc DA system could have been favoured by evolution when optimising for survival under the constraint of limited energy and synaptic connections or *capacity*. **(minimal learning)** [3]

3. if an agent is born or initialised with a set of innate *responses* where most of them are triggered by 'good' states, the probability of becoming trapped in an addiction to maladaptive behaviour is low, as the agent will preferentially seek and encounter more 'good' states than 'bad' ones. However, the further from birth or initialisation, the more 'bad' states will have been encountered by chance, the more (and stronger) *responses* will have been learnt to avoid their noxious effects, and the higher the probability of strongly reinforcing behaviours that lead to these 'bad' states; thus creating a maladaptive loop. This is why it is important to live in environments where 'bad' states occur with low probabilities. On the other hand, if *responses* to 'good' states are more strongly reinforced through life, the relative magnitude of the DA activity after a successful escape from a (current or imminent) 'bad' state to a 'good' state will favour the avoidance behaviour that enabled the escape over the approach behaviour that led to the 'bad' state, thus decreasing the probability of encountering it again in the future. Hence, the SNc DA system could itself be implementing the evolutionary pressure of optimising for survival by favouring agents initialised with the right balance (number and strength) of *responses* to 'good' and 'bad' states, **(initialisation)**

4. Lastly, R-TD error is likely not the only variable that SNc DA neurons encode. Midbrain DA neurons receive projections from multiple brain areas [185], some of which may indeed be encoding reward/value information (e.g. the habenula, the ventral pallidum, the rostromedial tegmental nucleus etc. [352]) indicating which states are 'good' or 'bad' for the agent. These could reduce the probability of reinforcing maladaptive behaviours by promoting DA activity for 'good' over 'bad' states. However, with dorsal striatal action circuits being the main input to SNc DA neurons [185], R-TD error could dominate as the quantity that they most represent (see Section A.3). **(other inputs)**

An interesting feature of R-TD learning is that it leads to a general tendency to repeat past actions, with only a small dependence on whether the states that triggered them were 'good', 'less good', or 'bad'. This is a common trait of animal behaviour [353], which in extreme cases manifests itself as stereotypical action patterns that are popularly referred to as 'habits' [354, 355]. This feature of R-TD learning can also explain why biological agents can take a long time to change or forget their learnt behaviours, as any lingering *responses* after changes in the environment will reduce the magnitude of the dip in DA (or negative R-TD error) used to forget them, and thus cause the agent to take longer to 'unlearn'. Past studies using RL to model biological behaviour have suggested several descriptive modifications to reward-based RL update rules to compensate for their fast-switching tendencies [356] and account for this value-independent drive to repeat [357, 339], whereas this falls naturally out of our framework.

[3] Note that the use of gradient descent when training artificial neural networks also favours solutions with minimal learning [351]. This similarity with the DA system could potentially explain why gradient descent has proved to be one of the most successful learning algorithms in machine learning research. It also implies that it could be a good way of modelling learning in biological agents.

The question of whether a SNc DA-like system implementing R-TD learning should be used when designing artificial agents is less clear. The stickiness of behaviour that it can lead to and the possibility of entering loops of self-defeating behaviour could lead to very inflexible and unstable agents. We believe the only case where it could be useful is if modulatory effects of R-TD error on action generation were to be implemented in the agent's action-selection policies, as with DA in the brain. Otherwise, incorporating solely the reinforcing effect of R-TD error would increase the chance of instabilities in training, without the benefit of learning trajectories with minimum learning and minimum norm forces. However, in scenarios where computational power or memory constraints are an issue, it could be worth experimenting with R-TD learning to see whether it leads to more efficient solutions than other training algorithms. Another possible use case would be in the context of evolutionary algorithms. Running several agents with a SNc DA-like learning system could be a principled way to find suitable initialisations for artificial agents. One can mimic evolutionary processes by selecting only the initialisations that perform best in a starting parallel run, and then exploring small deviations or combinations of their parameters (as in animal reproduction) and iterating the selection process multiple times. This could work better than exhaustive hyperparameter grid searches, since it can exacerbate the differences between initialisations through the above-described maladaptive run-off mechanisms, which increase in probability the longer an agent is simulated for.

Our key contribution with the R-TD framework is a proposal that most of SNc DA activity and function can be described by re-defining rewards as *responses* in the conventional updates of TD-RL, and that this can still lead to the efficient learning of useful behaviours (not yet demonstrated through simulation in Section A.5). In Section A.3, we describe how the corticostriatal and striatonigrostriatal circuits could be implementing the computations involved in this framework through known functional circuitry to trigger the reinforcing and gating functions of DA and enable learning through interaction.

A.3 Neural implementation: corticostriatal and striatonigrostriatal circuits

We propose a mapping of the R-TD framework to the brain's neurophysiology through the circuitry connecting the cortex, striatum and *substantia nigra pars compacta* (SNc), shown in Figure A.2. In this implementation, cortical neurons represent the salience of the current state $|S|_t$ and project to the striatum through cortico-striatal synapses with strength $q(S_t, A_t)$. Importantly, these cortico-striatal projections are segregated (as depicted by the colours in Figure A.2), leading to different cortical state representations being processed by different striatal neurons. Seminal anatomical research has also established the presence of closed loops and ascending (DLS-DMS)/descending (DMS-DLS) spirals of connectivity between dorsal striatum and SNc midbrain dopamine neurons (i.e., striatonigrostriatal loops) [282]. These loops/spirals form the largest proportion of input to the SNc DA neurons [185]. Interestingly, however, only the closed loops have been found to be functional [260]. As such, in our proposed mapping we put forward that these closed striatonigrostriatal loops perform the 'difference' (or time derivative $\frac{\partial}{\partial t}$) operation of the R-TD framework, the outcome of which is the R-TD error $\delta_t^{(r)}$ represented in the dopamine (DA) neurons of the SNc. This connectivity implies that DA neural activity should be diverse and dependent on the information present in the striatal neurons that project to them. The information in striatum in turn depends on the projections that it receives from cortex [186].

The precise implementation of the time-derivative operation through the striatonigrostriatal loops is more speculative (Figure A.3). Most of the intrastriatal circuitry is inhibitory [358],

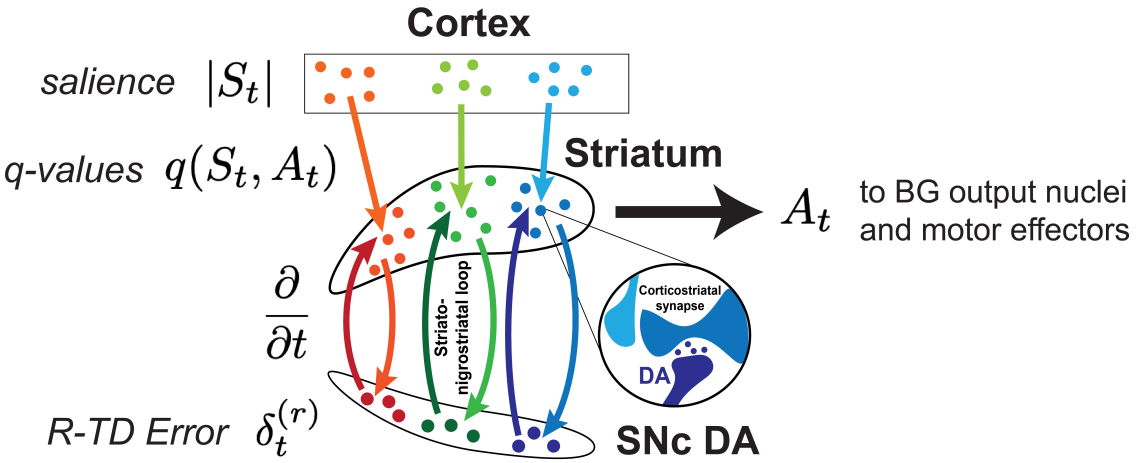


Figure A.2: Schematic of brain anatomy relevant to the response hypothesis. Cortical neurons represent the salience of the current state $|S|_t$. Cortico-striatal connectivity represents the learned q-values $q(S_t, A_t)$, which are in turn updated by dopaminergic signals from the SNc that represent the R-TD error $\delta_t^{(r)}$, i.e. a difference operation $\frac{\partial}{\partial t}$ between striatal activity over time. Colours indicate the segregated patterns of cortico-striatal connectivity that transfer to the striatonigrostriatal loops. The striatum projects to the output nuclei of the basal ganglia (BG) and downstream motor effectors to implement actions A_t.

and the closed-loop connectivity between striatal neurons and SNc DA neurons is known to pass through inhibitory GABA neurons in the *substantia nigra pars reticulata* (SNr) [322]. Using this knowledge, we designed a firing rate model based on leaky integrate-and-fire neurons [359] which could implement the computations of the R-TD hypothesis. We present a block diagram of the resulting dynamical system along with its anatomical mapping in Figure A.3, and a simplified version that collapses the feedback loop in Figure A.4. We termed it the 'integrator-derivative' model of striatonigrostriatal connectivity as the activity of the ith striatal neuron $r_s^{(i)}(t)$ is effectively an integral of the difference between its cortical input $r_c^{(i)}(t)$ weighted by the q-values and the activity of the $i-1$th striatal neuron $r_s^{(i-1)}(t)$. This should lead to the ith striatal neuron tracking the activity of its weighted cortical input if the activity of neighbouring striatal neurons is similar. In turn, SNc dopamine neurons track the activity of the striatal neurons that project to them (with some delay) and thus should reflect the moment-by-moment difference in weighted cortical inputs in their phasic activity,

$$r_{\mathrm{DA}}^{(i)}(t) = \frac{-1}{s+1}\frac{-1}{s+1} r_s^{(i)}(t) = \frac{1}{(s+1)^3 - 1}\left(q(S_t, A_t) r_c^{(i)}(t) - r_s^{(i-1)}(t)\right). \tag{A.7}$$

A simple way to think about this system is that, at each time point, any particular striatal neuron is receiving input from its cortical afferents and (inhibitory) neighbouring striatal neurons which determines its activity. If we assume that the activity of neighbouring striatal neurons is not too distinct, this will result in any particular striatal neuron reflecting its cortical input over time. The model then proposes that dopamine neurons reflect a delayed version of that striatal activity, causing the rapid phasic (c.f. tonic) changes in dopamine to represent a computation very similar to the R-TD error.

Taken together, our anatomically-relevant integrator-derivative model proposes that the cortex represents the salience of the different inputs to striatum, which are then weighted by the cortico-striatal synapses to represent action values in the striatum generating actions through the basal ganglia. The striatonigrostriatal loops compute the difference between consecutive striatal activations thus causing SNc DA activity to represent the R-TD error. This DA activity in turn

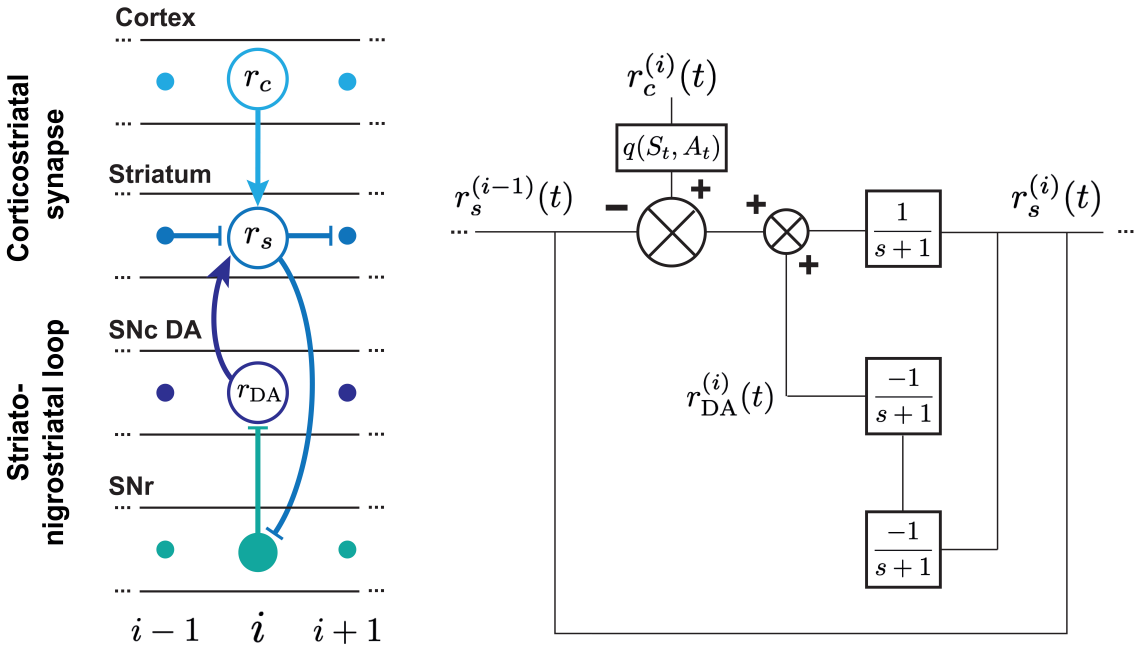


Figure A.3: Integrator-derivative model of striatonigrostriatal connectivity Left, cortico-striatal and striatonigrostriatal connectivity in three parallel channels $i-1$, i and $i+1$. Arrows with pointed heads indicate excitatory projections and those with flat heads indicate inhibitory projections. Right, complete block diagram of the 'integrator-derivative' model of striatonigrostriatal connectivity. The transfer function of striatal and dopamine neurons comes from a leaky integrate-and-fire model of neural activity [359].

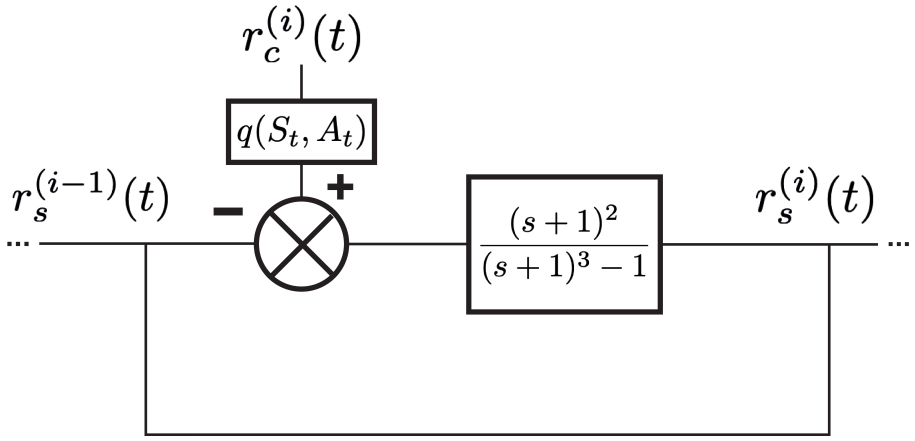


Figure A.4: Simplified block diagram of integrator-derivative model A simplified version of the block diagram for the 'integrator-derivative' model where the feedback loop implemented by the striatonigrostriatal connectivity has been collapsed.

promotes both faster/more vigorous action and reinforces the active cortico-striatal synapses (proportionally to their activation).

A.4 Experimental validation and predictions

This section is incomplete, but would like to focus on the predictions from the R-TD framework that can be validated through experimental studies. Since we believe it can explain everything that the TD-RPE hypothesis has already explained, here we focus on the observations that TD-RPE could not explain without considerable modifications. We list a few,

1. dips in dopamine levels for reward omissions being typically smaller than the rises upon unpredicted rewards
2. dopamine novelty responses [360, 335]
3. dopamine responses to noxious stimuli [347]
4. dopamine in working memory
5. dopamine in spontaneous behaviour [266]
6. dopamine reflecting motivation
7. different dopamine neurons responding to different 'rewards' (e.g., food vs. liquid)

With regards to the dopaminergic results presented in this thesis, the R-TD hypothesis could explain why dopamine release in the dorsomedial striatum (DMS) does not respond to rewards. This could be because DMS DA is triggered by changes in the activity of striatal neurons in the DMS, which are driven mostly by contralateral visual inputs, and as such would not change their activity in response to the water reward we provide the mice. Instead, they would reflect changes in the visual stimuli, such as the presentation or movement of the stimulus towards/away from the centre of the screen – something that we indeed observed in our data. As such, the DMS could be involved in generating contralateral orienting *reponses* to visual stimuli, similar to the contralateral bias that we observed when stimulating DMS DA optogenetically in Figure 3.14D-E. The R-TD hypothesis can also explain why we did not observe dorsolateral striatum (DLS) DA pre-outcome responses on 0-contrast trials or trials with the non-associated stimulus. This would be explained similarly to the 'tutor-executor' network through its proposed segregation of cortico-striatal connectivity into multiple pathways for different cortical representations (Figure A.5). In conclusion, the R-TD hypothesis can explain various features of dopaminergic heterogeneity that have required modifications to the original TD-RPE hypothesis to capture, but to formally verify its predictions further experiments are required.

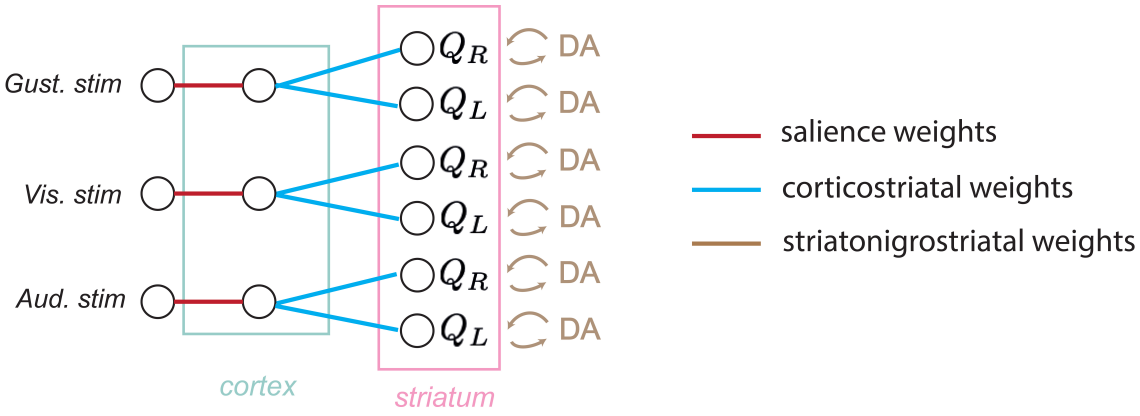


Figure A.5: Neural network model of brain anatomy relevant to the response hypothesis. Cortical representations of inputs from different modalities are weighted by segregated 'cortico-cortico' weights (red) and then associated with actions through the 'cortico-striatal' connectivity (blue) to similarly segregated striatal neurons. Dopamine neurons in the midbrain (brown) thus represent heterogenous R-TD errors corresponding selectively to the striatal neurons they receive inputs from.

A.5 Simulations

This section is also incomplete, but would like to simulate an R-TD agent in a simple environment (e.g., just left/right actions, with possible bifurcations at some points where you can choose up/down) and show that the hypotheses above are satisfied. The environment could be designed with two outcomes for actions, one 'good' and one 'bad', to assess the reward consumption of the agent over time. To implement the 'gating' effect of dopamine on the striatal activity representing action values, one could multiply the action values by the model's DA signals and this should act like a varying inverse temperature in the decision function, locking in learnt behaviour. Potential experiments to be performed with the R-TD agent are:

1. compare the relative size of weights and their learning trajectories with and without the DA gain modulation to verify the minimum norm hypothesis
2. evaluate the interactions between the fixed decision function inverse temperature and the effect of DA gain modulation on actions and learning
3. try varying the difference in size of outcome responses for the left and right actions and investigate resulting learning trajectories
4. train both a TD and R-TD agent on a task and then reverse its contingencies, comparing their respective adaptive performance
5. add more positive/negative outcomes (using up/down actions) and see what behaviours are reinforced

A.6 Conclusion

Although the *response* hypothesis has the potential to explain many experimental observations of the SNc DA system, it is likely not the sole variable governing DA function, as DA neurons receive projections from a multitude of other areas [352]. Some of these could be indeed calculating how 'good' or 'rewarding' a state is in order to preferentially reinforce behaviours that lead to 'good' states instead of 'bad' states. However, with the dorsal striatum being one of the major inputs to SNc DA neurons [185] we believe that this could be an important, if not the dominant, computation driving the activity of SNc DA neurons and thus governing learning in animals with a SNc DA system.

Appendix B
Alternative Models

Chapter contents

B.1 Introduction 106
B.2 Flexible Automated Toolbox for Trial-by-Trial Model Fitting 107
B.3 Minimal Deep Linear Recurrent Neural Network Model 108
B.3.1 Descriptive Fits of DLRNN Parameters 109
B.3.2 Goal-directed RL and Habitual Learning Rules for the DLRNN 110
B.4 Discussion and Conclusion 112

In this appendix I provide details of other approaches that I considered throughout my DPhil to explain the mouse data. These were implemented through a computational framework that can infer and evaluate the trial-by-trial evolution of decision-making model parameters. Here I applied it to the mouse data with a linear RNN that implements a drift diffusion model. First, using its inference capabilities, I obtained parameter trajectories that revealed the 3 main classes of learning trajectories. Second, I used the framework to evaluate a series of learning rules combining goal-directed and habitual mechanisms. Early in my investigations, I was keen on a model using habit values [357] to explain the perseverative behaviour of mice. However, given that I did not find a correspondence between habit value updating and our dopamine recordings (as this would predict most dopamine release to be action-locked [253]), I moved away from this modelling approach. This eventually led me to deep neural network modelling, which revealed that such perseverative behaviour can have other explanations such as saddle points in the learning dynamics.

B.1 Introduction

We present a computational framework implemented as a software toolbox capable of inferring trial-by-trial learning trajectories for the parameters of reasonably complex observation models. It builds on the work of [361], but leverages the automatic differentiation, just-in-time (JIT) compilation, and GPU acceleration capabilities of the Python library JAX [362] to flexibly perform efficient inference on different observation models that can be specified by the user. The framework additionally enables the guided design and evaluation of learning rules for the observation model parameters, thus allowing the principled testing of normative learning hypotheses.

We showcase the flexibility of the framework by applying it to data from our experiment. We capture evolving choice and reaction time (RT) statistics through trial-by-trial fits of a minimal Deep Linear Recurrent Neural Network (DLRNN) model of the sensory networks involved in the perceptual task; fit an architectural parameter (the 'depth' of the sensory network); and fit a rich learning rule for this DLRNN model. Our results show that the framework can capture the individual variability of the mice's learning trajectories, as well as more general features of the learning process observed across mice. We also find that the best-fitting model of those we consider is a 'deep' network architecture with reinforcement and habitual learning rules. More broadly, our overall approach takes a step towards linking modern theories of learning to large-scale data being

generated by systems neuroscience, offering a potential route to testing theories of learning in the brain and mind.

B.2 Flexible Automated Toolbox for Trial-by-Trial Model Fitting

To fit trial-by-trial performance over the full course of learning requires efficient inference procedures that can flexibly adapt to different observation models and learning rules. Inspired by and building on the PsyTrack package [361, 207], we provide a generic implementation of their inference method that efficiently fits trial-level parameters and the hyperparameters of different models that can be specified by the user. These inference procedures require first and second order derivatives (gradients and Hessians) of the observation model w.r.t. its parameters. Deriving these by hand is time consuming and, for complex models, technically challenging, presenting a major barrier to fitting a large variety of models. Here we use the automatic differentiation library JAX to automatically compute gradients and Hessians, thereby easing and speeding model creation and comparison. Further, since JAX supports JIT compilation and GPU acceleration, the inference can run on thousands of trials in just a few minutes.

Let $\mathbf{w}[n]$ be a vector containing the parameters at trial n (e.g., neural network weights, drift-diffusion model drift rates etc.) of a trial-level behavioral model designed such that the parameters change on long timescales to capture learning. Acting as a prior to constrain the fits, we make the assumption that parameters at adjacent timepoints are near each other, implemented by modeling transitions as a Gaussian walk,

$$p(\mathbf{w}[n+1] \mid \mathbf{w}[n]) = \mathcal{N}(\mathbf{w}[n+1]; \mathbf{w}[n], \Sigma), \tag{B.1}$$

where the diagonal matrix Σ encodes the variance for each parameter in $\mathbf{w}$. Note the updates $\Delta\mathbf{w}[n]$ do not have to be centered on 0, and could be centred on any function of the parameters at trial n [207]. The variances in Σ can be combined with other algorithm-/subject-level hyperparameters (e.g., learning rates, internal noise variance, etc.) and stored in a vector θ. We can then write the (unnormalized) posterior probability of observing the data in $\mathcal{D}$ (e.g., choices, reaction times, etc.) with a certain parameter trajectory $\mathbf{w} = \{\mathbf{w}[1], \mathbf{w}[2], \cdots, \mathbf{w}[N]\}$, given a certain hyperparameter setting θ as,

$$p(\mathcal{D}, \mathbf{w} \mid \theta) = p(\mathcal{D} \mid \mathbf{w}, \theta)p(\mathbf{w} \mid \theta) = p(\mathcal{D} \mid \mathbf{w})p(\mathbf{w} \mid \theta) = \prod_{n=0}^{N-1} p(\mathcal{D}[n+1] \mid \mathbf{w}[n+1])p(\mathbf{w}[n+1] \mid \mathbf{w}[n]), \tag{B.2}$$

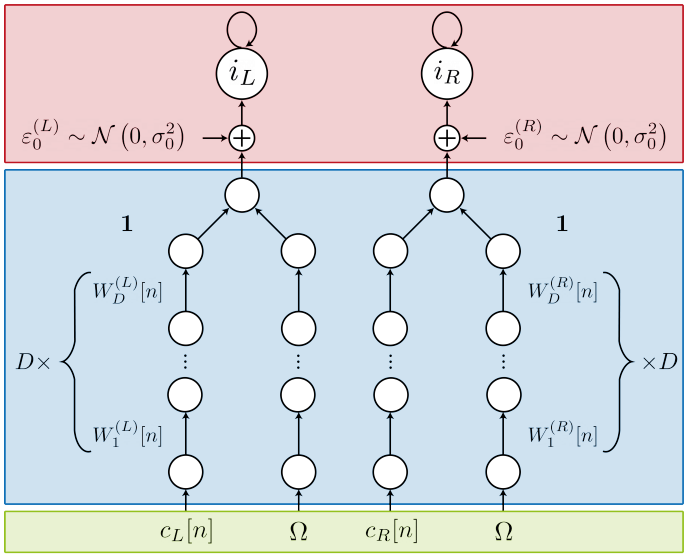
with $p(\mathbf{w}[1] \mid \mathbf{w}[0]) = p(\mathbf{w}[1]) = \mathcal{N}(\mathbf{w}[1]; \mathbf{0}, \Sigma_{\text{init}})$, where the variances in Σ_{init} are set to arbitrary large values.

We train the model using an algorithm structured around empirical Bayes [361] to find the optimal parameters $\mathbf{w}^*$, hyperparameters θ^*, and model evidence $p^*(\mathcal{D} \mid \theta)$. This is an iterative two-step algorithm involving 1. finding an estimate of $\mathbf{w}_{\text{MAP}}$ given a certain θ through gradient descent on the negative (unnormalized) posterior, and 2. optimizing the evidence $p(\mathcal{D} \mid \theta)$ w.r.t. θ, where the evidence is approximated using a decoupled Laplace approximation [363] of the (true) posterior and Bayes' rule, such that $p(\mathcal{D} \mid \theta) \approx \frac{\text{likelihood} \times \text{prior}}{\text{approximate posterior}} = \frac{p(\mathcal{D}|\mathbf{w},\theta)p(\mathbf{w}|\theta)}{\tilde{p}(\mathbf{w}|\mathcal{D},\theta)}$. The second step requires computing the Hessian of the (unnormalized) posterior to obtain its second-order Taylor expansion for the decoupled Laplace approximation. The gradient computations involved in both steps are greatly simplified by JAX, which performs them automatically and efficiently. We note that a naïve application of JAX results in inefficient code with large memory requirements. Our framework exploits the sparsity pattern in derivatives implied by the factorized structure of Eq. B.2 to remain efficient.

B.3 Minimal Deep Linear Recurrent Neural Network Model

With the software toolbox in hand, we next sought to develop observation models and learning rules that reflect 'reduced' or simplified versions of common network architectures and theories of learning, and which could be suitable for modeling our learning dataset. We first describe our network architecture, before turning to learning rules. Inspired by deep RL methods and prior work on race and drift-diffusion models of two-alternative perceptual decision-making, we consider the class of network architectures depicted in Fig. B.1. Here, two linear feed-forward processing streams (blue) convey monocular sensory input from each eye (green) to recurrent neurons that integrate inputs through time (red). These left and right processing streams race each other in integrating noisy evidence for each alternative, and the first to cross a threshold determines the response (choice and RT).

On each time step (separated by time dt), each processing stream $K \in \{L, R\}$ receives visual inputs c_K and an additional bias input Ω. Here, c_L and c_R are the raw contrast values of stimuli on each side of the screen (0, 0.25, or 0.5). These inputs are then individually amplified by a linear chain of D weights, resulting in an effective multiplication by a scalar perceptual weight w_K (for the c's) and bias weight b_K (for the Ω's). The outputs of the chains are then summed and fed to noisy integrating recurrent neurons (i_L and i_R). The first integrator to reach a threshold z determines the response of the agent on each trial. These parameters have a natural neural interpretation: c_K represents the sensory signal which may or may not be amplified by sensory circuits between the periphery and integrator, as parameterized by w_K. Ω represents a possible input to the processing streams that is unrelated to the visual stimuli (e.g., click at the beginning of each trial) and the bias weights b_K capture any processing of that input which influences the response (useful to capture behavior on 0-contrast trials). In addition, as the processing can be noisy, we include a noise source ε_0 that adds Gaussian noise to the integrator inputs at each timestep.



$$\text{s.t. } W_1^{(K)}[n] W_2^{(K)}[n] \cdots W_D^{(K)}[n] = \begin{bmatrix} w_K[n] & 0 \\ 0 & b_K[n] \end{bmatrix}$$

Figure B.1: Deep Linear Recurrent Neural Network (DLRNN) Architecture.

While simple, this model architecture can capture both choice and RTs, and retains aspects of more complex architectures such as deep layered structure and recurrence. Due to its linearity, depth does not change what the model architecture can represent, but can dramatically change learning trajectories as shown in [37] and as we show in Section B.3.2. Crucially, the exact likelihood of choice and RTs remains tractable under the model, as they depend solely on the network integrator dynamics. Linearity allows us to express the dynamics as the stochastic differential equation,

$$di_K = (w_K c_K + b_K \Omega)dt + \sqrt{v_K} dW_K, \quad K \in \{L, R\} \tag{B.3}$$

where i_K is the activation of the recurrent neurons, w_K is the effective perceptual weight, b_K is the effective bias weight, and v_k is the noise variance, with dW_K denoting a Wiener process for stream K. Each integrator evolves independently until hitting the threshold z. We set the noise to $v_R = v_L = \sigma_o^2 = 1$, as it determines the scale of the other parameters. The DLRNN is thus comparable to a race model with drift $a_K = w_K c_K + b_K \Omega$, variance $v_K = 1$, and threshold z. Defining t_K to be the first hitting time for integrator K, we can express the likelihood for choice C and RT T as the probability of the chosen-side integrator hitting the threshold at time T and the not-chosen ($\bar{C}$) integrator hitting later,

$$p(\text{choice} = C, \text{RT} = T) = p(t_C = T \text{ and } t_{\bar{C}} > T) = p(t_C = T)\left[1 - p(t_{\bar{C}} < T)\right], \tag{B.4}$$

with the separation of the joint dist. possible due to the independence of the integrators. The pdf and CDF of the first hitting times are those of an integrator undergoing Brownian motion with drift $a_K = w_K c_K + b_K \Omega$ and variance $v_K = 1$,

$$p(t_K = T) = \frac{z}{\sqrt{2\pi v_K T^3}} \exp\left[-\frac{(z - a_K T)^2}{2 v_K T}\right], \quad p(t_K < T) = \Phi\left(\frac{a_K T - z}{\sqrt{v_K T}}\right) + \exp\left[\frac{2 z a_K}{v_K}\right] \Phi\left(\frac{-a_K T - z}{\sqrt{v_K T}}\right). \tag{B.5}$$

To fit this DLRNN to a collection of N pairs of choices and RTs, $\mathcal{D} = \{(K_1, T_1), \cdots, (K_N, T_N)\}$, by independence, we calculate the probability of observing this sequence under the model as the product of the trial-level likelihoods in Eq. B.4.

B.3.1 Descriptive Fits of DLRNN Parameters

We applied the software toolbox from Section B.2 to fit this model trial-by-trial to the data from our experiment. Before introducing any learning prescriptions, we used the descriptive prior from Eq. B.1 to obtain smooth parameter trajectories over learning. Without update rules or external learning signals (e.g. reward prediction error), we only fit the effective perceptual weights w_K, bias weights b_K, and threshold z. The Ω parameter is absorbed by the fits of b_K, and is thus set to $\Omega = 1$. The resulting descriptive fits reveal interesting dynamics in the learning trajectory of the mice (Fig. B.2). Crucially, they show how each mouse's learning process is highly individualized. Generally, perceptual weights w_L/w_R start small and increase over learning. However, several mice develop strong unilateral perceptual preferences, where one of the weights grows quicker than the other. Bias weights b_L/b_R begin non-zero and oscillate more rapidly than the perceptual weights over learning, sometimes with $b_L > b_R$ and other times with $b_L < b_R$, capturing the mice's visual-stimulus-independent side biases. The RT's decrease much before the accuracy begins to increase, indicating that mice learn to act quickly before they have fully learnt the task. Lastly, we see that the threshold z remains mostly constant through learning. These observations allowed us to design tailored learning rules for the DLRNN in Section B.3.2.

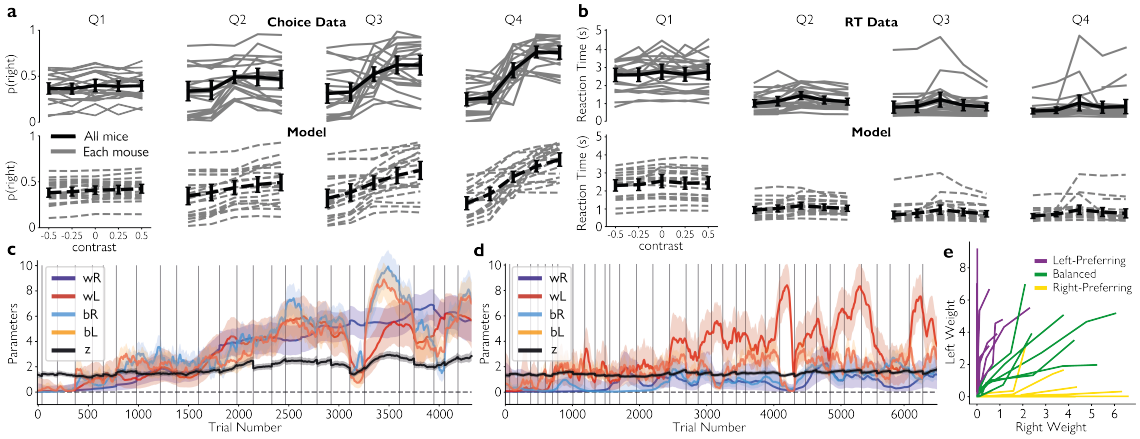
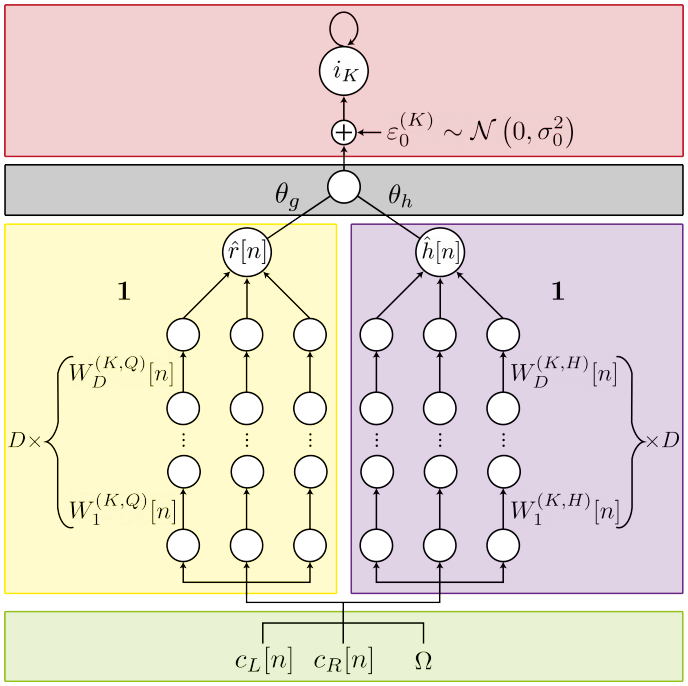


Figure B.2: Descriptive fits of DLRNN parameters. **a**, **b**: psychometric and RT curves, as a function of stimulus contrast, evolving over 4 learning quartiles for each mouse (gray) and averaged across all mice (black). **c**, **d**: trial-level evolution of DLRNN parameters over learning for a balanced mouse (c) and a left-preferring mouse (d). Vertical black lines are session boundaries. **e**: quartile-level evolution of left vs. right perceptual weights for each mouse, categorized by weight (im)balance.

B.3.2 Goal-directed RL and Habitual Learning Rules for the DLRNN

While the descriptive fit from the preceding section shows how general trends in behavior might be captured by changes in the parameters of the DLRNN, it leaves open the question of what learning rule might generate these parameter trajectories. Ultimately, theoretical progress depends on adjudicating between different learning rules. In this section we introduce several learning rules for the DLRNN architecture and fit them to trial-level data. Many learning rules could in principle operate in parallel in the brain, but here we focus on two: a goal-directed Q-learning rule that learns to predict rewards for specific stimuli and actions; and a habitual rule that learns to predict likely actions for specific stimuli. This is motivated by the repetitive unilateral behavioral patterns revealed by the descriptive fits. We formulate both learning rules as gradient descent on the feedforward pathways of the DLRNN architecture, but with differing target outputs.



$$\text{s.t. } \mathbf{Q}_n^{(K)} = W_1^{(K,Q)}[n]W_2^{(K,Q)}[n]\cdots W_D^{(K,Q)}[n] = \begin{bmatrix} Q_n(SL,K) & 0 & 0 \\ 0 & Q_n(SR,K) & 0 \\ 0 & 0 & Q_n(K) \end{bmatrix}$$

and

$$\text{s.t. } \mathbf{H}_n^{(K)} = W_1^{(K,H)}[n]W_2^{(K,H)}[n]\cdots W_D^{(K,H)}[n] = \begin{bmatrix} H_n(SL,K) & 0 & 0 \\ 0 & H_n(SR,K) & 0 \\ 0 & 0 & H_n(K) \end{bmatrix}$$

Figure B.3: RL-DLRNN model architecture.

With the input to the network defined similarly to before as a stimulus vector $\mathbf{s}_n = [c_L[n]\ c_R[n]\ \Omega]^\top$ (green in Fig. B.3), we define $\mathbf{Q}_n^{(K)}$ and $\mathbf{H}_n^{(K)}$ matrices for each action as the total weight of the

goal-directed (yellow) and habitual (purple) feedforward pathways respectively (see bottom of Fig. B.3). The pathways are constrained to be diagonal to maintain the independent processing of the monocular sensory input and the visual-circuit-independence of the bias input processing. The elements of the $\mathbf{Q}_n$ and $\mathbf{H}_n$ matrices can be interpreted as the goal-directed and habitual contributions to the drive for each action in response to the different stimuli (e.g., $Q_n(SL, K)$ captures the goal-directed contribution to action K in response to a left stimulus on trial n). The output of the goal-directed pathways on trial n is a prediction of the reward for each action $\hat{r}^{(K)}[n]$, whereas the habitual pathways output an estimate $\hat{h}^{(K)}[n]$ of the chosen action C_n (i.e., predict 1 if $C_n = K$ and 0 if $C_n \neq K$). The outputs are scaled by arbitrating constants, θ_g and θ_h (gray), to determine the integrator drift (red) with noise $\sigma_0^2 = 1$ as before.

Stochastic gradient descent on the squared error for the chosen-side goal-directed pathway $\|r_n - \hat{r}^{(C_n)}[n]^2\|$ and on both the chosen and non-chosen habitual pathways $\|\mathbb{1}\{K = C_n\} - \hat{h}^{(K)}[n]^2\|$ yields the trial-wise updates for $\mathbf{Q}$ and $\mathbf{H}$:

$$\Delta\mathbf{Q}_n^{(C_n)} = \alpha^{(Q)} D(r_n - \mathbf{1}\mathbf{Q}_n^{(C_n)}\mathbf{s}_n)\,\mathrm{diag}(c_L[n], c_R[n], \Omega)\mathbf{Q}_n^{(C_n)^{2-\frac{2}{D}}}, \tag{B.6}$$

$$\Delta\mathbf{H}_n^{(K)} = \alpha^{(H)} D(\mathbb{1}\{K = C_n\} - \mathbf{1}\mathbf{H}_n^{(K)}\mathbf{s}_n)\,\mathrm{diag}(c_L[n], c_R[n], \Omega)\mathbf{H}_n^{(K)^{2-\frac{2}{D}}}, \tag{B.7}$$

α denoting learning rate. In this way, only the chosen-side Q-values are updated (as only they are involved in the reward prediction for the chosen action) whereas both sides' H-values are updated (chosen increased, not-chosen decreased).

We hence have a total of 7 fixed parameters to fit in the full learning model: $\alpha^{(Q)}$, $\alpha^{(H)}$, θ_g, θ_h, Ω, z, and D. We initialized the $\mathbf{Q}_0$ and $\mathbf{H}_0$ matrices to arbitrary small values and fit these fixed parameters by minimizing the negative log-likelihood of the DLRNN model (Eq. B.4) for the sequence of choice and RT pairs of each mouse. To ensure that all the parameters were necessary to explain the data, we performed a nested model comparison (Fig. B.4) where we selectively set different combinations of the parameters to 0. We also compared the 'deep' $D = 2$ update to the more common $D = 1$ update to investigate the benefits of depth, and showed that the deep architecture was better at capturing the evolving behavior. Fitting the depth parameter, we found that it only provides marginal benefits over the update with $D = 2$.

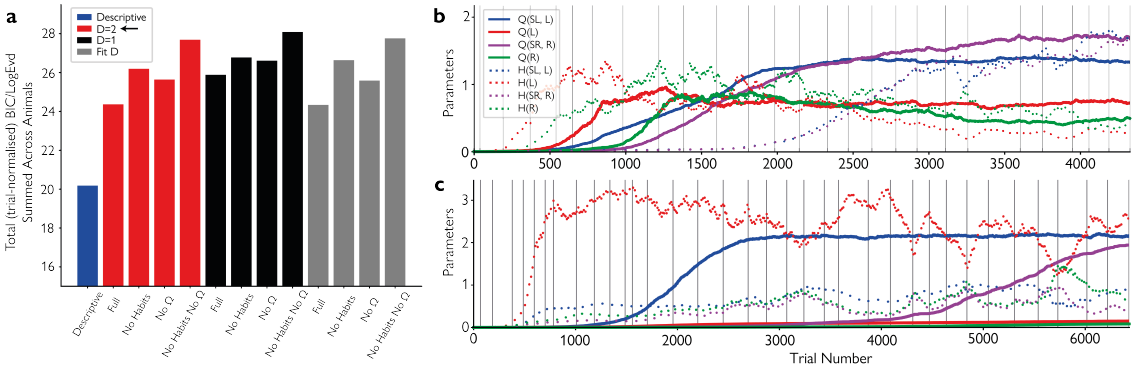


Figure B.4: Fits of RL-DLRNN model. a: comparison of the model evidence of the descriptive fits from Section B.3.1 with the BIC of different versions of the learning model. **b**, **c**: for the best learning model (i.e. $D = 2$ + Habits + Ω), trial-level evolution of the dominant DLRNN model parameters over learning for the same balanced mouse (b) and left-preferring mouse (c) as in Fig. B.2.

B.4 Discussion and Conclusion

We describe a framework capable of linking behavioural signatures of learning to potential network architectures and learning rules, providing trial-by-trial fits of learning trajectories to large empirical datasets. Our well-controlled learning paradigm in mice offers opportunities for testing diverse learning principles, and we hope our extensible software toolbox will help the community more rapidly iterate the theory-experiment loop. Finally, we propose a learning architecture and several learning rules that, while still far simpler than modern AI systems, incorporate features such as depth, recurrence, and gradient descent learning while preserving tractable inference. Fitting these models on thousands of trials requires just minutes, suggesting that future work will be able to consider more complex architectures and learning schemes.

Bibliography

[1] M. Twain, *The awful German language*. BVK, 1880.

[2] P. Fatás and J. A. Lasheras, "La cueva de altamira y su museo," *Cuadernos de arte rupestre*, vol. 7, pp. 25–35, 2014.

[3] F. J. Haydn, "The creation, hob.xxi:2: Pt. 1, awake the harp, the lyre awake!," 1798.

[4] Plato, *The Apology of Socrates*. 387-399 BCE.

[5] R. Etxepare and A. Irurtzun, "Gravettian hand stencils as sign language formatives," *Philosophical Transactions of the Royal Society B*, vol. 376, no. 1824, p. 20200205, 2021.

[6] G. Dewar, "Facial expressions for kids: Helping children read emotions," 2023.

[7] A. N. Kolmogorov, "On tables of random numbers," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 369–376, 1963.

[8] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[9] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[11] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[14] Z. Wu, L. Qiu, A. Ross, E. Akyürek, B. Chen, B. Wang, N. Kim, J. Andreas, and Y. Kim, "Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks," *arXiv preprint arXiv:2307.02477*, 2023.

[15] H. Choset, K. M. Lynch, S. Hutchinson, G. A. Kantor, and W. Burgard, *Principles of robot motion: theory, algorithms, and implementations*. MIT press, 2005.

[16] S. T. Piantadosi, "Modern language models refute chomsky's approach to language," *From fieldwork to linguistic theory: A tribute to Dan Everett*, pp. 353–414, 2023.

[17] A. Schapiro and N. Turk-Browne, "Statistical learning," *Brain mapping*, vol. 3, no. 1, pp. 501–506, 2015.

[18] R. M. Furr, *Psychometrics: an introduction*. SAGE publications, 2021.

[19] D. Green and J. Swets, *Signal detection theory and psychophysics*. Wiley, 1966.

[20] B. Inhelder and J. Piaget, *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*, vol. 84. Routledge, 2013.

[21] F. C. Keil, *Semantic and conceptual development: An ontological perspective*. Harvard University Press, 1979.

[22] R. S. Siegler, "Three aspects of cognitive development," *Cognitive psychology*, vol. 8, no. 4, pp. 481–520, 1976.

[23] J. M. Mandler and L. McDonough, "Concept formation in infancy," *Cognitive development*, vol. 8, no. 3, pp. 291–318, 1993.

[24] A. M. Saxe, J. L. McClelland, and S. Ganguli, "A mathematical theory of semantic development in deep neural networks," *Proceedings of the National Academy of Sciences*, vol. 116, no. 23, pp. 11537–11546, 2019.

[25] C. J. Brainerd, "The origins of number concepts," *Scientific American*, vol. 228, no. 3, pp. 101–109, 1973.

[26] J. Piaget, *Child's Conception of Number: Selected Works vol 2*. Routledge, 2013.

[27] C. R. Gallistel and R. Gelman, "Preverbal and verbal counting and computation," *Cognition*, vol. 44, no. 1-2, pp. 43–74, 1992.

[28] A. M. Leslie, R. Gelman, and C. Gallistel, "The generative basis of natural number concepts," *Trends in cognitive sciences*, vol. 12, no. 6, pp. 213–218, 2008.

[29] A. Newell and P. S. Rosenbloom, "Mechanisms of skill acquisition and the law of practice," in *Cognitive skills and their acquisition*, pp. 1–55, Psychology Press, 1981.

[30] C. R. Gallistel, S. Fairhurst, and P. Balsam, "The learning curve: implications of a quantitative analysis," *Proceedings of the National Academy of Sciences*, vol. 101, no. 36, pp. 13124–13131, 2004.

[31] W. Köhler, "The mentality of apes. kegan paul, trench, trubner & co," 1925.

[32] D. Durstewitz, N. M. Vittoz, S. B. Floresco, and J. K. Seamans, "Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning," *Neuron*, vol. 66, no. 3, pp. 438–448, 2010.

[33] M. Wertheimer, *Drei abhandlungen zur gestalttheorie*. Philosophische akademie, 1925.

[34] I. Krechevsky, "" hypotheses" in rats.," *Psychological Review*, vol. 39, no. 6, p. 516, 1932.

[35] S. Ohlsson, "Information-processing explanations of insight and related phenomena," *Advances in the psychology of thinking*, vol. 1, pp. 1–44, 1992.

[36] J. Kounios and M. Beeman, "The cognitive neuroscience of insight," *Annual review of psychology*, vol. 65, no. 1, pp. 71–93, 2014.

[37] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2014.

[38] A. Atanasov, B. Bordelon, and C. Pehlevan, "Neural networks as kernel learners: The silent alignment effect," *arXiv preprint arXiv:2111.00034*, 2021.

[39] A. T. Löwe, L. Touzo, P. S. Muhle-Karbe, A. M. Saxe, C. Summerfield, and N. W. Schuck, "Abrupt and spontaneous strategy switches emerge in simple regularised neural networks," *PLoS Computational Biology*, vol. 20, no. 10, p. e1012505, 2024.

[40] E. Thorndike, *Animal intelligence: Experimental studies*. Routledge, 2017.

[41] E. L. Thorndike, "A proof of the law of effect," *Science*, vol. 77, no. 1989, pp. 173–175, 1933.

[42] B. F. Skinner, *About Behaviorism: By BF Skinner*. Knopf, 1974.

[43] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[44] G. Tesauro *et al.*, "Temporal difference learning and td-gammon," *Communications of the ACM*, vol. 38, no. 3, pp. 58–68, 1995.

[45] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[46] P. I. Pavlov, "Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex," *Annals of neurosciences*, vol. 17, no. 3, p. 136, 1927.

[47] K. Samejima, Y. Ueda, K. Doya, and M. Kimura, "Representation of action-specific reward values in the striatum," *Science*, vol. 310, no. 5752, pp. 1337–1340, 2005.

[48] W. Schultz, "Neuronal reward and decision signals: From theories to data," *Physiological Reviews*, vol. 95, no. 3, pp. 853–951, 2015.

[49] A. Lak, M. Okun, M. M. Moss, H. Gurnani, K. Farrell, M. J. Wells, C. B. Reddy, A. Kepecs, K. D. Harris, and M. Carandini, "Dopaminergic and prefrontal basis of learning from sensory confidence and reward value," *Neuron*, vol. 105, no. 4, pp. 700–711, 2020.

[50] M. Fritsche, A. Majumdar, L. Strickland, S. Liebana Garcia, R. Bogacz, and A. Lak, "Temporal regularities shape perceptual decisions and striatal dopamine signals," *Nature Communications*, vol. 15, no. 1, p. 7093, 2024.

[51] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, pp. 9–44, 1988.

[52] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[53] G. Aston-Jones, C. Chiang, and T. Alexinsky, "Discharge of noradrenergic locus coeruleus neurons in behaving rats and monkeys suggests a role in vigilance," *Progress in brain research*, vol. 88, pp. 501–520, 1991.

[54] M. E. Hasselmo, "The role of acetylcholine in learning and memory," *Current opinion in neurobiology*, vol. 16, no. 6, pp. 710–715, 2006.

[55] K. Doya, "Metalearning and neuromodulation," *Neural networks*, vol. 15, no. 4-6, pp. 495–506, 2002.

[56] H. C. Blodgett, "The effect of the introduction of reward upon the maze performance of rats.," *University of California publications in psychology*, 1929.

[57] E. C. Tolman, "Cognitive maps in rats and men.," *Psychological review*, vol. 55, no. 4, p. 189, 1948.

[58] A. Pasupathy and E. K. Miller, "Different time courses of learning-related activity in the prefrontal cortex and striatum," *Nature*, vol. 433, no. 7028, pp. 873–876, 2005.

[59] Y. K. Takahashi, M. R. Roesch, R. C. Wilson, K. Toreson, P. O'donnell, Y. Niv, and G. Schoenbaum, "Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex," *Nature neuroscience*, vol. 14, no. 12, pp. 1590–1597, 2011.

[60] M. Blanco-Pozo, T. Akam, and M. E. Walton, "Dopamine-independent effect of rewards on choices through hidden-state inference," *Nature Neuroscience*, vol. 27, no. 2, pp. 286–297, 2024.

[61] N. D. Daw, Y. Niv, and P. Dayan, "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control," *Nature neuroscience*, vol. 8, no. 12, pp. 1704–1711, 2005.

[62] M. E. Walton, T. E. Behrens, M. J. Buckley, P. H. Rudebeck, and M. F. Rushworth, "Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning," *Neuron*, vol. 65, no. 6, pp. 927–939, 2010.

[63] M. El-Gaby, A. L. Harris, J. C. Whittington, W. Dorrell, A. Bhomick, M. E. Walton, T. Akam, and T. E. Behrens, "A cellular basis for mapping behavioural structure," *Nature*, pp. 1–10, 2024.

[64] P. Dayan, "Goal-directed control and its antipodes," *Neural Networks*, vol. 22, no. 3, pp. 213–219, 2009.

[65] P. Dayan and K. C. Berridge, "Model-based and model-free pavlovian reward learning: revaluation, revision, and revelation," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 14, pp. 473–492, 2014.

[66] H. H. Yin, B. J. Knowlton, and B. W. Balleine, "Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning," *European journal of neuroscience*, vol. 19, no. 1, pp. 181–189, 2004.

[67] M. Lengyel and P. Dayan, "Hippocampal contributions to control: the third way," *Advances in neural information processing systems*, vol. 20, 2007.

[68] R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, and Y. Niv, "Orbitofrontal cortex as a cognitive map of task space," *Neuron*, vol. 81, no. 2, pp. 267–279, 2014.

[69] K. J. Miller, M. M. Botvinick, and C. D. Brody, "Dorsal hippocampus contributes to model-based planning," *Nature neuroscience*, vol. 20, no. 9, pp. 1269–1276, 2017.

[70] J. P. Geerts, F. Chersi, K. L. Stachenfeld, and N. Burgess, "A general model of hippocampal and dorsal striatal learning and decision making," *Proceedings of the National Academy of Sciences*, vol. 117, no. 49, pp. 31427–31437, 2020.

[71] J. O'Keefe and L. Nadel, "The hippocampus as a cognitive map," 1978.

[72] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, no. 7052, pp. 801–806, 2005.

[73] J. C. Whittington, T. H. Muller, S. Mark, G. Chen, C. Barry, N. Burgess, and T. E. Behrens, "The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation," *Cell*, vol. 183, no. 5, pp. 1249–1263, 2020.

[74] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory.," *Psychological review*, vol. 102, no. 3, p. 419, 1995.

[75] K. Takehara-Nishiuchi and B. L. McNaughton, "Spontaneous changes of neocortical code for associative memory during consolidation," *Science*, vol. 322, no. 5903, pp. 960–963, 2008.

[76] T. Kitamura, S. K. Ogawa, D. S. Roy, T. Okuyama, M. D. Morrissey, L. M. Smith, R. L. Redondo, and S. Tonegawa, "Engrams and circuits crucial for systems consolidation of a memory," *Science*, vol. 356, no. 6333, pp. 73–78, 2017.

[77] T. Mauldin, "On the unification of physics," *The Journal of Philosophy*, vol. 93, no. 3, pp. 129–144, 1996.

[78] K. Friston, "The free-energy principle: a unified brain theory?," *Nature reviews neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[79] P. Dayan, "Twenty-five lessons from computational neuromodulation," *Neuron*, vol. 76, no. 1, pp. 240–256, 2012.

[80] M. V. Srinivasan, S. B. Laughlin, and A. Dubs, "Predictive coding: a fresh view of inhibition in the retina," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 216, no. 1205, pp. 427–459, 1982.

[81] P. Dayan and A. Jyu, "Uncertainty and learning," *IETE Journal of Research*, vol. 49, no. 2-3, pp. 171–181, 2003.

[82] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, G. Pezzulo, *et al.*, "Active inference and learning," *Neuroscience & Biobehavioral Reviews*, vol. 68, pp. 862–879, 2016.

[83] Y. Song, B. Millidge, T. Salvatori, T. Lukasiewicz, Z. Xu, and R. Bogacz, "Inferring neural activity before plasticity as a foundation for learning beyond backpropagation," *Nature Neuroscience*, vol. 27, no. 2, pp. 348–358, 2024.

[84] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[85] K. Friston, "A theory of cortical responses," *Philosophical transactions of the Royal Society B: Biological sciences*, vol. 360, no. 1456, pp. 815–836, 2005.

[86] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, "An internal model for sensorimotor integration," *Science*, vol. 269, no. 5232, pp. 1880–1882, 1995.

[87] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.

[88] J. M. Beck, W. J. Ma, R. Kiani, T. Hanks, A. K. Churchland, J. Roitman, M. N. Shadlen, P. E. Latham, and A. Pouget, "Probabilistic population codes for bayesian decision making," *Neuron*, vol. 60, no. 6, pp. 1142–1152, 2008.

[89] A. E. Stahl and L. Feigenson, "Observing the unexpected enhances infants' learning and exploration," *Science*, vol. 348, no. 6230, pp. 91–94, 2015.

[90] S.-J. Blakemore, D. M. Wolpert, and C. D. Frith, "Central cancellation of self-produced tickle sensation," *Nature neuroscience*, vol. 1, no. 7, pp. 635–640, 1998.

[91] C. Summerfield, T. Egner, M. Greene, E. Koechlin, J. Mangels, and J. Hirsch, "Predictive codes for forthcoming perception in the frontal cortex," *Science*, vol. 314, no. 5803, pp. 1311–1314, 2006.

[92] C. Summerfield, E. H. Trittschuh, J. M. Monti, M.-M. Mesulam, and T. Egner, "Neural repetition suppression reflects fulfilled perceptual expectations," *Nature neuroscience*, vol. 11, no. 9, pp. 1004–1006, 2008.

[93] P. Kok, J. F. Jehee, and F. P. De Lange, "Less is more: expectation sharpens representations in the primary visual cortex," *Neuron*, vol. 75, no. 2, pp. 265–270, 2012.

[94] E. K. Miller, L. Li, and R. Desimone, "A neural mechanism for working and recognition memory in inferior temporal cortex," *Science*, vol. 254, no. 5036, pp. 1377–1379, 1991.

[95] L. Li, E. K. Miller, and R. Desimone, "The representation of stimulus familiarity in anterior inferior temporal cortex," *Journal of neurophysiology*, vol. 69, no. 6, pp. 1918–1929, 1993.

[96] K. Grill-Spector, R. Henson, and A. Martin, "Repetition and the brain: neural models of stimulus-specific effects," *Trends in cognitive sciences*, vol. 10, no. 1, pp. 14–23, 2006.

[97] P. Berkes, G. Orbán, M. Lengyel, and J. Fiser, "Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment," *Science*, vol. 331, no. 6013, pp. 83–87, 2011.

[98] H. L. Hollingworth, "The central tendency of judgment," *The Journal of Philosophy, Psychology and Scientific Methods*, vol. 7, no. 17, pp. 461–469, 1910.

[99] O. Raviv, I. Lieder, Y. Loewenstein, and M. Ahissar, "Contradictory behavioral biases result from the influence of past stimuli on perception," *PLoS Computational Biology*, vol. 10, no. 12, p. e1003948, 2014.

[100] J. R. Saffran, R. N. Aslin, and E. L. Newport, "Statistical learning by 8-month-old infants," *science*, vol. 274, no. 5294, pp. 1926–1928, 1996.

[101] L. de Hoz, D. Barniv, and I. Nelken, "Prior unsupervised experience leads to long lasting effects in sensory gating during discrimination learning," *bioRxiv*, pp. 2022–09, 2022.

[102] L. Zhong, S. Baptista, R. Gattoni, J. Arnold, D. Flickinger, C. Stringer, and M. Pachitariu, "Distinct streams for supervised and unsupervised learning in the visual cortex," *bioRxiv*, pp. 2024–02, 2024.

[103] S. Zhang, Y. Tian, Q. Liu, and H. Wu, "The neural correlates of novelty and variability in human decision-making under an active inference framework," *bioRxiv*, pp. 2023–09, 2023.

[104] T. Isomura, K. Kotani, Y. Jimbo, and K. J. Friston, "Experimental validation of the free-energy principle with in vitro neural networks," *Nature Communications*, vol. 14, no. 1, p. 4547, 2023.

[105] R. Hodson, M. Mehta, and R. Smith, "The empirical status of predictive coding and active inference," *Neuroscience & Biobehavioral Reviews*, vol. 157, p. 105473, 2024.

[106] M. L. Anderson and T. Chemero, "The problem with brain guts: Conflation of different senses of" prediction" threatens metaphysical disaster," *Behavioral and Brain Sciences*, no. 3, p. 204, 2013.

[107] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[108] W. James, "The principles of psychology," *Henry Holt*, 1890.

[109] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.

[110] D. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press, 2005.

[111] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[112] J. L. McClelland and T. T. Rogers, "The parallel distributed processing approach to semantic cognition," *Nature reviews neuroscience*, vol. 4, no. 4, pp. 310–322, 2003.

[113] B. Millidge, T. Salvatori, Y. Song, R. Bogacz, and T. Lukasiewicz, "Predictive coding: Towards a future of deep learning beyond backpropagation?," *arXiv preprint arXiv:2202.09467*, 2022.

[114] L. Pinchetti, C. Qi, O. Lokshyn, C. Emde, A. M'Charrak, M. Tang, S. Frieder, B. Menzat, G. Oliviers, R. Bogacz, T. Lukasiewicz, and T. Salvatori, "Benchmarking predictive coding networks – made simple," in *The Thirteenth International Conference on Learning Representations*, 2025.

[115] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[116] V. Vapnik, A. Chervonenkis, *et al.*, "Theory of pattern recognition," 1974.

[117] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation*, vol. 1, no. 1, pp. 67–82, 1997.

[118] T. M. Mitchell, "The need for biases in learning generalizations," *Citeseer*, 1980.

[119] B. Neyshabur, R. Tomioka, and N. Srebro, "In search of the real inductive bias: On the role of implicit regularization in deep learning," *arXiv preprint arXiv:1412.6614*, 2014.

[120] A. N. Tikhonov *et al.*, "On the stability of inverse problems," in *Dokl. akad. nauk sssr*, vol. 39, pp. 195–198, 1943.

[121] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[122] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[123] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[124] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478*, 2021.

[125] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[126] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[127] B. Blum-Smith and S. Villar, "Machine learning and invariant theory," *Notices of the American Mathematical Society*, vol. 70, no. 8, pp. 1205–1213, 2023.

[128] D. Saad and S. A. Solla, "Exact solution for on-line learning in multilayer neural networks," *Physical Review Letters*, vol. 74, no. 21, p. 4337, 1995.

[129] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, "Exponential expressivity in deep neural networks through transient chaos," *Advances in neural information processing systems*, vol. 29, 2016.

[130] D. Saad and S. A. Solla, "On-line learning in soft committee machines," *Physical Review E*, vol. 52, no. 4, p. 4225, 1995.

[131] S. Pesme and N. Flammarion, "Saddle-to-saddle dynamics in diagonal linear networks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 7475–7505, 2023.

[132] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[133] L. Chizat, E. Oyallon, and F. Bach, "On lazy training in differentiable programming," *Advances in neural information processing systems*, vol. 32, 2019.

[134] T. Flesch, K. Juechems, T. Dumbalska, A. Saxe, and C. Summerfield, "Orthogonal representations for robust context-dependent task performance in brains and neural networks," *Neuron*, vol. 110, no. 7, pp. 1258–1270, 2022.

[135] E. Holton, L. Braun, J. Thompson, J. Grohn, and C. Summerfield, "Humans and neural networks show similar patterns of transfer and interference during continual learning," *psyArXiv preprint*, 2025.

[136] A. Jacot, F. Ged, B. Şimşek, C. Hongler, and F. Gabriel, "Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity," *arXiv preprint arXiv:2106.15933*, 2021.

[137] E. Abbe, E. B. Adsera, and T. Misiakiewicz, "Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics," in *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2552–2623, PMLR, 2023.

[138] D. E. Pohren, *Paco de Lucia and Family: The master plan.* Society of Spanish Studies, 1992.

[139] R. W. Tyler, "Basic principles of curriculum and instruction," in *Curriculum studies reader E2*, pp. 60–68, Routledge, 2013.

[140] A. J. Baroody and A. Dowker, *The development of arithmetic concepts and skills: Constructive adaptive expertise.* Routledge, 2013.

[141] R. Siegler and E. A. Jenkins, *How children discover new strategies*. Psychology Press, 2014.

[142] M. Akcaoglu, L. J. Jensen, and D. Gonzalez, "Understanding children's problem-solving strategies in solving game-based logic problems.," *International Journal of Technology in Education and Science*, vol. 5, no. 2, pp. 245–257, 2021.

[143] A. Yan, M. J. Lee, and A. J. Ko, "Predicting abandonment in online coding tutorials," in *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 191–199, IEEE, 2017.

[144] A. Dirzyte, A. Perminas, L. Kaminskis, G. Žebrauskas, Ž. Sederevičiūtė-Pačiauskienė, J. Šliogerienė, J. Suchanova, R. Rimašiūtė-Knabikienė, A. Patapas, and I. Gajdosikiene, "Factors contributing to dropping out of adults' programming e-learning," *Heliyon*, vol. 9, no. 12, 2023.

[145] N. Chitrakar and P. Nisanth, "Frustration and its influences on student motivation and academic performance," *International Journal of Scientific Research in Modern Science and Technology*, vol. 2, no. 11, pp. 01–09, 2023.

[146] J. J. Paton, M. A. Belova, S. E. Morrison, and C. D. Salzman, "The primate amygdala represents the positive and negative value of visual stimuli during learning," *Nature*, vol. 439, no. 7078, pp. 865–870, 2006.

[147] L. Ding and J. I. Gold, "The basal ganglia's contributions to perceptual decision making," *Neuron*, vol. 79, no. 4, pp. 640–649, 2013.

[148] H. Makino and T. Komiyama, "Learning enhances the relative impact of top-down processing in the visual cortex," *Nature neuroscience*, vol. 18, no. 8, pp. 1116–1122, 2015.

[149] C. M. Constantinople, A. T. Piet, P. Bibawi, A. Akrami, C. Kopec, and C. D. Brody, "Lateral orbitofrontal cortex promotes trial-by-trial learning of risky, but not spatial, biases," *Elife*, vol. 8, p. e49744, 2019.

[150] C. C. Rodgers, R. Nogueira, B. C. Pil, E. A. Greeman, J. M. Park, Y. K. Hong, S. Fusi, and R. M. Bruno, "Sensorimotor strategies and neuronal representations for shape discrimination," *Neuron*, vol. 109, no. 14, pp. 2308–2325, 2021.

[151] D. Lee, H. Seo, and M. W. Jung, "Neural basis of reinforcement learning and decision making," *Annual review of neuroscience*, vol. 35, no. 1, pp. 287–308, 2012.

[152] P. Le Merre, V. Esmaeili, E. Charrière, K. Galan, P.-A. Salin, C. C. Petersen, and S. Crochet, "Reward-based learning drives rapid sensory signals in medial prefrontal cortex and dorsal hippocampus necessary for goal-directed behavior," *Neuron*, vol. 97, no. 1, pp. 83–91, 2018.

[153] K. Iigaya, M. S. Fonseca, M. Murakami, Z. F. Mainen, and P. Dayan, "An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals," *Nature communications*, vol. 9, no. 1, p. 2477, 2018.

[154] K. J. Miller, M. M. Botvinick, and C. D. Brody, "Value representations in the rodent orbitofrontal cortex drive learning, not choice," *Elife*, vol. 11, p. e64575, 2022.

[155] B. A. Purcell and R. Kiani, "Neural mechanisms of post-error adjustments of decision policy in parietal cortex," *Neuron*, vol. 89, no. 3, pp. 658–671, 2016.

[156] T. E. Behrens, T. H. Muller, J. C. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-Nelson, "What is a cognitive map? organizing knowledge for flexible behavior," *Neuron*, vol. 100, no. 2, pp. 490–509, 2018.

[157] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex," *nature*, vol. 503, no. 7474, pp. 78–84, 2013.

[158] S. Reinert, M. Hübener, T. Bonhoeffer, and P. M. Goltstein, "Mouse prefrontal cortex represents learned rules for categorization," *Nature*, vol. 593, no. 7859, pp. 411–417, 2021.

[159] W. Sun, J. Winnubst, M. Natrajan, C. Lai, K. Kajikawa, M. Michaelos, R. Gattoni, C. Stringer, D. Flickinger, J. E. Fitzgerald, *et al.*, "Learning produces a hippocampal cognitive map in the form of an orthogonalized state machine," *bioRxiv*, pp. 2023–08, 2023.

[160] C. P. Burgess, A. Lak, N. A. Steinmetz, P. Zatka-Haas, C. B. Reddy, E. A. Jacobs, J. F. Linden, J. J. Paton, A. Ranson, S. Schröder, *et al.*, "High-yield methods for accurate two-alternative visual psychophysics in head-fixed mice," *Cell reports*, vol. 20, no. 10, pp. 2513–2524, 2017.

[161] R. Amo, S. Matias, A. Yamanaka, K. F. Tanaka, N. Uchida, and M. Watabe-Uchida, "A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning," *Nature neuroscience*, vol. 25, no. 8, pp. 1082–1092, 2022.

[162] P. Dayan and Y. Niv, "Reinforcement learning: the good, the bad and the ugly," *Current opinion in neurobiology*, vol. 18, no. 2, pp. 185–196, 2008.

[163] D. Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward is enough," *Artificial Intelligence*, vol. 299, p. 103535, 2021.

[164] W. T. Newsome, K. H. Britten, and J. A. Movshon, "Neuronal correlates of a perceptual decision," *Nature*, vol. 341, no. 6237, pp. 52–54, 1989.

[165] J. I. Gold and M. N. Shadlen, "The neural basis of decision making," *Annu. Rev. Neurosci.*, vol. 30, no. 1, pp. 535–574, 2007.

[166] M. Carandini and A. K. Churchland, "Probing perceptual decisions in rodents," *Nature neuroscience*, vol. 16, no. 7, pp. 824–831, 2013.

[167] O. Hikosaka, K. Nakamura, and H. Nakahara, "Basal ganglia orient eyes to reward," *Journal of neurophysiology*, vol. 95, no. 2, pp. 567–584, 2006.

[168] C. Padoa-Schioppa, "Neurobiology of economic choice: a good-based model," *Annual review of neuroscience*, vol. 34, no. 1, pp. 333–359, 2011.

[169] P. W. Glimcher, *Neuroeconomics: Decision making and the brain*. Academic Press, 2013.

[170] E. O. Neftci and B. B. Averbeck, "Reinforcement learning in artificial and biological systems," *Nature Machine Intelligence*, vol. 1, no. 3, pp. 133–143, 2019.

[171] G. Morris, A. Nevet, D. Arkadir, E. Vaadia, and H. Bergman, "Midbrain dopamine neurons encode decisions for future action," *Nature neuroscience*, vol. 9, no. 8, pp. 1057–1063, 2006.

[172] W. R. Stauffer, A. Lak, A. Yang, M. Borel, O. Paulsen, E. S. Boyden, and W. Schultz, "Dopamine neuron-specific optogenetic stimulation in rhesus macaques," *Cell*, vol. 166, no. 6, pp. 1564–1571, 2016.

[173] S. Soares, B. V. Atallah, and J. J. Paton, "Midbrain dopamine neurons control judgment of time," *Science*, vol. 354, no. 6317, pp. 1273–1277, 2016.

[174] M. Watabe-Uchida, N. Eshel, and N. Uchida, "Neural circuitry of reward prediction error," *Annual review of neuroscience*, vol. 40, no. 1, pp. 373–394, 2017.

[175] E. J. P. Maes, M. J. Sharpe, A. A. Usypchuk, M. Lozzi, C. Y. Chang, M. P. Gardner, G. Schoenbaum, and M. D. Iordanova, "Causal evidence supporting the proposal that dopamine transients function as temporal difference prediction errors," *Nature neuroscience*, vol. 23, no. 2, pp. 176–178, 2020.

[176] S. J. Gershman, J. A. Assad, S. R. Datta, S. W. Linderman, B. L. Sabatini, N. Uchida, and L. Wilbrecht, "Explaining dopamine through prediction errors and beyond," *Nature neuroscience*, vol. 27, no. 9, pp. 1645–1655, 2024.

[177] K. Nomoto, W. Schultz, T. Watanabe, and M. Sakagami, "Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli," *Journal of neuroscience*, vol. 30, no. 32, pp. 10692–10702, 2010.

[178] V. de Lafuente and R. Romo, "Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. 19767–19771, 2011.

[179] D. Bang, K. T. Kishida, T. Lohrenz, J. P. White, A. W. Laxton, S. B. Tatter, S. M. Fleming, and P. R. Montague, "Sub-second dopamine and serotonin signaling in human striatum during perceptual decision-making," *Neuron*, vol. 108, no. 5, pp. 999–1010, 2020.

[180] L. Ding, "Contributions of the basal ganglia to visual perceptual decisions," *Annual Review of Vision Science*, vol. 9, no. 1, pp. 385–407, 2023.

[181] I. Tsutsui-Kimura, H. Matsumoto, K. Akiti, M. M. Yamada, N. Uchida, and M. Watabe-Uchida, "Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task," *Elife*, vol. 9, p. e62390, 2020.

[182] M. M. Moss, P. Zatka-Haas, K. D. Harris, M. Carandini, and A. Lak, "Dopamine axons in dorsal striatum encode contralateral visual stimuli and choices," *Journal of Neuroscience*, vol. 41, no. 34, pp. 7197–7205, 2021.

[183] K. Schmack, M. Bosc, T. Ott, J. Sturgill, and A. Kepecs, "Striatal dopamine mediates hallucination-like perception in mice," *Science*, vol. 372, no. 6537, p. eabf4740, 2021.

[184] T. W. Bernklau, B. Righetti, L. S. Mehrke, and S. N. Jacob, "Striatal dopamine signals reflect perceived cue–action–outcome associations in mice," *Nature Neuroscience*, vol. 27, no. 4, pp. 747–757, 2024.

[185] M. Watabe-Uchida, L. Zhu, S. K. Ogawa, A. Vamanrao, and N. Uchida, "Whole-brain mapping of direct inputs to midbrain dopamine neurons," *Neuron*, vol. 74, no. 5, pp. 858–873, 2012.

[186] B. J. Hunnicutt, B. C. Jongbloets, W. T. Birdsong, K. J. Gertz, H. Zhong, and T. Mao, "A comprehensive excitatory input map of the striatum reveals novel functional organization," *eLife*, vol. 5, p. e19103, nov 2016.

[187] P. Zatka-Haas, N. A. Steinmetz, M. Carandini, and K. D. Harris, "Sensory coding and the causal impact of mouse cortex in a visual decision," *Elife*, vol. 10, p. e63163, 2021.

[188] J. N. Reynolds, B. I. Hyland, and J. R. Wickens, "A cellular mechanism of reward-related learning," *Nature*, vol. 413, no. 6851, pp. 67–70, 2001.

[189] C. D. Howard, H. Li, C. E. Geddes, and X. Jin, "Dynamic nigrostriatal dopamine biases action selection," *Neuron*, vol. 93, no. 6, pp. 1436–1450, 2017.

[190] B. T. Saunders, J. M. Richard, E. B. Margolis, and P. H. Janak, "Dopamine neurons create pavlovian conditioned stimuli with circuit-defined motivational properties," *Nature neuroscience*, vol. 21, no. 8, pp. 1072–1083, 2018.

[191] H. R. Kim, A. N. Malik, J. G. Mikhael, P. Bech, I. Tsutsui-Kimura, F. Sun, Y. Zhang, Y. Li, M. Watabe-Uchida, S. J. Gershman, *et al.*, "A unified framework for dopamine signals across timescales," *Cell*, vol. 183, no. 6, pp. 1600–1616, 2020.

[192] B. Engelhard, J. Finkelstein, J. Cox, W. Fleming, H. J. Jang, S. Ornelas, S. A. Koay, S. Y. Thiberge, N. D. Daw, D. W. Tank, *et al.*, "Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons," *Nature*, vol. 570, no. 7762, pp. 509–513, 2019.

[193] A. Lak, K. Nomoto, M. Keramati, M. Sakagami, and A. Kepecs, "Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision," *Current Biology*, vol. 27, no. 6, pp. 821–832, 2017.

[194] A. A. Hamid, J. R. Pettibone, O. S. Mabrouk, V. L. Hetrick, R. Schmidt, C. M. Vander Weele, R. T. Kennedy, B. J. Aragona, and J. D. Berke, "Mesolimbic dopamine signals the value of work," *Nature neuroscience*, vol. 19, no. 1, pp. 117–126, 2016.

[195] A. P. Chen, J. M. Malgady, L. Chen, K. W. Shi, E. Cheng, J. L. Plotkin, S. Ge, and Q. Xiong, "Nigrostriatal dopamine pathway regulates auditory discrimination behavior," *Nature Communications*, vol. 13, no. 1, p. 5942, 2022.

[196] M. Minsky, "Steps toward artificial intelligence," *Proceedings of the IRE*, vol. 49, no. 1, pp. 8–30, 1961.

[197] M. Minsky and S. A. Papert, *Perceptrons, reissue of the 1988 expanded edition with a new foreword by Léon Bottou: an introduction to computational geometry*. MIT press, 2017.

[198] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[199] A. Saxe, S. Nelli, and C. Summerfield, "If deep learning is the answer, what is the question?," *Nature Reviews Neuroscience*, vol. 22, no. 1, pp. 55–67, 2021.

[200] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[201] M. S. Advani, A. M. Saxe, and H. Sompolinsky, "High-dimensional dynamics of generalization error in neural networks," *Neural Networks*, vol. 132, pp. 428–446, 2020.

[202] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," *Advances in neural information processing systems*, vol. 31, 2018.

[203] S. S. Mannelli, F. Gerace, N. Rostamzadeh, and L. Saglietti, "Unfair geometries: exactly solvable data model with fairness implications," *arXiv preprint arXiv:2205.15935*, 2022.

[204] D. Kunin, J. Sagastuy-Brena, S. Ganguli, D. L. Yamins, and H. Tanaka, "Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics," *arXiv preprint arXiv:2012.04728*, 2020.

[205] B. Simsek, F. Ged, A. Jacot, F. Spadaro, C. Hongler, W. Gerstner, and J. Brea, "Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances," in *International Conference on Machine Learning*, pp. 9722–9732, PMLR, 2021.

[206] P. Dayan and N. D. Daw, "Decision theory, reinforcement learning, and the brain," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 8, no. 4, pp. 429–453, 2008.

[207] Z. Ashwood, N. A. Roy, J. H. Bak, and J. W. Pillow, "Inferring learning rules from animal decision-making," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3442–3453, 2020.

[208] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[209] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.*, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," *arXiv preprint arXiv:1712.01815*, 2017.

[210] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[211] J. C. Whittington and R. Bogacz, "An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity," *Neural computation*, vol. 29, no. 5, pp. 1229–1262, 2017.

[212] J. B. Heald, M. Lengyel, and D. M. Wolpert, "Contextual inference underlies the learning of sensorimotor repertoires," *Nature*, vol. 600, no. 7889, pp. 489–493, 2021.

[213] T. Salvatori, Y. Song, Y. Hong, L. Sha, S. Frieder, Z. Xu, R. Bogacz, and T. Lukasiewicz, "Associative memories via predictive coding," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3874–3886, 2021.

[214] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[215] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[216] G. E. Box, "Science and statistics," *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 791–799, 1976.

[217] J. S. Collina, G. Erdil, M. Xia, C. F. Angeloni, K. C. Wood, J. Sheth, K. P. Kording, Y. E. Cohen, and M. N. Geffen, "Individual-specific strategies inform category learning," *bioRxiv*, pp. 2024–09, 2024.

[218] I. B. Laboratory, V. Aguillon-Rodriguez, D. Angelaki, H. Bayer, N. Bonacchi, M. Carandini, F. Cazettes, G. Chapuis, A. K. Churchland, Y. Dan, *et al.*, "Standardized and reproducible measurement of decision-making in mice," *Elife*, vol. 10, p. e63711, 2021.

[219] M. Rosenberg, T. Zhang, P. Perona, and M. Meister, "Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration," *Elife*, vol. 10, p. e66175, 2021.

[220] M. Meister, "Learning, fast and slow," *Current opinion in neurobiology*, vol. 75, p. 102555, 2022.

[221] J. Bhagat, M. J. Wells, K. D. Harris, M. Carandini, and C. P. Burgess, "Rigbox: an open-source toolbox for probing neurons and behavior," *Eneuro*, vol. 7, no. 4, 2020.

[222] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[223] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, *et al.*, "Tslearn, a machine learning toolkit for time series data," *Journal of machine learning research*, vol. 21, no. 118, pp. 1–6, 2020.

[224] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[225] A. Syeda, L. Zhong, R. Tung, W. Long, M. Pachitariu, and C. Stringer, "Facemap: a framework for modeling neural activity based on orofacial tracking," *Nature Neuroscience*, vol. 27, no. 1, pp. 187–195, 2024.

[226] Z. V. Guo, N. Li, D. Huber, E. Ophir, D. Gutnisky, J. T. Ting, G. Feng, and K. Svoboda, "Flow of cortical activity underlying a tactile decision in mice," *Neuron*, vol. 81, no. 1, pp. 179–194, 2014.

[227] Z. C. Ashwood, N. A. Roy, I. R. Stone, I. B. Laboratory, A. E. Urai, A. K. Churchland, A. Pouget, and J. W. Pillow, "Mice alternate between discrete strategies during perceptual decision-making," *Nature Neuroscience*, vol. 25, no. 2, pp. 201–212, 2022.

[228] J. L. Seiler, C. V. Cosme, V. N. Sherathiya, M. D. Schaid, J. M. Bianco, A. S. Bridgemohan, and T. N. Lerner, "Dopamine signaling in the dorsomedial striatum promotes compulsive behavior," *Current Biology*, vol. 32, no. 5, pp. 1175–1188, 2022.

[229] B. Bathellier, S. P. Tee, C. Hrovat, and S. Rumpel, "A multiplicative reinforcement learning model capturing learning dynamics and interindividual variability in mice," *Proceedings of the National Academy of Sciences*, vol. 110, no. 49, pp. 19950–19955, 2013.

[230] D. M. Wolpert and J. R. Flanagan, "Motor learning," *Current biology*, vol. 20, no. 11, pp. R467–R472, 2010.

[231] B. Feulner, M. G. Perich, L. E. Miller, C. Clopath, and J. A. Gallego, "A neural implementation model of feedback-based motor learning," *Nature communications*, vol. 16, no. 1, p. 1805, 2025.

[232] J. S. Albus, "A theory of cerebellar function," *Mathematical biosciences*, vol. 10, no. 1-2, pp. 25–61, 1971.

[233] N. F. Parker, C. M. Cameron, J. P. Taliaferro, J. Lee, J. Y. Choi, T. J. Davidson, N. D. Daw, and I. B. Witten, "Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target," *Nature neuroscience*, vol. 19, no. 6, pp. 845–854, 2016.

[234] R. S. Sutton, *Temporal credit assignment in reinforcement learning*. University of Massachusetts Amherst, 1984.

[235] P. W. Glimcher, "Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis," *Proceedings of the National Academy of Sciences*, vol. 108, no. supplement_3, pp. 15647–15654, 2011.

[236] N. F. Parker, C. M. Cameron, J. P. Taliaferro, J. Lee, J. Y. Choi, T. J. Davidson, N. D. Daw, and I. B. Witten, "Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target," *Nature neuroscience*, vol. 19, no. 6, pp. 845–854, 2016.

[237] A. Pan-Vazquez, Y. S. Araujo, B. McMannon, M. Louka, A. Bandi, L. Haetzel, M. Faulkner, J. W. Pillow, N. D. Daw, and I. B. Witten, "Pre-existing visual responses in a projection-defined dopamine population explain individual learning trajectories," *Current Biology*, vol. 34, no. 22, pp. 5349–5358, 2024.

[238] O. Hikosaka, E. Bromberg-Martin, S. Hong, and M. Matsumoto, "New insights on the subcortical representation of reward," *Current opinion in neurobiology*, vol. 18, no. 2, pp. 203–208, 2008.

[239] E. S. Bromberg-Martin, M. Matsumoto, and O. Hikosaka, "Dopamine in motivational control: rewarding, aversive, and alerting," *Neuron*, vol. 68, no. 5, pp. 815–834, 2010.

[240] S. B. Flagel, J. J. Clark, T. E. Robinson, L. Mayo, A. Czuj, I. Willuhn, C. A. Akers, S. M. Clinton, P. E. Phillips, and H. Akil, "A selective role for dopamine in stimulus–reward learning," *Nature*, vol. 469, no. 7328, pp. 53–57, 2011.

[241] V. Pascoli, J. Terrier, A. Hiver, and C. Lüscher, "Sufficiency of mesolimbic dopamine neuron stimulation for the progression to addiction," *Neuron*, vol. 88, no. 5, pp. 1054–1066, 2015.

[242] A. Lak, W. R. Stauffer, and W. Schultz, "Dopamine neurons learn relative chosen value from probabilistic rewards," *Elife*, vol. 5, p. e18044, 2016.

[243] M. W. Howe and D. A. Dombeck, "Rapid signalling in distinct dopaminergic axons during locomotion and reward," *Nature*, vol. 535, no. 7613, pp. 505–510, 2016.

[244] W. Menegas, K. Akiti, R. Amo, N. Uchida, and M. Watabe-Uchida, "Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli," *Nature neuroscience*, vol. 21, no. 10, pp. 1421–1430, 2018.

[245] J. D. Berke, "What does dopamine mean?," *Nature neuroscience*, vol. 21, no. 6, pp. 787–793, 2018.

[246] A. Mohebi, J. R. Pettibone, A. A. Hamid, J.-M. T. Wong, L. T. Vinson, T. Patriarchi, L. Tian, R. T. Kennedy, and J. D. Berke, "Dissociable dopamine dynamics for learning and motivation," *Nature*, vol. 570, no. 7759, pp. 65–70, 2019.

[247] J. Cox and I. B. Witten, "Striatal circuits for reward learning and decision-making," *Nature Reviews Neuroscience*, vol. 20, no. 8, pp. 482–494, 2019.

[248] L. T. Coddington, S. E. Lindo, and J. T. Dudman, "Mesolimbic dopamine adapts the rate of learning from action," *Nature*, vol. 614, no. 7947, pp. 294–302, 2023.

[249] H. M. Bayer and P. W. Glimcher, "Midbrain dopamine neurons encode a quantitative reward prediction error signal," *Neuron*, vol. 47, no. 1, pp. 129–141, 2005.

[250] J. Y. Cohen, S. Haesler, L. Vong, B. B. Lowell, and N. Uchida, "Neuron-type-specific signals for reward and punishment in the ventral tegmental area," *nature*, vol. 482, no. 7383, pp. 85–88, 2012.

[251] W. Menegas, B. M. Babayan, N. Uchida, and M. Watabe-Uchida, "Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice," *elife*, vol. 6, p. e21886, 2017.

[252] K. Akiti, I. Tsutsui-Kimura, Y. Xie, A. Mathis, J. E. Markowitz, R. Anyoha, S. R. Datta, M. W. Mathis, N. Uchida, and M. Watabe-Uchida, "Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction," *Neuron*, vol. 110, no. 22, pp. 3789–3804, 2022.

[253] F. Greenstreet, H. M. Vergara, Y. Johansson, S. Pati, L. Schwarz, S. C. Lenzi, M. Wisdom, A. Gubanova, F. Marbach, L. Rollik, *et al.*, "Action prediction error: a value-free dopaminergic teaching signal that drives stable learning," *BiorXiv*, pp. 2022–09, 2022.

[254] R. S. Lee, Y. Sagiv, B. Engelhard, I. B. Witten, and N. D. Daw, "A feature-specific prediction error model explains dopaminergic heterogeneity," *Nature neuroscience*, vol. 27, no. 8, pp. 1574–1586, 2024.

[255] T. Akam and M. E. Walton, "pyphotometry: Open source python based hardware and software for fiber photometry data acquisition," *Scientific reports*, vol. 9, no. 1, p. 3521, 2019.

[256] M. Mahn, I. Saraf-Sinik, P. Patil, M. Pulin, E. Bitton, N. Karalis, F. Bruentgens, S. Palgi, A. Gat, J. Dine, *et al.*, "Efficient optogenetic silencing of neurotransmitter release with a mosquito rhodopsin," *Neuron*, vol. 109, no. 10, pp. 1621–1635, 2021.

[257] F. Sun, J. Zhou, B. Dai, T. Qian, J. Zeng, X. Li, Y. Zhuo, Y. Zhang, Y. Wang, C. Qian, *et al.*, "Next-generation grab sensors for monitoring dopaminergic activity in vivo," *Nature methods*, vol. 17, no. 11, pp. 1156–1166, 2020.

[258] T. Ljungberg, P. Apicella, and W. Schultz, "Responses of monkey dopamine neurons during learning of behavioral reactions," *Journal of neurophysiology*, vol. 67, no. 1, pp. 145–163, 1992.

[259] J. Morrens, Ç. Aydin, A. J. van Rensburg, J. E. Rabell, and S. Haesler, "Cue-evoked dopamine promotes conditioned responding during learning," *Neuron*, vol. 106, no. 1, pp. 142–153, 2020.

[260] P. Ambrosi and T. N. Lerner, "Striatonigrostriatal circuit architecture for disinhibition of dopamine signaling," *Cell reports*, vol. 40, no. 7, p. 111228, 2022.

[261] R. Bogacz, "Dopamine role in learning and action inference," *Elife*, vol. 9, p. e53262, 2020.

[262] W. van Elzelingen, J. Goedhoop, P. Warnaar, D. Denys, T. Arbab, and I. Willuhn, "A unidirectional but not uniform striatal landscape of dopamine signaling for motivational stimuli," *Proceedings of the National Academy of Sciences*, vol. 119, no. 21, p. e2117270119, 2022.

[263] S. N. Haber and B. Knutson, "The reward circuit: linking primate anatomy and human imaging," *Neuropsychopharmacology*, vol. 35, no. 1, pp. 4–26, 2010.

[264] A. K. Lahiri and M. D. Bevan, "Dopaminergic transmission rapidly and persistently enhances excitability of d1 receptor-expressing striatal projection neurons," *Neuron*, vol. 106, no. 2, pp. 277–290, 2020.

[265] J. A. Da Silva, F. Tecuapetla, V. Paixão, and R. M. Costa, "Dopamine neuron activity before action initiation gates and invigorates future movements," *Nature*, vol. 554, no. 7691, pp. 244–248, 2018.

[266] J. E. Markowitz, W. F. Gillis, M. Jay, J. Wood, R. W. Harris, R. Cieszkowski, R. Scott, D. Brann, D. Koveal, T. Kula, *et al.*, "Spontaneous behaviour is structured by reinforcement without explicit reward," *Nature*, pp. 1–10, 2023.

[267] L. T. Coddington and J. T. Dudman, "The timing of action determines reward prediction signals in identified midbrain dopamine neurons," *Nature neuroscience*, vol. 21, no. 11, pp. 1563–1573, 2018.

[268] R. Kawai, T. Markman, R. Poddar, R. Ko, A. L. Fantana, A. K. Dhawale, A. R. Kampff, and B. P. Ölveczky, "Motor cortex is required for learning but not for executing a motor skill," *Neuron*, vol. 86, no. 3, pp. 800–812, 2015.

[269] C. Drieu, Z. Zhu, Z. Wang, K. Fuller, A. Wang, S. Elnozahy, and K. Kuchibhotla, "Rapid emergence of latent knowledge in the sensory cortex drives learning," *Nature*, pp. 1–11, 2025.

[270] F. Zhang, L.-P. Wang, E. S. Boyden, and K. Deisseroth, "Channelrhodopsin-2 and optical control of excitable cells," *Nature methods*, vol. 3, no. 10, pp. 785–792, 2006.

[271] J. G. Bernstein, P. A. Garrity, and E. S. Boyden, "Optogenetics and thermogenetics: technologies for controlling the activity of targeted cells within intact neural circuits," *Current opinion in neurobiology*, vol. 22, no. 1, pp. 61–71, 2012.

[272] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the national academy of sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[273] W. Sun, M. Advani, N. Spruston, A. Saxe, and J. E. Fitzgerald, "Organizing memories for generalization in complementary learning systems," *Nature neuroscience*, vol. 26, no. 8, pp. 1438–1448, 2023.

[274] J. Sacramento, R. Ponte Costa, Y. Bengio, and W. Senn, "Dendritic cortical microcircuits approximate the backpropagation algorithm," *Advances in neural information processing systems*, vol. 31, 2018.

[275] J. Guerguiev, T. P. Lillicrap, and B. A. Richards, "Towards deep learning with segregated dendrites," *Elife*, vol. 6, p. e22901, 2017.

[276] B. Scellier and Y. Bengio, "Equilibrium propagation: Bridging the gap between energy-based models and backpropagation," *Frontiers in computational neuroscience*, vol. 11, p. 24, 2017.

[277] L. Bernáez Timón, P. Ekelmans, N. Kraynyukova, T. Rose, L. Busse, and T. Tchumatchenko, "How to incorporate biological insights into network models and why it matters," *The Journal of Physiology*, vol. 601, no. 15, pp. 3037–3053, 2023.

[278] U. Güçlü and M. A. Van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10005–10014, 2015.

[279] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, *et al.*, "Brain-score: Which artificial neural network for object recognition is most brain-like?," *BioRxiv*, p. 407007, 2018.

[280] B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, *et al.*, "A deep learning framework for neuroscience," *Nature neuroscience*, vol. 22, no. 11, pp. 1761–1770, 2019.

[281] E. Weinan, W. Ren, and E. Vanden-Eijnden, "Minimum action method for the study of rare events," *Communications on pure and applied mathematics*, vol. 57, no. 5, pp. 637–656, 2004.

[282] S. N. Haber, J. L. Fudge, and N. R. McFarland, "Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum," *Journal of Neuroscience*, vol. 20, no. 6, pp. 2369–2382, 2000.

[283] Y. Zhang, P. E. Latham, and A. Saxe, "A theory of unimodal bias in multimodal learning," *arXiv preprint arXiv:2312.00935*, 2023.

[284] A. M. Graybiel and S. T. Grafton, "The striatum: where skills and habits meet," *Cold Spring Harbor perspectives in biology*, vol. 7, no. 8, p. a021691, 2015.

[285] L. J. Wilod Versprille, C. McKenzie, F. E. Amorim, K. Yano, J. W. Dalley, and T. W. Robbins, "Dopamine dynamics in the nucleus accumbens reflect confidence in detecting the occurrence and non-occurrence of visual signals in perceptual decision-making," *bioRxiv*, pp. 2025–04, 2025.

[286] W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, and M. Botvinick, "A distributional code for value in dopamine-based reinforcement learning," *Nature*, vol. 577, no. 7792, pp. 671–675, 2020.

[287] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural networks*, vol. 2, no. 1, pp. 53–58, 1989.

[288] Z. Zhu and K. V. Kuchibhotla, "Performance errors during rodent learning reflect a dynamic choice strategy," *Current Biology*, vol. 34, no. 10, pp. 2107–2117, 2024.

[289] D. C. Schreiner and C. M. Gremel, "Orbital frontal cortex projections to secondary motor cortex mediate exploitation of learned rules," *Scientific reports*, vol. 8, no. 1, p. 10979, 2018.

[290] S. J. Gershman, "Uncertainty and exploration.," *Decision*, vol. 6, no. 3, p. 277, 2019.

[291] S. Pisupati, L. Chartarifsky-Lynn, A. Khanal, and A. K. Churchland, "Lapses in perceptual decisions reflect exploration," *Elife*, vol. 10, p. e55490, 2021.

[292] C. I. Jahn, J. Grohn, S. Cuell, A. Emberton, S. Bouret, M. E. Walton, N. Kolling, and J. Sallet, "Neural responses in macaque prefrontal cortex are linked to strategic exploration," *PLoS Biology*, vol. 21, no. 1, p. e3001985, 2023.

[293] A. Dickinson, "Actions and habits: the development of behavioural autonomy," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 308, no. 1135, pp. 67–78, 1985.

[294] W. Wood and D. Rünger, "Psychology of habit," *Annual review of psychology*, vol. 67, no. 1, pp. 289–314, 2016.

[295] L. J. Hammond, "The effect of contingency upon the appetitive conditioning of free-operant behavior," *Journal of the experimental analysis of behavior*, vol. 34, no. 3, pp. 297–304, 1980.

[296] C. D. Adams and A. Dickinson, "Instrumental responding following reinforcer devaluation," *The Quarterly Journal of Experimental Psychology Section B*, vol. 33, no. 2b, pp. 109–121, 1981.

[297] P. Alvarez and L. R. Squire, "Memory consolidation and the medial temporal lobe: a simple network model.," *Proceedings of the national academy of sciences*, vol. 91, no. 15, pp. 7041–7045, 1994.

[298] L. R. Squire and P. Alvarez, "Retrograde amnesia and memory consolidation: a neurobiological perspective," *Current opinion in neurobiology*, vol. 5, no. 2, pp. 169–177, 1995.

[299] D. Kumaran, D. Hassabis, and J. L. McClelland, "What learning systems do intelligent agents need? complementary learning systems theory updated," *Trends in cognitive sciences*, vol. 20, no. 7, pp. 512–534, 2016.

[300] R. S. Palais, "The principle of symmetric criticality," *Communications in Mathematical Physics*, vol. 69, no. 1, pp. 19–30, 1979.

[301] J. Sliwa, "Toward collective animal neuroscience," *Science*, vol. 374, no. 6566, pp. 397–398, 2021.

[302] M. C. Rose, B. Styr, T. A. Schmid, J. E. Elie, and M. M. Yartsev, "Cortical representation of group social communication in bats," *Science*, vol. 374, no. 6566, p. eaba9584, 2021.

[303] R. Báez-Mendoza, E. P. Mastrobattista, A. J. Wang, and Z. M. Williams, "Social agent identity cells in the prefrontal cortex of interacting groups of primates," *Science*, vol. 374, no. 6566, p. eabb4149, 2021.

[304] S. Finger, *Origins of neuroscience: a history of explorations into brain function*. Oxford University Press, 2001.

[305] D. Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward is enough," *Artificial Intelligence*, vol. 299, p. 103535, 2021.

[306] I. H. Witten, "An adaptive optimal controller for discrete-time markov environments," *Information and control*, vol. 34, no. 4, pp. 286–295, 1977.

[307] P. I. Pavlov, "Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex," *Annals of neurosciences*, vol. 17, no. 3, p. 136, 2010.

[308] A. Björklund and S. B. Dunnett, "Dopamine neuron systems in the brain: an update," *Trends in neurosciences*, vol. 30, no. 5, pp. 194–202, 2007.

[309] J. Olds and P. Milner, "Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain.," *Journal of comparative and physiological psychology*, vol. 47, no. 6, p. 419, 1954.

[310] G. Di Chiara and A. Imperato, "Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats.," *Proceedings of the National Academy of Sciences*, vol. 85, no. 14, pp. 5274–5278, 1988.

[311] S. R. Jones, R. R. Gainetdinov, M. Jaber, B. Giros, R. M. Wightman, and M. G. Caron, "Profound neuronal plasticity in response to inactivation of the dopamine transporter," *Proceedings of the National Academy of Sciences*, vol. 95, no. 7, pp. 4029–4034, 1998.

[312] J. N. Reynolds and J. R. Wickens, "Dopamine-dependent plasticity of corticostriatal synapses," *Neural networks*, vol. 15, no. 4-6, pp. 507–521, 2002.

[313] P. Calabresi, B. Picconi, A. Tozzi, and M. Di Filippo, "Dopamine-mediated regulation of corticostriatal synaptic plasticity," *Trends in neurosciences*, vol. 30, no. 5, pp. 211–219, 2007.

[314] M. G. Packard and B. J. Knowlton, "Learning and memory functions of the basal ganglia," *Annual review of neuroscience*, vol. 25, no. 1, pp. 563–593, 2002.

[315] C. K. Starkweather and N. Uchida, "Dopamine signals as temporal difference errors: recent advances," *Current opinion in neurobiology*, vol. 67, pp. 95–105, 2021.

[316] H. Jeong, A. Taylor, J. R. Floeder, M. Lohmann, S. Mihalas, B. Wu, M. Zhou, D. A. Burke, and V. M. K. Namboodiri, "Mesolimbic dopamine release conveys causal associations," *Science*, vol. 378, no. 6626, p. eabq6740, 2022.

[317] K. Bakhurin, R. N. Hughes, Q. Jiang, I. P. Fallon, and H. H. Yin, "Force tuning explains changes in phasic dopamine signaling during stimulus-reward learning," *bioRxiv*, 2023.

[318] C. M. Cannon and R. D. Palmiter, "Reward without dopamine," *Journal of Neuroscience*, vol. 23, no. 34, pp. 10827–10831, 2003.

[319] A. Caplin and M. Dean, "Axiomatic methods, dopamine and reward prediction error," *Current Opinion in Neurobiology*, vol. 18, no. 2, pp. 197–202, 2008. Cognitive neuroscience.

[320] X. Jin and R. M. Costa, "Start/stop signals emerge in nigrostriatal circuits during sequence learning," *Nature*, vol. 466, no. 7305, pp. 457–462, 2010.

[321] J. W. Barter, S. Li, D. Lu, R. A. Bartholomew, M. A. Rossi, C. T. Shoemaker, D. Salas-Meza, E. Gaidis, and H. H. Yin, "Beyond reward prediction errors: the role of dopamine in movement kinematics," *Frontiers in Integrative Neuroscience*, vol. 9, 2015.

[322] T. N. Lerner, A. L. Holloway, and J. L. Seiler, "Dopamine, updated: Reward prediction error and beyond," *Current Opinion in Neurobiology*, vol. 67, pp. 123–130, 2021. Neurobiology of Learning and Plasticity.

[323] T. Patriarchi, J. R. Cho, K. Merten, M. W. Howe, A. Marley, W.-H. Xiong, R. W. Folk, G. J. Broussard, R. Liang, M. J. Jang, H. Zhong, D. Dombeck, M. von Zastrow, A. Nimmerjahn, V. Gradinaru, J. T. Williams, and L. Tian, "Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors," *Science*, vol. 360, no. 6396, p. eaat4422, 2018.

[324] T. N. Lerner, C. Shilyansky, T. J. Davidson, K. E. Evans, K. T. Beier, K. A. Zalocusky, A. K. Crow, R. C. Malenka, L. Luo, R. Tomer, *et al.*, "Intact-brain analyses reveal distinct information carried by snc dopamine subcircuits," *Cell*, vol. 162, no. 3, pp. 635–647, 2015.

[325] A. Zolin, R. Cohn, R. Pang, A. F. Siliciano, A. L. Fairhall, and V. Ruta, "Context-dependent representations of movement in drosophila dopaminergic reinforcement pathways," *Nature neuroscience*, vol. 24, no. 11, pp. 1555–1566, 2021.

[326] J.-F. Poulin, J. Zou, J. Drouin-Ouellet, K.-Y. A. Kim, F. Cicchetti, and R. B. Awatramani, "Defining midbrain dopaminergic neuron diversity by single-cell gene expression profiling," *Cell reports*, vol. 9, no. 3, pp. 930–943, 2014.

[327] K. Tiklová, Å. K. Björklund, L. Lahti, A. Fiorenzano, S. Nolbrant, L. Gillberg, N. Volakakis, C. Yokota, M. M. Hilscher, T. Hauling, *et al.*, "Single-cell rna sequencing reveals midbrain dopamine neuron diversity emerging during mouse brain development," *Nature communications*, vol. 10, no. 1, p. 581, 2019.

[328] G. Heymann, Y. S. Jo, K. L. Reichard, N. McFarland, C. Chavkin, R. D. Palmiter, M. E. Soden, and L. S. Zweifel, "Synergy of distinct dopamine projection populations in behavioral reinforcement," *Neuron*, vol. 105, no. 5, pp. 909–920, 2020.

[329] P. Redgrave and K. Gurney, "The short-latency dopamine signal: a role in discovering novel actions?," *Nature reviews neuroscience*, vol. 7, no. 12, pp. 967–975, 2006.

[330] A. Klaus, J. Alves da Silva, and R. M. Costa, "What, if, and when to move: basal ganglia circuits and self-paced action initiation," *Annual review of neuroscience*, vol. 42, pp. 459–483, 2019.

[331] J. A. Da Silva, F. Tecuapetla, V. Paixão, and R. M. Costa, "Dopamine neuron activity before action initiation gates and invigorates future movements," *Nature*, vol. 554, no. 7691, pp. 244–248, 2018.

[332] A. A. Moustafa, S. Chakravarthy, J. R. Phillips, A. Gupta, S. Keri, B. Polner, M. J. Frank, and M. Jahanshahi, "Motor symptoms in parkinson's disease: A unified framework," *Neuroscience & Biobehavioral Reviews*, vol. 68, pp. 727–740, 2016.

[333] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, and A. E. Lang, "Parkinson disease," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–21, 2017.

[334] J. C. Horvitz, "Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events," *Neuroscience*, vol. 96, no. 4, pp. 651–656, 2000.

[335] J. C. Horvitz, T. Stewart, and B. L. Jacobs, "Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat," *Brain research*, vol. 759, no. 2, pp. 251–258, 1997.

[336] W. Schultz and R. Romo, "Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions," *Journal of neurophysiology*, vol. 63, no. 3, pp. 607–624, 1990.

[337] E. S. Bromberg-Martin and O. Hikosaka, "Midbrain dopamine neurons signal preference for advance information about upcoming rewards," *Neuron*, vol. 63, no. 1, pp. 119–126, 2009.

[338] A. Lak, W. R. Stauffer, and W. Schultz, "Dopamine prediction error responses integrate subjective value from different reward dimensions," *Proceedings of the National Academy of Sciences*, vol. 111, no. 6, pp. 2343–2348, 2014.

[339] R. Bogacz, "Dopamine role in learning and action inference," *Elife*, vol. 9, p. e53262, 2020.

[340] S. Kakade and P. Dayan, "Dopamine: generalization and bonuses," *Neural Networks*, vol. 15, no. 4-6, pp. 549–559, 2002.

[341] S. M. McClure, N. D. Daw, and P. R. Montague, "A computational substrate for incentive salience," *Trends in neurosciences*, vol. 26, no. 8, pp. 423–428, 2003.

[342] S. N. Haber, "The place of dopamine in the cortico-basal ganglia circuit," *Neuroscience*, vol. 282, pp. 248–257, 2014.

[343] Y. Zhang, K. M.-H. Larcher, B. Misic, and A. Dagher, "Anatomical and functional organization of the human substantia nigra and its connections," *eLife*, vol. 6, p. e26653, aug 2017.

[344] B. A. Sorg and P. W. Kalivas, "Effects of cocaine and footshock stress on extracellular dopamine levels in the ventral striatum," *Brain research*, vol. 559, no. 1, pp. 29–36, 1991.

[345] B. Sorg and P. Kalivas, "Effects of cocaine and footshock stress on extracellular dopamine levels in the medial prefrontal cortex," *Neuroscience*, vol. 53, no. 3, pp. 695–703, 1993.

[346] P. W. Kalivas and P. Duffy, "Selective activation of dopamine transmission in the shell of the nucleus accumbens by stress," *Brain research*, vol. 675, no. 1-2, pp. 325–328, 1995.

[347] F. Brischoux, S. Chakraborty, D. I. Brierley, and M. A. Ungless, "Phasic excitation of dopamine neurons in ventral vta by noxious stimuli," *Proceedings of the national academy of sciences*, vol. 106, no. 12, pp. 4894–4899, 2009.

[348] K. Thurley, W. Senn, and H.-R. Luscher, "Dopamine increases the gain of the input-output response of rat prefrontal pyramidal neurons," *Journal of neurophysiology*, vol. 99, no. 6, pp. 2985–2997, 2008.

[349] W. Matsuda, T. Furuta, K. C. Nakamura, H. Hioki, F. Fujiyama, R. Arai, and T. Kaneko, "Single nigrostriatal dopaminergic neurons form widely spread and highly dense axonal arborizations in the neostriatum," *Journal of Neuroscience*, vol. 29, no. 2, pp. 444–453, 2009.

[350] C. Hammen, "Stress and depression," *Annu. Rev. Clin. Psychol.*, vol. 1, pp. 293–319, 2005.

[351] B. Neyshabur, R. Tomioka, and N. Srebro, "In search of the real inductive bias: On the role of implicit regularization in deep learning," *arXiv preprint arXiv:1412.6614*, 2014.

[352] J. Tian, R. Huang, J. Y. Cohen, F. Osakada, D. Kobak, C. K. Machens, E. M. Callaway, N. Uchida, and M. Watabe-Uchida, "Distributed and mixed information in monosynaptic inputs to dopamine neurons," *Neuron*, vol. 91, no. 6, pp. 1374–1389, 2016.

[353] A. Dickinson, "Actions and habits: the development of behavioural autonomy," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 308, no. 1135, pp. 67–78, 1985.

[354] D. T. Neal, W. Wood, and J. M. Quinn, "Habits—a repeat performance," *Current directions in psychological science*, vol. 15, no. 4, pp. 198–202, 2006.

[355] W. Wood and D. T. Neal, "A new look at habits and the habit-goal interface.," *Psychological review*, vol. 114, no. 4, p. 843, 2007.

[356] Y. Niv, M. O. Duff, and P. Dayan, "Dopamine, uncertainty and td learning," *Behavioral and brain Functions*, vol. 1, no. 1, pp. 1–9, 2005.

[357] K. J. Miller, A. Shenhav, and E. A. Ludvig, "Habits without values.," *Psychological review*, vol. 126, no. 2, p. 292, 2019.

[358] D. A. Burke, H. G. Rotstein, and V. A. Alvarez, "Striatal local circuitry: a new framework for lateral inhibition," *Neuron*, vol. 96, no. 2, pp. 267–284, 2017.

[359] L. Lapicque, "Recherches quantitatives sur l'excitation electrique des nerfs," *J. Physiol*, vol. 9, pp. 620–635, 1907.

[360] W. Schultz, "Predictive reward signal of dopamine neurons," *Journal of neurophysiology*, 1998.

[361] N. A. Roy, J. H. Bak, A. Akrami, C. D. Brody, and J. W. Pillow, "Extracting the dynamics of behavior in sensory decision-making experiments," *Neuron*, vol. 109, no. 4, pp. 597–610, 2021.

[362] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs," 2018.

[363] A. Wu, N. A. Roy, S. Keeley, and J. W. Pillow, "Gaussian process based nonlinear latent structure discovery in multivariate spike train data," *Advances in neural information processing systems*, vol. 30, 2017.