Dopaminergic signalling during goal-directed behaviour in a structured environment



Marta Blanco-Pozo

Exeter College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2022

Dopaminergic signalling during goal-directed behaviour in a structured environment

Marta Blanco-Pozo
Exeter College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy
Trinity term 2022

Abstract

During flexible behaviour, dopamine is thought to carry reward prediction errors (RPEs), which update values and hence modify future behaviour. However, in real-world situations where the statistical relationships in the environment can be learned, continuously adapting values is not always the most efficient way of adapting to change. Moreover, the environment is not always fully observable, and observations may provide only partial information about the current state of the world. In such partial observable structured environments, as is found in many real-world situations, it is not well understood what kind of information dopamine conveys or its causal role in shaping adaptive behaviour. To probe dopamine's involvement in goal-directed behavioural flexibility in such environments, we measured and manipulated dopamine while mice solved a partially observable structured sequential decision task. In chapter 3 we show that mice solve such a task using state inference. In chapter 4, we recorded calcium activity from dopaminergic cell bodies in the ventral tegmental area and dopamine axonal projections in the ventral and dorsomedial striatum, as well as dopaminergic concentrations in the same striatal regions. Dopamine multiplexed a wide range of information. At different timescales dopamine signalling was consistent with carrying information about choice-specific RPEs, choice-independent reward history and lateralised movement signals. RPE computations were shaped by task structure and the inferred state of the task. Nonetheless, in chapter 5, we show that although dopamine responded strongly to rewards, optogenetically activating or inhibiting dopamine at the time of trial outcome had no effect on subsequent choice. However, in a different task context, we could show that the same stimulation had a substantial effect on animals' choices. Therefore, we conclude that when inference guides choice, rewards have a dopamine-independent influence on policy through the information they carry about the world's state.

A preprint of much of the experimental work described in this thesis can be found at:

Blanco-Pozo, M., Akam, T., & Walton, M. (2021) Dopamine reports reward prediction errors, but does not update policy, during inference-guided choice. *bioRxiv.* https://doi.org/10.1101/2021.06.25.449995

This research was funded by the Wellcome Trust Four-Year PhD Studentship Programme.

Acknowledgments

If someone had told me a few years ago that I would write this DPhil dissertation, I would not have believed it. This has been a real adventure, in which I not only grew scientifically, but also personally. It is an adventure I would definitely repeat, thanks to all the wonderful and amazing people I have crossed paths with during this journey. I need to thank so many people, that I am scared I will not be able to convey it with words.

I would like to begin by thanking my DPhil assessors Prof. Rafal Borgacz and Prof. Joe Paton. Thank you very much for your time and generosity in offering me such a wholesome experience during the viva, allowing me to fully discuss my project with you. I enjoyed it very much and it was interesting to see the different views and interpretations of our data. I would also like to thank them for catching the typos and for their suggestions to improve the thesis. I would also like to thank Prof. Rafal Borgacz together with Prof. Andrew Sharott for being my transfer and confirmation assessors, for their comments and suggestions, and for creating such a safe place for me to present my DPhil. It never felt like an exam, but rather a place to discuss my research with experts. I always had nice discussions.

Then, I would like to express my immense gratitude to my supervisors, Dr Thomas Akam, Prof. Mark Walton, and Prof. Tim Behrens. Thank you so much, Thomas, for allowing me to work on your 'science baby', the two-step task. Thank you for teaching me everything I know now. I started my DPhil with little to no knowledge in coding, computational modelling, and similar fields, and I finished it with coding being something I enjoy so much that I sometimes use it to 'procrastinate', a full chapter on behavioural modelling, and a full thesis of analysis and results that I would not have even dreamed about. So, thank you so much, Thomas, for all your time, patience, and teachings during this journey. I know that one of the reasons I am so happy with my DPhil is that I had the opportunity to closely work with you and your project, so I cannot reiterate enough how grateful I feel for that.

The other pivotal person in my successful DPhil journey is Prof. Mark Walton. Thank you so much, Mark, for always having time for me, listening, advising, and caring about me. So, thank you for being such a great mentor. Something that has helped me so much, at the same time it has always fascinated me, is your power to read my mind. Whenever anything was bothering me or running through my head, before I would come to you, you would have already come to me. Thank you very much for being so generous with me. I always felt so lucky and supported being in your lab, that it is now very hard to leave. Not to forget to thank you, Mark, so much for all your corrections and comments on the thesis, some parts of which I sent you very, very last minute and checking until the last moment if I successfully managed to submit the thesis. To both Thomas and Mark, I also need to thank them for always sharing and discussing their ideas with me. I owe my scientific growth to them. I would also like to thank Prof. Tim Behrens for, although me being the dopamine outlier, exposing me to other exciting (some hardcore computational) fields of research through the lab retreat and joint lab meeting. Also, thank you, Tim, for your time in discussing our results at the beginning of the project.

I would also like to thank the Walton lab in general for being a supportive group of people, creating such a great environment, and some of them even becoming close friends whom I will miss very much. In particular, I would like to thank the 'girls' gang', Beatriz Godinho, Raquel Pinacho and Merima Šabanović. They became such a big support in so many aspects that I cannot thank them enough, without forgetting all the fun and the undisclosable chats. Thank you so much, girls, for all the moments we shared together. I will miss you very much. Something I will also miss are the conversations fixing the world in the BNU with Beatriz and Juan Carlos Cerpa. Those conversations made the long training hours so much fun. On that note, I also want to thank Juan Carlos for putting up with me and always being ready to teach me self-defence. I will always remember when Beatriz joined the lab. You were (and are) such a bright light that is contagious. It has been difficult not to see you and chat with you every day in the lab since you left for 'fancy' London. Thank you, Beatriz, for everything and for making me feel quite unstoppable when we were together. Raquel, although she joined the lab later, thank you for becoming a big pillar and support for me. Merima, thank you for also becoming such a support, and your determination and strength are really inspiring.

Although Laura Grima left the lab some time ago, I will not forget when I first started in the lab. Thank you for being (and continuing to be) so kind to me. I learned a lot from you. I would also like to thank Lauren Burgeno for her energy and gossip. It made the long training hours more entertaining. Also, thank you so much Lauren for all the fun we had at the conference in Canada and for introducing me to everyone. It was very nice to finally meet many people in person. Finally, I want to thank Marios Panayi for being the stats guru when he was in the lab and for his help in introducing me to mixed models, which became a big part of the optogenetic results in my thesis.

Something that has also shaped my DPhil and allowed me to grow both personally and scientifically is being a part of the Cortex Club. I will not say being a part of the committee was not stressful at times, but it allowed me to develop management, social, and networking skills that will definitely help me in the future. Moreover, it allowed me to meet and work with amazing people. So, thank you very much, everyone.

Getting to where I am now would not have been possible without the initial support of Dr Johan Alsiö. Thank you very much for allowing me to come to Cambridge to work with you when I was just a final year undergraduate. That experience truly opened a door for me. Thank you very much Teresa Calafat, Marina Martínez, Maria Bacadare, Laura Lopez, and all the amazing people I shared the Cambridge experience with, which made it unforgettable and gave me the positive energy to continue in a foreign country far from family and friends. Afterwards, in London during my MSc, working with Dr Dimitar Kostadinov was invaluable. Thank you very much Dimitar for teaching me so much. I always felt heard and supported, and thank you for being so open with me with all your ideas. Since then, I have become extremely interested in the cerebellum, although I was studying the dopamine system during my DPhil, I became somewhat of the 'crazy cerebellum person' in the lab. But also, working with Dimitar showed me the kind of research I am truly interested in, circuits neuroscience.

Since London, a person whom I feel very fortunate to have had all this time is Tai-Ying Lee. Thank you, Tai, for always being there. Since we met in London and then moved together to Oxford (what a coincidence! Without even knowing it at the beginning), I feel we have always

had each other's back. Thank you very much for all the fun, your continuous support, kindness, and friendship.

I also feel extremely fortunate to have shared this journey at my side with my Tommy. Thank you for growing with me through this process, from London to Oxford, again, not planned, but it just happened. Thank you for being there for me, supporting me throughout, providing all the fun, caring about me, and in sum, for being such a great life partner. This adventure would not have been the same without you.

A group of people whom I also want to thank are my SHs, Marta Badia, Mireia Pons, Laura Pérez, Miriam Rivera, Adri Mallén, Hugo Ramos and Jordi Soldevila. Thank you for being such unconditional friends all this time, no matter the distance, thank you for always being ready to put up a fight for me. I admire you all on so many levels, and I feel extremely fortunate to have you, I really do not know what I would do without you. Marta, I also want to say thank you for coming to visit me on several occasions (such unforgettable times together) and always being there for me, despite you being busy. Thank you for always juggling time for me.

I also want to thank my family, my father, my mother, and my siblings, Jordi and Laura, to whom I dedicate this thesis. Thank you very much for your unconditional support and cheering throughout my life. Thank you for always being there for me, for your encouragement and strength, despite sometimes it is hard for you as it means me being far away. Thank you very much for providing me with everything, for all your moral support and love. I do not know how to thank you enough.

Finally, I would like to end this acknowledgement section, which is longer than I initially expected, by thanking my high-school biology teacher, Àngel Lillo. He was one of those teachers in high school who leaves a mark and guides you. There is something he wrote to me when I finished school, that at the end of my DPhil became very useful for me, and motivated me in a moment when I needed a push. He wrote to me: 'You have made the first step towards the science path, I wish it takes you far, but most importantly, I wish you enjoy it' (translated as much as I could). This last 'I wish you enjoy it', made me wake up in a low moment and realize the reason why I am in research, and gave me back the excitement of science. The fun in science is what has motivated me throughout, and I needed this small reminder.

I probably have missed many people I need to thank, so thank you to everyone who made this experience so amazing and fulfilling.

Table of Contents

1 Introduction 1

1.1 Evolution of animal psychology 2
1.1.1 Where do we come from: pick a side, association vs cognitive theories 2
1.1.2 From single-process to dual-system theories: habit vs actions 11

1.2 Behaviour through the lens of reinforcement learning 15
1.2.1 Rescorla-Wagner model and temporal-difference model in classical conditioning 16
1.2.2 Temporal-Difference Reinforcement Learning during decision-making problems 21
1.2.3 Planning through the acquisition of a model of the environment 25
1.2.4 Towards a computational framework of instrumental behaviour: model-free vs model-based and others 28

1.3 Dopamine system 39
1.3.1 A brief journey through dopaminergic anatomy 39
1.3.2 Differential effects of dopamine through firing patters and receptors distribution 42
1.3.3 Dissociation between dopaminergic activity and actual dopamine release 44

1.4 Dopamine and behaviour 48
1.4.1 The reward prediction error hypothesis 49
1.4.2 The motivation account of dopamine, a challenge to the RPE hypothesis? 51
1.4.3 Dopaminergic role in motor control 55
1.4.4 Internal model representation in the dopaminergic system 59
1.4.5 Diverse functions and interactions 67

1.5 Thesis aims 69

2 General methods 71

2.1 Subjects 72

2.2 Two-step task 73
2.2.1 Behavioural setup 73
2.2.2 Behavioural task 73
2.2.3 Behavioural training 75
2.2.4 Behavioural analysis 76
2.2.4.1 Logistic regression model 77
2.2.4.2 Lagged logistic regression model 78
2.2.5 Optimising the two-step design 79
2.2.5.1 Two-step task with reversal of reward and transition probabilities 81
2.2.5.2 Two-step task with reversal of reward probabilities ONLY 85

2.3 Surgery 87
2.3.1 General surgical protocol 87
2.3.2 GCaMP photometry surgery 88
2.3.3 dLight photometry 88
2.3.4 Optogenetic activation 89
2.3.5 Optogenetic inhibition 89

2.4 Mesolimbic and striatal dopamine recordings and manipulation 89

2.5 Analysis 90

2.6 Histology 90

3 Can animals use inference learning to guide behaviour in complex structured environments? 92

3.1 Introduction ... 93

3.2 Methods ... 95

3.2.1 Single RL behavioural models ... 96

3.2.1.1 Model-free ... 96

3.2.1.2 Model-based ... 97

3.2.1.3 Correlated inference ... 98

3.2.1.4 Reward-as-cue ... 99

3.2.1.5 Bayesian inference ... 100

3.2.1.6 Combined action values ... 102

3.2.1.7 Model fitting and comparison ... 103

3.2.2 Flexible RL behavioural model ... 105

3.2.2.1 Second-step value update ... 106

3.2.2.2 Action value update ... 107

3.2.2.3 Model fitting ... 109

3.3 Results ... 110

3.3.1 Mice behaviour in the two-step task is consistent with having an internal model of the task 110

3.3.2 Inference strategies outperform traditional RL models in explaining mice choice behaviour ... 114

3.3.3 Inference and not model-based with forgetting explains behaviour in the two-step task 119

3.4 Discussion ... 124

3.4.1 Rodent two-step task: model-based or inference ... 124

3.4.2 Asymmetric value updating ... 127

3.4.3 Conclusion ... 128

Tables ... 129

4 How do complex structured environments and animals' decision policies influence the dopamine system? ... 138

4.1 Introduction ... 139

4.2 Methods ... 142

4.2.1 Fibre photometry ... 143

4.1.1.1 Data acquisition ... 143

4.1.1.2 Data processing ... 143

4.1.1.3 Trial alignment ... 144

4.2.2 Linear regression model ... 145

4.1.1.4 Behavioural regressors ... 146

4.1.1.5 Inference reinforcement learning model value regressors ... 148

4.2.3 Exclusion criteria ... 148

4.3 Results ... 149

4.3.1 Mesostriatal dopamine activity and release have comparable dynamics ... 149

4.3.2 Dopamine reflects experienced and inferred value consistent with the RPE framework 154

4.3.3 Decision policies and action sequence are orthogonally reflected in the dopamine system 159

4.3.4 Dissociable influence of RPE and reward expectation in dopamine signalling ... 166

4.3.5 Other non-value related information also modulates dopamine 171

4.4 Discussion 173

4.4.1 Dopamine during latent state inference guided choice 174

4.4.2 RPE and motivation account coexistence in the same population signalling 175

4.4.3 Movement and other task parameters in dopamine 177

4.4.4 Dopamine differences across regions, activity and release 179

4.4.5 Conclusion 181

Supplementary figures 182

5 How does dopamine influence behaviour in complex structured environments? 184

5.1 Introduction 185

5.2 Methods 186

5.2.1 Behavioural task – The two-step task 187

5.2.2 Optogenetic manipulation 187

5.2.2.1 Intracranial self-stimulation (ICSS) assay 188

5.2.2.2 2-alternative forced choice task (bias task) 188

5.2.2.3 Activation and inhibition in the two-step task 189

5.2.3 Logistic regression model 189

5.3 Results 190

5.3.1 Dopamine stimulation promotes self-stimulation 190

5.3.2 Dopamine inhibition promotes avoidance behaviour 191

5.3.3 Dopamine stimulation does not recapitulate the effects of natural rewards despite reinforcing preceding actions 194

5.3.4 Dopamine inhibition neither promotes a change in policy 197

5.4 Discussion 198

5.4.1 Dopamine is reinforcing but does not substitute the effects of natural rewards 199

5.4.2 Optogenetic manipulation limitations 201

5.4.3 Conclusion 202

6 Discussion 203

6.1 Summary 204

6.2 Can inference accounts improve the understanding of flexible behaviour? 205

6.3 Rethinking the role of dopaminergic RPE in policy update 210

6.4 Dopamine a multiplex signal: RPE, motivation and movement 218

6.5 Positive and negative coding, an intrinsic property of the dopaminergic system? 221

6.6 Future directions 222

6.7 Conclusion 227

***REFERENCES* 229**

1 Introduction

The aim of this thesis is to investigate the psychological and neural processes underlying flexible behaviour. Specifically, we will focus on the dopaminergic system. Therefore, in this introduction, we will first examine foundational work investigating behavioural flexibility from psychological, computational, and neural perspectives. Secondly, we will describe the computational framework widely used to formally define reward-guided behaviour. Lastly, we will discuss the heterogeneity within the dopaminergic system, and its functional implications supporting flexible behaviour.

1.1 Evolution of animal psychology

Citing Carl Sagan, 'You have to know the past to understand the present'.

It is difficult to define the origin of the study of animal behaviour, but several landmarks have been of immense importance to guide us to the current state of thinking: i) Charles Darwin's theory of evolution (with the publication of *On The Origin of Species* (1859)), ii) Conwy Lloyd Morgan's 'trial-and-error' learning and association (1894), and iii) Edward Thorndike's experimental method (1898) (Warden, 1927). This does not mean that since the end of the 19th century, no developments occurred, but the opposite, that the study of animal behaviour expanded from this foundation. During the 20th century, two different schools of thoughts were in opposition: the associative accounts originated from Thorndike's ideas, and the cognitive map account from Tolman. Through studying the psychological processes underlying habitual and goal-directed behaviour, single-process theories turned into dual-systems, where both associative and cognitive accounts interact to generate behaviour. More recently, in order to computationally formalise aspects of habitual and goal-directed behaviour, the reinforcement learning framework is being used. Another dichotomy appeared, model-free versus model-based control. But again, this dichotomy diffused and merged with time.

1.1.1 Where do we come from: pick a side, association vs cognitive theories

The publication of *On The Origin of Species* (Darwin, 1859) posted a departure from previous philosophical and theological perspectives of explaining animals' behaviour. Darwin, together with other naturalists like Romanes (Romanes, 1886) and Lubbock (Lubbock, 1889), started

to consider human and animals' minds as comparable, differing in "degree but not kind". Animal behaviour was explained in terms of instincts or mental powers, reason, judgement and intelligence. However, at that time, the evidence used to support this was anecdotal.

> *'It would be easy to show that several distinct mental actions are commonly embraced by this term [instinct]; […] An action, which we ourselves require experience to enable us to perform, when performed by an animal, more especially by a very young one, without experience, and when performed by many individuals in the same way, without their knowing for what purpose it is performed, is usually said to be instinctive. But I could show that none of these characters are universal. A little dose of judgment or reason, as Pierre Huber expresses it, often comes into play, even with animals low in the scale of nature.'*
>
> *Darwin C., 1859. On The Origin of Species. Chapter VII Instinct.*(Darwin, 1859)

Against this approach of anthropomorphising and humanising animal behaviour through anecdotal evidence, Morgan argued for a more objective and systematic approach to studying *animal intelligence* (Morgan, 1886).

> *'In the study of Animal Intelligence, therefore, I would counsel—*
> *(a) The separation of fact [certain actions performed under certain external circumstances] from inference [which are drawn from the facts].*
> *(b) The distinguishing between objective and ejective [subjective] inference.*
> *(c) The elimination, so far as is practicable, of the ejective element.'*
>
> *Morgan C., 1886. II.-On The Study Of Animal Intelligence. Mind. Vol.os-XI, Issue 42 [p.185]*(Morgan, 1886)

Morgan's 'Canon of Interpretation' went away from interpreting animals' behaviour as governed by *reason* to explanations that focused on associations between a stimulus and a response through 'trial and error' learning, akin to what we currently call *instrumental learning*. However, for Morgan, this association was an association of ideas. Through experience, association occurs between sense-impressions of the world or as a result of actions. When later, a sense-impression is perceived, through memory, the linked idea of sense-experience (of another stimulus, action, concept…) is retrieved (Morgan, 1886, 1894).

> *A dog is described as "showing a deliberate design of deceiving" because he hobbled about the room as if lame and suffering pain from his foot. I would suggest that there was no pretence in this case, but a* ***direct association of ideas*** *between a hobbling gait and more pity and attention than usual.*
>
> *Morgan C., 1886. II.-On The Study Of Animal Intelligence. Mind. Vol.os-XI, Issue 42 [p.180]*(Morgan, 1886)

Another example that Morgan gives is an experiment from John Lubbock, who trained his black poodle to distinguish cards with words like 'Food' written in them and reach them when it wanted food in this case. This was another example of how animals associate a particular sense (e.g. word food) with the idea of a particular experience (e.g. obtaining food)(Morgan, 1894).

Although Morgan made a step forward in systematising the study of comparative psychology and departing from the more anthropomorphic view of animals' minds, it was Thorndike who diverged most radically from the prevailing views of the period, strongly criticising the use of anecdotal evidence and conferring animals with the ability of *reason*. Influenced by Lubbock's

experimental apparatus to study the behaviour of insects (Warden, 1927), Thorndike pioneered the use of controlled experiments to study *animal intelligence* (although probably in parallel, other laboratories at the time were implementing a similar approach to study comparative psychology (Warden, 1930)).

Thorndike constructed 'puzzle boxes' (precursor of Skinner box or operant box) where chickens, dogs, and cats (and later macaques) were closed in, and they had to find a way to get out (by pushing levers, strings…) to get a food reward. He described that animals were first using 'instinctive impulses' to try to escape the box until randomly they would perform an action that would open the door. After experiencing the same box multiple times and succeeding in escaping it, animals would take a shorter and shorter time to get out (Thorndike, 1898).

At that time, the current view of *animal intelligence* would have explained animals' successes in escaping the box by animals using reasoning (Romanes, 1886; Morgan, 1894) – animals connect or associate stimuli to *ideas* of actions that were performed in the past. But Thorndike, going against anthropomorphising animals' behaviour, explained it as animals connecting stimuli to the actual actions that freed them. He proposed that animals were not using reason, but just using instinctive behaviours like scratching in order to escape. In this set of random actions, by chance (not deliberately), one of the actions will open the door of the box. The animal will then form an *association* between being in the box (*sense-impression*) and the action (*impulse*) that liberates it (*successful movement* - to later get food) which is rewarding. Therefore, animals will strengthen connections between stimuli and responses that were rewarded (opening the door), while weakening the connections of the other

actions. After many exposures to the box, actions that previously result in opening the door were selected and mainly performed, while the others were suppressed, taking a shorter time for the animals to escape.

As a result of Thorndike's experiments, he postulated laws and hypotheses of behaviour and learning (Thorndike, 1898, 1911). Among these, was the 'Law of Effect':

> *'Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that when it recurs, they will be less likely to occur. The grater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.*
>
> *Thorndike, 1911. Animal Intelligence. Experimental Studies [p.244]*(Thorndike, 1911)

Although the concepts of *association*, *trial and error learning* and the use of controlled experiments were adopted from his predecessors, Thorndike's interpretation of them was different and drove a paradigm shift in the study of animal behaviour.

> *'The psychology of animal learning – not to mention that of child learning – has been and still is primarily a matter of agreeing or disagreeing with Thorndike, or trying in minor ways to improve upon him. […] all of us in America seem to have taken Thorndike, overtly or covertly, as our starting point'*
>
> *Tolman, 1938. The Psychological Review. Vol.45, No. 1 [p.11]*(Tolman, 1938)

It should be noted that in parallel to Thorndike's experiments, Ivan Pavlov was also conducting controlled experiments that led to comparable conclusions: learned behaviour is the result of an association or connection between stimuli. Pavlov's experiments show that dogs would salivate in response to a sound, which he called *conditioned reflex*. For Pavlov, animals had different types of *reflexes* – responses to external stimuli. An example of those is instincts. These *reflexes* are automatically produced in response to specific stimuli (e.g., a dog would salivate in response to food). Pavlov called the stimulus that *per se* causes a response an *unconditioned stimulus* (e.g. food). However, if a neutral stimulus like a sound is associated with an unconditioned stimulus like food which generates salivation; after learning the association that a sound precedes food, now the sound would directly generate the salivation. In this case, the sound will become a *conditioned stimulus*, and the salivation in response to this is the *conditioned reflex* (Pavlov, 1927). Although there are similarities between Thorndike's and Pavlov's theories, they were studying different mechanisms, instrumental and classical conditioning, respectively.

Going back to Thorndike's view on the study of animal intelligence, Edward Tolman was one of those that disagreed with him and his school of thought. Although he supported the use of controlled experiments to explain animal behaviour, Tolman switched back to humanising animals, despite Thorndike's efforts for the opposite. However, in contrast to Thorndike's predecessors, Tolman believed there was value in studying animals' behaviour from a human perspective in order to get 'hunches' which could later be formalised in an objective way.

> *'I in my future work intend to go ahead imagining how, if I were a rat, I would behave as a result of such and such a demand combined with such and such an appetite and such and such a degree of differentiation; and so on. And then, […] I*

shall try to figure out some sort of f_3 rules or equations. And [..] state these latter in some kind of objective and respectable sounding terms […].'

Tolman, 1938. The Psychological Review. Vol.45, No. 1 [p.24](Tolman, 1938)

More importantly, Tolman opposed Thorndike's (and others like Clark Hull) 'simple stimulus-response associations or connections'. Tolman, in something of a mocking manner, referred to them – the associationism or behaviourism school – as the 'telephone switching school', who believe that there are a defined set of actions that get connected to a certain stimulus by strengthening or weakening through experience after positive or negative outcomes, respectively. However, for Tolman and other self-referred as 'field theorists':

'*we assert that the central office is far more like a map control room than it is like an old-fashioned telephone exchange'*

Tolman, 1948. The Psychological Review. Vol.55, No. 4 [p.192](Tolman, 1948)

For Tolman, during learning, a cognitive-like map of the environment is constructed, which contains *routes* and *paths* and the *relationship between* them, controlling future actions (Tolman, 1948). To prove the cognitive map account, he and others performed a series of experiments, mainly on rats, involving exposing them to mazes and tracking their behaviour across many days.

For example, one influential set of experiments, which simple stimulus-response association could not straightforwardly explain, examined *latent learning*. In such experiments, different groups of rats were exposed to mazes. For example, in one experiment 3 different groups of rats were placed in a 14-unit T-alley maze, in which food could be found in a specific location

(Figure1.1.A). The HNR group was exposed to a maze without food in it, used as a control group. The HR group was a second control group which was exposed to the maze where a reward (i.e., food) could be found. This group after several days progressively reduced the number of errors performed and directly took the path that would get them to food. It was counted as an error when animals took the wrong path to get food. In the test group, HNR-R, in the first days, the maze did not have any food in it, therefore, rats were performing many 'errors' (these errors were fictitious as no food was yet in the maze). However, when in a later session food was introduced in the maze, as soon as they found the first reward, the errors decayed much faster than in both control groups. Decayed faster than the group that had food from the first day (Figure 1.1.B). This showed that animals were also learning, and building up a cognitive map of the maze in the absence of reinforcement (i.e. despite behaviour not being reinforced as food was not present in the maze), which allowed them, as soon as they were motivated to return to a specific place, to do it by using the already built map of the physical space. Another experiment supporting this conclusion involved exposing non-restricted rats to a maze with two paths, one of which led to food and the other to water. However, as soon as the rats were either food or water restricted, when returned to the maze (this time without either food or water present in the goal arms), they preferentially took the path that previously contained food or water, respectively (Tolman, 1948). These experiments emphasised that motivation was not needed for learning, but for performing what the experimenter wanted the animal to achieve.

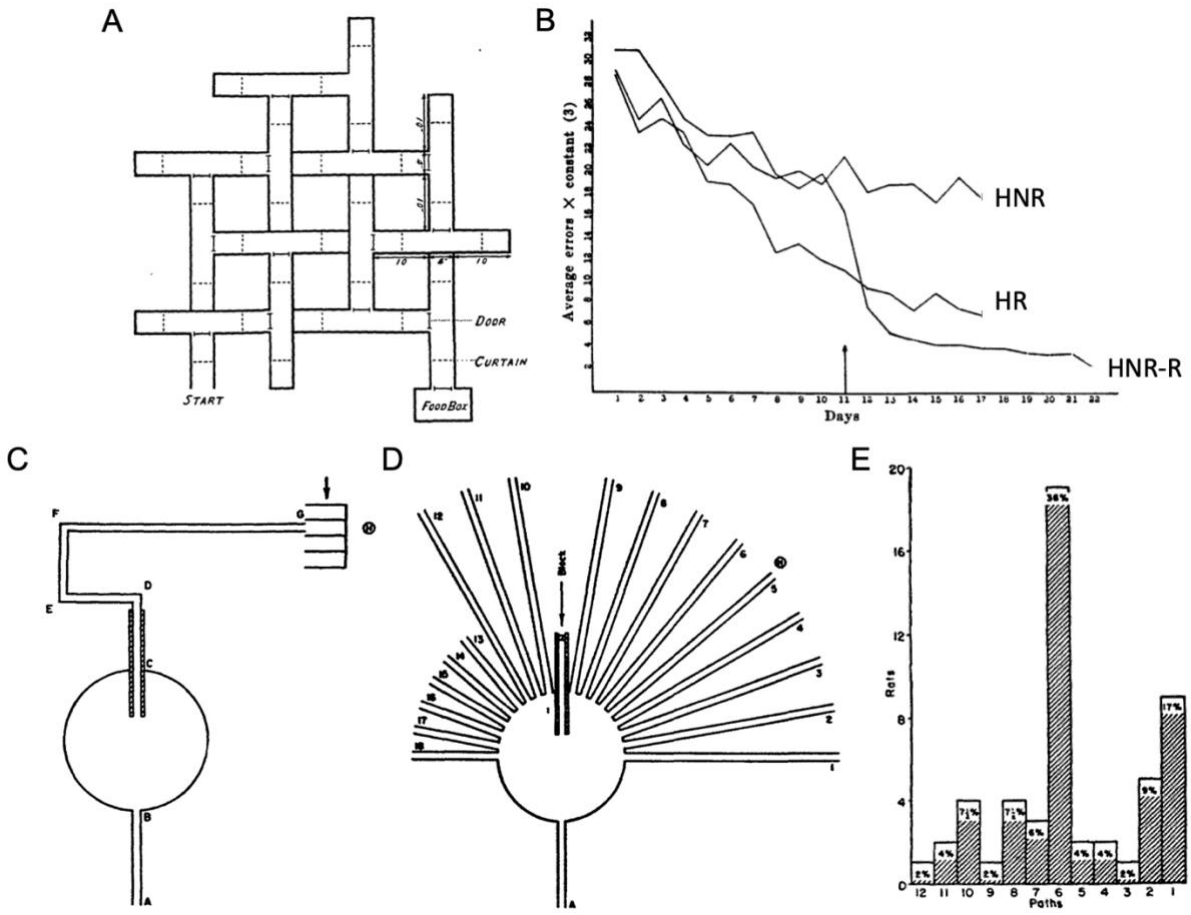


Figure 1.1. Experiments performed by Tolman and colleagues to support the cognitive map account. A) 14-unit T-alley maze. Rats were placed in the start location. Rats could explore the maze. In the groups that food was in the maze, this was placed in the Food Box location. B) Number of errors across times for 3 groups of rats placed in A. Error was considered when rats took the wrong path to get too the Food Box. HNR group, the maze never contained food. HR group, food was placed in the Food Box from the first dat. HNR-R group, food was introduced in the maze on day 11 (vertical arrow). C) Spatial maze used for initial training. Food could be found at the end (circled H). D) Spatial maze during test session. Circled H corresponds to the same spatial location as in C. E) Number of rats that chose each path in D. Figures from Tolman (1948) of experiments performed in previous studies.

The spatial orientation experiments were also very significant in supporting the 'cognitive map' hypothesis. In these experiments, rats were exposed to a specific maze layout to get reward (Figure 1.1C). Once learned, the maze was changed to include new routes (Figure 1.1D), and it was found that rats preferentially chose paths that spatially would lead them to the same or close spatial location to the previous maze (Figure 1.1E). This revealed that rats

were building a *flexible* spatial map than went beyond just the experienced paths (Tolman, 1948).

Despite initial Tolman's hard opposition to the stimulus-response school of thought, later, he postulated that there could be different kinds of learning, some of them reconciling with the 'association' school (Tolman, 1949). Others also aided in reconciling the different schools, and slowly, single-process theories were replaced by theories involving the different processes: Pavlovian conditioning, 'Thorndike's' S-R and 'Tolman's' cognitive maps (Rescorla and Solomom, 1967; Trapold and Overmier, 1972; Dickinson, 1985; Dickinson and Balleine, 1994; Dickinson and Pérez, 2018). The question then became: when are different learning processes used to guide behaviour and how could this be shown experimentally?

1.1.2 From single-process to dual-system theories: habit vs actions

Different kinds of experiments have been used to determine the processes guiding instrumental behaviour. One of these was *reinforcer devaluation*. The goal of these experiments is to test whether animals are learning the instrumental contingency and know the specific consequences of their actions (i.e., that pressing the lever causes water delivery). In such experiments, there is an initial instrumental training period where animals are trained to perform an action in order to get an outcome (e.g. press a lever for water reward). After the instrumental behaviour is established, that outcome is devalued (e.g., by giving the animal free access to that reinforcer so the outcome loses its reinforcing value). This is tested during an extinction session where the animal is placed in the initial instrumental context but now in absence of outcome delivery, and their responses are recorded. If the behaviour was simply

the result of a simple stimulus-response association, devaluation of the outcome in another context should have no effect on the level of responding in the extinction session. What would be learned/reinforced in this case is the association between a stimulus (the lever) and a response (pressing the lever), which is rewarding. On the other hand, if the behaviour is the result of the construction of a cognitive map of the environment, and the instrumental contingency is learned, after the devaluation procedure, the subject would decrease responding as that outcome is not valuable anymore.

Although it might have seemed quite straightforward to resolve the processes animals are using, different studies reached different conclusions. While some devaluation and incentive experiments supported the cognitive account (Miller, 1935; Tolman and Gleitman, 1949; Krieckhaus and George, 1968), some others failed to devaluate/incentivise the outcome, supporting the associative account (Tolman, 1933; Rozeboom, 1957; Dickinson et al., 1983). What could be the reasons for these contradictory results? Tolman (1933) suggested that both accounts could be correct and that the different stages of learning could have an impact on whether *associative* or cognitive interpretation explained behaviour. In this study, Tolman hinted that his negative results could have arisen because of animals being *fixated* due to overtraining, which foreshadows the habit vs goal-directed control.

The discrepancy in the conclusion of the different studies was due to the design. Dickinson and Nicholas (1983) showed that rats' behaviour in both devaluation and incentive experiments that appeared to be driven by their knowledge of the action to outcome mapping like in Krieckhaus and George (1968), might not be actually the case. After replicating Krieckhaus and George (1968) results, they tested whether rats actually had knowledge of

the consequences of their actions, and that pressing the lever delivered the reward. They tested that by delivering in different groups of rats the same outcomes as in the previous study, sodium or potassium solution vs water, however, in here, one outcome was delivered contingent on lever pressing, while the other was delivered independently/non-contingent on lever pressing. They tested rats' lever press behaviour during extinction after sodium appetite was induced (through injection of furosemide and desoxycorticosterone under anaesthesia). No difference in responding was found between the two groups that received the sodium solution contingent or non-contingent on lever pressing. The only difference found was between the sodium and the potassium groups (independently if these outcomes were delivered contingent or non-contingent on lever pressing). The groups that received sodium solution responded more than the two potassium groups. The fact that rats' response did not vary based on whether their behaviour produced the response or not, favoured the conditioning account against the cognitive one in explaining animal behaviour.

As mentioned by Dickinson and Nicholas (1983), their results could be explained by the two-process theory (Rescorla and Solomom, 1967; Trapold and Overmier, 1972). This theory postulates that there are two-processes guiding behaviour, a pavlovian and an instrumental process. Pavlovian conditioning controls instrumental responding by acting as a motivational drive of the conditioned response (Rescorla and Solomom, 1967; Trapold and Overmier, 1972; Dickinson and Balleine, 1994). In this case, the sodium solution acts as the unconditioned stimulus and other contextual cues, including the actual lever pressing, acquired the conditioned response. Therefore, when the animal is put in the context associated with a sodium solution, it acts as a motivational drive to press the lever, without the need of learning the instrumental contingency, and the subjects press the lever independently if it was

contingent or not to sodium solution delivery. On the other hand, animals trained with the potassium solution did not have the motivational drive when put in the context and press the lever at a significantly lower rate.

However, this does not mean that animals are not able to learn instrumental contingencies and that all animal behaviour is the result of a simple stimulus-response association. As Tolman suggested, both association and cognitive learning could be two possible processes that could manifest in different situations (Tolman, 1933). For example, Adams and Dickinson (1981) showed if the potential pavlovian influence of the contingent and non-contingent outcome was matched, the same amount of each of the two are delivered, then animals are able to learn the instrumental contingency. During extinction, the group that had the contingent reward devalued performed fewer responses than the group which had the non-contingent reward devalued. Another aspect affecting animals' ability to learn the action-outcome contingency is the reduction in the correlation between action and reward rate (Dickinson et al., 1983), which is a consequence of overtraining. Therefore, the effect of outcome devaluation can be affected by the reinforcement schedule (Chen and Amseli, 1980) or the amount of training, shifting goal-directed actions into habitual responding, insensitive to outcome devaluation (Adams, 1982; Dickinson et al., 1995), although in more complex tasks, the value of the outcome or reinforcer still stays present but in a smaller degree (Colwill and Rescorla, 1985a; Holland, 2004).

The current view of animal behaviour supports a dual-system theory in which instrumental behaviour is controlled by two parallel systems: goal-directed and habitual responding. This dual-system is separate from the two-process theory, which explains how pavlovian

conditioning controls instrumental behaviour through a motivational drive (Dickinson, 1985; Dickinson and Pérez, 2018). Through lesion and inactivation experiments, studies have shown that parallel circuits are involved in each process. For example, disrupting dorsolateral striatum (DLS) impaired habit formation (Yin et al., 2004, 2006). Conversely, goal-directed behaviour was impaired after lesioning or inactivation of dorsomedial striatum (DMS) (Yin et al., 2005). The medial prefrontal cortex (mPFC) was also a region that when lesioned, disrupted learning the action to outcome mapping (Balleine and Dickinson, 1998; Corbit and Balleine, 2003). However, once the instrumental contingency is learned, lesioning prelimbic disrupted goal-directed behaviour (Ostlund and Balleine, 2005), while lesioning infralimbic did not (Killcross and Coutureau, 2003). Other aspects such as motivation, vigorosity of actions and different aspects of Pavlovian behaviour were also controlled by prefrontal regions, ventral striatum and amygdala. Not only these regions were involved in the different behavioural responses in the rat, but also in their homologue regions in human and non-human primates (Cardinal et al., 2002; Balleine and O'Doherty, 2010).

1.2 Behaviour through the lens of reinforcement learning

As exposed in the previous section, there is a current interest in mathematically formalising the computations implemented at the behavioural, neuronal and circuitry levels. A framework discussed here is the use of reinforcement learning (RL) algorithms to explain behaviour and brain activity. In this section, we will go through some concepts in RL that have guided psychological and neuroscience research.

1.2.1 Rescorla-Wagner model and temporal-difference model in classical conditioning

Pavlovian conditioning, also known as *classical conditioning*, refers to the process in which a neutral stimulus (*conditioned stimulus,* CS) can generate a reflex response (*conditioned response*) such as salivary or fear conditioning or response, normally as a result of an appetitive or aversive stimulus (*unconditioned stimulus*, US) like food or an electric shock, respectively, when both stimuli (CS and US) are paired in time (Pavlov, 1927). Different theoretical models were proposed to explain the different results from classical conditioning experiments (Mackintosh, 1965; Kamin, 1968; Rescorla, 1969; Wagner, 1969; Rescorla and Wagner, 1972).

One of the most influential accounts, is the Rescorla-Wagner model (Rescorla and Wagner, 1972). Built on earlier linear operator models (e.g., Bush & Mosteller (1951)) using error based learning models, this model is based on updating a stimulus' *associative strength* or *behavioural influence* – reliability of the CS in predicting the US – when expected and actual value differ. The advantage of the model proposed by Rescorla and Wagner to previous theories was its capability to explain the differential reinforcing effects in classical conditioning experiments when stimulus compounds were presented. In these experiments they showed that prior learning of the association between a stimulus A and the US, affected the learning of the association of the US with a new stimulus X, when stimuli A and X were presented concurrently as a compound stimulus AX. The differential learning depended on previous experience and nature of the stimuli. *Blocking* effects (Kamin, 1968), reduced or no conditioned response to stimuli X, were obtained when prior conditioning between stimulus A and the US, followed by the conditioning between the compound AX and the same US,

impaired or *blocked* the conditioned learning of stimulus X. A reduced conditioned response to individual stimulus A and X was also observed when initially subjects were trained with a compound AX that predicted the US, and later each stimulus A or X were presented individually together with the US. During test sessions, there was a reduced conditioned response when A or X were presented, this phenomenon was called *overexpectation* (Rescorla, 1970). Another phenomenon that the Rescorla-Wagner model could explain is *overshadowing* effects, which happened when stronger intensity of stimulus A, reduced conditioning of less intense stimulus X when presented as a compound AX (Pavlov, 1927).

Let's imagine that on a conditioned trial stimulus compound AX is followed by Y (the US). The Rescorla-Wagner model has the assumption that the *associative strength* of the compound AX, V_{AX}, is the linear sum of the *associative strength* of each stimulus individually, V_A and V_X. Therefore:

$$V_{AX} = V_A + V_X \qquad \textit{Eq. 1.1}$$

On each conditioned trial of compound AX, the *associative strength* of each stimulus is updated as:

$$\Delta V_A = \alpha_A \beta_1 (\lambda_1 - V_{AX}) \qquad \textit{Eq. 1.2}$$

$$\Delta V_X = \alpha_X \beta_1 (\lambda_1 - V_{AX}) \qquad \textit{Eq. 1.3}$$

Where α_A and α_X are the learning rate of each stimulus, β_1 is the learning rate of the US, and λ_1 is the asymptotic level of responding to or *associative strength* of the US (Note that later, when the temporal-difference model is introduced, λ_1 gets replaced by reward, R). If a different outcome US_2 is delivered, then β_1 and λ_1 would change to β_2 and λ_2. The different

learning rates allows to capture the differential learning that happens depending on CS and US identity or intensity, and λ_1 captures the reinforcing effects of the US – how valuable or aversive is the US.

Despite the simplicity of the Rescorla-Wagner model, it was able to capture different phenomenon in classical conditioning experiments as described earlier. For example, because the *associated strength* of the compound AX is the linear combination of the *associative strengths* of stimulus A and X, if there was a prior exposure of A and the US, this prior exposure allows the subject to learn the *associative strength* of A, V_A. Therefore, when later the compound AX is presented followed by the same US, because the US was expected through the *associative strength* of A, there would not be update of V_{AX} or V_X, as $V_{AX} = V_A + V_X$, and consequently, $V_X = 0$. This explains *blocking* effects in classical conditioning.

However, this model had some limitations such as its inability to explain *latent inhibition* (Lubow and Moore, 1959), and higher order conditioning (Pavlov, 1927), among other phenomenon. To overcome these, others have modified the original Rescorla-Wagner model. Here I will focus on Sutton and Barto's extension of the Rescorla-Wagner model, the Temporal-Difference (TD) reinforcement learning model, which, among other properties, enables description of higher-order conditioning. Crucially, this comes about by incorporating the timing between stimuli within- but also between-trials into the model. Whereas the Rescorla-Wagner model is a *trial-level* model, updates happen at the trial level, the TD model is a *real-time* model, updates happen at each time-step. Furthermore, the model also added discounting and eligibility traces between time-steps (Sutton and Barto, 1998). Sutton and Barto reformulated the Rescorla-Wagner model *associative strength* vector update as:

$$w_{t+1} = w_t + \alpha\delta_t X(S_t) \qquad \textit{Eq. 1.4}$$

where w is the *associative strength* vector (i.e. vector of weights for each stimulus), α is the learning rate, δ is the prediction error, $X(S_t)$ is the vector with the values of stimuli or features in trial-type or *state* S, and t is the current trial. The prediction error, δ_t, is described as:

$$\delta_t = R_t - \hat{v}(S_t, w_t) \qquad \textit{Eq. 1.5}$$

where R_t is the prediction target or *associative strength* of the US and $\hat{v}(S_t, w_t)$ is the *aggregate associative strength* of the current trial-type or *state*.

The TD model replaces *trial* t for *time-steps* within- and between-trials. Each trial is a sequence of *states* and on each time-step, one *state* transitions to another. A *state* does not have a determined duration but refers to a specific moment in the environment described by a set of features (internal and external). For example, in a classical conditioning experiment trial, where first the CS is presented followed later by the US, the first CS presentation is one *state*, which transitioned to another *state* where the US is delivered. Although this is one way of understanding state, state can also be defined with other time-steps, like actual time, and therefore, between CS and US there might be multiple time-steps.

On each time-step t, TD RL model updates the *associative strength* vector update as:

$$w_{t+1} = w_t + \alpha\delta_t \boldsymbol{z_t} \qquad \textit{Eq. 1.6}$$

Note that the only differences with the Rescorla-Wagner model is that t refers to a time-step, $X(S_t)$ is replaced by z_t, a vector of eligibility traces, and the prediction error is now the TD error δ_t. We can imagine an eligibility trace as a short-term memory of previous events, indicating which features from the past contributed to the update (Figure 1.2). The TD error δ_t is defined as:

$$\delta_t = R_t + \boldsymbol{\gamma \hat{v}(S_{t+1}, w_t)} - \hat{v}(S_t, w_t) \qquad \textit{Eq. 1.7}$$

where γ is a discount factor, and $\hat{v}(S_{t+1}, w_t)$ and $\hat{v}(S_t, w_t)$ are the next step and current *aggregate associative strengths*. As can be observed, in comparison with the Rescorla-Wagner model, the TD error is the result not only the difference between the prediction target at the current time-step and the *associative strength* of the current *state*, but also includes the discounted future associative strength from the next time-step.

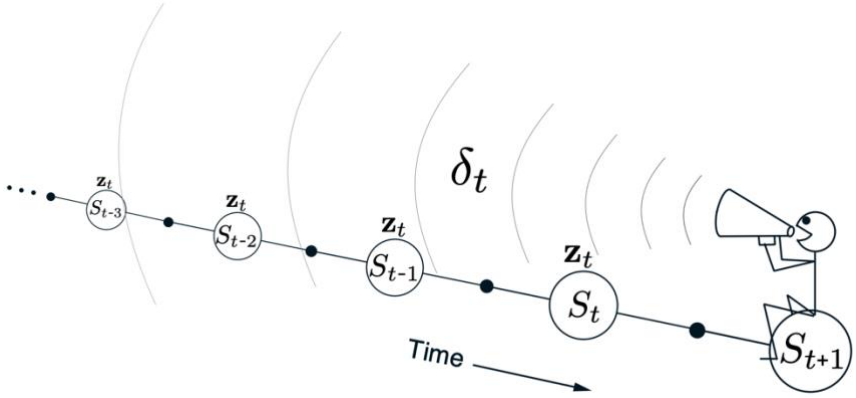


Figure 1.2*. Schematic of the learning updatates of the TD RL model. At each state, the associative strength vector update depends on the future state, and the current eligibility trace vector of previous states. Figure from Sutton and Barto (1998).*

The vector of eligibility traces also gets updated on each time-step following:

$$z_{t+1} = \boldsymbol{\gamma\lambda z_t} + X(S_t) \qquad \textit{Eq. 1.8}$$

where λ is the trace-decay parameter. Therefore, the update depends on the feature vector at the current state and the current eligibility trace vector that decays at a rate defined by $\gamma\lambda$. If the discount factor γ is 0, then the TD model turns into the Rescorla-Wagner model (Sutton and Barto, 1998).

1.2.2 Temporal-Difference Reinforcement Learning during decision-making problems

The TD RL model is a real-time model, which learns *state* values (or *aggregate associative strengths*) and future reward through the use of discounting and eligibility traces (Figure 1.3). In RL, this algorithm is called TD(λ), which is widely used in Markov problems (Sutton and Barto, 1998).

A Markov decision process (MDP) is a model to describe decision-making processes involving a sequence of events, situations or *states*, in which *actions* can be taken in each *state*, determining the immediate reward and the next possible states, and therefore, future rewards. In an MDP, there is the *agent* which take *actions* in the *environment*. As a result of the *action* taken by the *agent* when the *environment* was in a particular *state,* the environment will transition to a future *state* and deliver a numerical *reward* (Figure 1.3). Therefore, at each time-step *t*, the agent chooses an action A_t based on the *environment's state* S_t. The condition of a MDP is that the probability of *state* S_t and *reward* R_t, depends *only*

on the previous *state* S_{t-1} and previous *action* A_{t-1}. The goal is to maximise the long-run cumulative rewards.

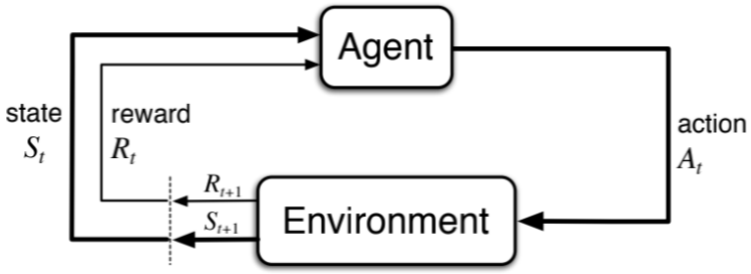


Figure 1.3. Markov-decision process. The agent interacts with the environment through actions (A_t), which based on the transition probabilty of action A_t in state S_t in the environment, will result in reward (R_{t+1}) and the next state (S_{t+1}). Figure from Sutton and Barto (1998).

TD RL does not require having a model of the environment, containing the statistical relationships between *states*, *actions,* and *outcomes*; instead, it learns through experience. Three different TD methods have been used to describe decision-making: Sarsa, Q-learning and actor-critic. While Sarsa and actor-critic are *on-policy* method, Q-learning (Watkins, 1989) is an *off-policy* TD method (Sutton and Barto, 1998). We can define *policy* as the probability of choosing and *action* in each *state*. Informally, *policy* is the strategy performed by the agent to interact with the environment. In *on-policy* methods the value estimates are computed based on the *agent's policy*, whereas in *off-policy* methods, the estimates are calculated based on the *optimal policy*, the one that maximises the cumulative future rewards, independently of the actual *policy* the *agent* is following.

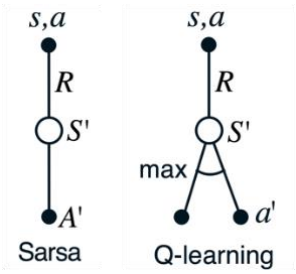


Figure 1.4. Diagrams for Sarsa and Q-learning TD control. From Sutton and Barto (1998).

Sarsa TD control is named like that because of the way it updates value estimates. It updates value estimates based on the *state* ***s***, and *action* ***a*** taken at the current time-step, the *reward* ***R*** obtained, the future *state* ***S'*** and the future chosen *action* ***A'*** (Figure 1.4). The TD error computed by Sarsa is essentially the same as Eq. 1.7, except that instead of learning state values, what is learned is the value of the state-action pair $Q(S_t, A_t)$. Therefore, the TD prediction error, δ_t, and state-action value $Q(S_t, A_t)$ update is calculated as:

$$\delta_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \qquad \textit{Eq. 1.9}$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\delta_t \qquad \textit{Eq. 1.10}$$

Q-learning (Watkins, 1989) updates follow the same procedure as Sarsa, but instead computes the estimates based on the optimal policy. Therefore, the TD error, δ_t, in Eq. 1.9, gets rewritten as:

$$\delta_t = R_{t+1} + \gamma \max_a Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \qquad \textit{Eq. 1.9}$$

The update is driven by the maximum state-action value in the next state, as the optimal policy is to choose the higher value option to maximise future reward (Figure 1.4). Both Sarsa and Q-learning use these updates to revise a lookup table of state-action pairs used to guide behaviour/actions (Sutton and Barto, 1998).

Similar to Sarsa, the actor-critic is an on-policy method learning values though the TD error, but also learns the *policy* (Figure 1.5). The actor defines the action selected at each *state*,

policy, and the critic, 'criticises' the actor choices by computing the current *state value* of the *environment* and computing the TD error, which updates the lookup table of states values from the critic, but it is also sent to the actor to update *policy* (Sutton and Barto, 1998).

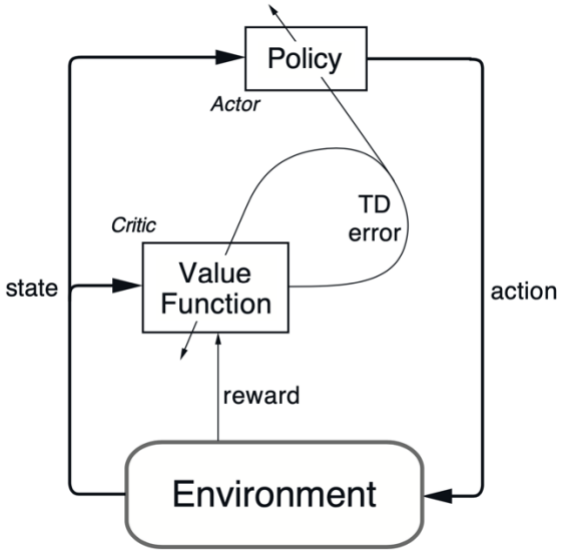


Figure 1.5. Actor-critic schematic. The critic learns the state value of the state of the environment based on the policy taken by the actor, and updates values using TD error. The actor, in turn, learns the policy to be taken in the environment to maximise future reward by receiving the state of the environment and the TD error from the critic. From Sutton and Barto (1998)

There are different policies that the agent can follow. However, the aim is to optimise choices to maximise future rewards. One policy to follow is to always choose the most valuable action in the current state if we look one-step ahead (Here, I have described one-step methods but the complexity can increase in other methods, going further down the tree: see Sutton and Barto, 1998). However, in new or dynamic environments this might not be optimal, as the policy could converge in suboptimal decisions, or it would take much longer to adapt to changes in the environment.

This is what is called, the exploration-exploitation trade-off, balancing between exploration – trying out alternative options – and exploitation – choosing the current best option. One strategy to balance exploration and exploitation is through an ε-greedy action-selection policy. During ε-greedy action-selection, there is a probability, ε, that a random action would be chosen. If ε = 0, then the policy would be completely 'greedy', and no exploration would occur. Another common strategy to balance exploration and exploitation is through a softmax decision rule (Sutton and Barto, 1998):

$$\pi_t(a|s) = P\{A_t = a \mid S_t = s\} = \frac{e^{\beta H_t(s,a)}}{\sum_k e^{\beta H_t(s,k)}} \qquad \textit{Eq. 1.10}$$

Where $\pi_t(a|s)$ is the probability of choosing *action* a in *state* s. $H_t(s,a)$ is the values of the parameters of the current policy, k is the number of other possible choices, and β is the inverse temperature. Through the softmax distribution, exploration does not occur randomly, but the probability of taking each action is weighted based on each action's value estimate. The degree to which the value estimates influence choices is determined by the inverse temperature β. Higher β promotes choosing actions with higher values, whereas lower β makes action selection less influenced by action value estimates. If $\beta = 1$, the softmax distribution becomes ε-greedy action-selection policy.

1.2.3 Planning through the acquisition of a model of the environment

The algorithms described in the previous section, Sarsa, Q-learning and actor-critic, are *model-free* methods, that do not require of a model of the environment. Model-free agents

are very cheap computationally, however, in some situations they might be slow at adapting in dynamically changing structured environments as many *episodes* or trials (i.e., real experience) are needed to detect changes and update value estimates. By contrast, *model-based* agents, instead, learn a model of the environment, including the transition probabilities between states, actions, and outcomes. Having a model of the environment allows an agent to simulate experience and therefore allows planning.

Model-based methods can still use TD error to update value estimates, but do so in conjunction with a model of the environment (i.e., the TD error is computed using the value estimate predicted through having a model of the environment and knowing the state transition matrix). While model-free and model-based methods can be implemented separately in different agents, in RL, there are also different architectures where both model-free and model-based algorithms work together to drive behaviour. An example of one of these architectures where model-free and model-based methods interact is the Dyna-Q architecture (Figure 1.6).

The Dyna-Q learns a model of the statistical relationship in the environment, but also uses Q-learning, together both influencing value and policy update. Real experience is used for learning a model of the environment, but also value and policy updating. At the same time, having a model of the environment, allows the Dyna-Q to simulate experience. This simulated experience by the model-based component, is also used to update value and policy. Therefore, value and policy get updated by both real and simulated experience

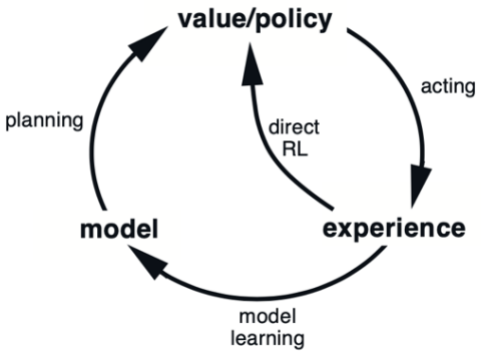


Figure 1.6. Diagram of integrated models such as Dyna-Q where model-free and model-based methods interact. Through experience, a model of the environment is learned. This model simulates experience. This simulated experience is used for planning, which in turn, updates value and policy. On top of this, experience also updates directly value and policy, in a model-free way. From Sutton and Barto (1998).

The models described assume that all the features of the environment are observable and can be learned. However, in many real-world situations that is not the case, and many situations are partially observable. Therefore, the inputs from the environment just give partial information about the state of the environment. In those cases, the environment instead of outputting states (Figure 1.3), emits observations which depended on the state of the environment. The state would then be a function of the history of observations and actions.

During these partially observable environments, before value or policy update, state-update should happen. At each time step, the state of the environment needs to be updated based on the current observations. A way of dealing with these is through a method called Partially Observable MDPs (POMDP). This is a Bayesian method that assumes that latent states, not accessible to the agent, control the environment's outputs. Therefore, the agent instead uses a *belief state* which is the distribution of hidden states given the event's history. This *belief*

state can then be updated on each step following Bayes' rule, given that the agent has a model of the statistical relationships in the environment (Sutton and Barto, 2018).

To conclude, the RL framework has been (and is currently) being used to explain behaviour from living creatures, both humans and animals. Different models have been presented, which capture different aspects of naturalistic behaviour, however, we will see in the next section that sometimes the distinction between models is not so clear. Finally, Bayesian approaches have become more prominent to capture environmental partial observability and explain not only behaviour but also neuronal activity.

1.2.4 Towards a computational framework of instrumental behaviour: model-free vs model-based and others

As seen in the previous sections, TD errors, also referred as reward prediction errors (RPEs) are used in order to update the value of states and actions when predictions differ from the actual outcome (Sutton and Barto, 1998). With the finding of RPE signals in dopamine activity (Houk *et al.*, 1995; Montague *et al.*, 1996; Schultz *et al.*, 1997) during pavlovian and instrumental tasks in non-human primates, there was a growing interest in formalising the computational principles underlying habit and goal-directed behaviours.

Doya (1999) proposed different learning algorithms are implemented within different brain circuits: supervised learning for the cerebellum, reinforcement learning (RL) for the basal ganglia, and unsupervised learning for the cerebral cortex. The interplay between these different *learning modules* could produce flexible goal-directed behaviour, supporting

different behavioural strategies, *reactive* (or model-free) and *predictive* (or model-based) *action selection.* Later, based on the proposed *modular control* for motor learning and the presence of *internal models* in the cerebellum (Wolpert and Kawato, 1998; Wolpert et al., 1998), Doya and colleagues showed computationally how complex tasks like in real-world scenarios, which can be dynamically changing or contain non-directly *observable* features, *hidden states*, could be solved. They presented that this could be implemented by decomposing the complex task into different *modules*, each with a *state prediction model* and an *RL controller*. With experience, the prediction errors obtained were updating both the prediction models and the values of the RL controllers, together with weighting the outputs of each RL controller to guide behaviour – perform actions (Doya et al., 2002).

To extend this idea into how the brain might have different controllers to support habit and goal-directed behaviour, Daw et al. (2005) took this multiple controller idea and computationally simulated two different devaluation experiments, a simpler - one action and one reward - and a more complex one - two actions and two different rewards - to capture the differential influence of goal-directed and habitual control that was previously shown experimentally (Colwill and Rescorla, 1985a; Holland, 2004). Based on previous lesion and inactivation experiments showing involvement of prefrontal cortex in goal-directed control (Balleine and Dickinson, 1998; Corbit and Balleine, 2003), and dopamine and dorsolateral striatum in habitual responding (Yin et al., 2004), they proposed two competing and interacting RL controllers simulating this prefrontal-dorsolateral circuitry guiding behaviour.

The 'prefrontal' system was a *model-based RL* agent which learned the structure of the task to inform their choices. It learned the *transition probabilities* between *states,* and the *action*

to *outcome* mapping in each *state*. This method can also be called *tree-search* as it builds a 'tree' or hierarchical map of the different states, actions and rewards, which allows to look ahead and predict the next state and outcome, allowing *planning*. For example, in the simulated simpler task, there were four different states, an initial state (S_0), a food delivered state (S_1), a food obtained state (S_3) and a no reward state (S_2), with different actions being performed in each state or outcome delivered in the terminal states S_2 and S_3. The *tree-search* agent, through experience, learned the state transitions and rewards through the actions taken in each state (Figure 1.7).

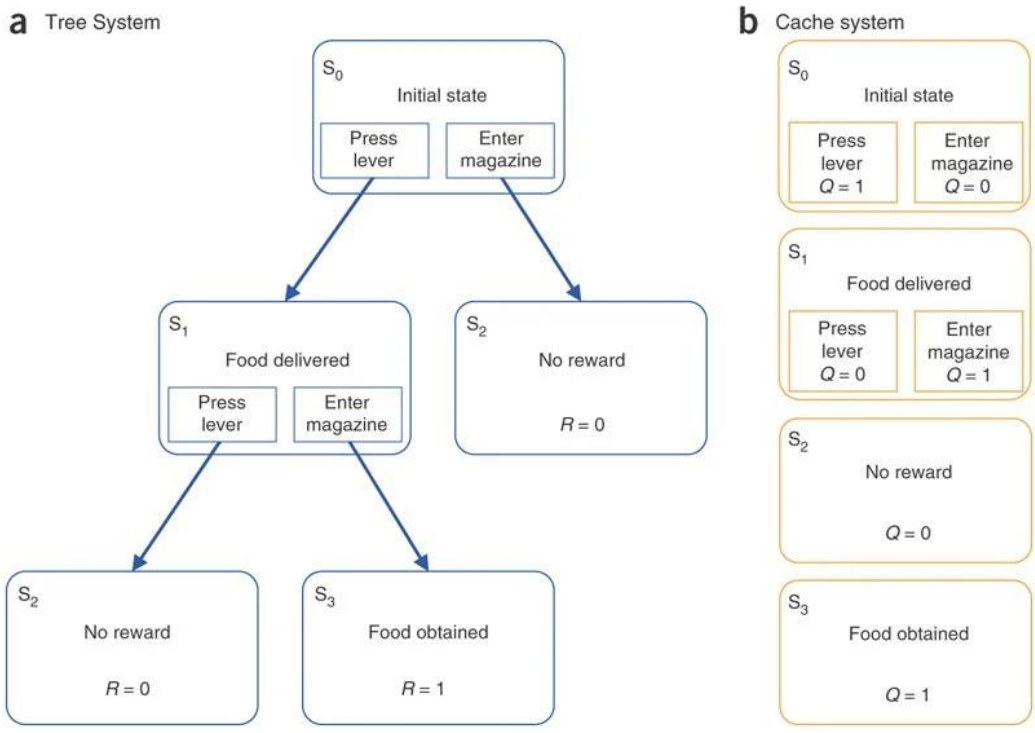


Figure 1.7. Simulated instrumental task in which there was one action (press lever) and one outcome (food delivered in a magazine). a) task representation by a tree-search controller. b) task representation by a cache controller. Figure from Daw et al. (2005).

In contrast, the other controller, the 'dorsolateral striatal' system, was a *model-free RL* agent, which did not construct a tree, but instead *cached* or stored values of each action in each state based on previous experience. The value of each action would be updated when deviations from the predicted value occur through RPEs. Crucially, the cached value of each action did not contain any information about the outcome. This dual controller produced an action in each state driven by the competition between both controllers based on their action value uncertainty in each state, and therefore, the system with less value uncertainty strongly guided behaviour.

After the initial training, they simulated the devaluation procedure by changing the outcome's *utility* (i.e. in the simpler task, changing the reward value to 0; and in the complex task, just set the value of one of the two outcomes to 0). After devaluation, although both systems experienced the lower utility of an outcome, only the tree-search system was able to use that information to update action values of other states – where reward is not directly obtained - as action values on each state depended on future states and rewards. The cache system as it only stored action values but not any information about the outcome, it could not update action values of the other states where the actions are performed. Therefore, while the tree-search system was sensitive to the devaluation procedure, the cache system was insensitive.

This dual-controller architecture was able to reproduce the degree of habitual vs goal-directed control as a function of training and complexity. In the simpler task, devaluation early in training produced devaluation-sensitive behaviour driven by the tree-search or model-based system, while devaluation later in training, behaviour was insensitive to devaluation (i.e. habitual) driven by the cache or model-free system. On the more complex task, both

devaluations early and late in training produced goal-directed - devaluation sensitive behaviour – driven by the model-based controller. These results reproduced previous devaluation experiments in rats (Colwill and Rescorla, 1985a; Holland, 2004).

To test the presence of these two controllers – model-free/model-based RL – simultaneously in the brain, Gläscher et al. (2010) designed a probabilistic Markov task (Figure 1.8A) – different states could be reached probabilistically through different actions – and perform functional magnetic resonance imaging (fMRI) in humans. In this task, there was an initial choice state where a fractal image was presented, and humans needed to choose between the left and the right button. The left button commonly transitioned to a state L1, another fractal image (i.e., on 70% of trials), and rarely transitioned to state L2 (30% of trials). Similarly, for the right button, this action commonly and rarely transitioned to other two states. In each of these second-step states, humans needed to choose again between the left and right buttons. Each action in each state probabilistically transitioned, commonly and rarely, to different reward states, where none, small or large monetary reward was given. The different rewards were explicitly told (i.e. that there were 3 different monetary outcomes: 0c, 10c and 25c), so no learning of the different rewards was needed, and the focus of this study was to find neural signatures of the different transition probabilities, supporting the building of an internal model of the task.

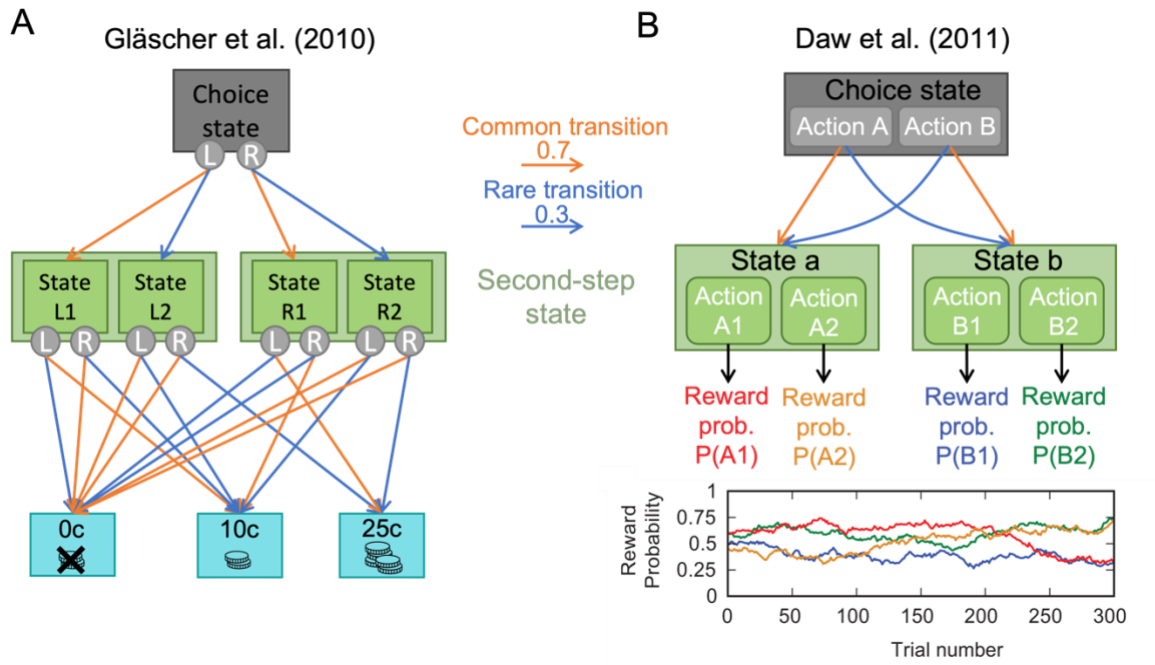


Figure 1.8. Multistep Markov probabilistic task from A) Gläscher et al. (2010) and B) Daw et al. (2011). A) There was an initial choice state, in were humans needed to choose between pressing the left (L) or right (R) button. Each action probabilistically transitioned to a second-step state. Again, in each second-step state, humans had to choose between L and R buttons. Each action in each state probabilistically transitioned to different reward states, where monetary reward was delivered (or not). Orange arrows – common transitions, probability of 0.7. Blue arrows – rare transitions, probability of 0.3 of happening. B) Modified version of A and where dynamic reward probabilities were introduced. There was a choice state where humans could choose between two images, action A and action B. Each action with different probability transitioned to one second-step state. Action A commonly transitioned to state a, while rarely transitioned to state b. The opposite happened for action B. In states a and b, two different images were persented, and humans needed to choose one. Each action at second-step state was probabilistically rewarded. Probability of reward of each action (A1, A2, B1 and B2) dynamically and independently changed across the session following Gaussian random walks. Bottom plot shows an example of the reward probabilities of each action across a session from Akam, Costa and Dayan (2015).

Behaviourally, they found that a hybrid RL agent – a weighted mixture of model-free and model-based RL – best explained subjects' behaviour. A correlate of model-based RL was localised in the intraparietal sulcus and lateral PFC. Blood oxygen level-dependent (BOLD) activity in these regions was correlated with *state prediction error* used by the model-based RL module to update the internal model of the task transition structure when there was a divergence between predicted and observed transitions. Additionally, RPE derived from the model-free component correlated with activity in the ventral striatum. This was the first evidence showing model-free and model-based signals simultaneously in different brain regions.

The probabilistic Markov task used by Gläscher et al. (2010) overcame many limitations of previous studies of goal-directed behaviour. As mentioned earlier, psychological research studying the dissociation between habits and goal-directed behaviour, relied mainly upon experiments like reinforcer devaluation. In these kinds of experiments, there was a single test session, an extinction session, to test whether behaviour was consistent with planning or if instead was habitual. The presence or absence of behaviour limited the computational description of it, plus the one-shot test limited the capability to look for neural correlates of this behaviour. As a consequence, this work together with the two-step task from Daw and colleagues (2011), explained below, have been very influential.

The two-step task in humans

Later Daw et al. (2011), studying how model-free and model-based controllers interact during learning to guide behaviour, revealed that these two systems are not fully independent but

that they are implemented in the brain in a more jointly or overlapping way. Daw et al. (2011) used a modified version of the multistep Markov decision task from Gläscher et al. (2010) into a two-step probabilistic decision task (Figure 1.8B). In this task, the transition structure was reduced while dynamic changes in reward probabilities were introduced. There was an initial choice state, in where subjects had to choose between two different fractal images, that we will call actions A and B. Each action with different probability transitioned to two second-step states, state a and state b. Action A commonly transitioned to state a, and rarely transitioned to state b, and the opposite for action B. Each second-step state contained two different fractal images to choose from, action A1 and A2 in state a, and action B1 and B2 in state b. The probability of getting rewarded after choosing A1, A2, B1 and B2 dynamically changed across the session following Gaussian random walks.

This task predicted differential behaviour depending on whether a model-free or a model-based strategy is followed. While a model-free strategy would repeat or switch actions based on the outcome received, reward or omission, respectively; a model-based strategy would promote stay or switching behaviour based not only on the outcome experienced in the previous trial, but also on the transition that occurred. For example, if in the previous trial the subject experienced a rare transition and got reward, while a model-free learner would repeat that same action, a model-based learner would switch to the other action as with higher probability would transition to the second-step state that was rewarded. Their results again showed that humans adopted a hybrid approach, using a combination of model-free and model-based strategies to guide behaviour. Both model-free and model-based value predictions correlated with BOLD activity between the second-step state and outcome delivery in the ventral striatum. Furthermore, the degree to which model-free or model-

based predictions influenced activity in the ventral striatum correlated with each subject's behaviour (i.e., subjects whose behaviour was more driven by a model-free approach, their activity in the ventral striatum reflected more model-free value predictions, whereas subjects with a higher model-based index, their ventral striatum activity reflected more model-based value predictions). They also found model-based predictions in ventromedial PFC during the same timepoints (between second-step and reward).

As mentioned, this task overcame many of the previous limitations that hindered the ability to study these processes over many trials. The differential behavioural predictions by each learning algorithm in the two-step task on a trial-by-trial basis, and the presence of many trials in which this behaviour could be manifested in a single session, but also over many sessions, allowed to study the computational algorithms implemented at the neuronal level.

There was a growing interest in mapping the computational algorithms implemented in the different brain regions guiding behaviour. Others also developed behavioural multi-step tasks in humans used to probe RL strategies in different situations (Simon and Daw, 2011; Huys et al., 2012; Wunderlich et al., 2012a). However, the two-step task has been widely and extensively adapted in many different studies. Studies using the two-step task in humans revealed that increased levels of dopamine promoted model-based choices (Wunderlich et al., 2012b), which could be the result of higher presynaptic dopamine levels in the ventral striatum resulting in decreased signalling of model-free signatures towards model-based in this area while increasing also model-based signals in lateral PFC (Deserno et al., 2015). Several studies have shown impairment of model-based behaviour in response to stressful situations (Otto et al., 2013b), in different neuropsychiatric patient groups (e.g., alcohol

dependence: Sebold et al., 2014, gambling disorder: Wyckmans et al., 2019), and during disruption of dorsolateral prefrontal cortex (Smittenaar et al., 2013). Moreover, decreased grey matter volumes in caudate and medial orbitofrontal cortex was also correlated with a preference towards model-based choices in obsessive-compulsive disorder, binge eating and methamphetamine consumption (Voon et al., 2014).

Other studies used versions of the two-step task to further interrogate the resulting behaviour, showing that humans actually use model-free strategies and are not just the result of learning action sequences (Morris and Cushman, 2019), with model-free or habitual control used to select goals and model-based behaviour used to achieve the selected goals (Cushman and Morris, 2015), that model-based behaviour is positively correlated with generalisation (Doll et al., 2015b) and with the presence of neural activity in visual and temporal cortices reflecting future decision paths (Doll et al., 2015a). These studies also revealed that the behavioural strategies used by humans depend on the instructions provided (Feher da Silva and Hare, 2020; Castro-Rodrigues et al., 2022), cognitive resources as a cost-benefit trade-off (Otto et al., 2013a; Kool et al., 2016, 2017) and cognitive control (Otto et al., 2015), but this differential model-free - model-based control evolves with age, with model-based behaviour being absent during childhood, appearing during adolescence and getting strengthened in adulthood (Decker et al., 2016).

Rodents adaptations of the two-step task

Not only has the two-step task accelerated the study of the neural correlates of habitual and goal-directed control in humans, but it has also been adapted for use in rodents, allowing

interrogation of these learning algorithms in tandem with neural recordings and manipulations. Miller et al. (2017) trained rats on a simplified version of the two-step task in which the rats only made a choice at the first and not the second step. Rats behaviour appeared to be dominated by model-based control. Muscimol inactivation of orbitofrontal cortex or dorsal hippocampus selectively disrupted model-based choices. Hasz and Redish (2018) adapted the full structure of the human two-step task, not the simplified version as in the previous study, in a spatial maze. However, instead of having probabilistic reward following second-step choices, they instead used delays to reward delivery (i.e., all trials were rewarded but at different delays) which, similarly to the original task, dynamically changed according to Gaussian random walks. As in the original task in humans, they found that rats' behaviour was consistent with the implementation of a hybrid RL strategy. Finally, Akam et al. (2021), also using the simplified version (Akam et al., 2015) but in mice and with added reversals in the transition probabilities, showed that the anterior cingulate cortex activity is modulated by future state and inactivation of these signals impairs model-based choices. Using the same task, Korn et al. (2021) found that blockage of the dopamine transporter impaired behaviour during reversals of the reward probabilities of the second-step state, but did not affect reversals in the transition probabilities.

Apart from a large amount of research using the two-step task in humans and the more recent adaptations in rodents, there has also been several computational and modelling studies to further explore the behavioural effects and possible behavioural strategies of the two-step task and versions of it (Akam et al., 2015; Kool et al., 2016; Feher da Silva and Hare, 2018) postulating that the model-free vs model-based dissociation might not be as straightforward as it was previously thought.

Akam *et al.* (2015a) exposed that model-free agents could produce behaviour that resembled model-based. Furthermore, they presented other behavioural strategies that could also look like the traditional model-based two-step behaviour. One of these strategies used Bayesian inference, which instead of continuously updating value, based on the observations (rewards) and actions, computed the *belief state*, and chose actions based on that. This type of strategies where inference guides behaviour are becoming popular approaches to explain complex flexible behaviour (Gallistel et al., 2001; Hampton et al., 2006; Behrens et al., 2007; Nassar et al., 2010, 2012; Costa et al., 2015; Jang et al., 2015; Rothenhoefe et al., 2017; Bartolo and Averbeck, 2020; Vertechi et al., 2020).

1.3 Dopamine system

1.3.1 A brief journey through dopaminergic anatomy

Dopamine is a monoamine neurotransmitter of the catecholamine family – together with norepinephrine and epinephrine. These neurotransmitters are synthesised endogenously from the aromatic amino acid tyrosine. Inside the neurons, tyrosine hydroxylase (TH) is the enzyme responsible of the step from tyrosine to dihydroxyphenylalanine (DOPA), next, aromatic L-amino acid decarboxylase transforms it into dopamine. Dopamine can be later converted into norepinephrine, and this one can also be then converted into epinephrine (Molinoff and Axelrod, 1971).

Neurons synthesising, releasing and metabolising dopamine are primarily located in the midbrain. The midbrain dopaminergic neurons comprise a diverse group divided in three different nuclei, nucleus A8 in the retrorubral area, nucleus A9 cells mainly located in the substantial nigra pars compacta (SNc), and nucleus A10 cells mainly in the ventral tegmental area (VTA) (German and Manaye, 1993). From these centres, four major dopaminergic projections have been described: the nigrostriatal, mesolimbic, mesocortical and inferotubular pathways. In rodents, dopaminergic neurons mainly from the SNc project to dorsal striatum, putamen and caudate nuclei, conforming the nigrostriatal pathway. This pathway has been related with motor control and goal-directed behaviour. From the VTA, dopamine neurons project to the ventral striatum, nucleus accumbens (NAc) and olfactory tubercle, amygdala and other limbic regions, named the mesolimbic pathway, but they also project to cortical regions, forming the mesocortical pathway. These pathways originating mainly from VTA dopaminergic neurons have been also named as the reward system for its role in reinforcement learning and decision-making (although not as reward-related the mesocortical pathway (Ellwood et al., 2017)). Finally, the tuberoinfundibular pathway, which regulates hormonal release, is formed by dopaminergic neurons in the hypothalamus, arcuate and periventricular nuclei (A12 and A14), projecting to the pituitary gland (Figure 1.9) (Baik, 2020).

Recent advances in genetic labelling approaches have uncovered the diversity and richness of dopaminergic projections, far from the crude four pathways framework. It was shown that within each region in the midbrain, different dopaminergic cells, characterised through their genetic profile, project to specific regions in the dorsal and ventral striatum and amygdala (Roeper, 2013; Poulin et al., 2018). This work evidenced the cellular heterogeneity within

region, but at the same time, the lack of clear distinction between regions both at the level of genetic profiling and projection target. While a population of dopaminergic neurons in VTA mainly project to NAc, and SNc neurons mainly project to dorsolateral striatum, a shared population of neurons in dorsolateral VTA and SNc project to ventromedial regions of dorsal striatum (Poulin et al., 2018).

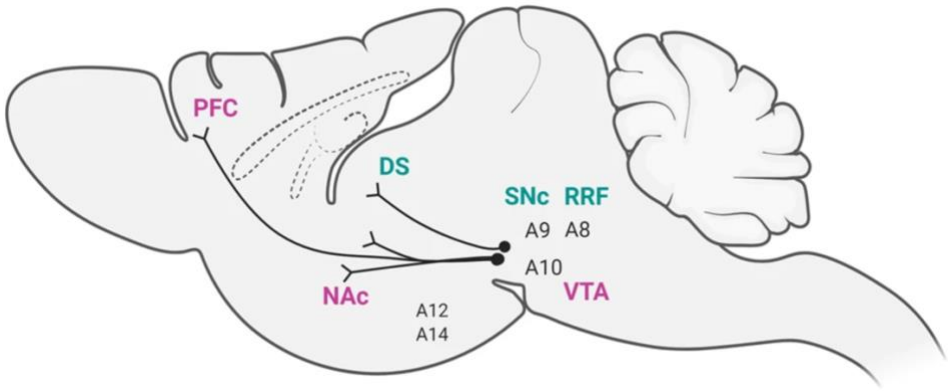


Figure 1.9. Dopamine neurons nuclei and projections. Midbrain dopaminergic cells are distributed in different nuclei: A8, A9, A10, A12 and A14, located mainly in the retrorubral field (RRF), substantia nigra pars compacta (SNc), ventral tegmental area (VTA), arcuate nucleus, and perventricular nucleus of the hypothalamus, respectively. These cells project to different brain regions conforming different dopaminergic pathways. Neurons mainly located in SNc project to dorsal striatum, nigrostratial pathway. Mesolimbic and mesocortical pathways are formed by dopaminergic cells mainly in VTA, projecting to ventral striatum, formed by the nucleus accumbens (NAc) and the olfactory tubercle, and projecting to cortical regions such as the prefrontal cortex (PFC). Lastly, A12 and A14 neurons in the hypothalamus project to the pituitary to form the tuberoinfundibular pathway. Figure from Baik (2020).

At the same time, midbrain dopaminergic neurons also receive inputs from very diverse brain regions. Although SNc and VTA receive projections from different areas, whereas SNc receive

monosynaptic inputs from somatosensory and motor cortex and subthalamic nucleus, VTA main inputs come from the lateral hypothalamus (Watabe-Uchida et al., 2012).

1.3.2 Differential effects of dopamine through firing patters and receptors distribution

Dopamine neurons exhibit spontaneous, slow, and irregular depolarising currents with a firing rate of 3-8Hz. This independent and continuous neuronal discharges is referred to tonic or pacemaker firing pattern (Grace and Bunney, 1984a; Paladini and Roeper, 2014). It is debated whether this pacemaker activity results in *tonic release.* Microdialysis studies have shown that dopamine concentrations in the synaptic space are around 20-30nM (but see recent studies (Oh et al., 2018; Yuen et al., 2021)), but these appear in response to phasic activity (Owesson-White et al., 2012), arguing that what it is considered as "tonic release" may instead be summed phasic activity (Hamid et al., 2016; Berke, 2018). These extracellular levels of dopamine diffuse from the synaptic cleft, allowing to act distant from its releasing site through volume transmission (Agnati et al., 1992; Rice, 2000). Dopaminergic tonic firing and sustained concentrations have been related to motivation or willingness to make an effort (Cagniard et al., 2006; Niv et al., 2007; Hamid et al., 2016).

Another mode of firing pattern is through burst firing, which are a sequence of spikes at high-frequency rate (2-8 action potentials at >15Hz, with decreasing amplitude and longer intervals) followed by periods without firing (Grace and Bunney, 1984b; Paladini and Roeper, 2014). This firing pattern produces phasic dopamine responses to excitatory stimuli, resulting in much higher phasic levels of dopamine release into the synaptic cleft (Garris et al., 1994). Phasic activity has been related to with reward learning (Schultz, 1998; Zweifel et al., 2009).

Although phasic and tonic dopamine activities are presented as separate models of action, it has been recently shown that phasic activity can influence tonic dopamine concentrations, and therefore, their behavioural effects can be more understood as a continuum (Owesson-White et al., 2012; Berke, 2018; Lohani et al., 2018).

The different modes of firing are not uniform across the midbrain dopaminergic neurons, with burst firing and baseline (tonic) activity levels specific of neuronal subtypes (Farassat et al., 2019; Otomo et al., 2020). Besides the heterogeneous firing patters which then result in diverse dopamine release patterns and concentrations into the postsynaptic site, it has also been suggested a differential effect of firing patterns on dopamine receptors (Dreyer et al., 2010; Marcott et al., 2014; Yapo et al., 2017; Martel and Gatti McArthur, 2020). Dopamine receptors (DR) are from the family of G protein-coupled receptors, which can be divided in two subtypes: D1- and D2-like DR. D1-like receptors subtypes, comprising D1 and D5 DR, contain the G_s alpha subunit which activates cAMP-dependent pathway through adenylyl cyclase. Conversely, D2-like receptors, comprising D2, D3 and D4, are coupled to the G_i subunit which instead, inhibit cAMP-dependent pathway (Missale et al., 1998). These two types of DR have different affinities to the ligand, dopamine, with D2 receptors having increased affinity to dopamine than D1 (Richfield et al., 1989; Rice and Cragg, 2008), although D2 affinity might be more complex that it has been thought (Yapo et al., 2017). Therefore, while burst activity has been suggested to impact both D1 and D2 receptors, tonic dopamine levels appear to preferentially exert its function through D2 receptors (Dreyer et al., 2010; Marcott et al., 2014; Yapo et al., 2017; Martel and Gatti McArthur, 2020).

D1 and D2 receptors subfamilies are distributed across many brain regions (Hall et al., 1994). In the striatum, D1 and D2 receptors are found in medium spiny neurons (MSNs). In dorsal striatum, D1- and D2-expressing MSNs have been shown to be involved in two opposing pathways. While dopamine activates the direct pathway through D1-expressing MSNs, which promotes ("Go") behaviour, dopamine also activates the indirect pathway through D2 MSNs which inhibit ("No-Go") behaviour (Kravitz et al., 2012; Soares-Cunha et al., 2016). However, some studies evidenced that this D1 vs D2 pathway distinction and opposing roles might not be so clear (Cazorla et al., 2014; Saunders et al., 2015). Similarly, in ventral striatum, D1 and D2 MSNs have been linked with reward and avoidance-related behaviour (Nakanishi et al., 2014), but again, the D1-direct pathway and D2-indirect pathway dichotomy might be a too simplistic view of the basal ganglia circuitry (Soares-Cunha et al., 2016; Kupchik and Kalivas, 2017). In parallel, in the different striatal regions, cholinergic interneurons firing pattern is differentially controlled by dopamine signalling through DR expression (Chuhma et al., 2014).

1.3.3 Dissociation between dopaminergic activity and actual dopamine release

As has been exposed in the previous sections, the dopaminergic system is far from being uniform, each region has different population subtypes, each with different projection targets, different DR distribution, different firing patters and different suggested functionality (which will be later exposed in more detail). A question arising on top of all this diversity is the relation between dopaminergic somatic and axonal activity, and between activity and actual dopamine release.

Using one of the new fluorescent sensors to optically track dopamine concentrations in the postsynaptic site, dLight (Patriarchi et al., 2018), Mohebi et al. (2019) revealed a prominent dissociation between phasic dopamine firing in VTA and dopamine concentrations in NAc. In this study, VTA dopamine cell firing was obtained through electrophysiological recordings, and dopamine concentrations were recorded using dLight and fibre photometry while rats performed a bandit task. Although there was a tight correlation between somatic bursts and striatal dopamine concentrations in response to cue delivery, they reported some striking differences. Not only dopamine modulation was not always tightly coupled across the trial – dopamine concentration increased when the trial was initiated, whereas no increase was present in somatic activity – but while striatal dopamine concentrations reflected value information that could not be explained through somatic firing (Mohebi et al., 2019). They suggested a differential role of dopaminergic activity and release for learning and motivation that will be discussed in the next sections. It needs to be noted though, that the recordings in this study were performed in different groups, and therefore, the relationship between cell activity and dopamine release was inferred.

However, it has been reported in SNc through simultaneous fibre photometry recordings of somatic and axonal calcium activity, that while differences were observed between somatic and axonal firing, this was the result of recording from a population which included several dopaminergic neuronal subtypes. Therefore, when each genetic-defined dopamine subpopulation was recorded separately, the same pattern of modulation appeared between somatic and axonal calcium activity during locomotion (Azcorra et al., 2022). Using electrophysiological recordings in VTA cell bodies, and fibre photometry recordings of somatic (VTA) and axonal (NAc) calcium activity and NAc dopamine concentrations, others have also

found an homogeneous RPE, and not value, coding in both dopaminergic activity (somatic and axonal) and release (Kim et al., 2020). Therefore, this is a current topic of debate, with several studies disagreeing between each other.

Potential uncoupled dynamics mechanisms

Although more studies are needed to determine the mapping between dopamine activity and release, we will now discuss potential mechanisms that could explain dopamine activity and concentrations uncoupled dynamics.

As previously seen, the dopaminergic pathway is very heterogeneous even within the same region, the receptors distribution is not uniform, which in turn, influence dopamine release. For example, D2 autoreceptors influence dopamine synthesis and reuptake (Sulzer et al., 2016). At the same time, dopamine transporters, which are responsible for reuptake of dopamine from the extracellular space, are also differentially distributed and dynamically regulated. This can also explain why somatic spiking from a defined neuron could result in different dopamine release in different projection targets (Nirenberg et al., 1997; Lammel et al., 2008; Condon et al., 2019; Bu et al., 2021). Similarly, distributed expression of active zone-like release sites and the proteins involved in the process also modulates the speed of dopamine release (Liu et al., 2018; Liu and Kaeser, 2019). Furthermore, action potential propagation is not uniform, and therefore it would also suggest compartmentalised dopamine release within the same neuron (Gentet and Williams, 2007).

Local microcircuitry can also influence dopamine release. Striatal dopamine release is bidirectionally modulated, increased and decreased, by cholinergic interneurons depending on dopaminergic firing frequency (Rice and Cragg, 2004; Zhang and Sulzer, 2004; Threlfell and Cragg, 2011). Moreover, cholinergic interneuron firing can modulate dopamine release independent of dopaminergic neuronal firing. Synchronised cholinergic interneuron firing can directly trigger striatal dopamine release (Threlfell et al., 2012). Besides, cholinergic interneurons through direct synaptic contact with dopaminergic axons promote dopamine release (Kramer et al., 2022), generating also action potentials in the dopaminergic axonal projections, both without the presence of dopaminergic somatic firing (Liu et al., 2022).

Independently from dopaminergic somatic spiking, striatal dopamine release can also be facilitated or depressed through glutamatergic synapses from the basolateral amygdala onto dopaminergic terminals in NAc (Floresco et al., 1998; Jones et al., 2010). Not only there is this complexity at the postsynaptic site controlling dopamine release, but there is also not a 1-to-1 mapping from dopamine concentrations to MSNs activation (Tritsch et al., 2012; Lee et al., 2021)

While there might be similarities between dopamine activity and concentrations into the postsynaptic site, different studies are further corroborating the complexity of the system, and more research is needed to uncover the mapping between dopamine activity and release (de Jong et al., 2022).

1.4 Dopamine and behaviour

Due to dopamine dysfunction during Parkinson's disease, dopamine was initially studied under the motor control framework (Carlsson, 1959; DeLong, 1983). However, later research using electrophysiological recordings in midbrain dopaminergic neurons provided evidence that dopamine responses could not be explained through a pure movement account nor a pure sensory one, and suggested a role in behavioural reactivity to events (Schultz, 1986; Blackburn et al., 1989).

At the same time, the dopaminergic system was also being linked with reward and reinforcement. It was shown that animals would press a lever to receive intracranial self-stimulation in dopamine-rich regions (Phillips et al., 1976; Corbett and Wise, 1980), and that the reinforcing and withdrawal effects of drugs of abuse such as cocaine (Weiss et al., 1992) or heroin (Kiyatkin, 1994) were mediated through the dopaminergic system. Dopamine was then presented as a substrate for reinforcement, attention and motivation (Koob, 1992; Schultz et al., 1993; Di Chiara, 1995). Later, all these ideas on dopamine and reward were formalised into the RPE theory of dopamine (Montague *et al.*, 1996; Schultz *et al.*, 1997). Since then, research has focused on further describing signals within an RPE framework. However, some studies have suggested that the unified RPE account of dopamine might not be the full picture, and dopamine could have other roles such as motivation or motor control (Berridge, 2007; Salamone and Correa, 2012; Coddington and Dudman, 2018a; Hughes et al., 2020). Furthermore, dopamine computations might integrate more complex information, an internal model of the environment, than previously thought.

1.4.1 The reward prediction error hypothesis

Schultz and colleagues used electrophysiology to record midbrain dopamine neurons while monkeys were performing a delayed response task. They showed that dopaminergic neurons responded to reward in higher proportion in area A10 (VTA), than A8 (retrorubral field) and A9 (SNc). However, this reward responses where only present during initial learning, and progressively disappeared. Dopaminergic neurons also responded to the two conditioned cues, the one signalling the correct lever, and the second one being a "go" cue to press the correct lever to obtain reward. When this study was published, they concluded that dopamine responded to *alerting* stimuli, postulating a role of dopamine for attention and motivation during learning (Schultz et al., 1993). It was few years later, when this study, and others (Romo and Schultz, 1990; Schultz and Romo, 1990; Ljungberg et al., 1992) were reinterpreted and fit into a predictive Hebbian learning model (Montague *et al.*, 1996), with dopamine responses being consistent with a RPE signal (Schultz et al., 1997). It was shown that when a reward was unexpected, dopamine would fire in response to it. When instead it has been learned that a stimulus predicts future reward, dopamine fire in response to the conditioned stimuli, but not when the reward is delivered. However, if reward was not delivered, dopamine neurons decreased their firing. This pattern of firing was taken as evidence of a signal in the brain representing the TD error, $\delta_t = r_t + \gamma \hat{V}_{t+1} - \hat{V}_t$, the error between predicted and actual outcome (Montague *et al.*, 1996; Schultz *et al.*, 1997).

The postulation of the RPE theory for dopamine set a milestone in this research field. Since then, myriad of studies further supported and extended this theory. Early on, it was exhibited that dopamine RPE could act as a teaching signal during learning to guide behaviour

(Hollerman and Schultz, 1998; Waelti et al., 2001). Later, through optogenetic activation and inhibition experiments this causal link between this error signal and learning was further supported (Witten et al., 2011; Steinberg et al., 2013; Chang et al., 2016; Stauffer et al., 2016; Keiflin et al., 2019). However, while early in learning there was a high correlation between dopamine responses and behaviour, this decreased across learning (Hollerman and Schultz, 1998), although RPE signalling was still present even if it did respect the behavioural policy that was being used (Bayer and Glimcher, 2005).

Another feature that was also soon described, was the asymmetric responding to positive and negative outcomes by midbrain dopamine firing. VTA dopamine activity reflected positive RPE, but lacked or more weakly reported negative RPE, shown in human (D'Ardenne et al., 2008), non-human primates (Bayer and Glimcher, 2005; Bayer et al., 2007) and rodents (Pan et al., 2005). This asymmetric effect of positive and negative RPEs is still currently being observed (Mohebi et al., 2019) and it is a topic of debate. It has been suggested that this could be a technical limitation in detecting activity suppression (Fiorillo et al., 2003; Pan et al., 2005). Another possibility is that another population of neurons are involved in the negative components of RL, another population of dopamine or non-dopaminergic neurons in the midbrain (Ungless et al., 2004; Fiorillo, 2013; Verharen et al., 2020) or through the interaction of other systems like serotonin signalling (Daw et al., 2002). This asymmetry, though, seems not homogeneously distributed throughout the dopaminergic system, with reports showing that ventral striatum do symmetrically report both positive and negative RPE (D'Ardenne et al., 2008; Hart et al., 2014).

Although an homogeneous and distributed RPE coding has been reported within midbrain dopamine neurons (Fiorillo et al., 2013; Eshel et al., 2016; Tian et al., 2016), some other studies highlight the heterogeneity in the system when specific projections (Lammel et al., 2011, 2012), or subregions (Cai et al., 2020) are recorded. Furthermore, recently Dabney et al. (2020) presented that individual neurons represent RPE, but together represent a probability distribution of future rewards, consistent with a distributional RL.

1.4.2 The motivation account of dopamine, a challenge to the RPE hypothesis?

Before the postulation of the RPE theory, some studies envisioned dopamine signalling for attention and motivation (Schultz and Romo, 1990; Koob, 1992; Schultz et al., 1993; Di Chiara, 1995). When the RPE theory was presented, research tended then to focus on the formal description of this signal (Montague *et al.*, 1996; Schultz *et al.*, 1997; Schultz, 1998; Waelti *et al.*, 2001; Eshel *et al.*, 2016; Tian *et al.*, 2016; Watabe-Uchida *et al.*, 2017), putting the motivation aspect aside.

However, dopamine has consistently been shown to be involved in reward seeking and response vigour (Evenden and Robbins, 1983; Taylor and Robbins, 1986; Ljungberg and Enquist, 1987; Ikemoto and Panksepp, 1999). Therefore, others suggested that in parallel to the RPE teaching signal, dopamine could still have a role in motivation – invigoration and willingness to make an effort (Salamone and Correa, 2002; McClure et al., 2003; Salamone et al., 2005; Niv et al., 2006, 2007). Computationally, it was suggested that tonic levels of dopamine could support this motivational role, controlling vigour, by using average reward rate as an opportunity cost (Niv et al., 2007). The output of this model was consistent with

previous experimental data, e.g. the relation between dopamine levels and speed of responding during effortful situations (Denk et al., 2005; Walton et al., 2006), or switching behaviour (Robbins and Everitt, 1982; Oades, 1985).

Later optogenetic experiments showed that activating dopamine reduced latency to initiate a trial (Hamid et al., 2016) or promoted movement initiation, while inhibiting dopamine prevented it (da Silva et al., 2018). It was also shown that NAc dopamine levels correlated with reward rate (Hamid et al., 2016; Mohebi et al., 2019). These studies support the motivational role of dopamine. However, it is still a matter of debate if this motivational signal is supported through tonic or phasic dopamine (Berke, 2018).

Supporting the motivational account, it is observed dopamine ramping – gradual increases of dopamine – as the subject becomes closer in space or time to natural reinforcers such as food (Roitman et al., 2004; Wassum et al., 2012; Syed et al., 2016) or drugs of abuse (Phillips et al., 2003). These signals have been viewed as indication of state value and reward expectation (Howe et al., 2013; Hamid et al., 2016; Mohebi et al., 2019), which could be used to build an internal model of the external world (Collins et al., 2016; Guru et al., 2020). However, others have advocated instead that ramping could still be explained through the RPE framework at each time step (Gershman, 2014; Morita and Kato, 2014; Kim et al., 2020).

In a bandit task, Hamid et al. (2016) recorded dopamine concentrations in NAc using FSCV. The RPE framework predicted a change in the dopamine peak response at outcome time with different reward expectation. If reward is received on consecutive trials, the first reward delivery will produce a higher peak than the second - reflecting the positive update in reward

expectation. During states with low reward expectation, the difference in the dopamine peaks between the first and the second reward would be bigger than in high reward expectation states. This would be the case because during low expectation states, the first reward would be highly unexpected, causing a big dopamine response – difference between predicted and actual outcome received is big – but at the same time, having received reward, increase the reward expectancy on the next trial, and therefore, if reward is delivered, as it was more expected, the dopamine peak is smaller, as predicted by the RPE model. During high reward expectancy state instead, the first reward is already highly expected, therefore, the difference between expected and actual outcome is small, which produces a small error signal, a small dopamine peak. This previous reward, increases even more the expectancy of receiving another reward, producing even a smaller error signal, however, because the background expectancy rate was already similar to the first reward, the difference in the peaks would be small. Alternatively, the value model did not predict a change in the magnitude of dopamine peaks, but instead predicted a change in dopamine baseline with reward expectation: increased value is reflected through increased baseline dopamine concentrations – big baseline change during low expectancy states, small baseline change during high expectancy state. Dopamine concentrations were consistent with the value model and not the RPE one. Therefore, Hamid et al. (2016) concluded that NAc dopamine concentrations reflected the value of the state which in turn provided a motivational signal or willingness to work. As described earlier, another study from the same lab, further showed that ramping activity was only present at the level of NAc dopamine concentrations and was absent in VTA dopamine spiking. While NAc concentrations correlated with reward rate, VTA somatic firing only reflected RPE-like signals. Therefore, they concluded that dopamine concentrations in NAc

independently carry phasic – not tonic – value and motivational signals, potentially through local circuitry modulation of dopaminergic axons (Mohebi et al., 2019).

Because RPE and value are tightly correlated, Kim et al. (2020) set up a virtual-reality experiment to generate different dopamine predictions for each account, and this study directly tested the two hypothesis. Using a virtual-reality task and recording VTA somatic spiking using tetrodes, and VTA somatic calcium activity, axonal calcium activity and dopamine concentrations in NAc using fibre photometry, Kim et al. (2020) concluded that dopamine ramping was present in dopaminergic somatic *and* axonal calcium activity *and* dopamine concentrations. Their dynamics were consistent with RPE signalling and not value. In this task, mice were on a virtual track where water reward was delivered at the end. On different trials, while animals were traversing the track, they were suddenly teleported to a closer location to the goal (i.e. where reward would be delivered). On another set of experiments instead the speed of traversing the track was modified. Both manipulations generated different dopamine dynamics across time for the RPE or value hypothesis. All recordings from somatic and axonal activity, and dopamine concentrations reflected dopamine modulation consistent with the RPE framework. Therefore, this study's results comprehensively refuted the value framework of dopamine (Hamid et al., 2016; Berke, 2018; Mohebi et al., 2019).

How to reconcile these two different perspectives is a matter of great debate. One possibility is that, while in Mohebi et al. (2019) dopamine signals were recorded during instrumental behaviour, Kim et al. (2020) could be considered a Pavlovian task, as no action was required from the animal to obtain reward. Second, the discrepancy in results could also be due to the

recording sites, and that different subregions (and/or different animal species) have a different code (Kim et al., 2020).

In conclusion, although there is evidence of dissociable RPE and value signals in the dopaminergic system, more studies are needed to reveal the relation between RPE and value (or motivation) representation across the dopaminergic system.

1.4.3 Dopaminergic role in motor control

The well-established relationship in Parkinson's disease between the loss of dopaminergic neurons and motor dysfunctions (Obeso et al., 2000; Lees et al., 2009) implies dopamine plays a role in motor control. As mentioned, early studies on dopamine supported this interpretation (Carlsson, 1959; DeLong, 1983; Schultz et al., 1983), and more recent studies are also challenging the unique RPE view on dopamine.

During sequence learning, it was exhibited that dopaminergic activity in SNc reflected start and stop movement signals, which were action-specific (Jin and Costa, 2010). NAc dopamine concentrations were also shown to be shaped by action initiation but this was conditioned on whether the movement would generate reward. Dopamine increased in response to the 'go' cue, which indicated that a movement was required to obtain reward, but was not modulated in response to the 'no-go' cue, which signalled to not move to obtain reward (Syed et al., 2016). These results displayed dopamine with an extra role different from reward, RPE or motivation.

Importantly, dopamine neurons are modulated by movement onset even during spontaneous movements in the absence of cues or rewards, such as those which are the result from animals adjusting their position in a head-fixed setup (Dodson et al., 2016). Such dopamine activity modulation was not homogeneous: most of the neurons recorded in susbtantia nigra pars lateralis (SNL) increased their firing prior to movement onset, SNc neurons instead phasically decreased their firing at movement onset, and VTA neurons, only few of the recorded neurons presented, similar to SNL, increased their activity prior to movement. Computational modelling showed that SNc movement modulation could then be reflected in DMS dopamine release, while VTA firing could be reflected on dopamine concentrations in NAc. This relation between midbrain firing and striatal dopamine release was impaired in a Parkinsonian mouse model (Dodson et al., 2016). Supporting this functional segregation across the midbrain and striatum, Howe and Dombeck (2016) found that neurons coding movement-related parameters like movement acceleration, were an independent subpopulation from those encoding reward. While SNc dopaminergic axonal projections mainly represented movement, VTA axons instead were mainly modulated by reward. This then resulted in a functional dorsal-ventral gradient in dopaminergic projections to the striatum – ventral striatum modulation correlated with reward, dorsal striatum axonal activity correlated with movement acceleration, and central regions presented a mixture of both (Howe and Dombeck, 2016). It has also been reported that during high motivation, higher modulation by movement velocity and/or acceleration was present in VTA dopaminergic neurons. This movement-related activity was independent from reward-related dopaminergic activity in the same cells (Puryear et al., 2010).

Some others have argued that what might have been thought as representing reward or RPE signals, are instead movement-related signals mistakenly been understood as RPE due to the lack of or inaccurate movement measures (Barter et al., 2015; Hughes et al., 2020). In SNc, in a Pavlovian task, although activity seemed to support the RPE framework, further analyses suggested it was better explained by movement kinematics (up, down, left and right), with horizontal movement being represented in a lateralised way – dopamine was modulated by contralateral movement – independently from action or movement value. Furthermore, optogenetic activation of these neurons promoted movement generation (Barter et al., 2015). In VTA, also in a Pavlovian task, dopamine neurons appeared to be better explained by the impulse vector – the force employed across time. They found three different subpopulations of VTA dopaminergic neurons that were differently modulated by the force exerted. Optogenetic manipulation modified animals' force direction. However, the relation between movement and dopamine modulation was only in trained animals (Hughes et al., 2020).

Although it might be the case that insufficient movement measures in previous studies have led to inappropriately definitive conclusions linking all dopamine to RPEs, it should also be noted that a vast range of studies have postulated a functional topography across midbrain and the striatum, as exposed earlier, with SNc and dorsal striatum representing movement kinematics and contralateral movement, while VTA and ventral striatum being modulated by reward-related computation (Howe and Dombeck, 2016; Parker et al., 2016; Lee et al., 2019; Moss et al., 2021). Furthermore, it has also been evidenced that dopamine can multiplex different types of signals, both movement and reward-related computations (Puryear et al., 2010; Parker et al., 2016; Engelhard et al., 2019; Kremer et al., 2020). Kremer et al. (2020)

made electrophysiological recordings of dopamine neurons in VTA while freely moving animals were performing a self-paced instrumental task. Dopamine neurons were modulated both by positive and negative RPE, together with movement-related parameters such as movement velocity, acceleration, and distance to the goal. While RPE signals were represented as a phasic response, sustained slow firing was modulated by these movement parameters. Strikingly, signal multiplexing occurred when animals were performing the task, but not outside (Kremer et al., 2020). This engagement-dependent multiplexed activity might explain conflicting results in terms on whether dopamine reflect reward and RPE or movement information.

Furthermore, another possibility is that all these signals are prediction errors but over different dimensions. Striatal neurons have been shown to be modulated by movement conditioned on the expected outcome (Kawagoe et al., 1998; Samejima et al., 2005; Hikosaka et al., 2006). Recently, it has been shown that dopamine concentrations dynamics in the tail of the striatum are consistent with conveying an action prediction error, the difference between the action taken and the expected action given the state (Greenstreet et al., 2022). It is hypothesised that movement or action prediction errors would be computed in parallel with RPE by different subpopulations, and the outputs of both systems would be combined to generate actions (Miller et al., 2019; Bogacz, 2020; Greenstreet et al., 2022).

Although there is some current debate on whether dopamine is only modulated by movement information, or whether can multiplex different types of signals, all these studies suggest that at least a proportion of dopamine neurons indeed show activity correlated with

movement information. It also evidences that importance of detailed movement measures and video analysis to uncover the role of dopamine in motor control.

1.4.4 Internal model representation in the dopaminergic system

We have described different studies supporting the RPE framework of dopamine, plus others challenging this view. Another question arising is: what source of information does dopamine has access to support its computations. The traditional view proposed dopamine signalling conveyed a RPE, computed in a model-free way, used as a teaching signal to update future predictions and behaviour (Montague *et al.*, 1996; Schultz *et al.*, 1997; Doya *et al.*, 2002; Daw *et al.*, 2005; Eshel *et al.*, 2016).

However, as described in previous section (chapter 1.1.3), real-world situations can be dynamically changing, can be ambiguous and/or structured, and therefore, learning the underlying structure or rules of the environment can improve behavioural policy. For example, the optimal path to go home might depend on the time of the day, therefore, knowing or *inferring* the current time, current *state* of the environment, would allow to optimally make a choice of what path to take. Human (Hampton et al., 2006, 2007; Behrens et al., 2007; Nassar et al., 2010, 2012; Zylberberg et al., 2018), non-human primates (Costa et al., 2015; Jang et al., 2015; Rothenhoefe et al., 2017; Sarafyazd and Jazayeri, 2019; Bartolo and Averbeck, 2021) and rodents (Gallistel et al., 2001; Gershman et al., 2010; Vertechi et al., 2020) can use an internal model of the external environment to guide their choices.

Originally, these model-based inferences or predictions guiding behaviour was conferred to cortical regions, while dopamine was thought to be a competing model-free controller (Doya et al., 2002; Daw et al., 2005; Gläscher et al., 2010). In humans, Daw et al. (2011) showed that BOLD activity in striatal regions reflected both model-free and model-based predictions. In the same task, high levels of dopamine in ventral striatum was associated with higher model-based choices (Deserno et al., 2015), and pharmacologically increasing dopamine levels promoted model-based decisions in humans (Wunderlich et al., 2012b). These studies support the idea that dopamine might has access to an internal model or representation of the external environment.

Inferences and belief states

In one early study, electrophysiological recordings of dopamine neurons in non-human primates showed evidence that dopamine RPE were conditioned on an internal model of the task structure. In this study, Nakahara et al. (2004) recorded dopaminergic neurons in SNc, areas A8 and A9, in non-human primates, while they were performing a memory-guided saccade task. Four different targets were presented on a screen and subjects had to make a saccade to the location where one of the targets was flashing. If a correct saccade was performed, a tone and reward, or only the tone was delivered. This task had an underlying hidden structure: on each block, only one target delivered reward with 100% probability, while the others did not. Importantly, the flashing target was determined in a pseudorandom order such that, on every four trials, every target would flash once. After 60 trials, the rewarded target changed, and again, this was unsignalled.

If the structure of the task was learned, reward prediction or expectancy would change based on the trial history. Previous unrewarded trials increase the probability of reward on the next one. Moreover, if the block structure is also learned, that every 60 trials only one target delivers reward, at the time of the flashing cue, reward can already be predicted. The authors found that dopaminergic neurons reflected RPE signals conditioned on a *hidden* variable of the task: the post-reward trials number. Specifically, the higher the post-reward trial number, the lower the firing rate of dopamine neurons at outcome time, reflecting the higher expectancy of reward – i.e., the difference between expected reward and actual reward delivery is smaller. Crucially, this is the *opposite* prediction of a standard model-free RPE. Moreover, this conditioned RPE signal on the hidden variable developed with experience, being the strongest during late training. In line with this explanation, dopamine neurons also responded differently to the cue based on the current block – dopamine firing increased only when the rewarded target in that block flashed (whether it was going to be rewarded or not was not signalled). Except from the first trials in each block, when reward was delivered, as it was fully predicted, no differences in dopaminergic firing rate was observed at outcome time. In sum, Nakahara et al. (2004) was one of the first reports showing that dopamine neurons computed a context- or state-dependent RPE, by inferring the state of the environment through tracking a hidden variable.

Dopamine value inference was further demonstrated by Bromberg-Martin et al. (2010). They trained monkeys on a reversal task in which one of two targets was rewarded with 100% probability while the other did not. Which target was rewarded changed in blocks in a way that was unsignalled to the subjects. The authors presented two different hypotheses for value learning. First, subjects could *learn by experience*, as TD learning would predict, and

therefore, on each trial, the value of the experienced target would be updated based only on the outcome (i.e., increased or decreased if the outcome was better or worse than expected, respectively). The second hypothesis was that subjects could *learn values by inference*. If the structure of the task was learned, that rewards are anticorrelated – i.e., if one target is rewarded, the other is not rewarded – then receiving reward on a trial would not only update the value of the experienced target, but also would update the value of the non-experienced one in the opposite direction. In fact, dopamine neuron firing in SNc was modulated by both *experienced* and *inferred* value. On the first trial after a block switch, dopamine neurons reflected the unexpected outcome (e.g. increased their firing at outcome when reward was delivered at the target that was previously not rewarded on the previous block – positive RPE). This did not only result in a change in behaviour (e.g. increased reaction time if the new unrewarded target appeared), but dopamine neurons reflected the *inferred value update* of the non-experienced target on the second trial (e.g. decreased firing if the new (inferred) unrewarded target appeared, and no change in firing rate was observed at outcome delivery, reflecting that outcome was fully expected) (Bromberg-Martin et al., 2010b). This has also been observed in human studies. BOLD striatal responses (Lohrenz et al., 2007; Chiu et al., 2008) and FSCV recordings of dopamine concentrations in Parkinson's patients (Kishida et al., 2016) reflected both experienced and counterfactual or fictive (i.e. inferred value of what would had happened if the other option was chosen) RPE during a sequential investment game.

Modelling how the dopamine system deals with ambiguity in the external environment, Starkweather et al. (2017) showed that dopamine cell firing was best explained through a TD model with *belief state*, the inferred hidden state distribution. Mice were trained in two

different tasks, a deterministic and a probabilistic task. In the deterministic task, reward was delivered with 100% probability, while in the probabilistic task with 90% probability. In both tasks, the interval between conditioned cue and reward delivery was variable. While in the deterministic task, longer intervals resulted in smaller dopamine neuronal activity in VTA, in the probabilistic task the opposite occurred, longer intervals resulted in higher firing rates. Dopaminergic modulation in the deterministic task could be explained through the traditional TD error because reward was delivered with 100% probability: the longer the interval, the higher the expectancy that reward would be delivered. But the traditional RPE computation could not explain the results in the probabilistic task. Instead, it could be explained through a TD error incorporating a *belief state* about whether the trial was an omission or a delayed reward trial. Therefore, in here, the hidden variable was the time interval. Dopamine activity RPE computations were conditioned on the time interval. The longer the interval, the higher the probability that it might be an omission trial, and therefore, if delayed reward is still delivered, higher dopamine reflects a more unexpected outcome. This study further confirmed that dopamine computations can be informed by hidden-state inferences (Starkweather et al., 2017).

Previously, it was shown in a Pavlovian task, that dopamine activity was sensitive to the timing of the reward predictions. However, the temporal specificity of the reward predictions were not very precise and declined with longer time interval durations (Fiorillo et al., 2008). This temporal prediction was impaired after NAc ipsilateral neurotoxic lesion, without impairing RPE computation involving amount size (Takahashi et al., 2016). Soares et al. (2016) demonstrated that dopamine system could assess time elapsed between cues to inform behaviour. Animals were presented two auditory cues separated by different time intervals.

Animals were asked to categorise the time intervals in short or long durations. Dopamine activity modulation in SNc reflected the different time intervals in a graded manner. Higher dopamine activity was observed with shorter time intervals. In incorrect trials, where mice misclassified the time interval duration, dopamine activity also reflected the miscoding of the time interval, exhibited with higher dopamine activity during longer intervals. Optogenetic activation of dopaminergic neurons slowed down the internal clock increasing the categorising of short interval for longer ones, and the opposite for optogenetic inhibition, although the effect was more subtle (Soares et al., 2016). This study was understood in the view that dopamine controls the internal clock of events. However, it could also be interpreted as dopamine having access to an internal model keeping track of the time elapsed, which is used as a variable informing the animals' belief state, is the task in the short interval state or the long one, which is used to guide behaviour.

Further supporting state inference, Babayan et al. (2018) further confirmed that dopamine RPEs in VTA and NAc were computed based on the belief of the current state of the task. Mice were trained in a task that could be in two different states, either in a big reward state, where a big water reward (10ul) was delivered, or in a small reward (1ul) state, were a small water reward was delivered. Blocks consisted of 5 trials in a certain state – 5 trials with big reward or 5 trials with small reward. The block type – big or small – was not signalled but could be inferred after the first trial. Additionally, however, rare blocks with intermediate reward sizes were introduced. In these intermediate reward blocks, while dopamine calcium activity reflected the traditional RPE in the first trial, on the second trial that was not the case. If reward size was lower than 5ul (mid reward size between big and small states), dopamine reflected the traditional RPE, bigger response with bigger rewards, consistent with a positive

RPE. But if reward was bigger than 5ul, this modulation was not followed, instead lower activity than would be predicted was observed. Dopamine modulation on the second trial could again only be explained if the animal's belief state is taken into account – if on trial 1 reward is smaller than 5ul, subject would *infer* that the environment is in the small reward state (1ul); therefore, any reward bigger than that, would generate a positive RPE, as reward would be bigger than expected in that state. If on trial 1 reward is bigger than 5ul, subjects would infer that the environment is in the big reward state; therefore, when on trial 2 reward is smaller than 10ul, a negative prediction error would signal a worse than expected outcome. It was concluded that dopamine value computations are influenced by hidden-state inference (Babayan et al., 2018).

We have seen that belief states can be driven by observable features in the environment, which are used to infer hidden features or variables in a structured external world. This observable features could be number of non-rewarded trials (Nakahara et al., 2004), a single unexpected outcome (Bromberg-Martin et al., 2010b), amount of time elapse since cue presentation (Soares et al., 2016; Starkweather et al., 2017), or initial reward size (Babayan et al., 2018). These variables were used to infer the hidden states of the task, which are used by the dopamine system to compute RPE according to it, based on the state subjects believe the task is in. For example, let's imagine we are at a town fair where there are different games, shooting and archery, to win a prize. Many people successfully win a prize from the shooting game but not the archery one. If we play the shooting, we will also have high expectation to win the prize. If we win it, our expectation would be match with the actual outcome, however, if we lose, that would be highly unexpected for us. Comparably, if instead we play the archery where we have low expectancy to win, not winning would be fully expected, but if we win, a

big positive RPE would be generated, as that was extremely unexpected. Perceptual and sensory confidence or uncertainty has also been shown to determine the belief state used to generate dopaminergic computations in uncertain or ambiguous situations (Lak et al., 2017, 2020).

Not only reward is encoded…

The presence of an internal model that dopamine has access to may not just affect reward responses, but also may reflect sensory prediction errors.

During a sensory preconditioning task in rats, it was also shown that VTA dopaminergic neurons firing reflected inferred value computation. In this task an initial cue A was followed by a cue B, and cue C was followed by cue D (A→B, and C→D), and in a different context, cue B was followed by reward while cue D did not. Dopamine firing increased when cue B was presented, consistent with RPE signalling. During probe test, rats responded more to cue A than D, reflecting transitive inference (i.e. A→B→reward). Importantly, dopamine firing increased at cue A presentation despite the fact that rats never experienced this cue being followed by reward. Therefore, dopamine modulation to the pre-conditioned cue reflected inferred value computation (Sadacca et al., 2016).

Further confirming the access by the dopaminergic system to internal models of the external environment, Takahashi et al. (2017) exhibited that VTA dopamine neurons signalled violations of reward identity, despite having the same value. They concluded that dopamine do not only signal deviations from expected and actual reward, but also predicts sensory

features (Takahashi et al., 2017). Furthermore, optogenetic activation of VTA dopamine neurons promoted (Sharpe et al., 2017, 2020; Keiflin et al., 2019) while inhibition impaired (Sharpe et al., 2017) stimulus-stimulus associations. But this was specific from VTA dopaminergic neurons and not SNc (Keiflin et al., 2019).

Furthermore, the presence and influence of an internal model over dopaminergic responses have also been observed in zebra finches where dopamine reflected prediction errors of song performance, and reflected disturbances of the produced and perceived song (Gadagkar et al., 2016). These suggest that the dopaminergic system might have a conserved, general role as a teaching signal, where reward is simply one particularly salient feature.

1.4.5 Diverse functions and interactions

Throughout this chapter, we have seen dopamine is functionally heterogeneous, with different studies suggesting a diverse range of functions, such as RPE, motivation and movement. Not only that, but dopamine has also been related to other functions, such as attention (Bromberg-Martin et al., 2010a), novelty seeking (Wittmann et al., 2008), uncertainty (Frank et al., 2009), risk-taking (Stopper et al., 2014) and credit assignment (Coddington and Dudman, 2019). But how a single signal can convey such a range of information? And if this is multiplexed (Engelhard et al., 2019; Kremer et al., 2020), how this can then be later decoded?

One possibility, as described earlier, is that specific population of dopaminergic neurons with projection-specific inputs and outputs are responsible for specific functions (Lammel et al.,

2008, 2011, 2012; Cohen et al., 2012; Beier et al., 2015; Menegas et al., 2015; de Jong et al., 2019; Keiflin et al., 2019; Cai et al., 2020). Another possibility is that the different roles of dopamine change not only depending on the task, whether an instrumental behaviour is required or not, but also depending on experience. Coddington and Dudman (2018) recorded projection-specific dopamine neuronal firing in mice in VTA and SNc in a Pavlovian task. They found that midbrain dopamine cell firing multiplexed sensory- and movement-related signals, and the summation of both explained dopamine modulation. But this combined activity changed with experience (Coddington and Dudman, 2018a).

This can also be understood as dopamine accessing an internal model of the environment, naïve animals do not know the structure of the external world, therefore, dopamine activity poorly reflects sensory- and movement-related activity. With experience, the statistical structure of the environment is learned (i.e. an internal model is built), and dopamine reflects it through its increased modulation of relevant sensory and motor features. The presence of an internal model of the external world can also explain the diversity in dopamine functionality, as based on the model and behavioural requirement, dopamine might reflect different observable and non-observable features.

Regarding demultiplexing signals, it has been suggested this could be performed at the level of the postsynaptic site through a range of mechanisms. Different firing modes, phasic and tonic, have been proposed to support this (Niv et al., 2007; Kremer et al., 2020). Different receptors distribution, as described in chapter 1.3.2, can respond differently to different firing patterns, and can also lead to different dopamine concentrations. Furthermore, the dopamine system is not isolated, but at the level of the postsynaptic site, cholinergic

interneurons interact with dopaminergic axons modifying their pattern of activity (Zhang and Sulzer, 2004; Rice and Cragg, 2008; Threlfell et al., 2012; Liu et al., 2022). At the same time, different neurotransmitters have been suggested to jointly support different computations. For example, Doya et al. (2002) proposed different neurotransmitters could support TD learning. Dopamine would convey the TD error, acetylcholine would regulate the learning rate, the temporal discount factor would be controlled by serotonin, and noradrenaline would control stochastic exploration (Doya et al., 2002).

1.5 Thesis aims

Real-world scenarios are rarely fully observable. In such situations, different behavioural strategies could be used in order to flexibly adapt to dynamic changes requiring a change in our behaviour in order to efficiently reach our goals. As exposed at the beginning of this introduction, model-free and model-based strategies within the RL framework have been shown to computationally describe different goal-directed behaviours. However, the strict dichotomy between these two models has been put in doubt, plus more strategies have been put forward to describe flexible behaviour. In this thesis, taking advantage of the features of the two-step task, we will explore the different behavioural strategies mice use to solve a partially observable, but structured, dynamically changing environment controlled by a hidden latent variable, allowing state inference.

RL strategies have been closely related to the dopamine system, through its role of conveying RPE used as a teaching signal to update state values and behavioural policy. Although a vast range of research has studied different features of this signal, little is known about i)

dopamine activity in complex tasks in which latent states interact with causal relationships in the environment and decision policies, ii) the relation between somatic and axonal dopaminergic activity, and this to striatal dopamine concentrations, and iii) the causal role of dopamine in the behavioural policy in such partially observable environments.

Therefore, the main aims of this thesis are to:

- Examine mice's RL strategies during a real-world-like scenario: a partially observable environment controlled by a hidden latent variable (Chapter 3).
- Record somatic (VTA) and axonal (NAc and DMS) dopaminergic activity, and dopamine concentrations in the same striatal targets during flexible behaviour in such structured environment (Chapter 4).
- Establish the causal role of VTA dopaminergic activity in choice behaviour in a structured environment with hidden states (Chapter 5).

2 General methods

In this thesis we trained mice in a reduced version of the two-step task. While mice were solving the task, we recorded dopamine somatic and axonal calcium activity in VTA, NAc and DMS, and dopamine concentrations in NAc and DMS. Finally, we optogenetically manipulated VTA dopaminergic activity at different timepoints in the two-step task. In this chapter, we will first describe the two-step task and the training protocol. This will be followed by the optimisation procedure of the task where different modifications of the two-step task were tested. Main surgical procedures will be described. A summary of the photometry recordings and optogenetic manipulations from chapter 4 and 5, respectively, will be explained. Finally, we will present the histological analysis of all the experiments.

2.1 Subjects

All procedures were performed in line with the UK Animal (Scientific Procedure) Act 1986 and in accordance with the University of Oxford animal use guidelines, and were approved by the local ethical review panel at the Department of Experimental Psychology, University of Oxford. A total of 55 C57BL6 female and male mice were used for the behavioural and photometry piloting. For the dopamine recordings and optogenetic manipulations, another 42 mice were used. From those, 12 dopamine transporter (DAT)-Cre heterozygous mice (7 females and 5 males) were used for the GCaMP photometry recordings, 6 WT C57BL/6 mice (3 females and 3 males) for the dLight recordings, 12 DAT-Cre mice (YFP: 2 females and 3 males; ChR2: 4 females and 3 males) for the optogenetic activation experiment, and 12 DAT-Cre mice (tdTomato: 3 females and 2 males; GtACR2: 4 females and 3 males) for the optogenetic inhibition experiment. Mice were aged 8-16 weeks at the start of behavioural training. Animals were typically housed in groups of 2-4 throughout training and testing, maintained in a 12h light/dark cycle with food provided *ad libitum.* Mice were water-restricted and kept above 85% of their initial body weight. The water restriction procedure during training consisted normally of 6 days of training where animals received a minimum of 1ml of water daily, followed by 1 rest day where they received 1h of water *ad libitum*. During testing, animals were tested daily up to a maximum of 28 days, after which a rest day was given with 1h of water *ad libitum*.

2.2 Two-step task

2.2.1 Behavioural setup

The task was run in custom-built 12x12 cm operant boxes (design files at https://github.com/pyControl/hardware/tree/master/Behaviour_box_small) controlled using pyControl (Akam et al., 2021a). Five nose poke ports were located on the back wall of the boxes; a central initiation port flanked by two choice ports 4 cm to the left and right and two second-step ports located 1.6 cm above and below the central poke. The second-step ports each had a solenoid to deliver water rewards. A speaker located above the ports was used to deliver auditory stimuli. Video was acquired from an FLIR Flea 3 camera positioned above each setup using a Bonsai based workflow (https://github.com/ThomasAkam/Point_Grey_Bonsai_multi_camera_acquisition) (Lopes et al., 2015).

2.2.2 Behavioural task

The behavioural task was adapted from the human two-step task (Daw et al., 2011). Each trial started with the central initiation port lighting up. Subjects initiated the trial by poking the illuminated port, which caused the choice ports to illuminate. On free-choice trials (75% of trials), both the left and right port lit up, allowing subjects to choose, while on forced-choice trials only one randomly selected choice port lit up, forcing animals to select that specific port. Poking a choice port was followed, after a 200ms delay, by the second-step port lighting up and 1s presentation of one of two auditory cues – second-step cue - (either a 5- or 12-kHz tone depending on whether the top or bottom second-step became active, counterbalanced across animals). Each choice port was commonly (80% trials) associated

with transitioning to one second-step state (up or down) and rarely (20% trials) to the other. The transition structure was fixed for each animal across all sessions, and was counterbalanced across animals – i.e. the task had 2 possible transition structures: transition type A, where left choice commonly led to the up second-step port, and right choice commonly led to the down second-step port; and the opposite for transition type B. The second-step port only became responsive to pokes after cue offset. Poking the second-step port triggered a 200ms delay, after which a 500ms auditory cue signalled whether the trial was rewarded or not (same 5- or 12-kHz tone as the second-step cue (counterbalanced across animals) pulsed with at 10Hz on rewarded trials, white noise on unrewarded trials). Reward was delivered at the offset of this cue. To ensure mice knew when they had made a nose poke, a click sound was presented whenever the subject poked a port that was active (e.g., the initiation port during the initiation state).

Reward probabilities for the two second-step ports changed in blocks. In non-neutral blocks, one second-step port had 80% reward probability and the other 20%, while in neutral blocks both second-step ports were rewarded with 50% probability. Block transitions from non-neutral blocks were triggered 5 to 15 trials after mice crossed a threshold of 75% “correct” choices (i.e., choosing the higher reward probability), computed as the exponential moving average with time constant of 8 free choice trials. Transitions from neutral blocks were triggered after 20-30 trials. An inter-trial interval of 2-4 s duration started once the subject remained out of the second-step port for 250ms after the trial outcome. Sessions typically lasted 90 min.

2.2.3 Behavioural training

Behavioural training took 4-6 weeks. Animals were put on water restriction 48 hours before starting training, and received 1 hour ad lib water access in their home cage 24 hours before starting training. On training days (typically 6 days/week) animals usually received all their water from the task, but were topped up outside the task as necessary to maintain a body weight >85% of their pre restriction baseline. On days off from training mice received 1 hour ad libitum water access in their home cage. Water reward size was decreased from 15ul to 4ul across training in order to increase the number of trials performed and blocks experienced on each session.

Training consisted of multiple stages with increasing complexity (Table 2.1). At the first training stage 1.1, only the second-step ports were exposed, with all other ports covered. Second-step ports were illuminated in a pseudorandom order with inter-trial interval of 2-4 seconds. Poking an illuminated port delivered reward with 100% probability, with no auditory cues. When animals obtained >50 rewards on this stage, transitioned to stage 1.2, where the auditory cues for second-step state and reward were introduced. Once animals were obtaining >70 rewards in a session, they were switched to stage 2 on the next session. At stage 2, the choice ports were introduced, but all trials were forced choice, such that only one choice port lit up and was active on each trial (responses in the non-illuminated port had no consequences). Mice were switched to stage 3 when they obtained >70 rewards on a single session. At stage 3, the initiation poke was introduced, and when they obtained >70 rewards on a single session, they moved to stage 4 on the next training session. Stage 4 consisted of multiple substages where progressively more free choice trials were introduced, and the

reward probabilities gradually changed until reaching the final task parameters. Mice moved to the next substage after two training sessions of 45 min or a single 90 min session, until they reached substage 4.6 where animals were training on a single 90 min session per day. Subjects were only moved to the final stage (full task) when they completed at least 5 blocks in a single session.

STAGE	START POKE	CHOICE POKES	REWARD POKES	TASK STRUCTURE
1.1			X	Reward probability 100%
1.2				Introduce auditory cues Reward probability 100%
2		X	X	Transition probability 80/20% Reward probability 100%
3	X	X	X	Transition probability 80/20% Reward probability 100%
4	X	X	X	Start introducing free choice trials and reward probabilities
SUBSTAGE	**Free choice trials**	**Non-neutral blocks reward prob.** (high / low)	**Neutral block reward prob.**	**Block change** (independent / dependent on behaviour)
4.1	0%	0.9/0.7	0.8	Independent – after 20-30 trials
4.2	25%	0.9/0.5	0.7	Independent – after 20-30 trials
4.3	25%	0.9/0.3	0.6	Independent – after 20-30 trials
4.4	25%	0.9/0.1	0.5	Independent – after 20-30 trials
4.5	50%	0.9/0.1	0.5	Dependent
4.6	75%	0.9/0.1	0.5	Dependent
4.7 – FINAL	75%	0.8/0.2	0.5	Dependent

Table 2.1. Description of training stages and task contingencies

2.2.4 Behavioural analysis

Four main analyses were performed on animals' choices, which others have used to behaviourally distinguish model-free from model-based behaviour (although note that as exposed earlier, this distinction is not so clear (Akam et al., 2015)). These were the stay probability analysis, a logistic regression and two different lagged regression models. The stay probability showed the probability of repeating a choice based on the outcome – reward or omission – and transition – common or rare – experienced on the previous trial.

2.2.4.1 Logistic regression model

The one-trial back logistic regression analysis predicted repeating choices (or "staying") as a function of the subsequent trial events, considering only free choice trials, implemented as a mixed effects model using the afex package (Singmann et al., 2018) in R programming language. The model formula was:

$$stay = correct + bias + transition * outcome + (random\ effects | subject)$$

We used orthogonal sum-to-zero contrasts, and likelihood ratio test to calculate p-values. The maximal random effect structure(Barr et al., 2013; Matuschek et al., 2017) with subject as a grouping factor was used.

The predictors were coded as:

- *Correct*: 3 level categorical variable indicating whether the previous choice was correct, incorrect or in a neutral block.
- *Bias*: Binary categorical variable explaining whether previous choice was left or right.
- *Outcome*: Binary categorical variable indicating whether the previous trial was rewarded or not.
- *Transition*: Binary categorical variable indicating whether previous trial had a common or a rare transition.

2.2.4.2 Lagged logistic regression model

The lagged logistic regression analysis assessed how subjects' choices were affected by the history of trial events over the last 12 trials. The regression predicted subjects' probability of choosing left, using the following predictors at lags 1, 2, 3-4, 5-8, 9-12 (where lag 3-4 etc. means the sum of the individual trial predictors over the specified range of lags).

We performed two different lagged regressions. The first included similar regressors as in the one trial back regression (see above). The predictors were effect coded as:

- *Choice at lag* n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by a left choice, and the opposite if followed by a right choice. This regressor controls for repeating previous choice.
- *Outcome at lag* n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by reward, and the opposite if followed by an unrewarded trial.
- *Transition at lag* n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by common transition, and the opposite if followed by a rare transition.
- *Transition x Outcome at lag* n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by common transition rewarded or rare transition unrewarded, and the opposite if followed by common transition unrewarded or rare transition rewarded.

The second type of lagged regression included the following predictors

- *Common Transition*: Rewarded at lag n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by a common transition and reward, 0 otherwise.
- *Rare Transition*: Rewarded at lag n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by a rare transition and reward, 0 otherwise.

- *Common Transition*: Unrewarded at lag n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by a common transition and no reward, 0 otherwise.
- *Rare Transition*: Unrewarded at lag n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by a rare transition and no reward, 0 otherwise.

The lagged regressions were fitted separately for each subject as a fixed effects model. Cross subjects mean and standard error for each predictor coefficient were plotted, significance of coefficients was assessed using a two-sided t-test comparing the distribution of the individual subjects' coefficients against zero, and performing Bonferroni multiple comparison correction.

2.2.5 Optimising the two-step design

The goal of this optimisation was to obtain mice behaviour that resembled humans' model-based or hybrid choice behaviour as shown in Daw *et al.* (2011) – differential effect of repeating choice based not only on the outcome experienced (reward or omission) but also the transition (common or rare). In this section we will start with versions of the task with in which both transition and reward probabilities reversed. Later we simplified it by "fixing" the transition probabilities (i.e., they remained constant across sessions for each animal), and only reversals of the reward probabilities occurred.

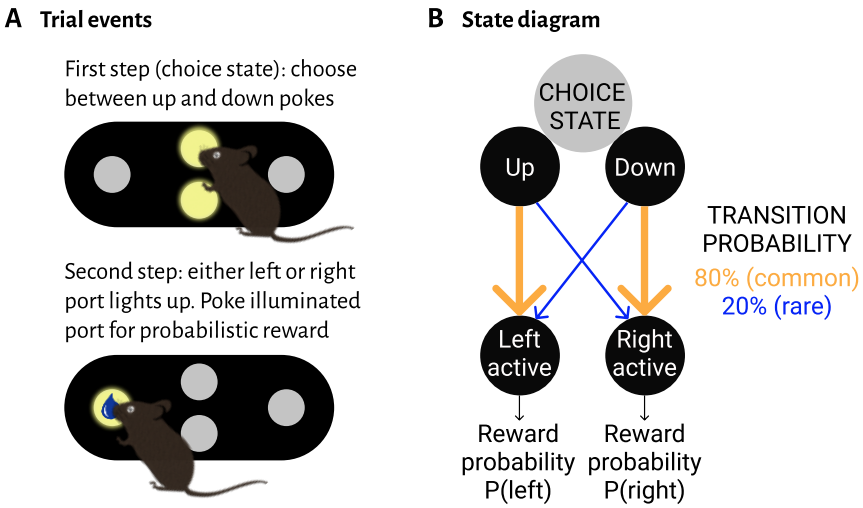


Figure 2.1. *Two-step task from* Akam et al. (2021b). *A) Trial events. B) State diagram. Each trial started with the choice state, were mice had to choose between the up and down port. Each choice with different probability, common and rare, transitioned to one second-step port, left or right. The second-step ports were probabilitistically rewarded, with probabilities that changed in blocks.*

We started from the version of the two-step task for mice developed by Akam et al. (2021b). In this paradigm, there is an initial choice state, where mice have to choose between a high or a low nose poke port, and each choice probabilistically transitions to a second-step state, left or right poke illuminated, which probabilistically delivers reward (Figure 2.1). However, Akam et al. (2021b) found that this implementation confounded motor and choice behaviour, as the motor action required to make a choice depended on the second-step reached on the previous trial.

To overcome this limitation, we introduced a central initiation poke (Figure 2.2). Therefore, the second-step reached on the previous trial should have less of an influence on the motor action performed on the next choice as subjects would always have to return first to a

common initiation state. Moreover, in order to aid state discrimination, each nose port had different sensory features (i.e. different shape), and two different auditory cues signalled different task states. The first auditory cue signalled the transition from choice to second-step state, which consisted of a 1s sinusoidal sound at different frequency depending on the second-step port illuminated (5- and 12-kHz). The next auditory cue signalled the outcome, after the second-step port was approached, a 500ms sound was delivered, pulsed sound at the same frequency as the previous cue for rewarded trials, and white noise for unrewarded trials.

2.2.5.1 Two-step task with reversal of reward and transition probabilities

Given the strong influence of motor parameters on task performance observed in previous experiments (Korn *et al.*, 2021; Akam, Rodrigues-Vaz, *et al.*, 2021), we investigated two slightly different sequential implementations of the two-step task to investigate how this affected the learning strategies (Figure 2.2). On one, the choice ports were the left and right ports, and the second-step reward ports were up and down ('UD' version). On the other version, the choice and second-step ports were reversed (choice ports: up/down → reward ports: left/right, 'LR' version). In both versions, the action – state transition probability and the reward probability reversed periodically based on the mice's performance. Block transition was triggered following a random delay of 5 to 15 trials after mice crossed a threshold of 75% correct (higher reward probability) choices, computed as the exponential moving average of the last 8 choice trials on non-neutral blocks. Transitions from neutral blocks were triggered after 20-30 trials.

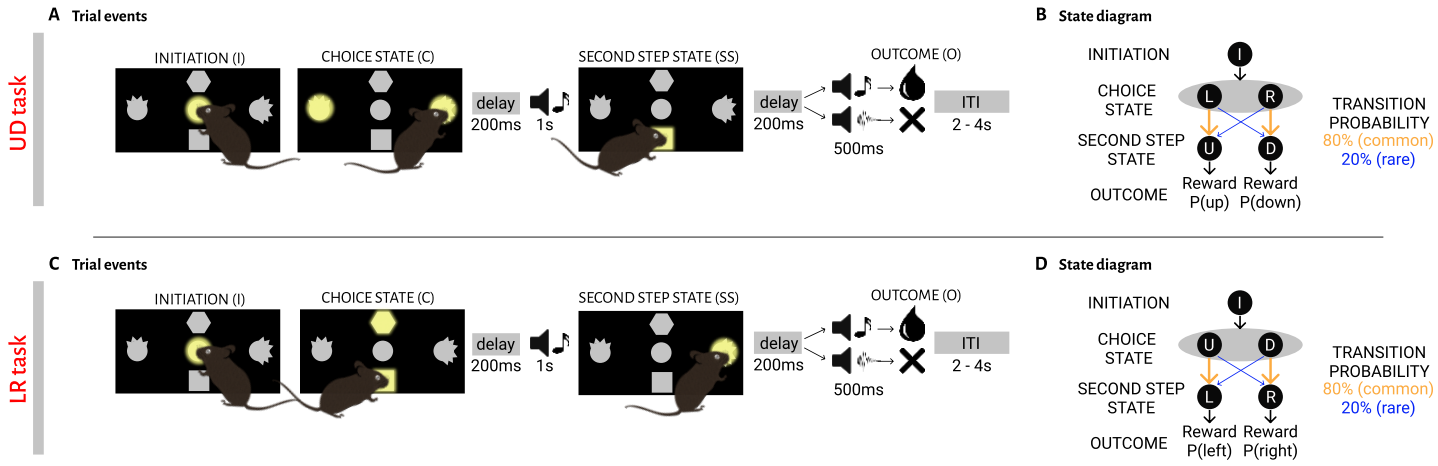


***Figure 2.2**. Two-step task with different layouts. A,B) UD task (N=9 mice). Choice state: left-right; Second-step state: up-down. C,D) LR task (N=9 mice). Choice state: up-down; Second-step state: left-right. Both transition and reward probabilities reversed in blocks.*

However, in neither version of the task did the mice exhibit behaviour consistent with a model-based agent (Figure 2.4) – there was not an effects of outcome interacting with transition influencing subsequent choices. Mice (N=9 for UD task, 9 for LR task) reached a similar behavioural plateau, performing ~300 trials (312.01±91.14 and 322.44±94.30 trials, on the UD and LR layout respectively), obtaining ~150 rewards (159.51±49.61 and 162.49±49.34 rewards on UD and LR layout respectively) and experiencing ~5 reversals (5.66±2.61 and 5.10±2.34 reversals on UD and LR layout respectively) on each session (Figure 2.3). There were no differences between groups on any of these measures on each session (trials: $F(1,16)=0.118, p=0.736$; rewards: $F(1,16)=0.034$, $p=0.855$; blocks: $F(1,16)=0.534, p=0.475$).

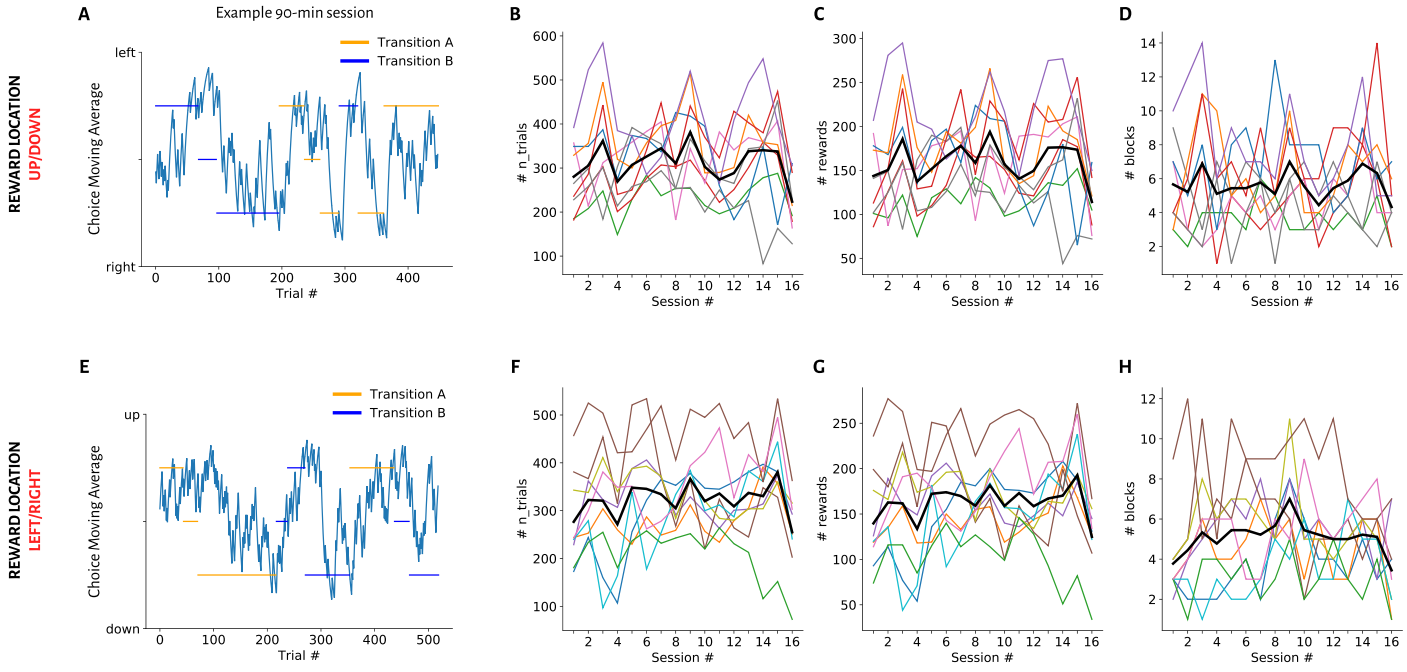


Figure 2.3. *Behavioural performance on the different two-step task layouts. A-D) UD layout. E-H) LR layout. A,E) Example 90-min session. Blue trace, exponential moving average of animal's choices. Horizontal orange and blue lines denote correct choice on that block and the transition state. B,F) Number of trials performed across sessions. C,G) Number of rewards received across sessions. D,H) Number of blocks experienced on each session.*

However we found that each group (UD and LR) developed a different behavioural strategy (Figure 2.4). We first performed a "stay probability" analysis, which examines the probability a choice was repeated on the next trial as a function of the state transition (common or rare) and trial outcome (rewarded or not) of the previous choice. An ANOVA showed a main effect of trial type ($F(3,48)=13.648, p<0.0001$), main effect of choice direction (i.e. whether choice is in the horitzontal or vertical ports) ($F(1,16)=13.105, p=0.002$) and a significant trial type x choice direction interaction ($F(3,48)=7.574, p<0.0001$). Whereas in the UD layout mice's likelihood of repeating choices were primarily influenced by the outcome and not by the

transition probability (consistent with model-free RL) (Figure 2.4A), the LR stay probability presented a pattern of behaviour that appeared more complex than a simple model-free strategy (i.e., repeat rewarded choices, switch after omissions) in which transition interacted with reward (Figure 2.4E).

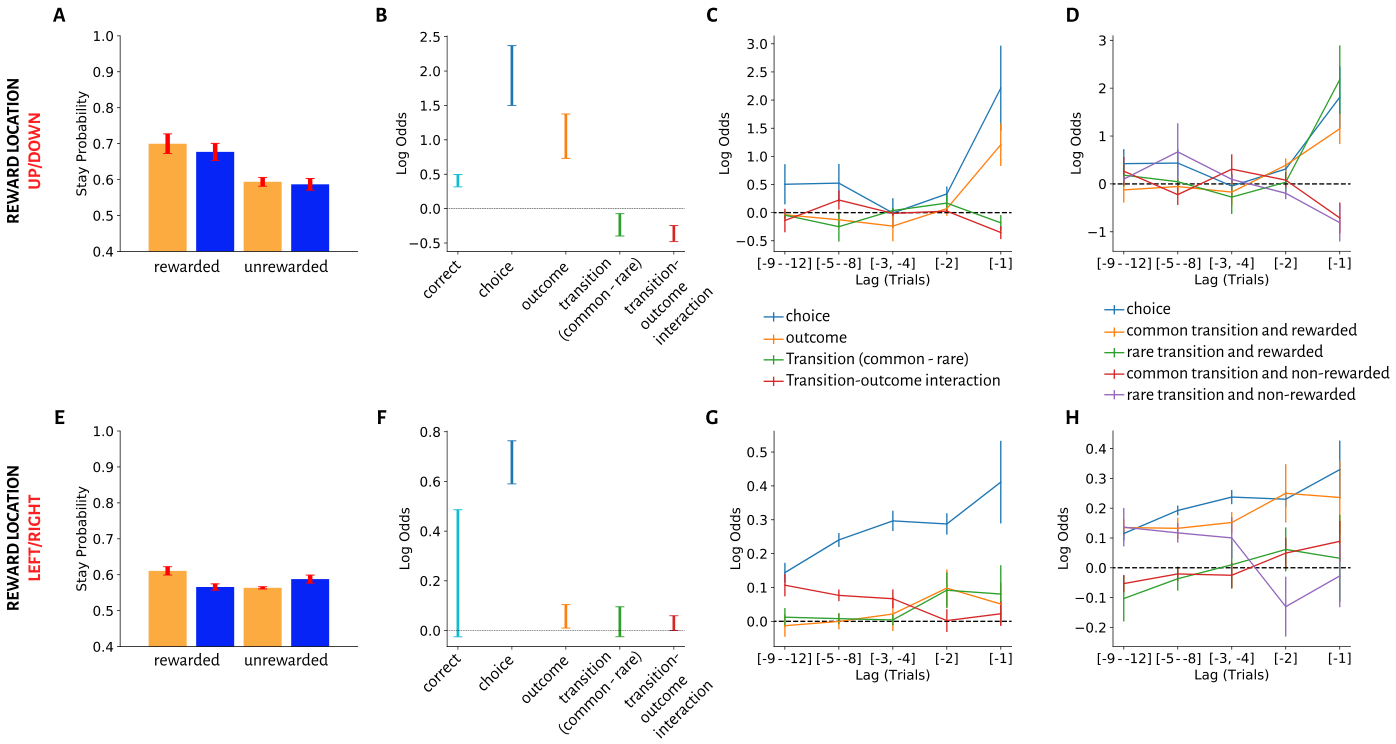


Figure 2.4. Behavioural analysis on the two-step task with changes in transition and reward probabilities on the different layouts. A-D) UD layout. Orange, common transition; Blue, rare transition. E-H) LR layout. A,E) Stay probability analysis showing the probability of repeating choice as a function of state transition (common or rare) and trial outcome (rewarded or not). B,F) One trial back logistic regression. C,D,G,H) 12-trials back logistic regression.

To better understand the strategies the animals were using to make their choices, we next performed a logistic regression based on the previous trial. Surprisingly, neither transition nor transition-outcome interaction coefficients were significantly different from 0 in either the UP *or* LR configuration. Instead, there was a main effect of choice (Figure 2.4B,F)

demonstrating that animals tended to repeat previous choice. In the UD layout, there was also a major impact of outcome on animals' subsequent choice behaviour.

We further analysed the impact of trial history on the choice behaviour of the current trial. In the UD layout, lagged regression on the last 12 trials showed that outcome and choice on the previous trial were the major influence over choice behaviour on the current trial, and no major impact of previous trial history (Figure 2.4C,D). Behaviour resembled a biased choice repetition, which could be considered a model-free strategy. By contrast, choice behaviour on the LR layout did not obviously match the predictions of either a model-free or a model-based agent. Trial history analysis suggested that previous choices had a major impact on animals' current choice (Figure 2.4G), with them exhibiting a strong likelihood of repeating choices that were rewarded and came from a common transition (Figure 2.4H). Nonetheless, overall, it was difficult to categorise the behavioural strategy that animals were following.

2.2.5.2 Two-step task with reversal of reward probabilities ONLY

In order to try to promote behaviour that was sensitive to both the trial outcome *and* transition probabilities, we therefore decided to focus on the UD version but move to a simplified version in which the transition probabilities were fixed, like in the original two-step task from Daw et al. (2011) (the reward probabilities still reversed as before).

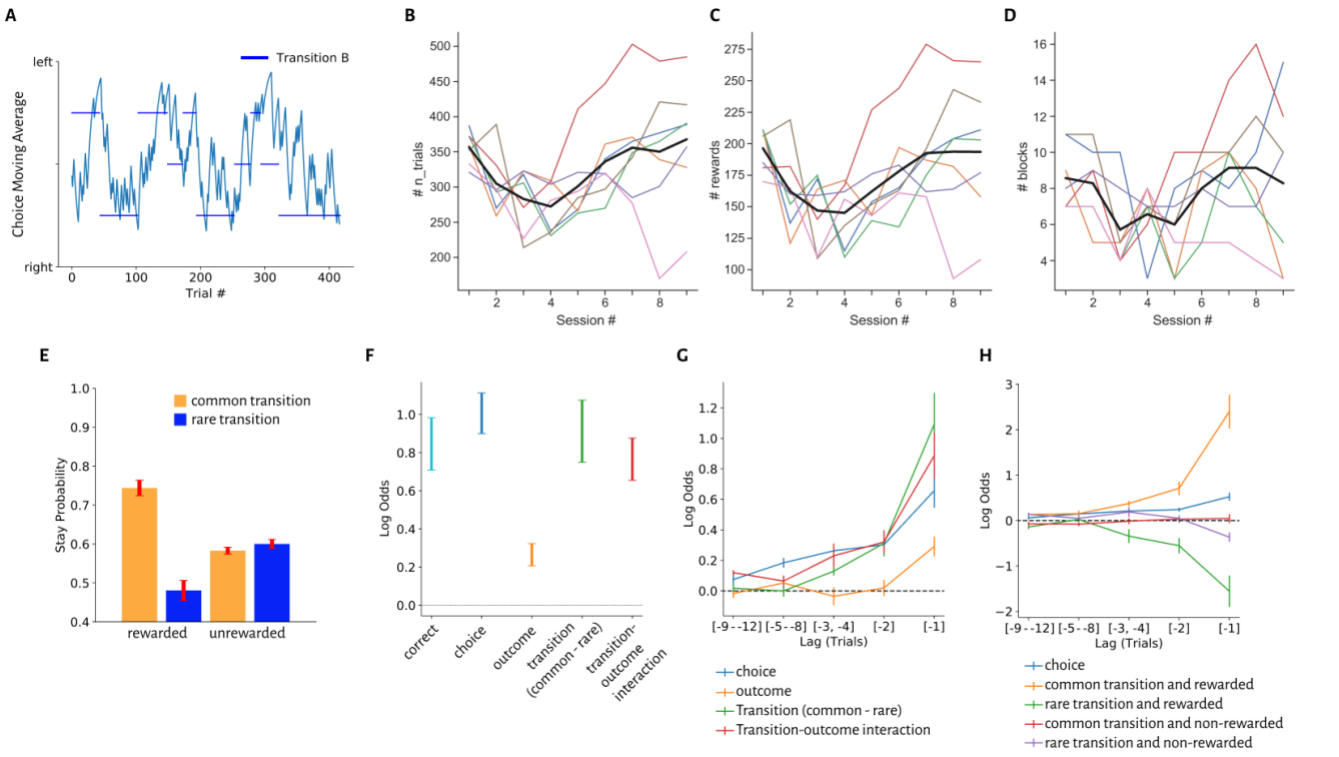


Figure 2.5. Behavioural performance and analysis of the two-step task with fixed transition probabilities. A) Choice moving average across and example session. B) Number of trials across sessions. C) Number of rewards obtained across sessions. D) Number of blocks experienced across sessions. B-D) Black trace, mean number; coloured traces shown individual animals. E) Stay probability. F) Logistic regression. G-H) Lagged regression.

After several sessions on the full version on the two-step task with fixed transition probabilities, mice (n=7) performed 367.86 ± 86.00 trials, received 193.57 ± 51.50 rewards and experienced 8.29 ± 4.68 blocks on a single session (Figure 2.5A-D). Choice behaviour was consistent with mice using model-based RL as a component of their strategy (Daw et al., 2011), with trial outcome – reward or no reward – and state transition – common or rare – interacting to determine subsequent choice ($F(3,18)=41.947, p<0.001$) (Figure 2.5E, H). Importantly, logistic regressions using either the previous trial or recent trial history to predict choice also showed a strong effect of both the transition-outcome interaction and state

transition on choices (Figure 2.5F, G). The behavioural patterns will be described in more details in chapter 3 with more animals.

Therefore, we used this version of the task as the main task of this study (task described in section 2.2.2 above in more detail).

2.3 Surgery

For the fibre photometry recordings (chapter 4) and the optogenetic manipulation experiments (chapter 5), animals were intracranially injected with viral constructs and fibre optics.

2.3.1 General surgical protocol

Mice were anaesthetised with isoflurane (3% induction, 0.5-1% maintenance), and injected with buprenorphine (0.08 mg/kg), meloxicam (5 mg/kg) and glucose saline (0.5ml). Marcaine (max 2 mg/kg) was injected into the scalp and before placing mice into the stereotaxic frame. Mice were maintained at ~37 degrees with rectal probe and heating blanket (Harvard Apparatus). Different viral constructs were intracranially injected and fibre optics were implanted (details for each experiment are described below). After surgery, mice were given additional doses of meloxicam each day for 3 days after surgery, and were monitored carefully for 7 days post-surgery.

2.3.2 GCaMP photometry surgery

Mice were intracranially injected with 1µl of saline containing 1:10 dilution of AAV1.Syn.Flex.GCaMP6f.WPRE.SV40 (titre: 6.22 x 10^{12} vg/ml, Penn Vector Core) and 1:20 dilution of AAV1.CAG.Flex.tdTomato.WPRE.bGH (AllenInstitute864) (titre: 1.535 x 10^{12} vg/ml, Penn Vector Core) at 2nl/s in VTA (antero-posterior (AP): -3.3, medio-lateral (ML): ±0.4, dorso-ventral (DV): -4.3 from bregma) in one hemisphere for mesolimbic dopamine recordings in VTA and NAc and VTA/SNc (AP: -3.1, ML: ±0.9, DV: -4.2 from bregma) in the other hemisphere for DMS recordings. Three 200 µm diameter ceramic optic fibres were implanted chronically in each animal in VTA (AP: -3.3, ML: ±0.4, DV: -4.3 from bregma) and NAc (AP: +1.4, ML: ±0.8, DV: -4.1 from bregma) in the same hemisphere, and DMS (AP: +0.5, ML: ±1.5-1.7, DV: -2.6 from bregma) in the contralateral hemisphere. The hemispheres recorded from each region were counterbalanced across mice.

2.3.3 dLight photometry

Mice were intracranially injected with 500 nl of saline containing 1:5 dilution of pAAV5-CAG-dLight1.1 (titre: 1.4 x 10^{12} vg/ml, addgene) and 1:5 dilution of pssAAV-2/5-hSyn1-chI-tdTomato-WPRE-SV40p(A) (titre: 4.9 x 10^{11} vg/ml, ETH Zurich) at 2nl/s in NAc (AP: +1.4, ML: ±0.8, DV: -4.1 from bregma) and DMS (AP: +0.5, ML: ±1.5/1.7, DV: -2.6 from bregma) in opposite hemispheres. Two 200 µm diameter ceramic optic fibres were implanted chronically targeting the injection sites.

2.3.4 Optogenetic activation

Mice were injected bilaterally with 500 nl/hemisphere of saline containing either AAV2-EF1a-DIO-EYFP (titre: >1 x 10^{12} vg/ml, UNC Vector Core) (YFP group) or rAAV2/Ef1a-DIO-hchR2(E123t/T159C)-EYFP (titre: 5.2 x 10^{12} vg/ml, UNC Vector Core) (ChR2 group) at 2 nl/s in VTA (AP: -3.3, ML: ±0.4, DV: -4.3 from bregma). Two 200 µm diameter ceramic optic fibres were implanted chronically targeting the injection sites at a 10-degree angle.

2.3.5 Optogenetic inhibition

Mice were injected bilaterally with 500 nl/hemisphere of saline containing 1:10 dilution of either ssAAV-1/2-CAG-dlox-tdTomato(rev)-dlox-WPRE-bGHp(A) (titre: 7.9 x 10^{12} vg/ml, ETH Zurich) (tdTomato group) or AAV1-hSyn1-SIO-stGtACR2-FusionRed (titre: 1.9 x 10^{13} vg/ml, addgene) (GtACR2 group) at 2 nl/s in VTA (AP: -3.3, ML: ±0.4, DV: -4.3 from bregma). Two 200 µm diameter ceramic optic fibres were implanted chronically targeting the injection sites at a 10-degree angle.

2.4 Mesolimbic and striatal dopamine recordings and manipulation

Fibre photometry recordings of dopaminergic calcium activity (GCaMP) and release (dLight) were obtained using pyPhotometry (Akam and Walton, 2019) using a 465 and 560 nm LED dual-system. All recordings involved the calcium- or dopamine-dependent fluorophore together with a control one to correct for artifacts. Details on the setup and data processing will be described in chapter 4.

Optogenetic manipulation experiments, both activation and inhibition, involved bilateral VTA dopaminergic neurons stimulation using two 465 nm LEDs (Plexon Plexbright). Manipulations occurred at different timepoints in the two-step task. But also, control experiments were performed, an intracranial self-stimulation essay (ICSS) and a 2-alternative forced choice task (2-AFC) for the activation and inhibition experiments, respectively, to ensure our manipulation parameters were sufficient to elicit the expected behaviour. More details on the stimulation parameters and setup will be described in chapter 5.

2.5 Analysis

All behavioural, photometry and optogenetic manipulation analysis was performed using custom Python and R code and will be described in the following chapters (chapters 3-5).

2.6 Histology

Mice were terminally anesthetised with sodium pentobarbital and transcranial perfused with saline then 4% PFA solution. 50 µm coronal brain slices were cut, covering striatum and VTA, and immunostained with anti-GFP and anti-TH primary antibodies, and 488 and Cy5 secondary antibodies. For the animals used on the optogenetic inhibition experiment, only anti-TH and Cy5 primary and secondary antibodies, respectively, were used. An Olympus FV3000 microscope was used to image the slices and determine the fibres placement (Figure 2.6).

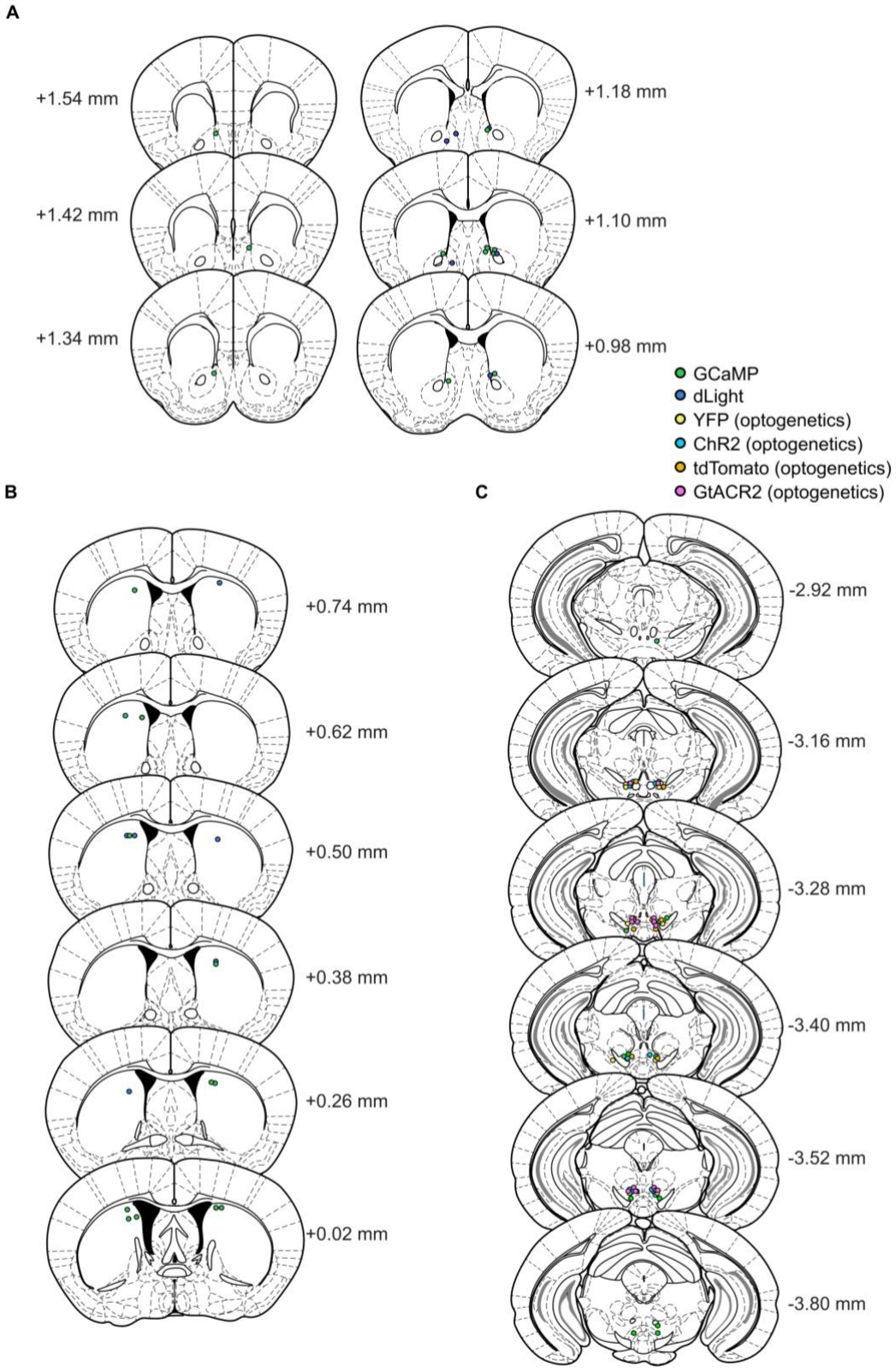


Figure 2.6. *Fibre placement in A) NAc, B) DMS, and C) VTA. Green, GCaMP6f; dark blue, dLight1.1; yellow, YFP animals from the optogenetics activation experiment; light blue, ChR2; orange, tdTomato animals from the inhibition experiment; pink, GtACR2 animals.*

3 Can animals use inference learning to guide behaviour in complex structured environments?

Our day-to-day is filled with continuous decisions on how to act. These decisions are based in our goal, and the state of the environment. In real-world scenarios, the state of the environment (i.e., stimuli and features of a specific condition of the environment) is not always observable (i.e. there is not stimuli informing of the current state). In such situations, one option is to infer the state of the world by tracking a latent or hidden variable (non-directly observable) through observable features in the environment (e.g., stimuli and rewards). In this chapter, we will explore possible behavioural strategies mice could perform to solve a modified version of the two-step task which allows for inference to guide behaviour. We found that an inference-based RL strategy best explained mice behaviour in a modified version of the two-step task which allows for state inference.

3.1 Introduction

The continually changing external world requires humans and animals to continuously adapt actions to changes in the environment. During habitual behaviour, an action or sequences of actions are endowed with a value (high or low) based on the outcome (good or bad), which is the basis of a model-free learner. When a change in the environment contingencies happens (e.g. a path closure during spatial navigation), a model-free learner, through trial-and-error, will update the value of the different actions/sequence of actions (Sutton and Barto, 1998). However, a more flexible strategy, what a model-based agent would do, is to store an internal model of the world. Storing a predictive model of the consequences of each action, allows the model-based learner to simulate the outcome of actions, and quickly update the internal model when the outcome differs from the predicted one (Tolman, 1948; Daw and Dayan, 2014; Doll et al., 2015a).

Outcome devaluation tests have been extensively used in different species to dissociate value catching or model-free behaviour from model-based behaviour (Adams and Dickinson, 1981; Colwill and Rescorla, 1985b; Baxter et al., 2000; Gottfried et al., 2003; Graybiel, 2008; Tricomi et al., 2009). However, this kind of test just reveals planning in a single moment or few trials, limiting the computational and physiological description of this behaviour. Later work on humans showed through a multi-step decision-making task, the two-step task, that model-free could be dissociated from model-based behaviour in a stable manner and over hundreds of trials (Daw et al., 2011), making this task ideal for physiological recordings.

The two-step task, and modified versions of it, have been widely used in humans to study model-free and model-based behaviour, interacting with the dopamine system (Wunderlich et al., 2012b; Deserno et al., 2015) or cortical areas (Smittenaar et al., 2013), during stress (Otto et al., 2013b), or neuropsychiatric disorders (Sebold et al., 2014; Voon et al., 2014; Wyckmans et al., 2019). More recent adaptations of the two-step task in rodents have allowed neural interrogation of the circuitry through neural recordings and manipulations (Miller et al., 2017; Akam et al., 2021b). Although this task was designed to dissociate model-free from model-based behaviour, computational and modelling studies have revealed that this dissociation is not as simple, and many other strategies (different from the standard model-free and model-based approaches) can explain the behavioural effects (Akam et al., 2015; Kool et al., 2016; Feher da Silva and Hare, 2018).

Model-free and model-based RL, without or with the aid of an internal model respectively, through *value learning*, are continuously updating the value of states and actions when predictions differ from actual outcomes. However, the changes occurring in the external environment can follow structured dynamics. For example, during peak hours we may know that a path despite being longer because is less busy, is faster than the shorter one, while the shorter one is faster during non-peak hours. In this situation, instead of relearning the values of the different paths all the time through value learning, we can perform *state inference*. We can *infer* what is the state of the environment – is it peak hour or not – and adapt our actions directly based on that – if peak hour then take the longer path, otherwise, the shorter. Through experience, we can learn *latent* or *hidden states* of the environment (i.e., not directly observable features), which allows prior learned values and behaviour or *choice policies* to be

reused when the same situation is reencountered. Inference during reward-guided tasks has been shown extensively in humans (Hampton et al., 2006, 2007; Behrens et al., 2007; Nassar et al., 2010, 2012; Zylberberg et al., 2018), and in non-human primates (Costa et al., 2015; Jang et al., 2015; Rothenhoefe et al., 2017; Sarafyazd and Jazayeri, 2019; Bartolo and Averbeck, 2020). There is also some evidence suggesting that rodents could also use state inference to guide behaviour (Gallistel et al., 2001; Gershman et al., 2010; Vertechi et al., 2020).

In this chapter, we tested whether mice could learn latent variables in the environment allowing them to perform state inference as opposed to value learning in the two-step task. To do so, we trained mice on a simplified version of the two-step task (Akam et al., 2015) in which the reward probabilities are anti-correlated and reversed periodically. This feature introduces an inferable latent variable, thus allowing inference to guide behaviour. We used different approaches to explore and evaluate the possible behavioural strategies mice could perform to solve the task, in order to maximise rewards. The best model reproducing mice behaviour learned a single latent variable tracking the anticorrelation of the reward probabilities. Therefore, in the simplified version of the two-step task (Akam et al., 2015), mice behaviour was consistent with learning an internal model of the task structure, using inference to guide their choices.

3.2 Methods

In this section, we will describe the behavioural modelling of the two-step task. A total of 18 mice were used (these subjects were used in the next chapter for the dopaminergic

recordings. For more details on the subjects, two-step task and analysis of choice behaviour, see chapter 2: general methods.

Model fitting was performed using two different approaches: i) compare goodness-of-fit to subjects' choices for a set of different RL agents with different strategies, that we called single RL behavioural models, and ii) creating a flexible RL behavioural model which contained the different learning strategies as different components and test the influence of each component on explaining mice's behaviour.

3.2.1 Single RL behavioural models

We evaluated a total of 197 RL models using Bayesian Information Criterion (BIC) score, and k-fold cross-validation. The individual models were from 5 different classes: i) model-free, ii) model-based, iii) hybrid (MF + MB), iv) Inference (correlated inference and/or reward-as-cue) and v) Bayesian Inference. We fitted the different models using maximum likelihood without priors. Except for the correlated inference model, all the other models were implemented as described in previous literature (Watkins, 1989; Ito and Doya, 2009; Daw et al., 2011; Akam et al., 2015).

3.2.1.1 Model-free

The model-free strategy (Watkins, 1989; Daw et al., 2011; Akam et al., 2015) updated value $V(s)$ of second-step state and value $Q_{MF}(c)$ of the chosen action as:

$$V(s) \leftarrow (1 - \alpha)V(s) + \alpha r$$

$$Q_{MF}(c) \leftarrow (1 - \alpha)Q_{MF}(c) + \alpha((1 - \lambda)V(s) + \lambda r)$$

Where α is the learning rate, λ is the eligibility trace parameter and r is the outcome.

Variations of this model were whether it included and asymmetric effect form rewards and omissions and/or forgetting. The outcome asymmetry was implemented through different learning rates for rewards (α_r) and omissions (α_{nr}) (Cazé and Van Der Meer, 2013). To capture forgetting of the non-experienced states (Ito and Doya, 2009), both the non-experienced second-step value, $V(ns)$, and the non-chosen action value, $Q_{MF}(nc)$, decayed towards neutral value as:

$$V(ns) \leftarrow (1 - \text{forget})\, V(ns) + 0.5\, forget$$

$$Q_{MF}(nc) \leftarrow (1 - \text{forget})\, V(ns) + 0.5\, forget$$

Where $forget$ is the forgetting rate, ranging between 0 and 1.

3.2.1.2 Model-based

The model-based strategy (Daw et al., 2011; Akam et al., 2015) updated value $V(s)$ of the second-step reached, and both first-step action values $Q_{MB}(a)$ as:

$$V(s) \leftarrow (1 - \alpha)V(s) + \alpha r$$

$$Q_{MB}(a) = \sum_{s} P(s|a)V(s)$$

Where α is the learning rate, r is the outcome, $P(s|a)$ is the transition probability of reaching second-step s after taking action a (i.e. 0.8 or 0.2).

Like the model-free model, variations of this model included asymmetric learning from rewards and omissions through two different learning rates based on the outcome (α_r and α_{nr}, learning rate for rewarded and unrewarded trials, respectively), and/or forgetting of the non-experienced second-step value, $V(s)$, as:

$$V(ns) \leftarrow (1 - \text{forget})\, V(ns) + 0.5\, forget$$

Where $forget$ is the forgetting rate, ranging between 0 and 1.

3.2.1.3 Correlated inference

The inference strategy used a single variable B to track the anticorrelated reward probabilities at the two second-step states, representing subjects' belief about which second-step state currently had high reward probability. This was updated as:

$$B \leftarrow (1 - \alpha)B + \alpha S \qquad \text{if previous trial rewarded}$$

$$B \leftarrow (1 - \alpha)B + \alpha(1 - S) \qquad \text{if previous trial not rewarded}$$

Where S is a variable which takes value 1 if the up second-step state was reached on the trial and 0 if the down second-step state was reached.

The variable B determined the value of both the second-step states and first-step actions as:

$$V_{Inf}(up) = B$$

$$V_{Inf}(down) = (1 - B)$$

$$Q_{inf}(a) = \sum_{s} P(s|a) V_{inf}(s)$$

Where $V_{Inf}(up)$ and$V_{Inf}(down)$ are the values of the up and down second-step states, α is the learning rate, and $P(s|a)$ is the transition probability of reaching second-step s after taking action a (i.e. 0.8 or 0.2).

A variation of this model, the asymmetric correlated inference, included different learning rates, $\alpha_r$and α_{nr}, for rewarded and not rewarded trials.

3.2.1.4 Reward-as-cue

We considered another strategy in which the most recent trial's outcome acts as a discriminative stimulus for the state of the task, termed 'Reward-as-cue' in (Akam et al., 2015). This strategy has fixed habits (here expressed as first-step action values) deployed depending on the previous trial's outcome and second-step state where it was obtained:

If previous trial rewarded:

$$Q_{RC}(a \rightarrow s) = z$$

$$Q_{RC}(a \rightarrow s') = 0$$

If previous trial not rewarded:

$$Q_{RC}(a \rightarrow s) = 0$$

$$Q_{RC}(a \rightarrow s') = (1 - z)$$

Where $Q_{RC}(a \rightarrow s)$ is the first-step action that commonly leads to the second-step state reached on the previous trial, $Q_{RC}(a \rightarrow s')$ is the first-step action that commonly leads to the second-step state not reached on the previous trial, and z is a variable between 0 and 1 which determines the relative strength with which reward and reward omission affect the next trial choice. If z = 0.5, the model is symmetric and both rewards and omissions update influence next trial choice with the same strength. if $z > 0.5$ and $z < 0.5$, the model is asymmetric and rewards or omissions, respectively, have a stronger influence on next trial choice.

3.2.1.5 Bayesian inference

The Bayesian inference strategy (Akam et al., 2015), similar to the correlated inference strategy, used a single variable *B* to track the belief of the state of the task – whether reward is high on the up or down port. This single variable *B* ranged between 0 and 1, and captured the probability the task was in a particular state (up or down is good). $B < 0.5: down\ is\ good; B > 0.5: up\ is\ good.$ *B* was updated using Bayes rule as:

$$B \leftarrow \frac{p(s,r \mid up) \times\ p_{up}}{p(s,r)}$$

Where $p(s,r \mid up)$ is a matrix of the probability of the second-step s reached and outcome r given state "up is good". $p(s,r \mid up)$ is defined as:

$p(s,r \mid up)$		Outcome experienced	
		Rewarded	Unrewarded
Second-step state experienced	Up	0.8	0.2
	Down	0.2	0.8

B was also updated based on the probability a reversal occurred as:

$$B \leftarrow (1 - pr) \times p_{up} + pr\ \times (1 -\ p_{up})$$

Where pr is the probability a block reversal occurred.

The variable B determined the value of both the second-step states and first-step actions as:

$$V_{B_inf}(up) = p_{up}$$

$$V_{B_inf}(down) = (1 - p_{up})$$

$$Q_{B_inf}(a)\ = \sum_{s} P(s|a)V_{B_inf}(s)$$

Where, as in the correlated inference strategy, $V_{B_inf}(up)$ and$V_{B_inf}(down)$ are the values of the up and down second-step states, $Q_{inf}(a)$ are the value of the actions, and $P(s|a)$ is the transition probability of reaching second-step s after taking action a (i.e. 0.8 or 0.2).

Because this model does not have a learning rate parameter, inspired by Funamizu et al. (2012), the asymmetry between positive and negative outcomes was captured by modifying the observations that the model received from the task, specifically, the observations from unrewarded trials through an asymmetry parameter z. Therefore, $p(s, r \mid up)$ and $p(s, r \mid down)$ were defined as:

$p(s, r \mid up)$		Outcome experienced	
		Rewarded	Unrewarded
Second-step state experienced	Up	0.8	$0.2 + 0.3z$
	Down	0.2	$0.8 - 0.3z$

$p(s, r \mid \text{down})$		Outcome experienced	
		Rewarded	Unrewarded
Second-step state experienced	Up	0.2	$0.8 - 0.3z$
	Down	0.8	$0.2 + 0.3z$

Note that although the difference between this model and the correlated inference is the way the variable *B* gets updated, the resulting second-step and action value updates from both models are in the same direction (Figure 3.1). Therefore, the correlated inference model could be considered an approximation of the Bayesian Inference model (see Hampton et al. (2007) which they used a similar model to our correlated inference to capture the anticorrelation between choice values).

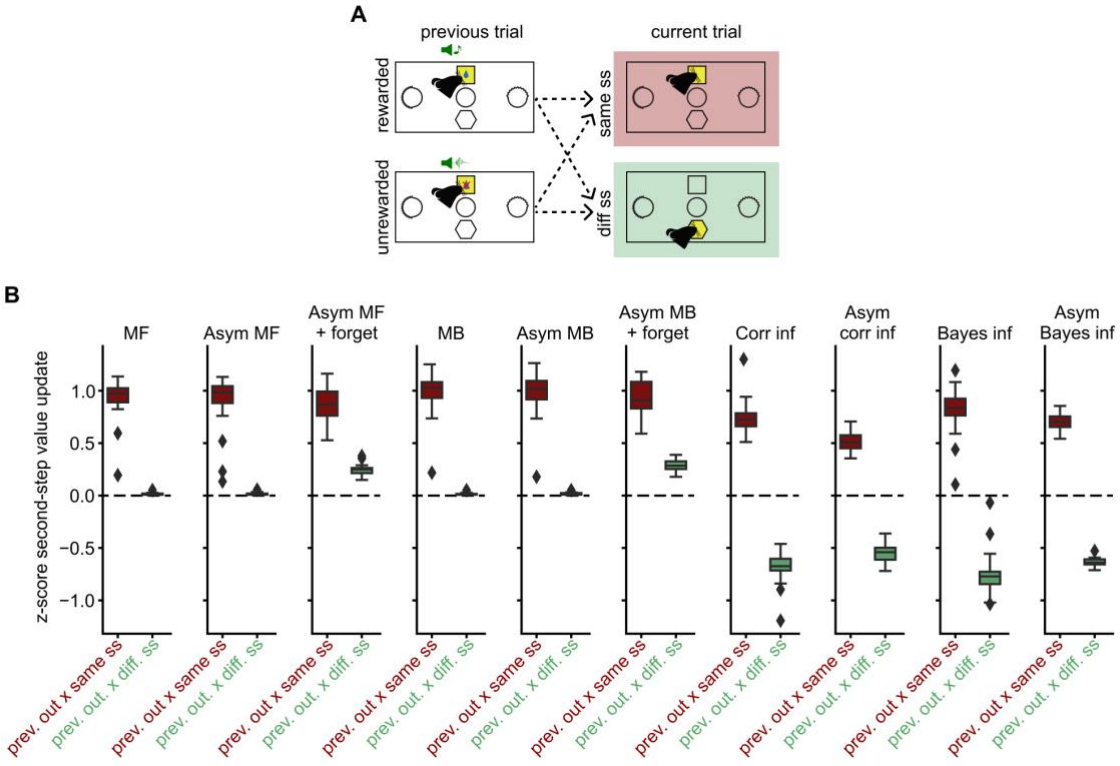


Figure 3.1. Direction of the second-step value update. A) Schematic of the different events influencing the second-step value update: influence of previous trial outcome on the value of same (dark red) or different (green) second-step reached on the current trial. B) Direction of the second-step value update as a function of whether the current trial was the same or different from the previous trial (value update when previous trial was rewarded – value update when previous trial was unrewarded), for each model class, using subject's fits. Value update was z-scored for comparison across models. MF, model-free; MB, model-based; Corr inf, correlated inference; Bayes inf, Bayesian inference; Asym, asymmetric effect of rewards and omissions; + forget, forgetting of the non-experienced second-step.

3.2.1.6 Combined action values

A set of different candidate models were created by combining one of more of the above strategies in a weighted sum with (optionally) bias and perseveration parameters, to give net action values:

$$Q_{net}(a) = \sum_i w_i Q_i(a) + K(a)$$

Where w_i is the weight assigned to strategy i whose first step action values are given by $Q_i(a)$, and $K(a)$ is the modifier to the value of first-step action a due to any bias or perseveration terms included in the model. In models where bias was included, this increased the value of the left action by an amount determined by a bias strength parameter on all trials. In models where multi-trial perseveration was included, this increased the value of the first step action chosen on the previous trial by an amount determined by a perseveration strength parameter over the exponential moving average of choices determined by and alpha multi-trial perseveration parameter. Note that even the models where only a single strategy was included, contained a weight parameter w_i.

The combined action values determined choice probabilities via a softmax decision rule as:

$$P(a) = \frac{e^{Q_{net}(a)}}{\sum_a e^{Q_{net}(a)}}$$

As mentioned in the introduction (section 1.2.2), in the softmax decision rule, the degree to which the value estimates influence choices is controlled through an inverse temperature parameter. In here, the softmax temperature is captured through the weight of each strategy, w_i.

3.2.1.7 Model fitting and comparison

We generated a total of 197 different individual models from the following classes:

Model-free (MF)*:* These models used only the model-free strategy, and varied with respect to whether they had different learning rates for positive and negative outcomes (asymmetric updates), included forgetting of the non-experienced states, and whether they contained perseveration or multi-trial perseveration and/or bias.

Model-based (MB)*:* These models used only the model-based strategy, and varied with respect to whether they had different learning rates for positive and negative outcomes (asymmetric updates), included forgetting of the non-experienced second-step state, and whether they contained perseveration or multi-trial perseveration and/or bias.

Hybrid (MF + MB)*:* These models used both the model-free and model-based strategy, and varied with respect to whether they had different learning rates for positive and negative outcomes (asymmetric updates), included forgetting of the non-experienced states, and whether they contained perseveration or multi-trial perseveration and/or bias.

Inference*:* These models used the correlated inference and/or the Reward-as-cue strategies, and varied with respect to whether they learned from both rewarded and non-rewarded outcomes symmetrically or asymmetrically (the correlated inference model through different learning rates for rewards and omissions), and whether they included perseveration or multi-trial perseveration and/or bias.

Bayesian inference: These models used only the Bayesian inference strategy, and varied with respect to whether they had an outcome asymmetry in updating the *belief state*, and included perseveration or multi-trial perseveration and/or bias.

Each model was fit separately to data from each subject using maximum likelihood without priors. The optimisation was repeated 50 times starting with randomised initial parameter values drawn from a Beta distribution (α=2, β=2) for unit range parameters, Gamma distribution (α=2, β=0.4) for positive range parameters, and Normal distribution (σ=5) for unconstrained parameters.

Model comparison used the Bayesian Information criterion, which was computed per subject, and summed across subjects. To further evaluate the different models, we performed k-fold cross-validation across subjects. The data was split into 10 folds, and the optimisation was repeated 15 times.

To compare data simulated from the model with real data, we simulated the same amount of sessions for each animal (28.9 $\pm$ 1.6 sessions, mean $\pm$ standard deviation) of their average number of trials per session (351.6 $\pm$ 93.0 trials, mean $\pm$ standard deviation), using parameters values from each animal's fits.

3.2.2 Flexible RL behavioural model

We created a flexible RL agent which contained the different behavioural strategies we tested as components. The flexible RL model had 2 ways of updating the second-step state, by learning i) independent or ii) inferred values; and 3 ways of updating action values, through i) model-free, ii) model-based or iii) inference RL. We fitted this model using maximum likelihood without priors.

3.2.2.1 Second-step value update

The *independent second-step value update* module updated second-step values in the traditional RL way. It updated ONLY the second-step state reached (experienced), $V_{trad}(s)$, while the value of the non-experienced second-step state, $V_{trad}(ns)$, decayed towards neutral value as:

$$V_{trad}(s) \leftarrow (1 - \alpha)V_{trad}(s) + \alpha_r r \quad \text{if previous trial rewarded}$$

$$V_{trad}(s) \leftarrow (1 - \alpha)V_{trad}(s) + \alpha_{nr} r \quad \text{if previous trial not rewarded}$$

$$V_{trad}(ns) \leftarrow (1 - \text{forget})\, V_{trad}(ns) + 0.5\, forget$$

Where α_r and α_{nr} are the learning rate for rewarded and unrewarded trials, respectively, r is the outcome, and $forget$ is the forgetting rate.

The *inferred second-step value update* module updated BOTH second-step state values on each trial in a *correlated* manner. It used a single variable B to track the anticorrelated reward probabilities at the two second-step states, representing subjects' 'belief' about which second-step state currently had high reward probability. This was updated as:

$$B \leftarrow (1 - \alpha)B + \alpha_r S \quad \text{if previous trial rewarded}$$

$$B \leftarrow (1 - \alpha)B + \alpha_r(0.5z + (1 - z)(1 - S)) \quad \text{if previous trial not rewarded}$$

Where α is a learning rate, S is a variable which takes value 1 if the up second-step state was reached on the trial and 0 if the down second-step state was reached, and z is the asymmetry parameter, that captures the degree of the asymmetric update between rewarded and unrewarded trials. If z is 0, there is no asymmetry, and both rewards and omissions, increase

or decrease, respectively, the value in the same degree. If z is 1, omission trials update the value of the second-step towards neutral value.

The variable B determined the value of both the second-step states as:

$$V_{Inf}(up) = B$$

$$V_{Inf}(down) = (1 - B)$$

Where $V_{Inf}(up)$ and$V_{Inf}(down)$ are the values of the up and down second-step states.

3.2.2.2 Action value update

In order to update the value of each action (left or right), there *flexible* RL contained 3 different strategies: *model-free, model-based* and *inference*.

Model-free

The model-free strategy updated value $Q_{MF}(c)$ of the chosen action using a weighted sum of the independent, V_{trad}, and inferred, V_{Inf}, second-step values, while also forgetting value $Q_{MF}(nc)$ of the non-chosen action as:

$$V(s) \leftarrow w \times V_{trad}(s) + (1 - w)V_{Inf}(s)$$

$$Q_{MF}(c) \leftarrow (1 - \alpha)Q_{MF}(c) + \alpha((1 - \lambda)V(s) + \lambda r)$$

$$Q_{MF}(nc) \leftarrow (1 - forget)Q_{MF}(nc) + 0.5\, forget$$

Where w is the weight controlling whether independent or inferred second-step values are used, λ is the eligibility trace parameter, r is the outcome and $forget$ is the forgetting rate.

Model-based

The model-based strategy used independent second-step values, V_{trad}, and both first-step action values $Q_{MB}(a)$ were updated as:

$$Q_{MB}(a) = \sum_{s} P(s|a)V_{trad}(s)$$

Where $P(s|a)$ is the transition probability of reaching second-step s after taking action a (i.e. 0.8 or 0.2).

Inference

The inference strategy used inferred second-step values, V_{Inf}, and first-step actions were updated as:

$$Q_{Inf}(a) = \sum_{s} P(s|a)V_{Inf}(s)$$

Where $P(s|a)$ is the transition probability of reaching second-step s after taking action a.

The *flexible* RL model combined then the action values of the 3 strategies (model-free, model-based and inference) in a weighted sum with bias and multi-trial perseveration parameters, to give net action values:

$$Q_{net}(a) = \sum_{i} w_i Q_i(a) + K(a)$$

Where w_i is the weight assigned to strategy i whose first step action values are given by $Q_i(a)$, and $K(a)$ is the modifier to the value of first-step action a due to any bias or multi-

trial perseveration terms included in the model. Bias increased the value of the left action by an amount determined by a bias strength parameter. Multi-trial perseveration increased the value of the first-step action chosen on the previous trial by an amount determined by a perseveration strength parameter over the exponential moving average of choices determined by an alpha multi-trial perseveration parameter.

The combined action values determined choice probabilities via a softmax decision rule as:

$$P(a) = \frac{e^{Q_{net}(a)}}{\sum_a e^{Q_{net}(a)}}$$

3.2.2.3 Model fitting

We followed the same procedure as with the single model fitting. The *flexible* RL model was fit separately to data from each subject using maximum likelihood without priors. The optimisation was repeated 50 times starting with randomised initial parameter values drawn from a Beta distribution (α=2, β=2) for unit range parameters, Gamma distribution (α=2, β=0.4) for positive range parameters, and Normal distribution (σ=5) for unconstrained parameters.

We also simulated choices from the model using the same amount of sessions for each animal (28.9 $\pm$ 1.6 sessions, mean $\pm$ standard deviation) and the average number of trials per session (351.6 $\pm$ 93.0 trials, mean $\pm$ standard deviation), using each animal's fits.

To confirm that the behavioural output from the model-based and the inference components could be distinguished, we removed one of these two components by setting their parameters to 0, fitted this reduced model to subjects' choices, simulated behaviour using the fitted values, and fitted now the full model to the reduced simulated data, to confirm that the removed component did not explain variance of the reduced simulated data. We perform the same procedure but instead of removing on of the components, the model-based or the inference components were the only ones present in the reduced model, by setting the parameters of the rest of the models to 0, to confirm that just the incorporated model explains the variance of the reduced simulated data even if the full model is fitted.

3.3 Results

3.3.1 Mice behaviour in the two-step task is consistent with having an internal model of the task

We trained mice (n=18) on a sequential decision-making task, adapted from the human two-step task developed by Daw and colleagues (2011) (Figure 3.2A,B). Subjects tracked which option was currently best, performing 348.70 $\pm$ 93.90 (mean $\pm$ SD across subjects) trials and completing 8.32 $\pm$ 3.24 blocks per session (Figure 3.2C). Though both model-based and model-free strategies can track the correct option, the task is designed so that the different strategies update their preferred first-step action differently as a function of the transition (common or rare) and outcome (rewarded or not) on each trial. A model-free strategy would reinforce actions based just on the outcome, so would repeat first-step choices that are rewarded independently of the state transition. By contrast, a model-based strategy

understands the causal structure of the task, knows the transition probabilities of each choice, and therefore, it would repeat choices based not only on whether it was rewarded or not but also, depending on whether that outcome came after a common or a rare transition. This strategy is expressed behaviourally by repeating choices that were rewarded after a common transition, but crucially, also repeating unrewarded choices after a rare transition; while switching choice after unrewarded common transitions and also switching despite being rewarded choice if it came from a rare transition (Daw et al., 2011).

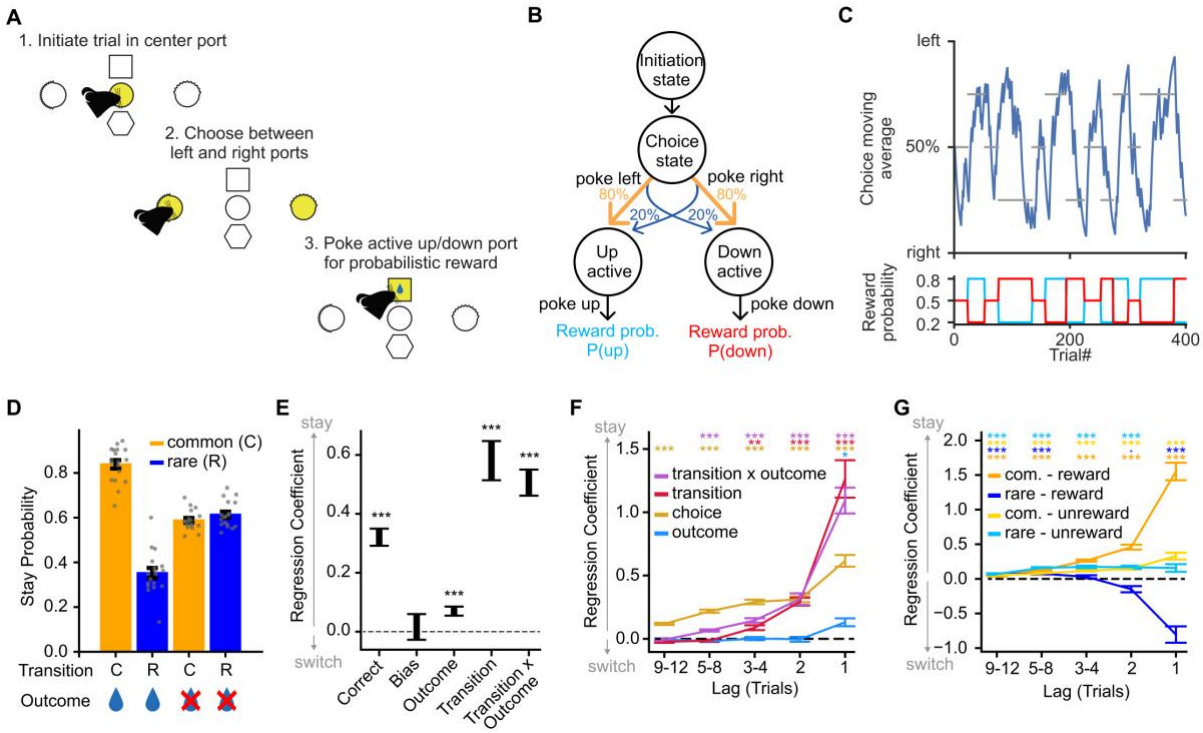


Figure 3.2. *Two-step task behaviour. A) Diagram of trial events. B) Diagram of the task state space, reward and transition probabilities. C) Example behavioural session. Top panel: Moving average of choices (blue trace) with reward probability blocks shown by grey bars, indicating by their vertical position whether the reward probability was higher for the state commonly reached for the left or right choice or a neutral block. Bottom panel: Reward probability at the up (blue) and down (red) ports. D) Probability of a choice being repeated as a function of the subsequent state transition (common or rare) and trial outcome (rewarded or unrewarded). Dots show individual subjects, error bars cross subject*

*SEM. E) Mixed effects logistic regression predicting repeating choice as a function of the trial outcome (rewarded or not), state transition (common or rare), the transition-outcome interaction, and whether the choice was to the correct (high reward probability) option. F) Lagged logistic predicting choice as a function of the history of different trial types defined by choice, transition, outcome, and transition-outcome interaction. G) As F) but as a function of transition and outcome *p<0.05, **p<0.01, ***p<0.001*

We found that rewards did not simply reinforce the choice made, but the trial outcome (rewarded or not) interacted with the state transition (common or rare) to determine the probability of repeating choice. Rewards following common transitions strongly promoted staying while rewards following rare transitions strongly promoted switching ($P < 0.001$, mixed effects logistic regression). This pattern of behaviour is consistent with following a model-based strategy, however, unlike the standard model-based agent, the state transition had no effect on mice stay probability for non-rewarded trials (Figure 3.2D).

It has been shown that two-step tasks with a strong contrast between good and bad options (like in our setup - 80% vs 20% reward probabilities), just looking at the stay probability can give a misleading picture of the influence of trial events on subsequent choice, due to correlations across trials (Akam et al. 2015). We, therefore, used a logistic regression analysis including correct and bias as additional regressors to control for the effect of repeating correct choices or a particular side (left or right), to quantify the influence of the transition, outcome and their interaction on the subsequent choice. This confirmed our results from the stay probability analysis – transition and outcome interact to influence the next choice (Figure 3.2E).

To assess how the extended trial history affected choices, we used a lagged logistic regression analysis which predicted repeating choice as a function of events over the last 10 trials. We further confirmed that the effects of transition by outcome interaction persisted over multiple trials (Figure 3.2F). However, the effect of transition is not symmetric for rewarded and unrewarded trials, which was also manifested in the stay probability analysis. Getting reward after a common transition on previous trials had a high positive influence on the current choice, indicating that previous rewards interacting with experiencing a common transition increased the likelihood of making that same choice again. On the other side, as expected, experiencing a rare transition followed by reward several trials back had a strong negative influence on the current choice, indicating that animals are less likely to repeat those choices even if they got reward. However, no apparent effect of transition was observed in unrewarded trials (Figure 3.2G).

These data show that mice are able to solve this simplified version of the two-step task, performing many trials, and experimenting many block switches. Choice behaviour is consistent with model-based RL where subjects evaluate actions – choose left or right – based on the transition probabilities linking each first-step action to second-step state, and the value of these second-step states, and not only the trial outcome. However, the influence of transition and outcome on subsequent choices is asymmetric, with a major influence of transition on rewarded trials, and minimal on unrewarded trials.

3.3.2 Inference strategies outperform traditional RL models in explaining mice choice behaviour

In the previous section, we have seen that choice behaviour is consistent with mice acquiring an internal model of the task and using this model to guide their future actions. A crucial feature of our adapted version of the two-step task is the anticorrelated nature of the reward probabilities – if one second-step state is rewarded with high probability, the other is rewarded with low probability. This feature of the task allows for inference learning to guide behaviour. Based on the information acquired in each trial (second-step reached and outcome obtained), subjects can infer the current state of the task – whether reward is high in the up or the down port. Therefore, instead of learning independent values for each second-step state (i.e. what standard RL models do for updating action and state values), the agent can learn cached values of each second-step state of the task (Figure 3.3). During state inference, rewards shape subjects' *beliefs* about the current latent state (i.e. non-observable feature) – is the task in the state where the up port is highly rewarded, or the state where the down port is highly rewarded – determining which set of second-step cached values to use to guide choices. The agent can then adapt to reversals by changing their beliefs about the current state, without needing to relearn values.

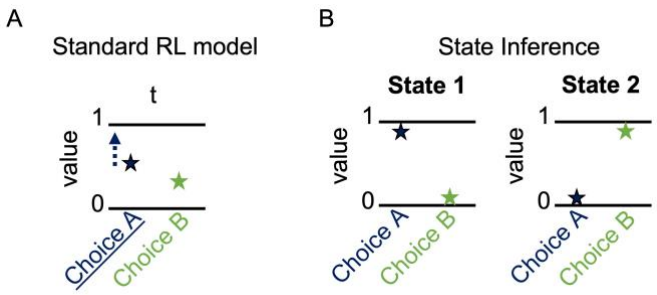


Figure 3.3. Schematic of value update in the standard RL models vs in state inference. A) In the standard RL models, on each trial, the chosen action gets updated based on the outcome received. But no update happens in the non-chosen action. B) During state inference, based on the subject's belief of the state of the environment (State 1 or 2), the value of both actions gets updated accordingly.

To examine what behavioural strategy mice were performing to give rise to this choice behaviour – specifically whether it is consistent with the traditional model-based RL or are instead performing state inference – we created different RL agents implementing different strategies to solve the task and performed model fitting on animal's behaviour (Figure 3.4A, Table 1).

All the RL models tested fell into 6 different categories: model-free RL, model-based RL, hybrid RL, Bayesian inference, correlated inference, and reward-as-cue. The first three categories are standard RL strategies. The model-free RL just uses trial outcomes to update the value of the chosen choice. The model-based RL has knowledge of the task structure and is used to update action values. In this case, on each trial, the value of the second-step reached is updated based on the outcome obtained, but then the first-step action values are updated based on the transition probabilities and the second-step values. Finally, the hybrid RL updates first-step action values through a weighted combination of the model-free and model-based strategies.

We have three other categories performing inference. The Bayesian inference strategy is a popular approach (Gallistel et al., 2001; Behrens et al., 2007; Nassar et al., 2010; Costa et al., 2015; Rothenhoefe et al., 2017; Vertechi et al., 2020) to explain behaviour through inference.

This strategy uses Bayes rule to update the probability the task is in a particular state (i.e. state up is good – highly rewarded -, or state down is good). This probability is then used to update action values taking into account the transition probabilities. However, because of the anticorrelation of the reward probabilities, an approximation of the full Bayesian model and simpler strategy which also involves inferring the value of the non-experienced state is what we called the correlated inference RL model. The correlated inference RL has knowledge of the anticorrelated nature of the reward location. Therefore, on each trial, the value of the second-step state reached is updated based on the outcome obtained, and at the same time, the value of the non-experienced second-step state is inferred in an anticorrelated fashion as $1 - V_{\text{experienced second-step}}$. This second-step value was then used to update action values through either model-free, model-based or a hybrid strategy. Finally, the reward-as-cue (Akam et al., 2015) is another simple strategy in which the outcome of the previous trial is used as a discriminative stimulus to determine the state of the task, and reinforce first-step action that commonly leads to the rewarded second-step state, or to the non-experienced second-step state if the previous trial was unrewarded. The correlated inference and the reward-as-cue strategy were also combined, named mixed inference strategy.

Model comparison supported an inference-based account over standard RL models (Figure 3.4B). The overall best-fitting model was the asymmetric Bayesian inference learning strategy with bias and multi-trial perseveration. This model favoured learning from positive than negative outcomes (the parameter determining asymmetry between rewarded and non-rewarded trials, z, was close to 1 - Table 2). Behaviour simulated from the best model closely recapitulated subjects' data (Figure 3.4C). Also, closely fitting subject's behaviour – the second-best model – was the mixed inference strategy, the mixture of correlated inference

and reward-as-cue models with asymmetric updating from rewards and omissions. Rewards rather than omissions, more strongly updated the latent variable – through the correlated inference RL – and promoted one or the other next-trial choice – through the reward-as-cue strategy (Table 2).

The Bayesian inference model explicitly inferred which discrete state the reward probabilities were in – up is good, or down is good. The mixed inference model (correlated inference + reward-as-cue) used instead an approximation of the Bayesian update. The correlated inference component approximated the state inference through a continuous latent variable which captured the anticorrelation structure of the reward probabilities.

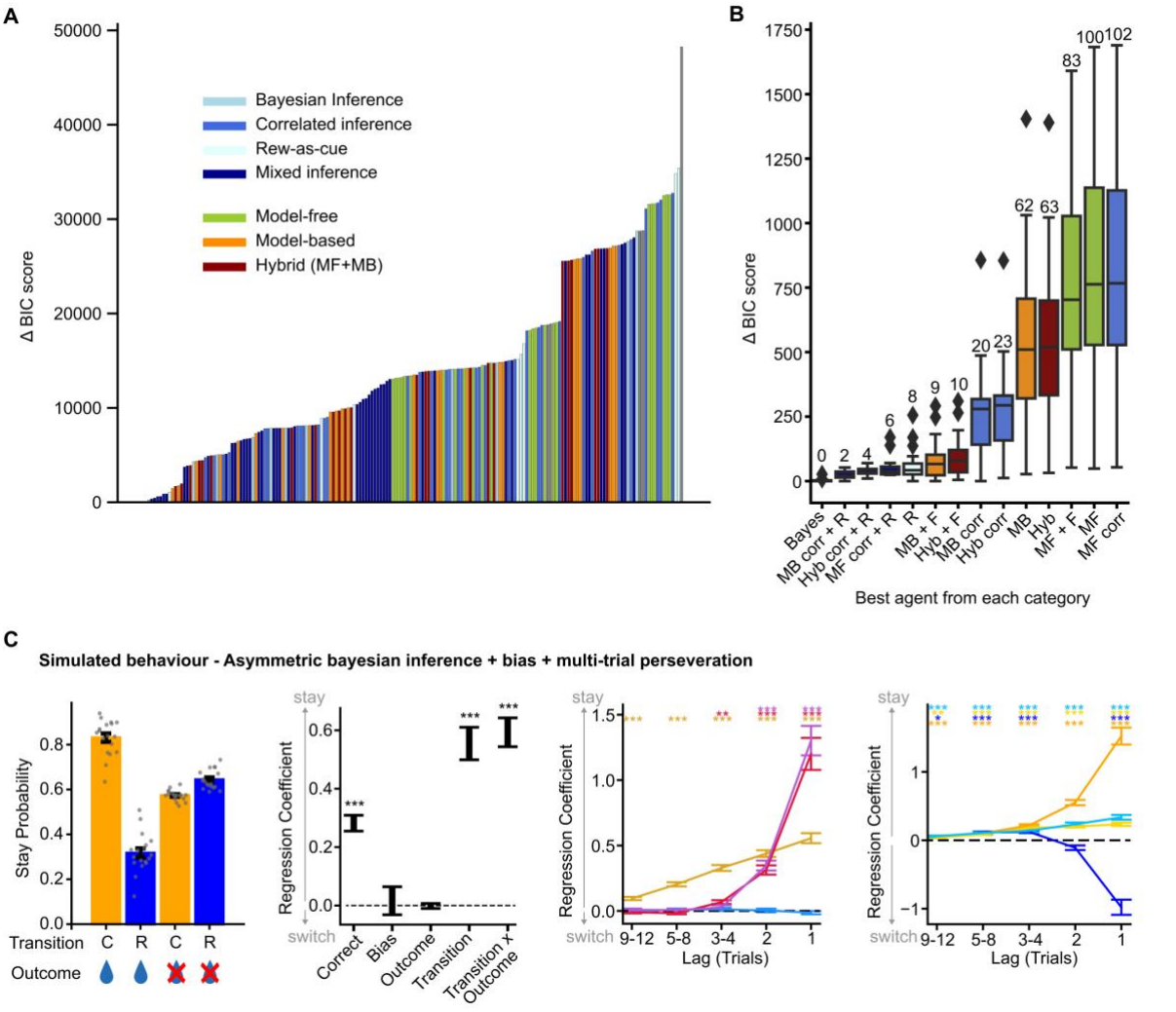


Figure 3.4. *Model comparison and simulated behaviour. A) BIC score model comparison of all the models tested (see Table 1 for details of each model) in comparison to the best fitting one. Colours indicate the model's strategy. B) BIC score model comparison per subject of the best model from each category. Bayes, Bayesian inference model; R, reward-as-cue model; MB corr, model-based correlated inference model; Hyb corr, hybrid (model-free + model-based) correlated inference model; MF corr, model-free correlated inference model; F, added forgetting of unexperienced state; MB, standard model-based models; Hyb, standard hybrid (model-free + model-based) model; MF, standard model-free model. Numbers on top of each boxplot reflects the rank of each model (see Table 1). C) Stay probability and logistic regression analysis of data simulated for the best fitting model, the asymmetric Bayesian inference strategy with choice bias and multi-trial perseveration. *p<0.05, **p<0.01, ***p<0.001.*

Altogether, inference learning strategies fit the data better than standard RL models which learn values independently for the two second-step states, irrespective of whether these were model-free, model-based or a mixture of both, including models with asymmetric learning rates for positive and negative outcomes, or forgetting of the non-experienced second-step state (Figure 3.4A,B, Tables 1 & 2). This suggests that subjects were learning and updating a single latent variable updated in opposite directions by reward or omissions at the up and down ports which in turn determined the relative value of the two first-step actions – i.e., if the model had recently received more reward at the up port, it favoured the first-step action which commonly led to that port.

These data support that mice, instead of learning independent action and second-step values, understand the anticorrelation of the reward probabilities – if one second-step state has high reward probability, the other has low – and use their current belief about the state of the reward probabilities to guide first step choices consistent with the causal structure of the task. Finally, this value updating happening more strongly on rewarded than unrewarded trials explain the behavioural results showing that animal's behaviour after unrewarded trials was not influenced by transition type (Figure 3.2D, G).

3.3.3 Inference and not model-based with forgetting explains behaviour in the two-step task

As seen in the previous section, inference RL strategies provide a better fit – smaller BIC score – on mice choice behaviour. However, although simulating behaviour from the asymmetric Bayesian inference model (also from the mixed inference model) closely reproduced stay

probability behaviour and the effects of the different events' trial history (Figure 3.4D-G), simulating behaviour from the best fitting model from the standard RL category, the standard model-based RL model with added forgetting parameter, also resembled mice's choices. The reason why this might be the case is because if the forgetting rate is very high (i.e. on each trial the non-reached second-step state value is completely forgotten) – as is the case (forgetting parameter fits close to 1 – Table 2) – the resulting contrast between second-step values are very similar to the values following the inference approach, and therefore, producing similar choice behaviour.

Therefore, to compare the influence of the inference and the model-based with forgetting strategy on the actual behaviour and confirm whether an inference-based strategy is more closely related to animals' behaviour than the standard RL strategies, we created a flexible RL model containing both strategies.

Because the first best models included an inference value update strategy to learn the single latent variable of the reward location for the long-run value tracking, but also had a one-trial back effect in promoting first-step action, we included these two features in a single inference strategy. Therefore, the *inference strategy* updated second-step value using a single latent variable tracking the anticorrelation of reward probabilities – like the correlated inference RL – but this single latent variable is updated differently on rewarded and unrewarded trials based on a parameter z, capturing the asymmetric update on rewarded and unrewarded trials – like the reward-as-cue model. Note that the crucial aspect of this strategy is that both second-step state values get updated on each single trial in an anticorrelated fashion. The *model-based RL model with forgetting* instead learned independent values for each second-

step state, and therefore, on each single trial, only the second-step state experienced got updated based on the obtained outcome. However, the non-experienced second-step state value decayed towards neutral value – adding a forgetting component. The asymmetric learning from rewards and omissions was captured through different learning rates for positive and negative outcomes. Both models, the inference and the model-based with forgetting, then updated first-step action values using the second-step values together with the transition probabilities linking first-step action to second-step state.

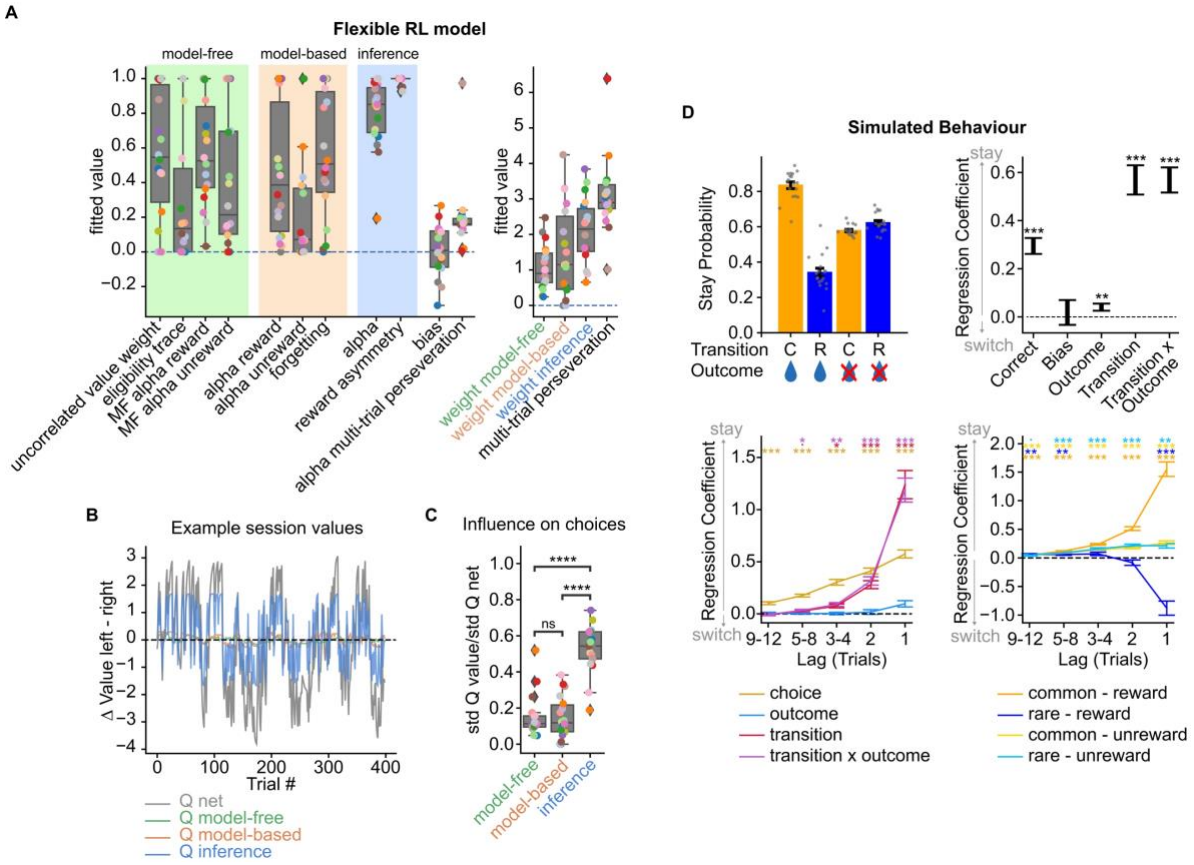


Figure 3.5. Behavioural modelling. A) Flexible RL model fitted values for each component. Orange – model-free component; green – model-based component; blue – inference component. B) Q values derived from the model based on an example session mouse choices for each component. C) Influence of each model component on mice choices, computed as the mean standard deviation of the Q values

*across each session (as in B) across animals. In A and C, each dot represents a single animal. D) Simulated behaviour from the flexible RL model. ·p<0.1, *p<0.05, **p<0.01, ***p<0.001, Bonferroni correction.*

In summary, this flexible model had a model-free component, a model-based component learning independent values for each second-step state and forgetting of the non-visited second-step state, and an inference model-based component learning correlated values of the second-step states. This model also included bias and multi-trial perseveration (Figure 3.5A).

The flexible RL model was able to explain animals' choice behaviour well (Figure 3.5D), and when all models compete, the inference component is capturing most of the variance in animals' choices (Figure 3.5B,C).

To further ratify that we could distinguish the behaviour produced by the inference component from the model-based with forgetting, we first removed each component from the model, fitted these reduced models to animals' choices, simulated the behaviour from the reduced models, and fitted the full model on the simulated behaviour from the reduced models, to confirm that the removed component did not explain the simulated behaviour from the reduced model. Secondly, we did the same, but instead of removing the model-based or the inference component, these were the ONLY components of the model, and again, we asked whether the simulated behaviour was just explained by the component present in the reduced model. Figure 3.6 confirms that we could dissociate the behaviour produced by the model-based vs the inference model, as only the components present in

each reduced model explained the simulated behaviour when the full model – containing all the components – was fitted.

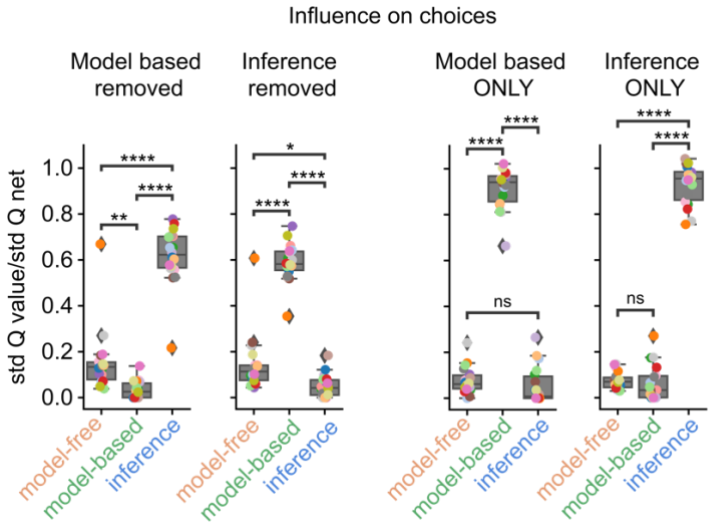


*Figure 3.6. Influence of each model component on simulated behaviour from models where the model-based (first panel), or the inference (second panel) were removed; or models where the model-based (third panel) or the inference (fourth panel) components were the only bits present in the reduced RL model. When a component is removed, that component DOES NOT capture value variance across the session. When the component is the only bit present in the model, is the only component that explains variance in choices. ·p<0.1, *p<0.05, **p<0.01, ***p<0.001, Bonferroni correction.*

These results from the flexible RL model further demonstrated that an inference strategy, sensitive to the transition structure and using the anticorrelated nature of the reward probabilities for value updating, better-explained mice behaviour than the traditional model-based RL learning independent values even if it contains forgetting of the non-experienced state.

3.4 Discussion

In this chapter, we presented an adapted version of the two-step task for mice which allowed for inference to guide behaviour. Mice's behaviour reflected both the causal and correlation structure of the task. Unexpectedly, we found an asymmetric effect of rewards and omissions on future choices. Finally, although previous research using the same reward probability structure as in our setup described animals' behaviour through model-based RL, here we presented evidence that an inference strategy outperformed the traditional model-based RL strategy, even if forgetting is included.

3.4.1 Rodent two-step task: model-based or inference

As discussed earlier, the two-step task from Daw and colleagues (2011) offered a computational framework to study goal-directed behaviour, importantly, over many several trials allowing to study the neural correlates of behaviour. This lead to widespread interest in adapting the two-step task for animal research.

The first published study of the two-step task in animals (Miller et al., 2017) used the simplified version of the two-step task from Akam et al. (2015). In this study (Miller et al., 2017), rats had to choose between two actions which with different probability transitioned to two different second-step states. Unlike in Daw et al's original version, in this simplified version, there was no choice at second-step, and rats simply had to approach the active second-step port for reward to be delivered. Reward was probabilistic at both second-step ports, with probabilities changing in blocks and in an anticorrelated fashion. This task

structure, although not the physical layout, is the same as in our two-step task for mice. Miller et al. (2017) concluded that rats followed a model-based RL strategy to solve the task.

Another version of the two-step task in rats, but with choice at second-step and reversals with deterministic rewards, also found that behaviour could be explained by a model-based strategy, although they discussed that some level of model-free behaviour could also be involved (Dezfouli and Balleine, 2019). Other studies with other versions of the two-step task for rats, found that like humans, rats used a hybrid model-free – model-based approach determining future choices (Hasz and Redish, 2018; Groman et al., 2019). Akam et al. (2021) using the reduced version (but with reversals also in the transition probabilities) but in mice also found that mice behaviour was better explained through a hybrid approach.

However, apart from the model-free, model-based and hybrid RL distinction exposed in these studies (and in the original two-step study), animals could also be solving the task using state inference, which can lead to behaviour which resembles the classical model-based RL (Akam et al., 2015). The reduced version of the two-step task in which the reward probabilities change in blocks in an anticorrelated fashion further justifies testing the inference strategy as one could describe the task as being in two possible different states: state 1 – up port is good; state 2 – down port is good.

Inference can be implemented in different ways. One standard way is through Bayesian inference. This has been used to explain human (Hampton et al., 2006; Behrens et al., 2007; Nassar et al., 2010, 2012; Vertechi et al., 2020), non-human primates (Costa et al., 2015; Jang

et al., 2015; Rothenhoefe et al., 2017; Bartolo and Averbeck, 2020), and rodent (Gallistel et al., 2001; Vertechi et al., 2020) behaviour. This strategy uses Bayes rule to update the probability the environment is in a certain state, guiding the subject's belief the world is in a particular state and behaving according to that belief.

Others have used an approximation of the full Bayesian inference approach where Bayesian update is not needed, named *counterfactual* updating (Montague et al., 2006; Hampton et al., 2007; Lohrenz et al., 2007; Tobia et al., 2014; Van Hoeck et al., 2015; Zylberberg et al., 2018; Santo-Angles et al., 2021). Here, value updating occurs like in the classical TD RL, but this update happens in both experienced and non-experienced states. The experienced state gets updated through a RPE signal, and the non-experienced state through a *fictive prediction error* signal. This fictive signal captures what would have been the outcome if the opposite state had been chosen/reached. However, this kind of approximation is normally used in setups where the counterfactual outcome is explicitly given (e.g. to study regret signals) or is inferred through the task structure (e.g. explicitly given as instructions to humans).

In our setup, fictitious signals are out of scope as the animal cannot know what would have happened if the opposite state was reached. Instead, we used what we named *correlated inference*. The implementation is similar to counterfactual updating in the sense that both experienced and non-experienced states are updated. However, no fictitious signal is needed, making this strategy even simpler. What it is learned is a single latent variable tracking the reward probabilities and updating value following the correlated structure. Therefore, the non-experienced state value = 1 – experienced second-state value.

In our study, we found that both Bayesian and correlated inference fitted better mice's behaviour than the classical model-based RL even when forgetting of the non-experienced state is included. Although we observed marginal involvement of model-free RL, this is the result of including multi-trial perseveration into our model, and therefore, a hybrid account involving correlated value update together with a degree of model-free RL is in line with our conclusions.

3.4.2 Asymmetric value updating

Rewards and omissions were both observable events in the two-step task, signalled by discrete auditory cues, which could be used to infer the latent or unobservable variable, the reward probabilities, of the state of the task. However, our results support the idea that mice were only primarily using rewards rather than omissions to update state values, and consequently, to update their belief about the latent state of the task. First, we observed through mice stay probabilities and the effect of trial history on the probability of repeating choice or switching based on previous trial events, that mice preferentially used rewards rather than reward omissions to guide their future behaviour. Second, the best model, the Bayesian inference model, needed the asymmetric component to outperform the simpler models. Third, the best RL model from each category contained an asymmetric updating and/or an asymmetric effect on the next trial choice.

This asymmetry observed was unexpected given that a previous adaptation of the two-step task in rats did not report this asymmetry (Miller et al., 2017). This discrepancy could be the result of differences in training protocols, task structure or species (Jones et al., 2017). This

different learning from rewards and omissions have also been reported in other tasks in mice (Cieślak et al., 2018). However, this is unlikely to be an effect unique to mice. Similar asymmetries have also been observed in humans during two-step tasks (Sebold et al., 2014; Wyckmans et al., 2019). But also in patients (Frank et al., 2004) and healthy volunteers (Behrens et al., 2007; Niv et al., 2012; Lefebvre et al., 2017) in other probabilistic reward-guided tasks. It has been suggested some of these may reflect a differential influence of positive and negative prediction errors over striatal direct and indirect pathways (Frank et al., 2007, 2009; Krugel et al., 2009).

Theoretically, asymmetric learning rates for positive and negative RPE can improve performance in two-armed bandit tasks (Cazé and Van Der Meer, 2013). In mice, asymmetric learning rates were shown to control exploratory behaviour based on the background reward probability in a 5-armed bandit task. In poor environments, with low reward probability, the learning rate for positive outcomes increased while decreased for negative outcomes (Ohta et al., 2021). However, the direction of the asymmetry and the effect of background reward rate may be task-dependent (Gershman, 2015). In our task, further examination on neutral blocks and changes in reward rate across a session may untangle whether the asymmetry we observe can be the result of differential exploratory rates, or whether learning rates also change within blocks.

3.4.3 Conclusion

In summary, we have observed that mice behaviour reflected the causal and correlation structure of our reduced two-step task. Second-step values were updated using inference.

This was consistent with Bayesian state inference or an approximation of it, correlated inference. The correlated inference approach updated the second-step value using a single latent variable controlling the state of the reward probabilities, and therefore, on each trial, both second-step states, experienced and non-experienced got updated in an anticorrelated manner. Second-step values were then used to choose the appropriate action using the task structure, the different transition probabilities from first-step action to second-step state. Finally, we also observed a striking asymmetric effect of outcome, while rewards had a big impact on subsequent choices, omissions did not or in a much lesser degree.

Tables

				Learning rate		Reward-as-cue		Kernels		
Rank	**Model Category**	**BIC**	**Forget**	**Sym.**	**Asym.**	**Sym.**	**Asym.**	**Bias**	**Persev.**	**Multi-Persev.**
0	Bayesian inference	140101.34			■			■		■
1	Bayesian inference	140362.60			■					■
2	Mixed inference (MBC)	140469.42			■		■	■		■
3	Mixed inference (MBC)	140548.93		■			■	■		■
4	Mixed inference (HC)	140711.08			■		■	■		■
5	Mixed inference (HC)	140723.18		■			■	■		■
6	Mixed inference (MFC)	140992.39			■		■	■		■
7	Mixed inference (MFC)	140999.87		■			■	■		■
8	Reward-as-cue	141182.55		■			■	■		■
9	Model-based	141590.81	■		■			■		■
10	Hybrid (MF+MB)	141805.52	■		■			■		■
11	Model-based	141867.80	■		■					■
12	Hybrid (MF+MB)	142073.50	■		■					■

13	Mixed inference (MBC)	143848.53									
14	Mixed inference (HC)	143970.43									
15	Hybrid (MF+MB)	144008.16									
16	Bayesian inference	144394.69									
17	Model-based	144459.12									
18	Hybrid (MF+MB)	144526.49									
19	Hybrid (MF+MB)	144544.82									
20	MB Correlated inference	144849.01									
21	Mixed inference (MBC)	145003.96									
22	Hybrid (MF+MB)	145051.16									
23	Hybrid Correlated inference	145068.17									
24	Bayesian inference	145137.17									
25	Model-based	145149.48									
26	MB Correlated inference	145159.60									
27	Mixed inference (HC)	145224.88									
28	Hybrid Correlated inference	145389.39									
29	Mixed inference (MFC)	146368.95									
30	Mixed inference (HC)	146399.30									
31	Model-based	146579.22									
32	Mixed inference (MBC)	146648.68									
33	Mixed inference (MFC)	146790.23									
34	Mixed inference (MBC)	146821.27									
35	Mixed inference (HC)	146863.35									
36	Bayesian inference	146978.55									
37	Model-based	147405.69									
38	Mixed inference (MBC)	147551.61									

39	Mixed inference (HC)	147680.54									
40	Hybrid Correlated inference	147918.24									
41	MB Correlated inference	147930.87									
42	Bayesian inference	147948.45									
43	Mixed inference (HC)	147953.74									
44	Mixed inference (MBC)	147957.11									
45	Mixed inference (MBC)	147959.20									
46	Model-based	147965.78									
47	Bayesian inference	147966.86									
48	Mixed inference (HC)	147974.87									
49	Hybrid (MF+MB)	148007.05									
50	Mixed inference (MFC)	148148.69									
51	MB Correlated inference	148209.46									
52	MB Correlated inference	148210.40									
53	Hybrid Correlated inference	148217.08									
54	Bayesian inference	148229.72									
55	Model-based	148241.58									
56	Mixed inference (MFC)	148255.58									
57	Hybrid (MF+MB)	148284.72									
58	Hybrid Correlated inference	148314.01									
59	Reward-as-cue	148995.43									
60	MB Correlated inference	149016.94									
61	Hybrid Correlated inference	149129.12									
62	Model-based	149671.55									
63	Hybrid (MF+MB)	149682.46									
64	Model-based	149776.48									
65	Hybrid (MF+MB)	149815.09									

66	Model-based	150022.51									
67	Hybrid (MF+MB)	150023.82									
68	Model-based	150123.43									
69	Hybrid (MF+MB)	150145.86									
70	Reward-as-cue	150449.50									
71	Mixed inference (MFC)	150482.13									
72	Mixed inference (HC)	150714.18									
73	Mixed inference (MFC)	151029.58									
74	Mixed inference (MBC)	151101.79									
75	Mixed inference (MFC)	151490.51									
76	Mixed inference (HC)	151909.48									
77	Mixed inference (MFC)	152106.51									
78	Mixed inference (MBC)	152221.67									
79	Mixed inference (MFC)	152580.80									
80	Mixed inference (HC)	152611.44									
81	Mixed inference (MBC)	152948.27									
82	Mixed inference (MFC)	153139.76									
83	Model-free	153158.95									
84	Model-free	153253.21									
85	Model-free	153271.91									
86	Model-free	153345.20									
87	Model-free	153457.04									
88	Hybrid Correlated inference	153483.12									
89	Model-free	153517.90									
90	Model-based	153589.59									
91	Hybrid (MF+MB)	153598.39									
92	MB Correlated inference	153888.76									

93	Mixed inference (MBC)	153927.68									
94	Hybrid (MF+MB)	153970.40									
95	Hybrid (MF+MB)	154009.16									
96	MB Correlated inference	154010.41									
97	Mixed inference (HC)	154039.84									
98	Model-based	154043.05									
99	Model-based	154102.13									
100	Model-free	154115.34									
101	Hybrid Correlated inference	154131.88									
102	MF Correlated inference	154196.29									
103	Bayesian inference	154204.80									
104	Model-free	154205.24									
105	MF Correlated inference	154257.60									
106	Hybrid Correlated inference	154263.81									
107	Model-free	154273.02									
108	Model-based	154324.33									
109	Hybrid (MF+MB)	154340.56									
110	Model-free	154357.99									
111	MF Correlated inference	154359.84									
112	MF Correlated inference	154418.43									
113	Model-free	154621.49									
114	MB Correlated inference	154621.87									
115	Hybrid (MF+MB)	154854.01									
116	Model-free	154860.06									
117	Hybrid (MF+MB)	154861.15									
118	Bayesian inference	154879.49									
119	Model-based	154940.30									
120	Model-based	154959.66									
121	Mixed inference (MBC)	155051.17									
122	MB Correlated inference	155124.80									
123	Mixed inference (HC)	155154.05									

124	Hybrid Correlated inference	155250.35								
125	Reward-as-cue	155252.78								
126	Reward-as-cue	155799.42								
127	Reward-as-cue	156907.79								
128	MF Correlated inference	158274.65								
129	Model-free	158325.34								
130	Model-free	158491.35								
131	Model-free	158561.97								
132	MF Correlated inference	158635.13								
133	MF Correlated inference	158854.28								
134	Model-free	158884.89								
135	Control	158903.27								
136	Control	159003.42								
137	Model-free	159080.31								
138	Model-free	159151.08								
139	MF Correlated inference	159265.63								
140	Hybrid (MF+MB)	165659.27								
141	Mixed inference (HC)	165665.47								
142	Hybrid (MF+MB)	165694.02								
143	Hybrid (MF+MB)	165766.56								
144	Model-based	165817.93								
145	Model-based	165920.96								
146	Model-based	165923.02								
147	Hybrid Correlated inference	166064.16								
148	Mixed inference (MFC)	166335.01								
149	Mixed inference (MBC)	166336.95								
150	MB Correlated inference	166734.38								
151	Hybrid (MF+MB)	166940.09								
152	Hybrid (MF+MB)	166969.09								
153	Mixed inference (MFC)	166972.69								
154	Mixed inference (HC)	166994.82								

155	Hybrid (MF+MB)	167002.03		■						
156	Model-based	167062.90	■	■						
157	Model-based	167257.04			■					
158	Model-based	167264.02		■						
159	Hybrid Correlated inference	167346.48		■						
160	Mixed inference (MFC)	167436.14			■	■				
161	Mixed inference (MBC)	167594.46		■		■				
162	Bayesian inference	167728.20		■				■		
163	MB Correlated inference	167947.95		■						
164	Mixed inference (MFC)	168147.86		■		■				
165	Bayesian inference	168835.61		■						
166	Control	168847.38		■				■	■	
167	Control	168878.81		■					■	
168	MF Correlated inference	171200.53			■			■		
169	Model-free	171680.06	■	■				■		
170	Model-free	171713.00		■				■		
171	Model-free	171738.62			■			■		
172	MF Correlated inference	171834.53		■				■		
173	MF Correlated inference	172139.22			■					
174	Model-free	172598.88	■	■						
175	Model-free	172669.69		■						
176	Model-free	172694.82			■					
177	MF Correlated inference	172856.97		■						
178	Reward-as-cue	174946.62		■		■		■		
179	Reward-as-cue	175508.35		■		■				
180	Control	188360.39		■				■		

Table 1*. All ranked models, BIC score. Black squares indicate the parameters each model included.*

Best fitting model from each category	BIC	weight	Rew.-as-cue asym	reversal prob.	Asym. z	Sym. α	Asym. α (+/-)	λ	forget	Kernels
Asymmetric Bayesian inference	140101.34	14.21 ± 12.85		0.34 ± 0.05	0.94 ± 0.10					bias: 0.01 ± 0.14 mult.persv.: 3.11 ± 0.45 alpha m.p: 0.20 ± 0.10
MB asymmetric correlated inference + reward-as-cue	140469.42	0.80 ± 0.35	weight: 2.06 ± 0.94 z: 0.98 ± 0.07				0.44 ± 0.14 / 0.13 ± 0.09			bias: 0.01 ± 0.15 mult.persv.: 2.99 ± 0.52 alpha m.p: 0.24 ± 0.13
Hybrid asymmetric correlated inference + reward-as-cue	140711.08	MF: 0.25 ± 0.24 MB: 0.60 ± 0.42	weight: 10 ± 0.98 z: 1.00 ± 0.00				0.53 ± 0.19 / 0.24 ± 0.17	0.54 ± 0.43		bias: 0.01 ± 0.15 mult.persv.: 3.04 ± 0.47 alpha m.p: 0.23 ± 0.12
MF asymmetric correlated inference + reward-as-cue	140992.39	3.91 ± 13.67	weight: 2.52 ± 0.92 z: 0.95 ± 0.08				0.52 ± 0.20 / 0.42 ± 0.20	0.27 ± 0.26		bias: 0.01 ± 0.14 mult.persv.: 3.17 ± 0.44 alpha m.p: 0.23 ± 0.10
Asymmetric reward-as-cue	141182.55		weight: 2.75 ± 0.93 z: 0.92 ± 0.09							bias: 0.01 ± 0.11 mult.persv.: 3.48 ± 0.48 alpha m.p: 0.21 ± 0.07
MB asymmetric + forgetting	141590.81	4.98 ± 1.45					0.71 ± 0.14 / 0.16 ± 0.08		0.83 ± 0.13	bias: 0.01 ± 0.14 mult.persv.: 2.98 ± 0.61 alpha m.p: 0.23 ± 0.19
Hybrid asymmetric + forgetting	141805.52	MF: 0.37 ± 0.41 MB: 4.81 ± 1.41					0.68 ± 0.14 / 0.14 ± 0.06	0.45 ± 0.44	0.82 ± 0.16	bias: 0.01 ± 0.14 mult.persv.: 2.92 ± 0.65 alpha m.p: 0.20 ± 0.20
MB asymmetric correlated inference	144849.01	1.55 ± 0.48					0.90 ± 0.10 / 0.32 ± 0.08			bias: 0.01 ± 0.15 mult.persv.: 3.13 ± 0.64 alpha m.p: 0.32 ± 0.21

Hybrid asymmetric correlated inference	145068.17	MF: 0.13 ± 0.11 MB: 1.47 ± 0.46					0.93 ± 0.10 / 0.34 ± 0.09	0.23 ± 0.31		bias: 0.01 ± 0.15 mult.persv.: 3.10 ± 0.65 alpha m.p: 0.33 ± 0.22
Model-based	149671.55	1.55 ± 0.49				0.69 ± 0.11				bias: 0.00 ± 0.16 mult.persv.: 3.13 ± 0.51 alpha m.p: 0.37 ± 0.19
Hybrid	149682.46	MF: 0.25 ± 0.13 MB: 1.32 ± 0.44				0.78 ± 0.11		0.17 ± 0.22		bias: 0.01 ± 0.15 mult.persv.: 3.13 ± 0.51 alpha m.p: 0.37 ± 0.19
MF asymmetric + forgetting	153158.95	3.06 ± 0.94					0.56 ± 0.19 / 0.26 ± 0.20	0.49 ± 0.15	0.40 ± 0.14	bias: 0.01 ± 0.11 mult.persv.: 2.37 ± 1.06 alpha m.p: 0.49 ± 0.37
MF	154115.34	1.26 ± 0.40				0.46 ± 0.08		0.53 ± 0.08		bias: 0.01 ± 0.12 mult.persv.: 3.06 ± 0.46 alpha m.p: 0.39 ± 0.18
MF correlated inference	154196.29	1.25 ± 0.40				0.42 ± 0.08		0.55 ± 0.09		bias: 0.01 ± 0.12 mult.persv.: 3.09 ± 0.47 alpha m.p: 0.38 ± 0.18

***Table 2**. Best model from each category.*

4 How do complex structured environments and animals' decision policies influence the dopamine system?

It is widely accepted that dopamine activity reflects RPEs which are used as a teaching signal to update values and policy. However, some studies are challenging this view. Furthermore, it is not known how dopamine is influenced by latent state inference when this guides behaviour, and how it interacts with causal relationships in the environment. Using fibre photometry, we recorded calcium activity in VTA, NAc and DMS, together with dopamine concentration in the same striatal regions. In this chapter we analysed the influence of inference-guided behaviour in a structured task on dopamine somatic and axonal calcium activity, and dopamine concentrations. Dopamine reflected both experienced and inferred state value. Not only that, but dopamine multiplexed a wide range of signals such as RPE, motivation and movement, through dynamics at different timescales. Despite some few differences, axonal activity and dopamine concentrations within each striatal region were closely comparable.

4.1 Introduction

The traditional view of dopamine activity postulates that dopamine signalling encodes reward prediction errors (RPEs). This signal is used as a teaching signal to update the value estimates when predicted and actual outcomes differ (Schultz et al., 1997; Eshel et al., 2016), informing and guiding behaviour (Steinberg et al., 2013). Gradual increases of activity are also observed in dopamine when animals are approaching natural reinforcers such as food (Roitman et al., 2004; Wassum et al., 2012; Syed et al., 2016), or drugs of abuse like cocaine (Phillips et al., 2003). While some studies envision these ramps of activity as animals get closer to the goal as representing state value and reward expectation (Howe et al., 2013; Hamid et al., 2016; Mohebi et al., 2019) and could be used to build a cognitive map of the external world (Collins et al., 2016; Guru et al., 2020), others argue these ramps can arise just as a consequence of a temporal-difference RPE at each time step (Gershman, 2014; Morita & Kato, 2014; Kim *et al.*, 2020). Furthermore, as evidenced by the motor dysfunctions present in Parkinson's disease (Obeso et al., 2000; Lees et al., 2009), dopamine also plays a major role in motor control (Dodson et al., 2016; Howe and Dombeck, 2016; Parker et al., 2016; da Silva et al., 2018; Lee et al., 2019; Hughes et al., 2020).

In addition to this debate on what information does dopamine convey – RPE vs value/reward expectation vs motor information – a second important issue concerns what source of information dopamine neurons may be using to calculate RPEs. Although dopamine signalling has been suggested to convey model-free RPE signal (Montague *et al.*, 1996; Schultz *et al.*, 1997; Steinberg *et al.*, 2013; Eshel *et al.*, 2016), using fMRI, Daw et al. (2011) observed that activation in the ventral striatum, a rich-dopamine region, could not be explained by a pure

model-free value computation, but instead reflected both model-free and model-based predictions weighted by their influence on humans' choice behaviour in the two-step task. As it is not possible to directly relate BOLD signal changes to dopamine, it remains unclear the extent to which dopamine activity is itself directly influenced by such model-based predictions. However, increasing evidence is pointing toward the presence of a richer set of environmental information in the dopamine activity (Sadacca et al., 2016; Takahashi et al., 2017; Yves et al., 2018; Engelhard et al., 2019), allowing for stimulus-stimulus associations to be learned without endowing value learning (Sharpe et al., 2017, 2020).

As seen in chapter 3.1, mice behaviour in our adapted two-step task can be explained through the presence of an internal model of the task which allows the animals to learn a single latent variable tracking the location of the highly rewarded second-step state. This latent variable permits animals to use inference to guide their future choices. Several studies have reported that dopamine activity can reflect inferences about latent states, through RPE signals reflecting the subject's belief of the current state of the task (Nakahara et al., 2004; Starkweather et al., 2017; Babayan et al., 2018; Gershman and Uchida, 2019) or decision confidence (Lak et al., 2017, 2020), and inferring the value of non-experienced state (Bromberg-Martin et al., 2010a). But little is known about dopamine dynamics in more complex tasks where latent states interact with causal relationships (e.g. when the best path depends on the time of the day).

Furthermore, while classical models have assumed that uniform dopamine signals are broadcast to terminal regions (Schultz, 2002), there is increasing evidence that the dopamine system is heterogeneous with respect to molecular properties (Poulin et al., 2018), projection

targets (Fallon and Moore, 1978; Lynd-Balta and Haber, 1994; François et al., 1999; Lerner et al., 2015), and functional modulation (Matsumoto and Hikosaka, 2009; Brown et al., 2011; Howe and Dombeck, 2016; Parker et al., 2016; Tsutsui-Kimura et al., 2020) throughout the mesostriatal circuitry. Therefore, we also do not know much about whether all these different signals described earlier are differently represented across the projection targets. A final complication is that patterns of dopamine release can be locally regulated at the terminals by local circuitry such as cholinergic interneurons (Threlfell et al., 2012; Reynolds et al., 2022), meaning that release has the potential to be decoupled from spiking activity (Mohebi et al., 2019).

In this chapter, we addressed several of these questions: i) how does dopamine reflect latent states interacting with decision policies, ii) can all these signals coexist together, iii) do these signals differ at the somatic and axonal level across striatal targets, and iv) are there differences between axonal dopamine activity and dopamine concentrations in different striatal regions.

As mentioned in the previous chapters, the two-step task is an ideal task to tackle all these questions as it provides a framework to study reward-guided behaviour in a stable manner across hundreds of trials and many sessions, making it ideal for physiological recordings. Therefore, we recorded dopamine population calcium activity in VTA, and dopamine axonal activity and receptor binding in NAc and DMS, using fibre photometry while mice were performing the two-step task. We found that dopamine activity and release in these regions tracked the latent state variable that controlled both reward probabilities. At points in a trial when the animal received new information, dopamine correlated with an RPE signal that

reflected knowledge of the task structure. This coexisted with a separate sustained influence of recent reward rate throughout the trial, and signals also correlated with lateralised movements. These patterns were seen across target regions, though with some important differences with respect to specific task variables and release dynamics across regions and heterogeneity in encoding across the mediolateral axis in DMS.

4.2 Methods

Using fibre photometry, we recorded dopamine calcium activity (GCaMP6f) in VTA and its axonal projections in NAc. We also recorded from midbrain (between VTA and SNc) axonal projections in DMS. Recordings of the three regions were performed on the same animals, although recordings were not performed simultaneously. For these recordings 12 dopamine transporter (DAT)-Cre heterozygous mice (7 females and 5 males) were used.

In another cohort of animals, 6 WT C57BL/6 mice (3 females and 3 males), we recorded dopamine concentrations in the same striatal regions (NAc and DMS) using a dopamine sensor (dLight1.1). Again, recordings on each region occurred on alternating days.

All these recordings were performed when animals were solving the two-step task. For details on subjects, two-step task design and surgery protocol, see chapter 2: general methods.

4.2.1 Fibre photometry

4.1.1.1 Data acquisition

Dopamine calcium activity (GCaMP) and release (dLight) were recorded at 130Hz sampling rate using pyPhotometry (Akam and Walton, 2019). The optical system comprised a 465 nm and a 560 nm LED, a 5 port minicube and fibre optic rotary joint (Doric Lenses), and two Newport 2151 photoreceivers. Time division illumination with background subtraction was used to prevent crosstalk between fluorophores due to the overlap of their emission spectra, and changes in ambient light from affecting the signal. Synchronisation pulses from pyControl onto a digital input of the pyPhotometry board were used to synchronise the photometry signal with behaviour (Akam and Walton, 2019). A total of 458 sessions were recorded across animals and regions, which resulted in a total of 161,472 trials.

4.1.1.2 Data processing

Photometry signals were pre-processed using custom Python code. A median filter (width 5 samples) was first used to remove any spikes due to electrical noise picked up by the photodetectors. Afterwards, a 5Hz low-pass filter was used to denoise the signal. To motion-correct the signals, we band-passed the denoised signals between 0.001 Hz and 5 Hz, and used a linear regression to predict the GCaMP or dLight signal using the control fluorophore (tdTomato) signal. The predicted signal due to motion was subtracted from the denoised signal. To correct for bleaching of the fibre and fluorophores, signals were detrended using a double exponential fit to capture the temporal dynamics of bleaching: a first fast decay and a second slower one. Finally, the pre-processed dopamine signal was z-scored for each session to allow comparison across sessions and animals with different signal intensities.

4.1.1.3 Trial alignment

Photometry signals were aligned across trials by linearly time-warping the signal at the two points in the trial where timings were determined by subject behaviour; between initiation and choice, and between the second-step port illuminating and being poked. Activity at other time periods was not warped. The warping procedure did not introduce data distortion, and the same results are obtained without time-warping and just time-locking the data to single events (Figure 4.1).

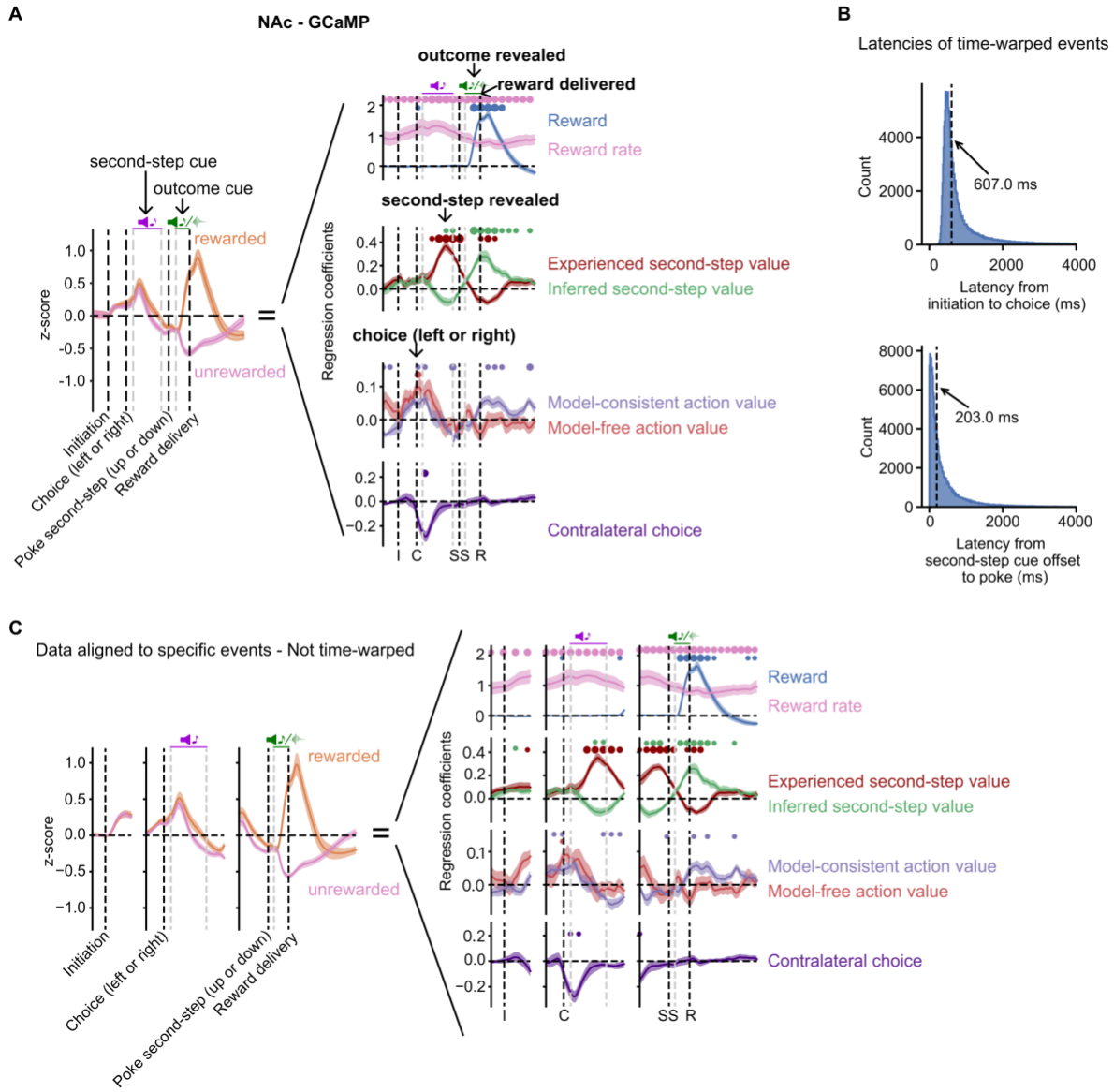


Figure 4.1. Signal comparison between time-warping and single events time-locking approaches. A,C) NAc axonal activity. A) Time-warped signal. Left – Mean z-scored activity. Right – Influence of each behavioural regressor on NAc axonal dopamine activity. Although the value of each regressor was the same across the whole trial, it only explained dopamine activity at the time the information became available. E.g. the reward regressor just explains dopamine when the outcome cue is played, first explaining a first positive modulation of the signal in response to the outcome cue, and a second response to the actual reward delivery. B) Latencies of the two time-warped events. Top – latencies (in ms) between initiation (poke centre) to choice (poke left or right). Bottom – latencies (in ms) between second-step cue offset and poking second-step (poke the illuminated port, up or down). C) As A, but data aligned to specific events, without warping the different latencies, from initiation to choice, and from second-step cue offset and second-step poke.

4.2.2 Linear regression model

To disentangle what behavioural variables influenced dopamine activity or released, we used linear regression to predict trial-by-trial dopamine activity at each timepoint in a trial as:

$$y(i,t) = \sum_{p} \beta_p(t) X_p(i) + \beta_o(t) + \varepsilon(i,\ t)$$

where $y(i,t)$ is the calcium z-scored activity on trial i at timepoint t, $\beta_p(t)$ is the weight for predictor p at timepoint t, $X_p(i)$ is the value of the predictor p on trial i, $\beta_o(t)$ is the intercept at timepoint t, and $\varepsilon(i,\ t)$ is the residual unexplained variance.

The linear regression was fit separately for each subject to obtain the coefficient time-courses $\beta_p(t)$. For each predictor we plot the mean and standard error across subjects. Statistical significance of each predictor at each timepoint was assessed using a t test comparing the

distribution of coefficients across subjects with zero, with Benjamini-Hochberg correction for comparison of multiple timepoints.

4.1.1.4 Behavioural regressors

One of the linear regressions used contained the following predictors:

- *Reward*: +0.5 if current trial is rewarded, -0.5 otherwise
- *Good Second-Step*: +0.5 if the second step reached on the current trial has high reward probability, -0.5 if low reward probability, 0 if neutral block.
- *Correct Choice*: +0.5 if subject current trial choice commonly leads to the high reward probability second-step port, -0.5 if to the low reward probability second-step port, 0 if neutral block.
- *Repeat Choice*: +0.5 if same choice as previous choice, -0.5 if different choice to previous trial.
- *Direct Reinforcement Action Value Update*: +0.5/-0.5 if current choice is the same as the previous choice and the previous trial was rewarded/not-rewarded, 0 if different choice from previous trial.
- *Model-Consistent Action Value Update*: +0.5/-0.5 if current choice commonly leads to the previous second-step when it was rewarded/not-rewarded, -0.5/+0.5 if current choice rarely leads to the previous second-step when it was rewarded/not-rewarded.
- *Previous Reward, Same Second-Step*: +0.5 if current second-step is the same as in the previous trial and previous trial was rewarded, -0.5 if same second-step and previous trial was unrewarded, 0 if current second-step different from second-step on the previous trial.

- *Previous Reward, Different Second-Step*: +0.5 if current second-step is different from previous second-step and previous trial was rewarded, -0.5 if different second-step and previous second-step was unrewarded, 0 if current second-step is the same as the second-step on the previous trial.
- *Common Transition*: +0.5 if a common transition occurs on the current trial, -0.5 if a rare transition.
- *Reward Rate*: exponential moving average of the recent reward rate (tau=10 trials).
- *Contralateral Choice*: +0.5 if current choice is in the contralateral side from the recording site, -0.5 if current choice is in the ipsilateral side.
- *Up Second-Step:* +0.5 if current second-step is up, -0.5 if current second-step is down.

Another linear regression used the same predictors as above except predictors *Previous Reward, Same Second-Step*, *Previous Reward, Different Second-Step* and *Model-Consistent Action Value Update*, which were replaced with predictors split by outcome as:

- *Same vs Different Second-Step, Previously Rewarded*: +0.5/-0.5 if current second-step is the same/different as in the previous trial and the previous trial was rewarded, 0 if the previous trial was not rewarded.
- *Same vs Different Second-Step, Previously Non-Rewarded*: +0.5/-0.5 if current second-step is the same/different as in the previous trial and the previous trial was not rewarded, 0 if the previous trial was rewarded.
- *Model-consistent Action Value Update from Rewarded Trials*: +0.5/-0.5 if current choice commonly/rarely leads to the previous second-step and the previous trial was rewarded, 0 if the previous trial was not rewarded.

- *Model-consistent Action Value Update from Unrewarded Trials*: -0.5/+0.5 if current choice commonly/rarely leads to the previous second-step and the previous trial was not rewarded, 0 if the previous trial was rewarded.

4.1.1.5 Inference reinforcement learning model value regressors

Other linear regressions used the value derived from the best fitting RL model, the inference reinforcement learning model. These regressions had the above-described *Reward* & *Reward Rate* regressors with the following additional regressors:

- *Previous Reward:* +0.5 if previous trial was rewarded, -0.5 otherwise
- *Second-Step Value:* The value $V_{Inf}(s)$ of the second-step reached on the current trial from the inference model.
- *Chosen Action Value*: The value $Q_{Inf}(c)$ of the first-step action chosen on the current trial from the inference model.

4.2.3 Exclusion criteria

From the 12 GCaMP mice implanted with 3 fibres each in VTA, NAc and DMS, just one fibre in NAc from one animal was misplaced into the anterior commissure, which was excluded for analysis. In DMS, although we obtained signal from the 12 fibres, the two with the most medial coordinates were excluded from the main analyses as they presented a negative modulation of reward – described in the results section.

From the 6 dLight mice implanted with 2 fibres each in NAc and DMS, again one NAc fibre was misplaced, into the ventricle, which was also excluded.

Single sessions with large artefacts were also excluded (46 sessions, ~9% of the total). These artefacts were due to a series of causes: large step change in recorded signals introduced through a malfunctioning rotary joint or a patch cord being unplugged from the fibre, or complete signal loss in one of the channels due to discharged battery.

4.3 Results

4.3.1 Mesostriatal dopamine activity and release have comparable dynamics

To assess how latent states in the environment interacting with decision policies influence both dopamine activity and striatal concentrations, we recorded dopamine activity in VTA, NAc and DMS using GCaMP6f (n=12 DAT-cre mice), and dopamine receptor binding in NAc and DMS using dLight1.1 (n=6 wild-type mice) while mice were performing the two-step task to get water rewards (Figure 4.2A). Each animal in the GCaMP group was implanted with 3 optic fibres, in VTA and NAc ipsilaterally, and DMS in the contralateral hemisphere; in the dLight group, 2 optic fibres were implanted in NAc and DMS in opposite hemispheres and recorded dopamine concentrations through dLight fluorescence, pan-neuronally expressed in striatum. Recordings were made from only one site per session and recording sites were alternated across sessions.

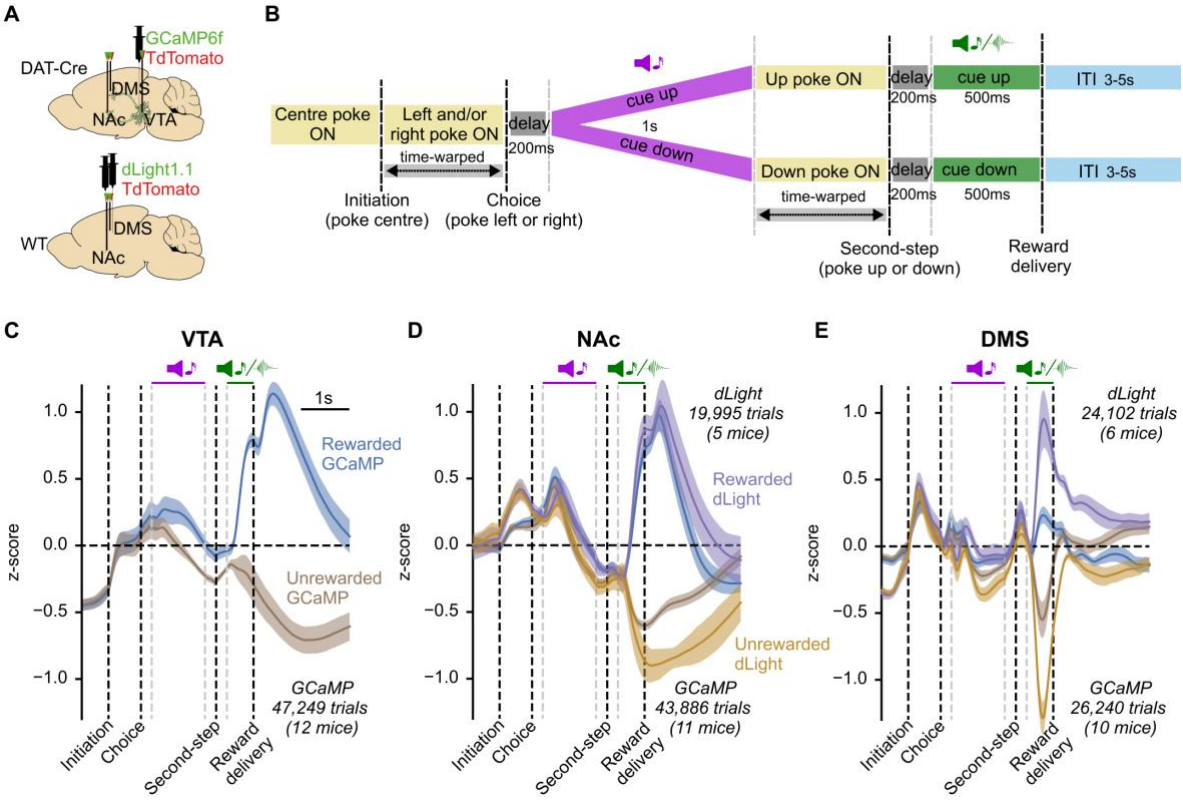


Figure 4.2. Dopamine activity and release are modulated across the trial. A) Injection and implant schematic. For calcium photometry, DAT-Cre mice were injected bilaterally with GCaMP6f and TdTomato in VTA, and optic fibres were implanted in VTA and NAc in one hemisphere, and DMS in the other hemisphere. For dLight photometry, WT mice were injected with dLight1.1 and TdTomato in NAc and DMS in different hemispheres, and optic fibres implanted at the same sites. B) Trial timeline. Trials were aligned by warping the intervals between the trial initiation and choice, and between the second-step port lighting up and being poked. C-E) Mean z-scored calcium signal on rewarded and unrewarded trials with shaded areas indicating cross-subject standard error. C) VTA dopamine activity (n: 12 mice, 121 sessions, 47,249 trials). D) NAc dopamine activity (n: 11 mice, 108 sessions, 43,886 trials) and concentrations (n: 5 mice, 64 sessions, 19,995 trials). E) DMS dopamine activity (n: 10 mice, 74 sessions, 26,240 trials) and concentrations (n: 6 mice, 91 sessions, 24,102 trials). Blue, GCaMP rewarded trials; Brown, GCaMP unrewarded trials; Purple, dLight rewarded trials; Yellow, dLight unrewarded trials.

Dopamine activity and concentrations fluctuated dynamically across the trial as mice initiated the trial, made their initial choice, received information about the second-step reached, and discovered the trial outcome (Figure 4.2 C-E). As expected, reward responses were prominent in all signals, increased after positive outcome, and decreased after negative (omission) outcome. However, reward responses in DMS calcium activity were relatively weaker, and asymmetrical in the opposite direction (stronger negative response to omissions, and weaker positive response to rewards) (Figure 4.2E). Moreover, while VTA and NAc dopamine also responded to reward delivery as well as the reward-associated cue, DMS was primarily modulated by the reward cue. Dopamine axonal activity (GCaMP) and release (dLight) also followed largely similar patterns. However, the dynamic range of outcome signals was substantially greater when measured using dLight than GCaMP, particularly in DMS.

As seen in Figure 4.2C-D, somatic dopamine activity dynamics in VTA is closely comparable to activity in their axonal projections to NAc. Similarly, NAc axonal activity and dopamine concentrations differ only (at least from this rough comparison, we will see later, how activity and release differ in response to task variables other than reward) in the response magnitude on unrewarded trials (although this could also be the result of the different indicators, and not an actual difference), and between initiation and choice. Dopamine concentrations increase as animals move from poking the central port to moving to the left or right choice port, which is not as pronounced in dopamine activity, consistent with a previous report (Mohebi et al., 2019). Similar modulation between initiation and choice is also present in both activity and concentrations in the DMS.

DMS dopamine activity and concentrations' pattern of modulation differed a bit more from the VTA to NAc somatic and axonal activity and release. While mesolimbic dopamine activity and concentrations were modulated both at the time of outcome cue and outcome delivery, DMS responses were more strongly modulated by the auditory cues than the actual water delivery or omission. Another striking difference in DMS is the magnitude of the outcome response between rewarded and unrewarded trials. As mentioned earlier, DMS activity presented and weaker and asymmetric response to reward and omission, but that was not the case in DMS dopamine release, which strongly responded to the rewarded (positively) and unrewarded (negatively) auditory cue.

To resolve other influences on dopamine activity and concentrations, we used a linear regression to model dopamine fluctuations across trials and time-points. We aligned activity across trials (Figure 4.1A, 4.2B) and ran a separate regression analysis for each timepoint in the aligned data, using predictors which varied from trial-to-trial but took the same value for all timepoints in each trial. The time courses of predictor loadings across the trial, therefore, reflect when each predictor explained variance in the dopamine activity or concentrations.

In this section, we will focus on the influence of reward in our regression model, and will discuss the other task-related predictors in the next sections (See supplementary figure 4.1 for the full regression model). Consistent with the average traces in figure 4.2, dopamine activity was positively modulated by trial outcome when outcome information became available, with some differences in dynamics and scale across regions (Figure 4.3A-E). The fact that the reward regressor did not explain any dopamine modulation prior to the outcome cue, further validated the suitability of the regression model.

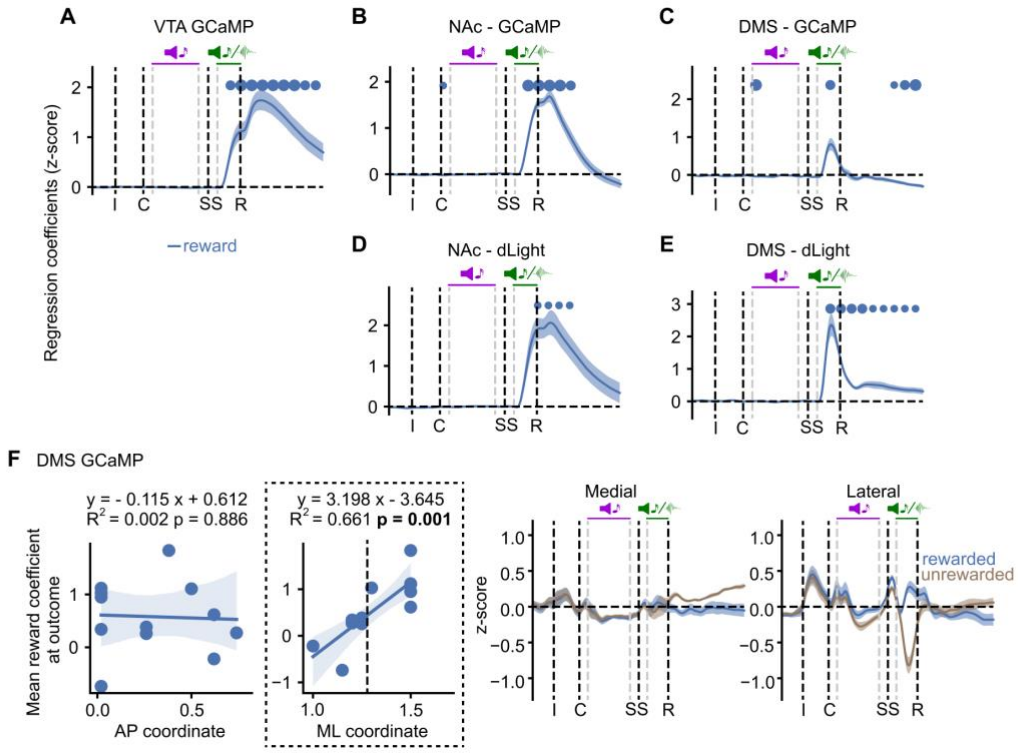


Figure 4.3. Influence of reward information on dopaminergic activity and concentrations. A-E) Regression coefficients predicting dopamine activity - GCaMP (A-C) and concentrations - dLight (D-E) based on outcome information. F) Medio-lateral gradient in reward modulation in DMS activity measured through GCaMP. Left, Correlation between mean reward coefficient at the time of outcome cue (from C) per subject and optic fibre's anteroposterior (AP) or mediolateral (ML) coordinate. Right, mean-z-scored activity split by outcome for subjects with medial or lateral placement determined by the dashed vertical line in the correlation plot. I, initiation; C, choice; SS, second-step; R, reward delivery.

Regressor coefficients from the reward predictor in dopamine activity in DMS differed across animals. A closer examination of this variability in individual mice's responses to reward revealed a mediolateral gradient in reward responsivity in DMS dopamine activity (Figure 4.3F). Most medial dopaminergic axons in DMS presented an inverse modulation to reward

(i.e. negative modulation to rewards, positive modulation to omissions), whereas the reward regressor coefficient increases with the mediolateral axis (Figure 4.3F).

4.3.2 Dopamine reflects experienced and inferred value consistent with the RPE framework

An important feature of our adapted two-step task is the anticorrelated structure of the reward probabilities. Therefore, experiencing reward on one second-step port gives information not only about the value of the experienced second-step port, but also enables the value of the non-experienced second-step port to be inferred. This feature allows for inference learning to guide behaviour, which the behavioural modelling indicates the mice are using to solve the task efficiently (See chapter 3). Here, we looked for evidence in the dopamine signal to support the conclusion that the mice were inferring a single latent variable that controlled both reward probabilities.

To do this, we examined the influence of two regressors that coded the previous trial's outcome (rewarded or not) either on trials where the second-step state was i) the *same* as that reached on the previous trial (i.e., direct experience can be used to update value estimates), or (ii) *different* from the previous trial (i.e., mice have to infer a value change from the previous trial's outcome) (Figure 4.4A). Based on the RPE hypothesis:

$$RPE = r + \gamma V_{new\ state} - V_{old\ state}$$

if each trial is composed of a discrete sequence of 3 different states (initiation state → second-step state → outcome state), at a state transition from one state to another within a trial, the value of the new or current state (e.g. value of the second-step state when we are in the second-step state) contributes positively to the RPE, while the value of the old state (e.g.

value of the second-step state when we are in the outcome state) contributes negatively. Therefore, if previous trial was rewarded and on the current trial the same second-step state is reached, we would expect the value update of the second-step state to be represented by a *positive modulation* of the dopamine signal at the time the second-step state is revealed, representing the value of the new or current state (value of the second-step state), followed by a *negative modulation* when it exits the second-step state and enters the outcome state, representing the value of the old state (i.e. now we are in the outcome state, and therefore, the negative modulation is the value of the old state which was the second-step state) (Figure 4.4B). This biphasic signal is a signature of an RPE computation (Behrens et al., 2008; Rutledge et al., 2010; Niv et al., 2012).

We have different predictions within the RPE framework if instead a different second-step state (i.e. the previously non-experienced second-step state) is reached in the current trial. If the dopamine system is learning *independent* values for each second-step state, we would expect no influence on dopamine if a different second-step state is reached, as no update from the previous trial would have occurred (Figure 3.1, 4.4C). However, if instead, dopamine had access to the latent variable tracking the reward probabilities, and it performs state value *inference*, we would expect a mirror image from the same second-step value. This would be the case, as if on the previous trial reward was received, dopamine would increase the value of the experienced state, while decreasing the value of the non-experienced second-step state. Consequently, when the different second-step is reached, outcome would negatively influence dopamine at the time the second-step state is revealed, reversing at outcome time, as predicted by RPE computation (Figure 3.1, 4.4D).

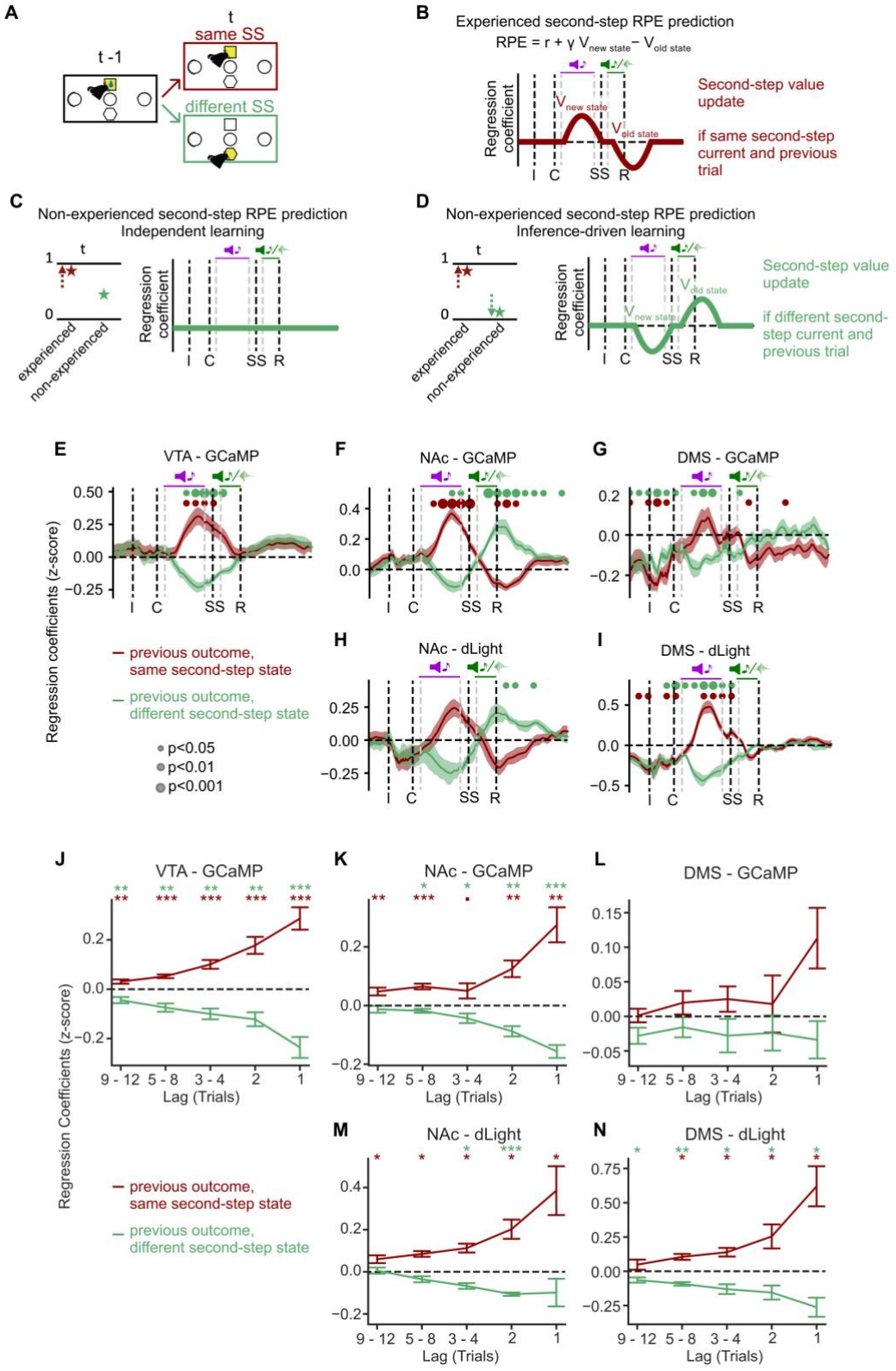


Figure 4.4. *Dopamine is modulated by experienced and inferred value. A) Regressors schematic. Dark red, effect of previous outcome if the same second-step state (SS) is reached; Green, effect of previous outcome if a different second-step state is reached. B) Dopamine modulation prediction for the*

*previous outcome – same second-step state regressor based on RPE hypothesis. C-D) Dopamine modulation predictions for the previous outcome – different second-step state regressor based on RPE hypothesis: C) if dopamine learns independent values, no update of the different second-step state would have occurred, D) if dopamine has access to the latent variable and infers the value of the different (non-experienced) second-step state. E-I) Regression coefficients predicting dopamine activity - GCaMP (E-G) and concentrations - dLight (H-I) based on second-step value update that would have ocurred on the previous trial. J-N) Lagged linear regression predicting dopamine response (500ms at the end of the second-step cue, baseline substracted) to the second-step cue as a function of the extended history of trial events over the previous 12 trials. ·p<0.1, *p<0.05, **p<0.01, ***p<0.001 after Benjamini-Hochberg correction for multiple lags.*

When the second-step state was the same, outcome on the previous trial *positively* modulated dopamine activity at the time the second-step was revealed on the current trial (Figure 4.4E-I). This is predicted by both independent value learning at the two second-step states and by inference about a single latent variable that controls both reward probabilities. However, when the second-step state was *different* from the previous trial, reward on the previous trial had a *negative* influence on dopamine at the time the second-step state was revealed on the current trial. This is not predicted by independent value learning – because reward at one second-step state does not affect the value of the other second-step state – but *is* predicted by inference of a single latent variable: evidence one state is good is also evidence the other is bad. This pattern of modulation is observed in all the regions, in both dopamine activity and concentrations. In NAc, but, unexpectedly, in neither other region, loading on these regressors reversed sign at outcome time. This is exactly as predicted for an RPE signal.

To further confirm that this modulation by inferred value was not the result of confounding effects of the trial history of past events, we performed a lagged regression predicting mean dopamine photometry signal at the end of the second-step cue using the same regressors as for the one-trial back photometry regression (Figure 4.4J-N). This confirmed our initial analyses, showing that the dopamine circuitry is modulated not just by experienced but also inferred state values that have an influence that extends across multiple trials (though less clearly in the DMS axonal activity).

Behaviourally, we observed that animals were treating rewarded and unrewarded trials differently, with omission trials having negligible effects on future behaviour. We therefore tested whether this asymmetric effect of reward and omissions was also reflected in dopamine. To do so, we modified our two second-step value regressors to compare the effect of reaching the same or different second-step state if previous trial was rewarded or not. Regressor loadings predicting dopamine activity in VTA, and activity and concentrations in NAc showed that dopamine was more strongly driven by value updates as a result of reward than omissions (Figure 4.5A,B,D). Strikingly, while this was also the case in predicting DMS dopamine concentrations, activity was instead more strongly modulated by updates from omissions (Figure 4.5C,E), in line with the response to reward observed in the average traces (Figure 4.5E).

In summary, here we have seen that not only is mice's behaviour consistent with inference learning as opposed to independent value learning, but mesostriatal dopamine activity and release were also influenced by inferred state value.

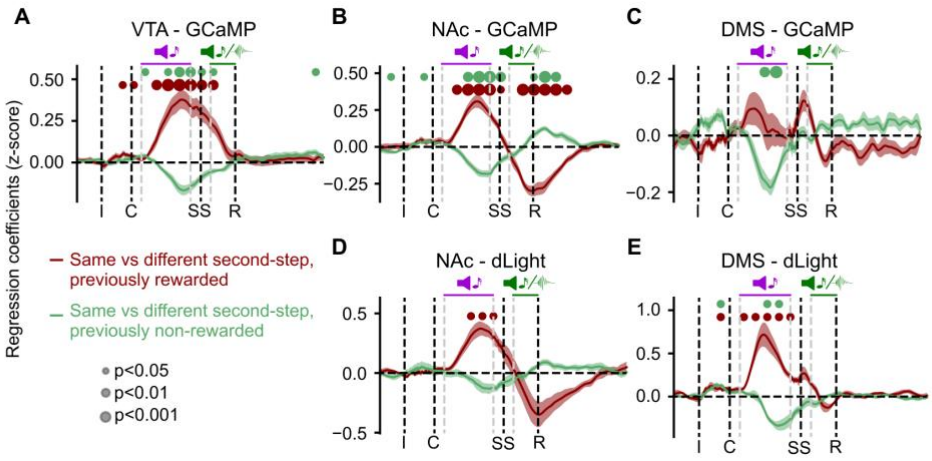


Figure 4.5. Asymmetric second-step state value update based on previous outcome. A-E) Effect of reaching the same vs different second-step state based on if previous trial was rewarded (dark red), or unrewarded (green). A) VTA dopamine calcium activity (GCaMP). B) NAc axonal activity. C) DMS axonal activity. D) NAc dopamine concentrations (dLight). E) DMS dopamine concentrations.

4.3.3 Decision policies and action sequence are orthogonally reflected in the dopamine system

We next assessed how different decision policies inform dopamine responses. To do so, we used two regressors in our one-trial back photometry regression which captured first-step choice value based on different decision policies, without or with a model of the task. Note this regression includes additional predictors to capture variance associated simply with repeating the previous choice or choosing the high value option (see Supplementary Figure 4.1, 4.2) to ensure that what we are observing can be attributed to an outcome-specific update.

A *direct-reinforcement* regressor coded the previous trial outcome (rewarded or not) only on trials where the current choice was the same as the previous choice (Figure 4.6A) – i.e., the

direction in which model-free direct reinforcement of the previous choice by trial outcome would update the value of the first-step action chosen on the current trial. This regressor loaded positively around choice time in VTA and NAc activity, but not DMS (red lines, Figure 4.6C-E), consistent with action value estimates used by mesolimbic dopamine incorporating a simple model-free update over observable events.

A second *model-consistent* predictor coded the direction in which the previous trial's outcome would affect the value of the action chosen on the current trial, given a correct model of the transition structure linking first-step actions to second-step states (Figure 4.6B). This loaded positively around choice time in VTA, NAc and DMS calcium activity, consistent with the dopamine system having access to value information about the chosen action that respected the task structure (Figure 4.6C-E). Dopamine concentrations, although suggestive, modulations by model-based predictions did not reach significance (Figure 4.6F-G). Similarly to second-step value update coding in NAc where the signal reverses at outcome time, this is also observed for the model-consistent action value update, again consistent with RPE computation.

As for the second-step value update regressor, to determine whether there was also an asymmetric action value update based on previous outcome, we split the model-consistent regressor based on whether previous trial was rewarded or unrewarded. All regions were mainly driven by regressors coding updates after rewarded trials, and not unrewarded trials. Not only this was the case in dopamine activity, but also dopamine concentrations (Figure 4.7A-E).

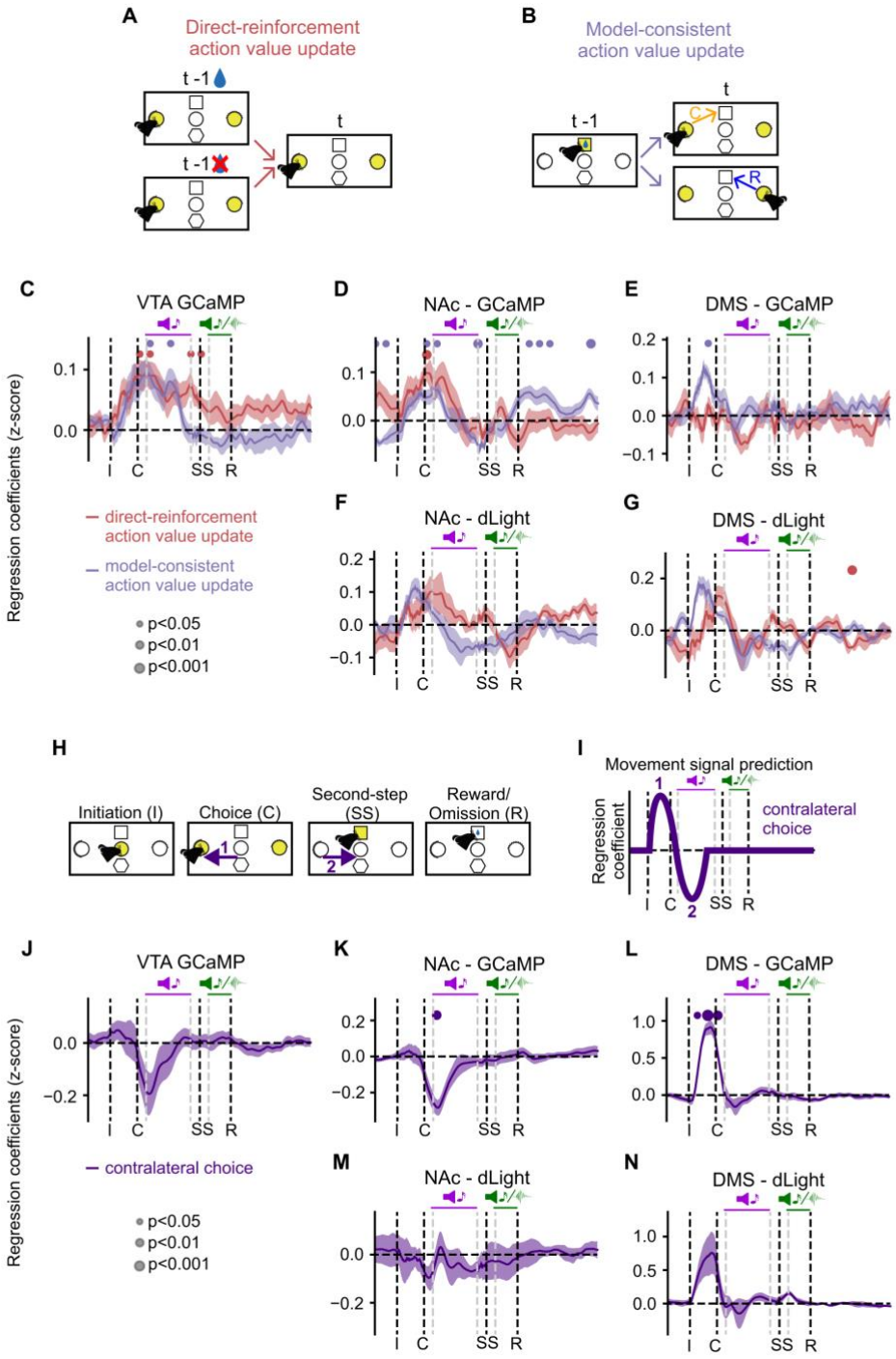


Figure 4.6. Dissociable effects of action value computation and action sequence in the dopamine signalling. A-G) Regressors reflecting the effect of action value update if different strategies are used – without or with the presence of an internal model. H-N) Contralateral choice regressor capturing the effect on dopamine signalling if a contralateral choice from the recording site is made on the current trial as opposed to the ipsilateral choice. A) Schematic of the direct-reinforcement action value

update regressor, coded as positive if the same choice is made as in the previous trial which was reward, negative if it was not rewarded, and 0 if a different choice is made. B) Schematic of the model-consistent action value update, coded as positive if the choice on the current trial commonly leads to the same second-step port as the previous trial was rewarded, negative if the previous trial was unrewarded, and the opposite if the current choice rare leads to the same second-step as the previous trial. C) VTA dopamine calcium activity (GCaMP). D) NAc axonal activity. E) DMS axonal activity. F) NAc dopamine concentrations (dLight). G) DMS dopamine concentrations. H) Schematic of the multi-step sequence: 1. Movement from centre to choice (left or right) and 2. Movement from side to centre. I) Dopamine prediction of the influence of contralateral choice if dopamine reflect a pure movement signal. J) VTA dopamine calcium activity (GCaMP). K) NAc axonal activity. L) DMS axonal activity. M) NAc dopamine concentrations (dLight). N) DMS dopamine concentrations.

In parallel to action value computation, dopamine signalling reflected action movement. Our linear regression also contained another regressor, a *contralateral choice* regressor, which captured whether the choice made (left or right) on the current trial was on the contralateral vs ipsilateral side from the recording site, independently of choice or state value. Each trial requires a left/right movement first from centre to side to make the choice, then the converse movement direction to go from the side to the second-step port (as well as a contribution of up/down) (Figure 4.6H). Dopamine activity that simply reflects contralateral movement would therefore generate *positive* loading on the contralateral choice regressor between initiation and choice as animals move from the centre to side, followed by *negative* loading as they return to the centre following choice (Figure 4.6I).

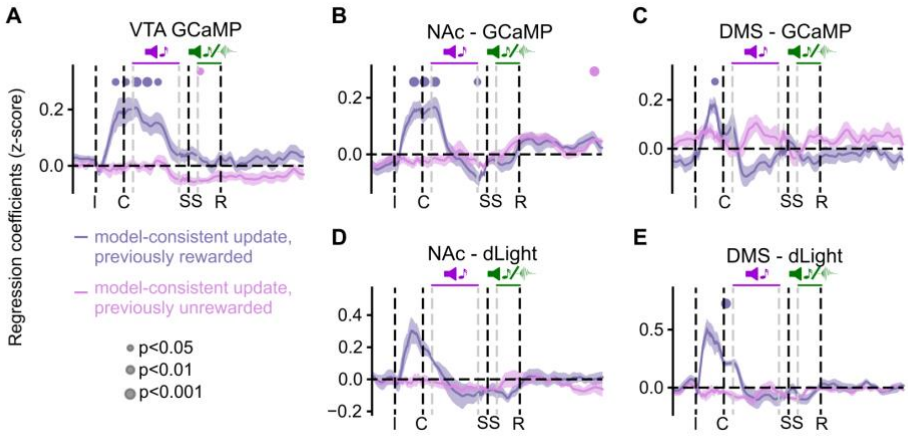


Figure 4.7. Asymmetric effect of model-consistent action value update from rewarded and unrewarded trials. A) VTA dopamine calcium activity (GCaMP). B) NAc axonal activity. C) DMS axonal activity. D) NAc dopamine concentrations (dLight). E) DMS dopamine concentrations. Purple, update after rewarded trials; Pink, update after unrewarded trials.

Strikingly, activity in dopamine axons in NAc and DMS responded differentially to lateralised movements at different points in the trial: the first movement from centre to left/right choice port or the second movement from choice port to second-step ports). Contralateral choice *positively* modulated calcium activity in DMS terminals only during the first movement (Figure 4.6L). By contrast, there was a negative influence on NAc activity only around the second movement (Figure 6.6K). This is inconsistent with a pure movement signal in each region, as the predicted biphasic modulation (Figure 4.6I) is not observed in a single region, but rather suggests a lateralised response action sequence. In VTA, a similar modulation as in NAc was observed, although this did not reach corrected significance at any timepoint (Figure 4.6J). This pattern was replicated with DMS dLight, but, surprisingly, not with NAc dLight (Figure 4.6M, N). Note that this regressor do not capture all the movement information from animals,

and therefore, we cannot discard actual movement from being encoded in the dopamine signalling.

Because of this lateralised action signal, which particularly influences dopamine activity in DMS, we then asked whether lateralised actions interact with decision strategies and state value updates. To do this, we split the previous regressors – model-free, model-consistent, and second-step state value update – based on whether the current choice is contralateral or ipsilateral. Note that these results are not definite and only suggestive as more data would be required to further explore this effect (i.e. regressors were strongly correlated). There was no difference in value coding based on lateralised choice in VTA and activity and concentrations in NAc. But in DMS, while dopamine concentrations regressors also did not differ based on whether the choice was contralateral or ipsilateral from the recording site, dopamine activity regressors mainly explained dopamine modulation from updates from contralateral choices and not ipsilateral ones (Figure 4.8).

These results show that both mesolimbic and dorsomedial striatal dopamine reflect the value computations at the level of choice using information from an internal model of the task structure. This computation was dissociable from a signal tracking the action sequence. The action sequence is tracked across regions, first DMS reflects the initial action movement from initiation to choice, and afterwards, NAc tracks the second action movement from choice to second-step state. Furthermore, DMS activity but not concentrations report contralateral value updates.

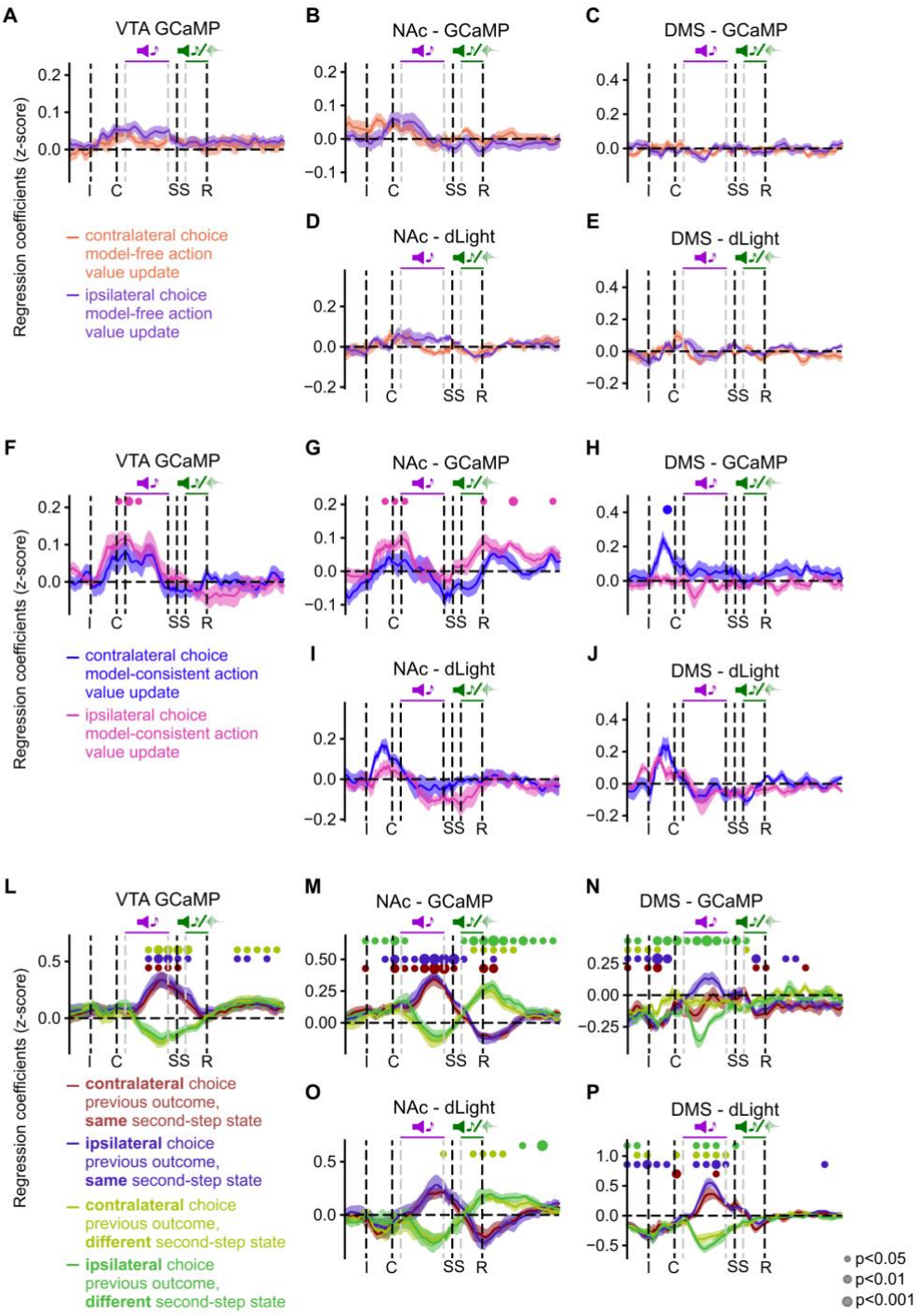


Figure 4.8. Value updates interacting with contralateral or ipsilateral choice. A-E) Model-free action value update split by contralateral or ipsilateral choice on the current trials across regions and indicators. F-J) Model-consistent action value update split by contralateral and ipsilateral choice across regions and indicators. L-P) Second-step state action value update split by contralateral and ipsilateral choice across regions dopamine activity and concentrations.

4.3.4 Dissociable influence of RPE and reward expectation in dopamine signalling

The above analysis used predictors expressing specific hypotheses about how the previous trial's events affected value information in the dopamine system. A complementary picture can be obtained using trial-by-trial estimates of decision variables from the best-fitting behavioural model to predict dopamine fluctuations. We used the same linear regression framework but with model values for the first-step choice and second-step state as predictors. As seen in chapter 3.1, the best model explaining animals' behaviour used a single latent variable to track which option is best, updated in opposite directions by reward at the two second-step states. The second-step value is just this variable interacted with the second-step state reached on the trial, while the chosen action value is this variable interacted with the chosen action. Both values, therefore, respect the anticorrelated reward probabilities, and the chosen action value combines this with the action-state transition structure.

Second-step state value had a strong positive influence on VTA and NAc calcium activity and concentrations at the time the second-step was revealed, which subsequently reversed at outcome time, consistent with RPE signalling (Figure 4.9A,B,D). The time-course was more complex in DMS, with the regressor coefficients having peaks following both the second-step cue and second-step port entry (Figure 4.9C,E). Chosen action value had a strong positive influence on activity in all 3 regions around the time of the choice, which then reversed when the second-step was revealed, again consistent with RPE in all three regions, and striatal dopamine concentrations. However, most prominently in NAc.

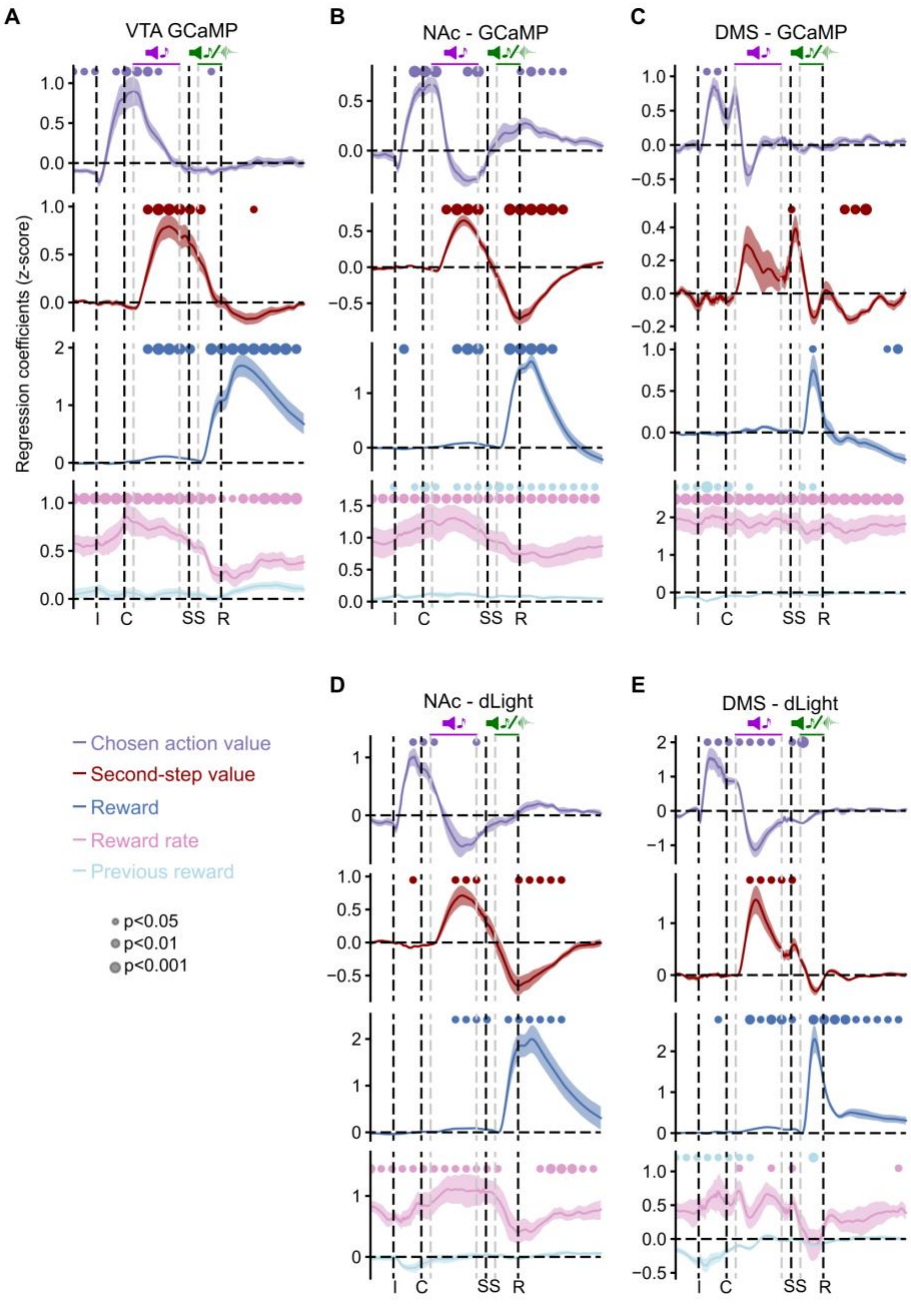


Figure 4.9. *Dissociable influence of value and reward rate on dopamine. A-E) Coefficients from linear regression predicting dopamine activity using decision variables from the correlated inference model. A) VTA dopamine calcium activity (GCaMP). B) NAc axonal activity. C) DMS axonal activity. D) NAc dopamine concentrations (dLight). E) DMS dopamine concentrations. Purple, chosen action value: behvaioural model value of chosen first-step action; Dark red, second-step value: behvaioural model value of second-step reached on current trial; Blue, reward: current trial outcome; Pink, reward rate: exponential moving average with tau = 10 trials; Light blue, previous reward: previous trial outcome.*

To investigate the different value reporting in DMS activity versus concentrations, with activity tracking value during contralateral movements but with concentrations tracking both ipsilateral and contralateral movements, we split the chosen action value and second-step value regressor based on whether the current trial was a contralateral or ipsilateral choice. Again, while VTA and NAc calcium activity, and NAc and DMS concentrations were modulated by regressors reporting action and state value on both contralateral and ipsilateral movements, regressor mainly loaded during contralateral movements only in DMS dopamine activity (Figure 4.10).

These data confirm that value information, which respected the task causal and correlation structure, strongly influenced dopamine responses to both the first-step action and second-step choice.

As both reward rate and value have been shown to correlate with dopamine activity (Hamid et al., 2016; Mohebi et al., 2019), we additionally included recent reward rate as a predictor, calculated as an exponential moving average with a time constant of 10 trials (qualitatively similar results were obtained using time constants of 5-15 trials, data not shown). Reward rate had a strong positive influence on dopamine in all 3 regions activity and concentrations (Figure 4.9A-E), accounting for separate variance from the prominent influence of action and state values. Unlike the influence of value and reward, which were tightly time-locked to when information became available, reward rate positively influenced activity at all timepoints, with only weak modulation across the trial.

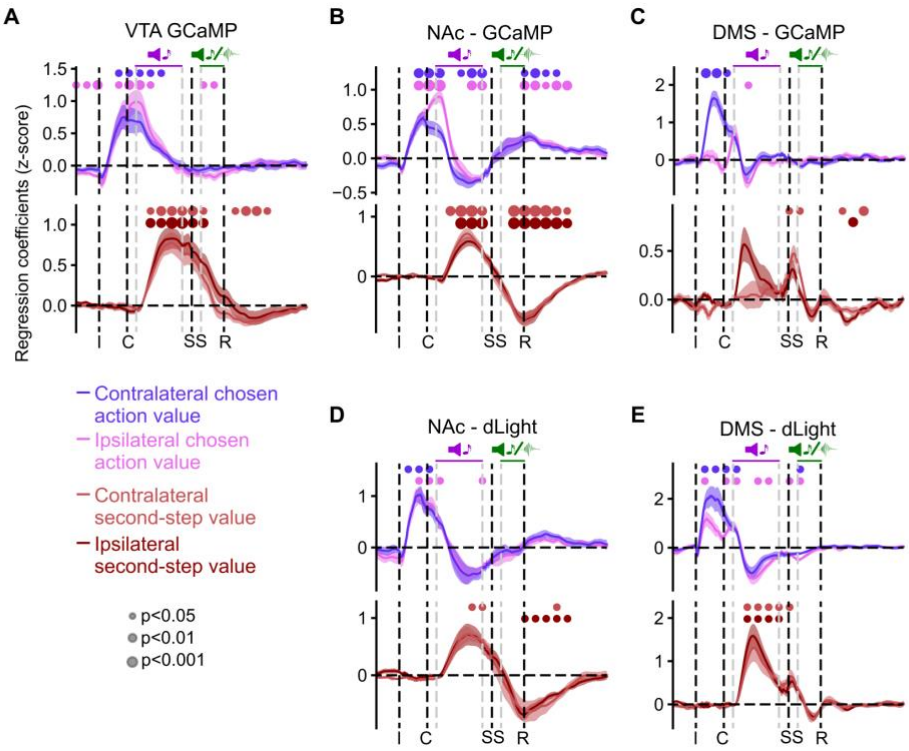


Figure 4.10. Contralateral and ipsilateral values across regions. Influence of chosen action value when on the current trial a contralateral (purple) or ipsilateral (pink) choice is performed on VTA dopamine calcium activity (A), NAc axonal dopamine calcium activity (B), DMS dopamine axonal calcium activity (C), NAc dopamine concentrations (D), and DMS dopamine concentrations (E). Also, influence of second-step value based on the chosen side, contralateral (light red) or ipsilateral (dark red) on dopamine signalling.

To understand how this effect was manifested in the photometry signals themselves, we divided up the data into periods when the reward rate was high (>0.7 rewards / trial), mid (between 0.4 – 0.7 rewards / trial) or low (<0.4 rewards / trial). As can be observed in Figure 4.9, this constant positive influence across the whole trial of reward rate on dopamine signalling observed in the linear regression was observed as a baseline shift throughout the trial based on recent reward rate (Figure 4.11).

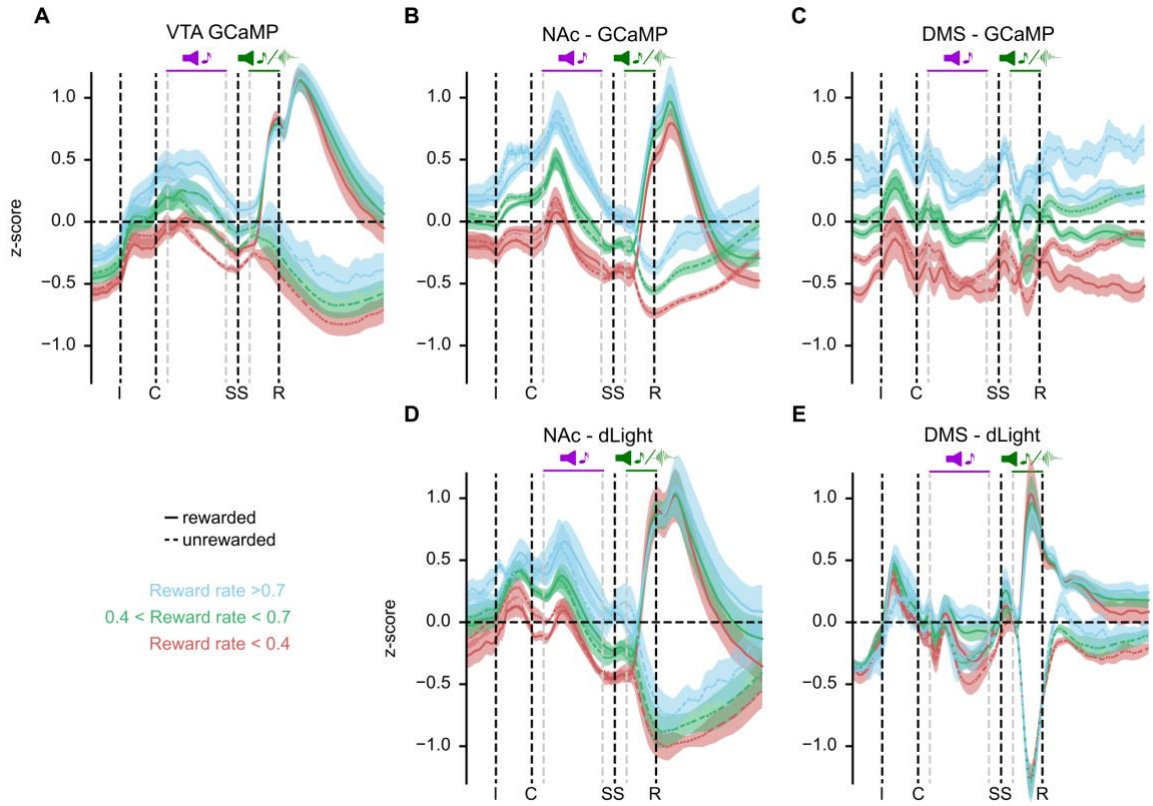


Figure 4.11. *Sustained dopamine signals corresponding to recent reward rate. A-E) Mean z-scored dopamine activity in VTA (A), NAc (B), and DMS (C), and dopamine concentrations in NAc (D) and DMS (E) split by rewarded and unrewarded trials, and recent reward rate. Blue – high reward rate > 0.7 rewards/trial (exponential moving average with tau = 10 trials). Green, medium reward rate (between 0.4 and 0.7). Red, low reward rate (less than 0.4). Continuous line, rewarded trials; dashed line, unrewarded trials.*

Altogether, these data show that while action and state values influenced phasic dopamine activity in a manner consistent with RPE signalling, recent reward rate explained separate variance in dopamine activity on a slow timescale. Therefore, both RPE and reward expectation framework can coexist in the same dopamine population signalling, as a phasic or sustained response, respectively.

4.3.5 Other non-value related information also modulates dopamine

In addition to value or movement information being conveyed by dopamine, the behavioural richness of the two-step task allowed us to test whether other important task parameters could also be influencing dopamine signalling.

First, we asked whether the nature of the transition experienced on the current trial was reflected in dopamine signalling. Common transition positively modulated dopamine calcium activity in VTA at the time of the second-step cue, when the time when transition information becomes available, but also, negatively modulated somatic activity at outcome time (Figure 4.12A). This is consistent with a prediction error signal. Although this could be considered a state prediction error signal, another possibility is that common transitions could be endowed with a higher value than rare ones. Activity and dopamine concentrations in NAc reflected a similar pattern as in VTA, however, while modulation at second-step did not reach statistical significance, the negative modulation at outcome time was stronger (Figure 4.12B, D). Finally, in DMS, there was again some evidence for a divergence between the patterns of signals recorded using GCaMP in the axons versus dLight for concentration, although it did not reach significance after multiple comparison correction. While DMS dopamine axonal signalling was *negatively* influenced by common transition (i.e., more strongly modulated by rare transitions), dopamine release had the opposite modulation, (i.e., common transition positively modulated dopamine concentrations in DMS) (Figure 4.12C, E).

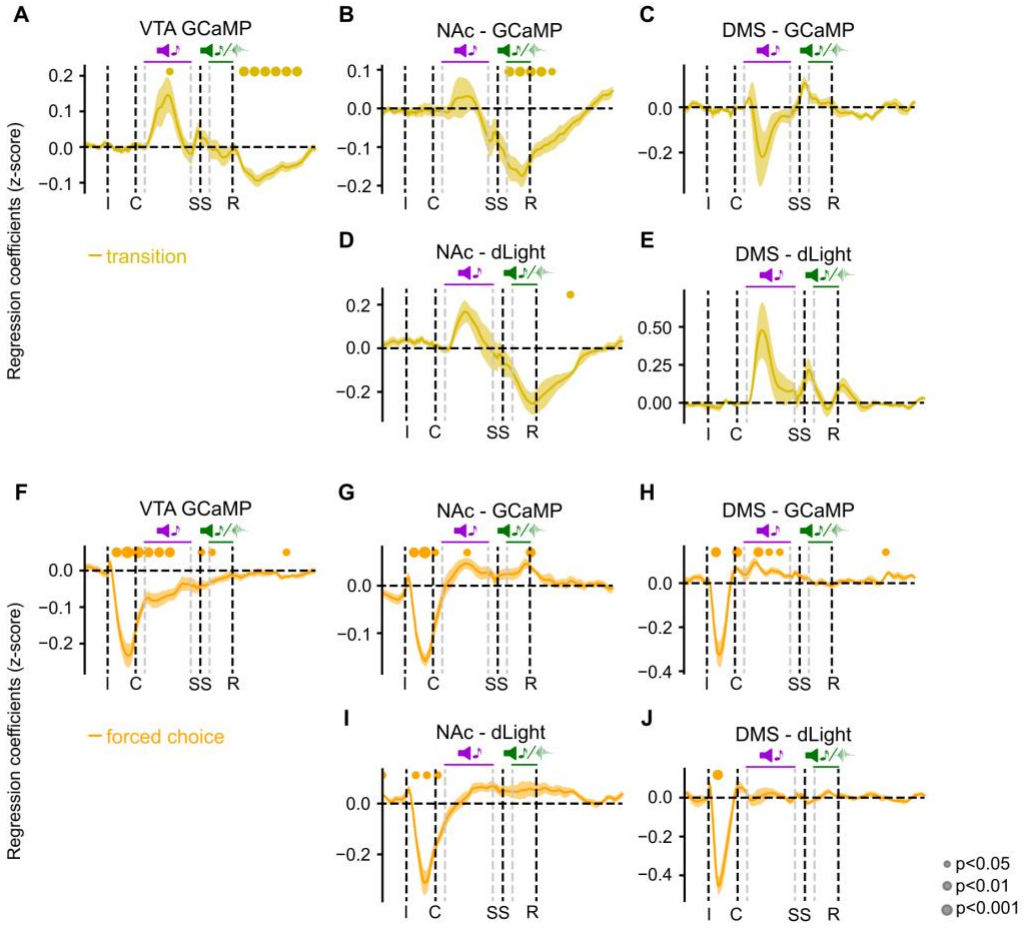


Figure 4.12. Dopamine is also modulated by non-value related information. A-E) Coefficients for the transition regressor, which captured the influence for experiencing a common transition versus a rare transition on the current trial on dopamine calcium activity in VTA (A), NAc (B), and DMS (C), and dopamine concentrations in NAc (D) and DMS (E). F-J) Coefficients for the forced choice regressor, which captured the influence of experiencing a forced choice trial versus a free choice one on the current trial on dopamine calcium activity in VTA (F), NAc (G), and DMS (H), and dopamine concentrations in NAc (I) and DMS (J).

The lack of statistical significance across regions at the time when transition occurred but not at outcome time, could be the result of the variability in the timing of movement when the animal detected the transition that actually occurred. This could be the case, because although the second-step auditory cue reveals whether the up or down illuminated, in some trials during rare transitions, animals first move towards the port that the chosen port

commonly transitioned to, and then they correct to the actual second-step transitioned to. However, further analysis, together with video analysis of animals' actual movement on each trial, would help answer these questions.

Finally, we examined whether there was an effect of forced trials, when animals were presented with one illuminated choice port as an option. This demonstrated that all regions were sharply negatively modulated by forced trials at the time between initiation and choice (Figure 4.12F, J).

These results indicate that population calcium activity from dopaminergic neurons in VTA and their axonal projections to NAc and DMS multiplex a variety of task-related parameters.

4.4 Discussion

We recorded dopamine activity from neuronal cell bodies in VTA, and their axonal projections to NAc and DMS, together with dopamine concentrations in these striatal regions while mice were performing a multi-step sequential decision-making task.

Moment-by-moment dopamine fluctuations were strongly influenced by both the value of actions taken and states encountered in a way consistent with RPE signalling. This modulation respected the task structure and reflected the presence of a single latent variable controlling the reward probabilities. Consequently, dopamine was not only modulated by experienced value but also inferred value predictions.

In parallel to this phasic signal in line with the RPE framework, we also observed other signals at different timescales. Sustained dopamine signalling correlated with reward rate, consistent with the reward expectation/motivation framework of dopamine. Furthermore, phasic dopamine reflected lateralised movement sequence.

Although there were some differences between dopamine activity and concentrations, particularly in DMS, both signals had comparable dynamics across the two-step task.

4.4.1 Dopamine during latent state inference guided choice

The information dopamine carries was classically presented as a model-free signal. This signal responded to reward predictive cues and unexpected outcome (Montague *et al.*, 1996; Schultz *et al.*, 1997; Steinberg *et al.*, 2013; Eshel *et al.*, 2016) in a model-free manner, without the presence of an internal model of environment structure.

Recent and growing evidence is exposing the complexity of dopamine dynamics, responding in a way consistent with an internal model of task structure, tracking the statistical relationships in the environment (Sadacca et al., 2016; Takahashi et al., 2016; Sharpe et al., 2017, 2020; Yves et al., 2018; Engelhard et al., 2019). Having an internal model of the environment allows to infer not directly experienced events or features of the environment, which inform dopamine responses (Bromberg-Martin et al., 2010b; Wimmer et al., 2012; Starkweather et al., 2017; Babayan et al., 2018; Gershman and Uchida, 2019).

The data in this chapter build on the previous evidence of dopamine reflecting inference learning. Rewards obtained in one second-step state *positively* influenced dopamine

response on the next trial if the same second-step state was reached, reflecting state value update of the experienced state. But at the same time, they *negatively* influenced dopamine if the *non-experienced* second-step was reached, reflecting *inferred* state value. This is inconsistent with learning independent values for the two states, as reward in one state would then not affect the value of the other, but is consistent with the inference that reward obtained at one second-step state makes obtaining reward at the other less likely.

At the same time, the unobservable latent state, the anticorrelated reward probabilities, interacted with subjects' decision policies. Dopamine was *positively* modulated at action time, reflecting action value respecting the task structure, and it *negatively* modulated dopamine in the transition from the choice state to the next one, displaying the signatures of an RPE modulation.

4.4.2 RPE and motivation account coexistence in the same population signalling

Dopamine signalling has long been conceptualised as a signal conveying RPE when predicted and estimated outcome differ (Schultz et al., 1997; Eshel et al., 2016), which could be used for policy update (Sutton and Barto, 1998; Steinberg et al., 2013). However, recently, the RPE framework of dopamine has been questioned.

Dopamine activity and concentrations ramp up (gradually increase) as positive reinforcer becomes closer, physically or in time (Roitman et al., 2004; Wassum et al., 2012; Collins et al., 2016; Syed et al., 2016; Guru et al., 2020). Dopamine ramping has been recently interpreted as a signature of motivation. This motivational account argues that dopamine ramping is the

result of value expectation, as it is correlated with reward rate, and cannot be explained under the RPE framework (Howe et al., 2013; Hamid et al., 2016; Mohebi et al., 2019).

However, others argue that ramping can arise as a consequence of RPE at each time step (Gershman, 2014; Morita and Kato, 2014; Kim et al., 2020). Environmental manipulations in a virtual environment in which both the RPE and the motivation accounts were tested, supported dopamine ramps being explained through the former one (Kim et al., 2020).

A key signature of RPE is that the value of encountered states contributes at two timepoints with opposite signs – first positively, when the state is entered, then negatively, when entering the next state (Behrens et al., 2008; Rutledge et al., 2010; Niv et al., 2012). We observed this biphasic influence of the values of both first-step actions and second-step states in both calcium activity in VTA, NAc and DMS, and in dopamine concentrations in NAc and DMS. This biphasic pattern was most prominent in NAc, with VTA and DMS signals both showing substantially larger initial positive modulation than later negative modulation. Intriguingly, when evaluating only the influence of the most recent outcome on second-step state value, rather than the extended history, the biphasic pattern was prominent in NAc but absent in VTA. Differences between the VTA and striatal signals could reflect local modulation of terminals, for example by cholinergic neurons (Threlfell and Cragg, 2011; Kramer et al., 2022; Liu et al., 2022; Reynolds et al., 2022), or alternatively, the VTA signal may be influenced by calcium activity in dendrites that is partially decoupled from spiking due to somatic inhibition.

In addition, we observed the average reward rate over recent trials had a strong positive influence on dopamine activity in all regions. Importantly, this accounted for an entirely

separate portion of variance from the value of trial events. Value-related signals were tightly time-locked to specific events, whereas reward rate had sustained effects across the trial from before initiation to after outcome. This appears broadly consistent with theoretical proposals that tonic dopamine represents average reward rate, which acts as the opportunity cost of time in average-reward RL (Niv et al., 2007). Correlation between reward rate and NAc dopamine concentrations is consistent with recent reports (Hamid et al., 2016; Mohebi et al., 2019), but that with VTA calcium is unexpected given previous reports of no correlation with VTA spikes (Mohebi et al., 2019).

Therefore, RPE and motivation frameworks are not confronting theories of dopamine function, but both signals can coexist in the same population coding.

4.4.3 Movement and other task parameters in dopamine

Apart from value signals, dopamine is also involved in motor control (Dodson et al., 2016; Howe and Dombeck, 2016; Parker et al., 2016; da Silva et al., 2018; Lee et al., 2019). In line with previous reports, DMS dopamine activity was positively modulated by contralateral choice (Parker et al., 2016; Lee et al., 2019). However, in our task, on each trial, the first-step action movement from initiation to choice was followed by a second lateralised movement from choice back to the centrally located second-step port. The second lateralised movement did not evoke an inverse response as animals change the direction of movement. Therefore, against previous reports, this suggests that the lateralised activity did not simply reflect movement, but rather may reflect action sequence initiation (Jin and Costa, 2010). This lateralised activity conveying movement sequence initiation is further supported by the fact,

that the second lateralised movement, was actually present in NAc activity (and suggested in VTA).

Lee *et al.* (2019) concluded that this contralateral movement selectivity in DMS axonal projections could not be explained through the RPE account, and instead, simply reflects movement direction. Our data further adds to this RPE versus movement dichotomy. DMS dopamine activity and concentrations were strongly modulated by contralateral choice, reflecting the direction of movement of the first-step action. This separate contribution of movement direction from RPE value information was also present in NAc activity in the second lateralised movement, but not concentrations. The lateralised response present in NAc activity (and VTA although not significant) is unexpected as previous reports suggest that these regions do not respond to lateralised actions (Parker et al., 2016; Moss et al., 2021).

Separately, DMS activity, but not concentrations, RPE value coding was also lateralised. DMS axonal activity reflected action and second-step state value which resulted from contralateral movements. DMS activity was modulated by contralateral action value, whereas no modulation was present if an ipsilateral choice was performed. Similarly, second-step state value was only present in DMS activity after contralateral movement (after ipsilateral initial first-step choice, performing a contralateral movement from choice to second-step port). This movement direction interacting with action and state value was only present in DMS activity (consistent with Moss *et al.*, (2021), even in a no-movement task, but responding to contralateral stimuli), but not in DMS concentrations, and neither in NAc (activity and concentrations) nor VTA.

SNc and DMS dopamine activity and concentrations have been shown to increase before movement onset (Dodson et al., 2016; da Silva et al., 2018). Although we did not observe this anticipatory signal, further analysis using actual mice position and movement would confirm or reject this in our multi-step decision-making task. One reason why we did not observe it despite aligning activity to action initiation, could be due to animals not staying still in the initiation port until it became active, but were, majorly, constantly moving.

4.4.4 Dopamine differences across regions, activity and release

The relation between dopamine somatic and axonal spiking, together with its relation to axonal dopamine release is a topic of current discussion, with little and sparse evidence about this relation, due to technical limitations that current developments are postulated to overcome (de Jong et al., 2022).

One report showed a marked contrast between VTA somatic dopamine firing and NAc dopamine concentrations, with dopamine concentrations correlating with reward rate but not somatic spiking (Mohebi et al., 2019). In this report, however, while somatic activity was obtained through electrophysiological recordings, dopamine concentrations were recorded using dLight photometry readings. A recent study found in SNc dopamine projections, that if photometry recordings were performed in genetically defined neuronal subtypes, then somatic and axonal activity exhibited similar patterns of modulation (Azcorra et al., 2022).

We performed photometry recordings from somatic dopamine activity in VTA, and their axonal projections in NAc. Dopamine somatic and axonal activity had a markedly similar pattern dynamics, although some differences were present. These observed differences could

be explained by the fact that NAc axonal projections were a partial pool from the recorded VTA dopamine neurons, and/or local axonal modulation through cholinergic interneurons (Threlfell and Cragg, 2011; Kramer et al., 2022; Liu et al., 2022; Reynolds et al., 2022).

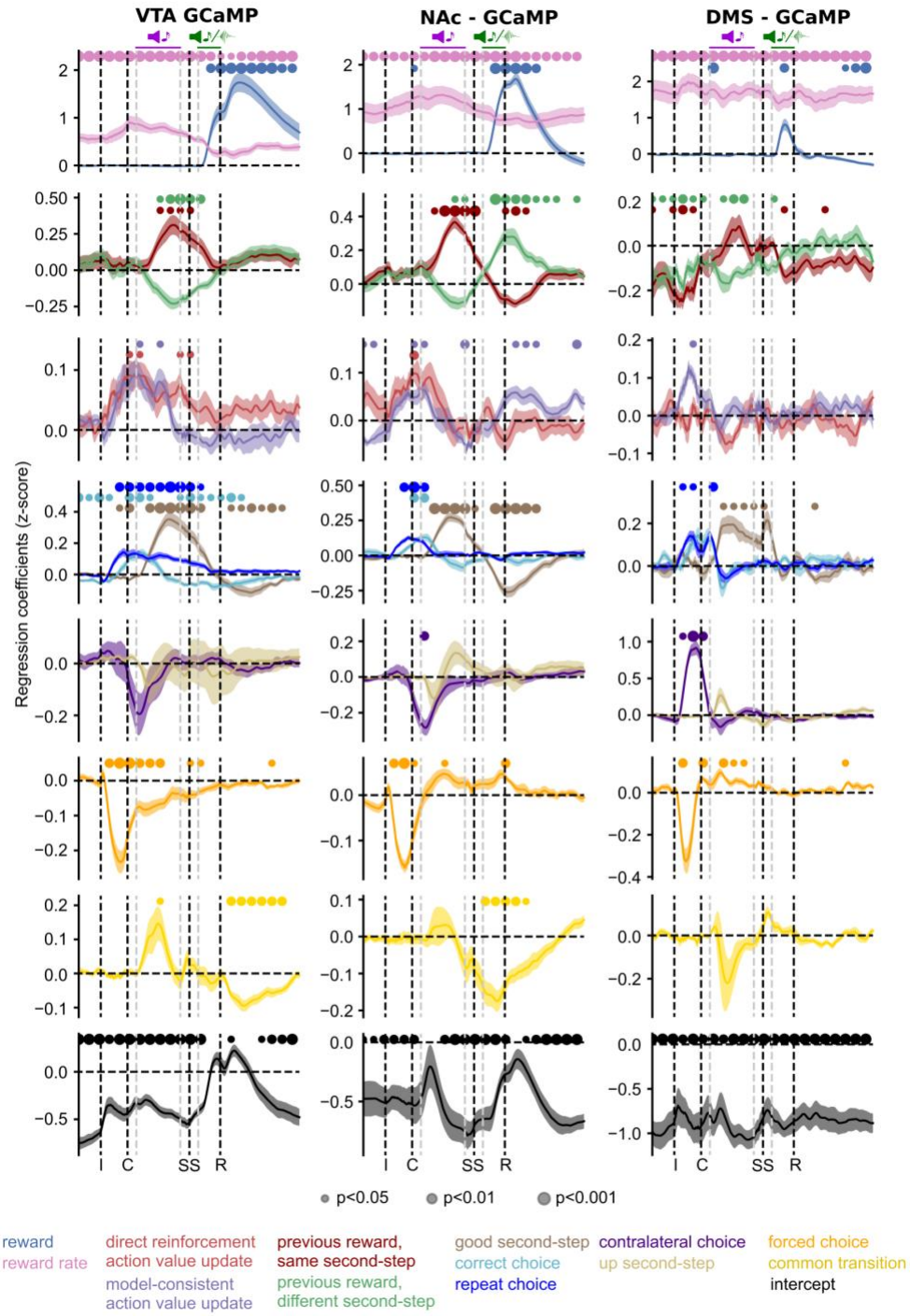
Cholinergic modulation in striatum could also explain the displayed differences between striatal axonal activity and dopamine concentrations in both NAc and DMS. In NAc the main differences were i) the prominent increase of dopamine concentrations from initiation to choice (as previously reported (Mohebi et al., 2019)); ii) lack of contralateral movement coding between choice to second-step port in dopamine concentrations; and iii) stronger negative responding to reward omissions in dopamine concentrations than activity. In DMS, instead, the observed differences were observed in i) the asymmetric response to rewards and omissions, with activity being stronger negatively modulated by omissions; ii) even stronger negative modulation to omission in dopamine concentrations (like in NAc concentrations); iii) RPE responses to contralateral movements only in axonal activity, whereas contralateral and ipsilateral value was reported in concentrations; and iv) opposite modulation by transition type, activity was positively modulated by rare transitions, whereas concentrations were positively modulated by common transitions instead.

Note that although some of these differences could be explained through cholinergic modulation modifying dopamine release, some differences like the different magnitude of signals can also be due to dopamine release and reuptake mechanisms (Liu and Kaeser, 2019; Liu et al., 2021) or differences in the indicators, GCaMP (Chen et al., 2013) and dLight (Patriarchi et al., 2018), dynamics.

4.4.5 Conclusion

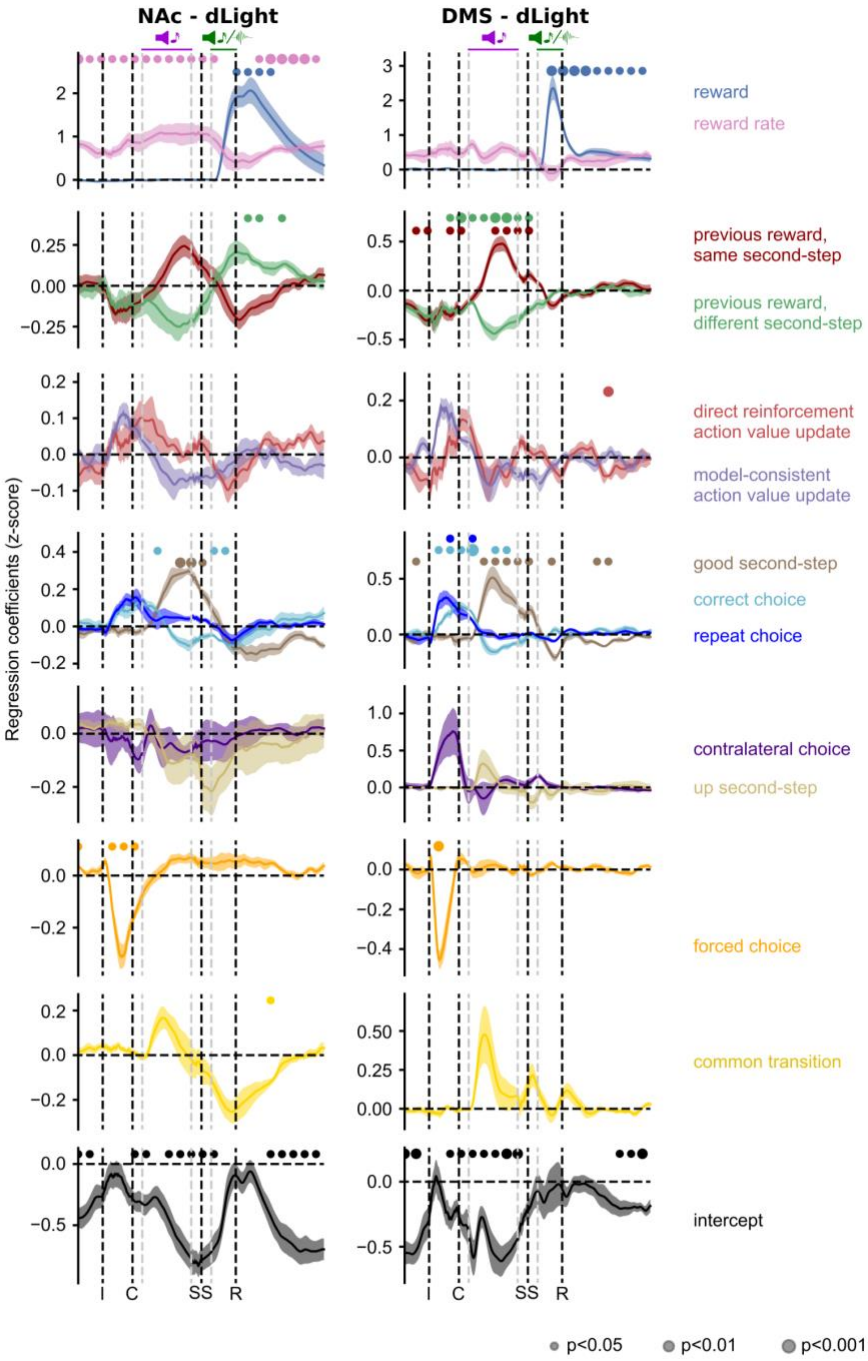
Dopamine reports a rich set of behavioural and task related parameters. At different timescales we observed that dopamine is modulated by RPE, motivation and movement signals. While RPE and a lateralised action sequence signal are correlated with phasic dopamine responses. A sustained signal that lasts the whole trial is correlated with reward rate. To sum up, dopamine signalling multiplexes different types of information.

Supplementary figures



Supplementary Figure 4.1. *Regression coefficients from all the predictors used in the behavioural logistic regression model predicting VTA, NAc and DMS dopamine activity (GCaMP recordings). Shaded area indicates cross subject standard error. Shaded area indicates cross subject standard error. Dots represents statistical significance, two-sided t-test comparing the cross-subject distribution against 0,*

after Benjamini-Hochberg multiple comparison correction, small dot: p < 0.05, medium dot: p < 0.01, large dot: p < 0.001.



Supplementary Figure 4.2*. Regression coefficients from all the predictors used in the behavioural logistic regression model predicting NAc and DMS dopamine receptor binding (dLight recordings). Shaded area indicates cross subject standard error. Dots represents statistical significance, two-sided t-test comparing the cross-subject distribution against 0, after Benjamini-Hochberg multiple comparison correction, small dot: p < 0.05, medium dot: p < 0.01, large dot: p < 0.001.*

5 How does dopamine influence behaviour in complex structured environments?

Dopamine activation has been shown to reinforce behaviour, while the opposite happens if dopamine activity is inhibited instead. Although it is accepted that dopamine activation can substitute natural reinforcers such as water, and can promote similar behaviour, one question arising is whether the effects of natural reinforcers are mediated through dopamine. In chapter 3, we observed that rewards in the two-step task have a big influence on the next trial mice choice behaviour. In chapter 4, VTA dopaminergic neuronal activity was strongly modulated by outcome, rewards and omissions, not only at the time of outcome but also reflecting action and state value updates on the next trials. In the current chapter, we will show that neither VTA dopaminergic neuronal population activation nor inhibition has an influence on next trial choice behaviour.

5.1 Introduction

Dopamine RPE signals are claimed to serve as a teaching signal to guide behaviour (Sutton and Barto, 1998; Steinberg et al., 2013). Causal manipulations of dopamine activity has further supported this conception that dopamine activation is rewarding and reinforces actions that support this increase in dopamine (Hamid et al., 2016; Sharpe et al., 2017; Keiflin et al., 2019). However, in all these previous reports, either action and outcome delivery occurred in the same location, or stimulation happened at action time. Therefore, it is unknown what are the differential effects of dopamine in the different states: action and outcome states. Most critically, during state inference, situations where rewards provide information about the *state* of the world, we do not know if their effect on subsequent behaviour is mediated by the RPEs they evoke (putatively carried by dopamine), or through influences on subject's beliefs.

In chapter 3, we saw that animals' choice behaviour respected the two-step task's structure. Thus, rewards not always reinforced choice (i.e. rewards after rare transition promoted *switching* to the other choice). In chapter 4, we observed correlational evidence that dopamine reports RPEs and reward information. However, this cannot establish whether this signal was causally responsible for updating mice's choice policy. Therefore, here we asked whether the reward effects that we see behaviourally were actually mediated by dopamine. Furthermore, as a consequence of the multi-step nature of the two-step task, in which the action and outcome states are spatially separated, it allows to dissociate the effects of dopamine manipulation at action and outcome time.

To assess this, we manipulated dopamine activity in two cohorts (activation and inhibition group) of mice. The activation group expressed either channelrhodopsin (ChR2, N=7) or a control fluorophore (EYFP, N=5), and the inhibition group expressed an anion-conducting channelrhodopsin (stGtACR2, N=7) or a control opsin (tdTomato, N=5) targeted at VTA dopamine neurons bilaterally. We determined brief stimulation parameters that would either reinforce (activation) or perturb (inhibition) action choices. We then tested the same animals on the two-step task to examine the effects of optogenetic manipulation just following either the first step choice or the second step action when the outcome was signalled. Activation of dopamine neurons at choice reinforced previous actions, while inhibition was able to bias animals in order to avoid inhibition. Importantly, neither dopamine stimulation nor inhibition was able to recapitulate or block, respectively, the effects of natural rewards. These data demonstrate that i) dopamine stimulation does not simply mimic a natural reward, ii) dopamine reinforces actions only when activation is immediately preceded by the action, and iii) during state inference behaviour, dopamine has a limited causal influence on choices through the information in conveys about the state of the environment.

5.2 Methods

We performed two types of optogenetic manipulation: activation and inhibition. For the optogenetic activation 12 DAT-Cre mice (YFP: 2 females and 3 males; ChR2: 4 females and 3 males) were used. Another 12 DAT-Cre mice (tdTomato: 3 females and 2 males; GtACR2: 4 females and 3 males) were used for the optogenetic inhibition experiment.

We optogenetically manipulated dopamine activity while mice were solving the two-step task. For details on the animals, two-step task and surgery, see chapter 2: general methods.

5.2.1 Behavioural task – The two-step task

As in the previous chapters, the task used for the optogenetic manipulation of dopamine neurons was the two-step task in mice. This is a multi-step probabilistic decision-making task. In each trial, a choice between the left and right ports needs to be done. Each choice probabilistically, commonly and rarely, to one of the second-step ports, up and down. The transition structure was the opposite for each choice port (e.g. left choice commonly, while right rarely transitioned to the up second-step port). Each second-step port was probabilistically rewarded with probabilities that changed in blocks.

Animals were pre-trained on the task before surgery was performed. After surgery, mice continued the training for ~1 month, which was the time for the opsin to be expressed in the neurons.

5.2.2 Optogenetic manipulation

VTA dopamine neurons were stimulated bilaterally using two 465 nm LEDs (Plexon Plexbright) connected to 200 µm 0.66 NA optic fiber patch cords. Optogenetic activation stimulations lasted for 200 ms and consisted of 5 x 10-ms pulses at 25 Hz, 25% duty cycle and ~8-10mW optical power at the fibre tip. Optogenetic inhibition stimulations consisted of 1s continuous light at ~5mW optical power at the fibre tip.

5.2.2.1 Intracranial self-stimulation (ICSS) assay

As a positive control, we performed an intracranial self-stimulation (ICSS) assay in which mice were presented with either 4 or 2 nose poke ports, one of which triggered optical stimulation when poked. A minimum 1-second delay was imposed between stimulations. Mice were tested on ICSS on 40-60min sessions for 4 days. This control experiment was performed before commencing with stimulation sessions during the two-step task.

5.2.2.2 2-alternative forced choice task (bias task)

Mice (tdTomato and GtACR2) were presented with 3 ports, a central initiation port, and a left and right ports. Mice initiated each trial by poking the illuminated centre port, which triggered the left and right ports to lit up on free choice trials (50% of the trials were forced choice trials where just one choice port illuminated, and mice had to approach it). When animals poked one of the illuminated choice ports, after a 200ms delay, a 500ms outcome cue was triggered (5- or 12-kHz tone – left or right frequency tone counterbalanced across animals – pulsed with at 10Hz on rewarded trials, white noise on unrewarded trials), after which, reward was delivered on rewarded trials (4ul sucrose). Reward probability of the left and right ports was 50%. After 3 days where animals showed a bias towards one of the choice ports (for at least 2 consecutive days) – preferred choice, 1s continuous light stimulation was introduced at the time of the outcome cue whenever the initially preferred choice port was chosen (on both free- and forced-choice trials). After 4 days, optical stimulation was switched to be paired with the other choice port for an extra 4 more days. Each day animals were tested in a single 60min session. This assay was performed at the end of the experiment, after all the testing sessions on the two-step task.

5.2.2.3 Activation and inhibition in the two-step task

We had three types of sessions: baseline, after choice stim, and outcome stim sessions. During baseline sessions, no stimulation was delivered. On after choice stim sessions, stimulation was delivered 200 ms after the first-step choice, at the time of second-step cue onset. Finally, on outcome stim sessions, stimulation happened at outcome-cue onset. Stimulation occurred on 25% of the trials, under the constraints that (i) the trial after stimulation was always a free choice trial, and (ii) there were always at least 2 non-stimulated trials after each stimulation. The stimulation sessions were interspersed with baseline sessions. Session type was counterbalanced across animals.

5.2.3 Logistic regression model

For the analysis of two-step optogenetic manipulation, we used the same regression model as in the behavioural analysis (see Chapter 2: general methods), but also added stimulation and group, and its interaction with trials events as addition predictors, giving the formula:

$$stay = correct + bias + transition * outcome * stim * group + (random\ effects || subject)$$

The extra stim and group predictors were coded as:

- *Stim*: 3 level categorical variable indicating whether previous trial was non-stimulated, stimulated after choice, or stimulated at outcome time.
- *Group:* Binary categorical variable indicating whether the data is from control or manipulation subject. YFP and ChR2 for the stimulation experiment, and tdtomato and GtACR2 for the inhibition experiment.

We performed a single regression for the stimulation experiment including both experimental groups (YFP & ChR2) and both stimulation times (after choice & outcome). We did the same for the inhibition experiment (tdtomato & GtACR2). To ensure stimulated and non-stimulated trials had matching histories, we only included trials where stimulation could have potentially been delivered, i.e., excluding the 2 trials following each stimulation where stimulation was never delivered.

As a follow-up analysis, we performed separate regressions per group (YFP, ChR2, tdtomato & GtACR2) and stimulation time (after choice & outcome).

5.3 Results

5.3.1 Dopamine stimulation promotes self-stimulation

First, we verified that we were able to activate dopamine neurons by confirming, in line with previous reports, that brief dopamine activation was rewarding and promoted repeating choices (Steinberg et al., 2013; Hamid et al., 2016; Coddington and Dudman, 2018b; de Jong et al., 2019; Keiflin et al., 2019; Pan et al., 2021). To do so, we bilaterally injected animals with channelrhodopsin (ChR2), a light-gated channel – ChR2 group – or a control opsin – YFP group – into VTA, and delivered optical stimulation through two optical fibres placed over the same injection site (Figure 5.1A). These two groups were going to be later tested on the effect of optic stimulation during the two-step task, but first, they were exposed to an ICSS assay using the same stimulation parameters.

Animals were presented with either 2 or 4 ports (Figure 5.1B), in a different configuration from the two-step task. Only one of those ports delivered stimulation. Our brief stimulation parameters (25Hz, 5 pulses) were sufficient to promote and maintain intracranial self-stimulation – repeating poking the port that delivered optical stimulation contingent on nose-poking – in the ChR2 but not YFP groups (Two-sided t(10) = -3.107, p = 0.011, 95% CI [-414.07, -68.18], Cohen's D = 1.819) (Figure 5.1C).

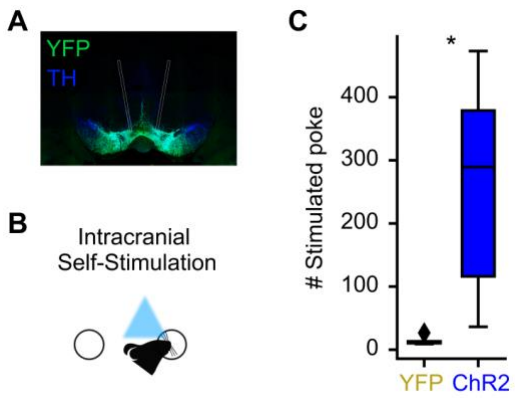


Figure 5.1. *Activation of dopamine neurons is rewarding. A) Histology of a ChR2 mouse stained for TH (blue) and YFP (green). Fibres were placed at 10 degrees into VTA. B) Schematic of the intracranial self-stimulation assay. C) Mean number of entries to a nose-poke delivering optical stimulation in the YFP vs the ChR2 group. Error bars show cross-subject standard error. *p<0.05 two-sided t-test.*

The results on the ICSS assay confirmed that our viral approach and stimulation parameters were able to activate dopamine neurons and that this activation promoted repeating actions, consistent with stimulation working as a positive RPE. This was our first positive control confirming the suitability of our approach to activate the dopamine system.

5.3.2 Dopamine inhibition promotes avoidance behaviour

In another cohort of animals, we tested whether we were also able to inhibit dopamine neuronal activity. This other group of animals were also tested in the two-step task. We injected bilaterally in VTA either and anion-conducting channhelrhodopsin – GtACR2 group –

or a control virus – tdTomato group. Inhibition was induced through optical stimulation using two optical fibres placed at the same injection sites (Figure 5.2A).

To demonstrate that we could effectively inhibit dopamine activity, we designated a task to test switching or avoidance behaviour on the stimulated site – behaviour expected as a result of dopamine inhibition (Hamid et al., 2016; Parker et al., 2016). Animals were presented with 3 ports, an initiation central port and 2 choice ports on the left and right side. Animals initiated a trial by poking the illuminated centre port, which triggered the choice state where animals had to choose between the left or right port on free-choice trials. Both left and right ports delivered water reward with 50% probability (Figure 5.2B). In this task there was not a correct action, as both choices had the same reward probability. Animals learned the reward contingency, which resulted on a bias behaviour – animals preferentially chose one side, e.g. mainly chose left. After 3 consecutive sessions where animals showed the same preferred side (>50% of choices were to one side), stimulation on the preferred side was introduced, contingent on nose-poking it. Optical stimulation drastically changed GtACR2 group behaviour towards the non-preferred side, avoiding the light stimulation, and thus avoiding dopamine inhibition. The tdTomato group did not show any switch of their choice preference. For 4 consecutive sessions, the GtACR2 switch their bias towards the previously non-preferred side, avoiding the preferred that was now stimulated. Afterwards, optical stimulation was reversed to the other side (the current preferred side, initially non-preferred one). Again, this drastically switched GtACR2 animals towards the non-stimulated side, avoiding dopamine inhibition (Figure 5.2C).

This validated that our stimulation parameters were sufficient to cause animals to switch their choice policies to avoid the side where they received stimulation. This was our positive control corroborating that our optical stimulation successfully inhibited dopamine activity.

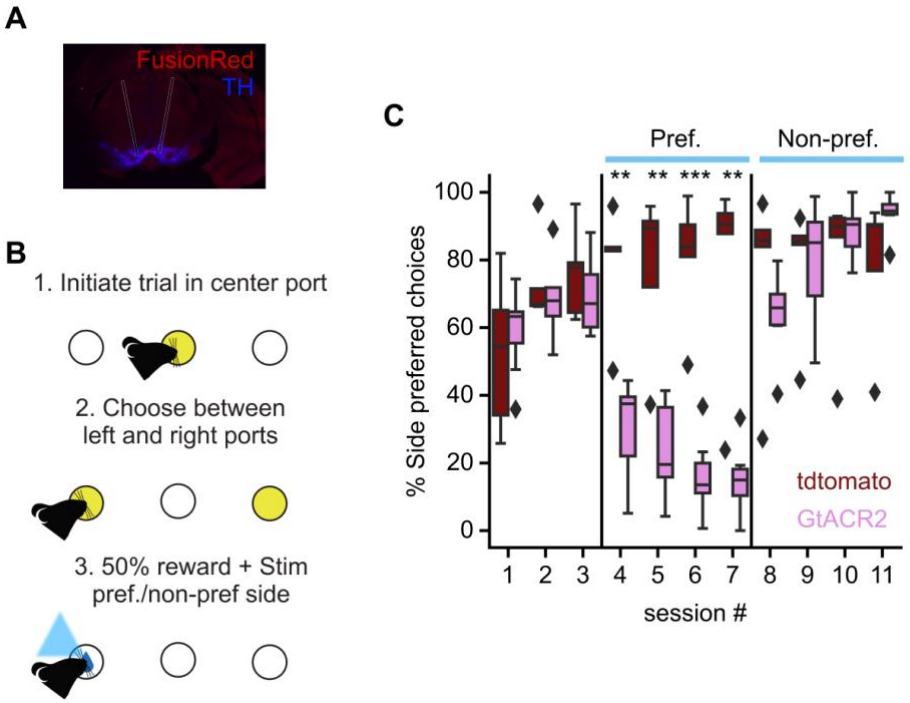


*Figure 5.2. Inhibiting dopamine activity promotes switching behaviour. A) Histology of a GtACR2 mouse stained for TH (blue) and FusionRed (red). Fibres were placed at 10 degrees into VTA. B) Schematic of the bias switching assay. C) Mean across subjects of the percentage of choices to the initial preferred side in the tdTomato (red) and GtACR2 (pink) groups. From session 4 to 7, optical stimulation was delivered when poking their initial preferred choice (preferred side in sessions 1-3 when stimulation was not delivered). From session 8 to 11 the side of optical stimulation was reversed, now to their initial non-preferred side. Error bars show cross-subject standard error. *p<0.05, **p<0.01, ***p<0.005, two-sided t-test with Bonferroni multiple comparison correction.*

5.3.3 Dopamine stimulation does not recapitulate the effects of natural rewards despite reinforcing preceding actions

Once we confirmed we can reproduce previous reports (Steinberg et al., 2013; Hamid et al., 2016; Coddington and Dudman, 2018b; de Jong et al., 2019; Keiflin et al., 2019; Pan et al., 2021) showing that dopamine activation promotes repeating actions, we then examined the effect of activating dopamine in the two-step task in the same animals. As mentioned before, the multi-step nature of the task, allows to dissociate the effects of dopamine stimulation at action and outcome time.

In the two-step task, we optically stimulated dopamine neurons at two different timepoints in the task – (i) after the first-step choice, at the time of second-step cue onset, or (ii) at outcome cue presentation (Figure 5.3A) (25% stimulated trials, stimulation timepoint fixed for each session, counterbalanced across sessions and mice, N=6-8 sessions per mouse and stimulation type).

Using a logistic mixed effects model, which included all data from both groups (YFP and ChR2) and 3 stimulation types (i.e. non-stimulated trials, stimulation after first-step choice, and stimulation at outcome), we found a significant stimulation type by group effect ($p = 0.038$). To further explore this effect, we performed separate logistic mixed effect regressions for each group and each stimulation type (after first-step choice, and stimulation at outcome).

In line with the ICSS results, in the ChR2 group, stimulating dopamine neurons after the first-step choice was reinforcing; it significantly increased the probability of repeating that choice on the next trial ($p=0.0082$, mixed effect logistic regression, Figure 5.3B). This stimulation did

not significantly interact with either the transition or outcome in its effect on next trial choice (all p>0.069). This was our second positive control that our stimulation was rewarding.

Stimulating dopamine neurons at the time of trial outcome – where we observed large increases and decreases in dopamine following a reward or omission respectively – allows us to ask whether the effects of natural rewards are mediated by dopamine. If stimulation just reinforces previous action, we expect a positive loading on the stimulation regressor. If stimulation recapitulates the effect of natural rewards, we expect positive loading in the transition by stimulation interaction (as stimulation would have the effect of reward, like the positive loading of the transition by outcome interaction in the natural behaviour – Chapter 3, Figure 3.2E). Finally, if stimulation interacts not just with the transition but also the outcome that occurred, then we expect to see a positive loading on the 3-way interaction.

Strikingly, stimulating dopamine neurons at the time of trial outcome had no significant influence on the subsequent choice (Figure 5.3E); it did not reinforce the preceding first-step choice (effect of stimulation: p=0.191), nor did it act like a reward by interacting with the state transition (p=0.561), nor with the trial outcome to influence on the subsequent choice (p=0.433), nor the two interacting (i.e. outcome interacting with transition to influence subsequent choice: p=0.560). No effect of stimulation was found in the YFP group for either of the stimulation types (all p>0.42). To further confirm the null result of the effect of stimulation at outcome time on the ChR2 group, we calculated Bayes factor, which demonstrates that the manipulation results provide “strong evidence” (using the classification proposed by Jeffreys (1961)) against dopamine stimulation recapitulating the behavioural consequences of rewards in this task.

By contrast, while stimulation after the first step choice had no effect on the latency to initiate the next trial (Figure 5.3C, Two-sided paired t(6) = 0.347, p = 0.740, 95% CI [-30.25, 40.25], Cohen's D = 0.034), stimulation at outcome time significantly reduced the latency to initiate the next trial (Figure 5F, Two-sided paired t(6) = 4.228, p = 0.0055, 95% CI [20.98, 78.6], Cohen's D = 0.369). Again, there was no effect on this latency in the YFP group (Figure 5.3F, Two-sided paired t(4) = 0.816, p = 0.460, 95% CI [-62.7, 114.9], Cohen's D = 0.302, stimulation after choice; Figure 5.3F, Two-sided paired t(4) = 1.713, p = 0.162, 95% CI [-9.87, 41.67], Cohen's D = 0.141, stimulation at outcome time).

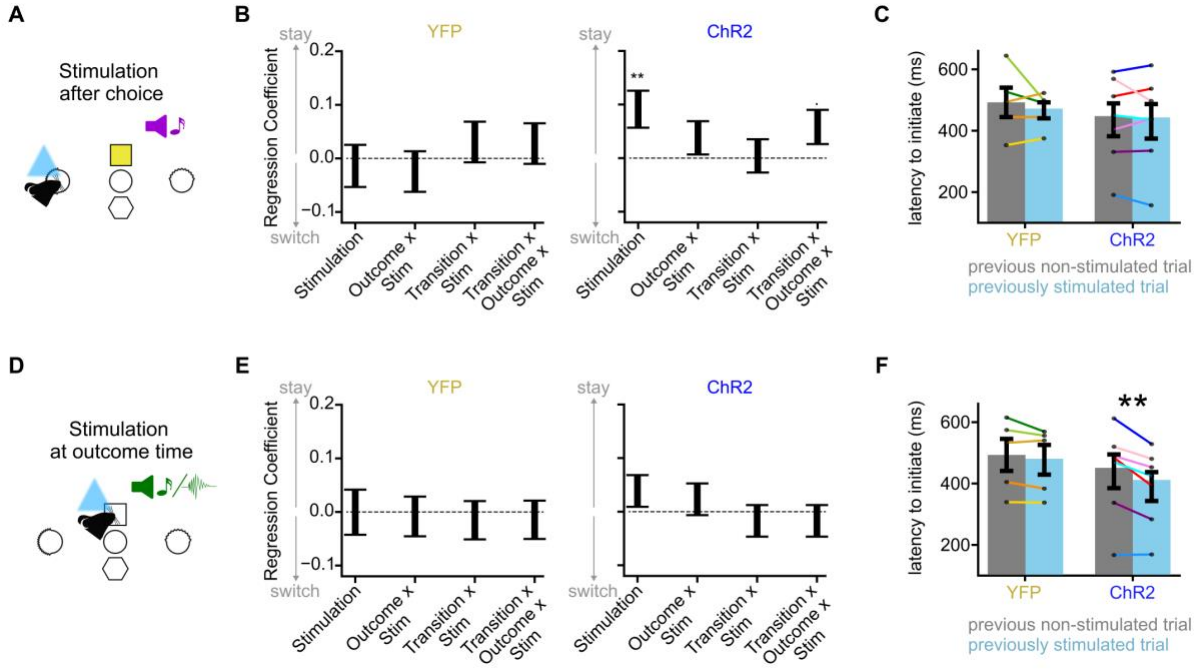


*Figure 5.3. Dopamine stimulation does not recapitulate natural reward effects. A) Schematic of optogenetic stimulation 200ms after choice. B) Mixed effects logistic regression predicting repeating the same choice on the next trial based on whether previous trial delivered stimulation and its interaction with trial events on sessions when stimulation happened after choice. Left, YFP group; right, ChR2 group. •p<0.1, **p<0.01. C) Mean latency to initate following stimulated and non-stimulated trials on sessions when stimulation happened after choice. Dots indicate individual animals, error bars show standard error. D, E, F) as A, B, and C, respectively, for sessions where*

*stimulation was delivered at outcome time. **p<0.01, two-sided paired t-test. N: YFP: 5 animals, 13,079 trials (choice-time stimulation sessions), 15,419 trials (outcome-time stimulation sessions); ChR2: 7 animals, 20,817 trials (choice-time stimulation sessions), 23,051 trials (outcome-time stimulation sessions).*

Therefore, while stimulation of VTA dopamine neurons was sufficient to reinforce immediately preceding actions, it completely failed to recapitulate the behavioural consequences of natural rewards at outcome, despite rewards driving the largest dopamine signals observed in the task.

5.3.4 Dopamine inhibition neither promotes a change in policy

To further corroborate that the reward effects observed in the behaviour are not dopamine dependent, we repeated the previous experiment, but instead of activating dopamine neurons, we optogenetically inhibited them. We used the same animals as in section 5.3.2, in which we demonstrated our optical stimulation parameters were able to promote avoidance behaviour, consistent with a negative RPE.

We again first performed a logistic mixed effects model, which included all data from both groups (tdTomato and GtACR2) and the 3 stimulation types (i.e. non-stimulated trials, stimulation after first-step choice, and stimulation at outcome). No significant effect of stimulation type by group, nor interacting with other trial events (i.e. transition and outcome) were observed (all $p > 0.283$). Confirming the null result, Bayes Factor calculation indicated "extreme evidence" against dopamine inhibition blocking the effects of rewards and omissions.

Likewise, neither stimulation type had an effect on the latency to initiate the next trial on the GtACR2 group (stimulation after choice: Two-sided paired t(6) = -0.835, p = 0.436, 95% CI [-128.07, 62.92], Cohen's D = 0.073; stimulation at outcome: Two-sided paired t(6) = -0.695, p = 0.513, 95% CI [-162.1, 90.38], Cohen's D = 0.086), nor the tdTomato group (stimulation after choice: Two-sided paired t(4) = 0.96544, p = 0.389, 95% CI [-67.91, 140.31], Cohen's D = 0.1268; stimulation at outcome: Two-sided paired t(4) = -1.074, p = 0.343, 95% CI [-108.24, 47.84], Cohen's D = 0.1104).

5.4 Discussion

We manipulated dopamine activity in the two-step task where rewards do not simply reinforce preceding actions, but rather affect subsequent choices in a manner that respects the task structure. To ask whether the rewards effects were causally mediated by dopamine activity, we focused on disrupting – either activating or inhibiting – signalling at outcome time when reward and omissions are delivered.

While rewards strongly influenced subsequent choices, stimulating dopamine neurons at the time of trial outcome had no effect on animal's next choice. This was despite positive controls verifying that, in the same subjects, the same stimulation was sufficient to i) reinforce a choice when delivered shortly after it, ii) induce robust self-stimulation in a different context, and iii) speed subsequent trial re-initiation when stimulating at outcome. Dopamine inhibition also failed to block the effects of rewards and omissions.

5.4.1 Dopamine is reinforcing but does not substitute the effects of natural rewards

It is assumed that dopamine signals carry information about rewards in order to maintain, update and guide future behaviour (Schultz et al., 2003; Steinberg et al., 2013). However, our findings do not align with this canonical idea. All the previous reports could not dissociate the effect of dopamine modulation at action and outcome time, and therefore, it was assumed that rewards guide future behaviour through dopamine signalling. However, using the two-step task, we have shown for the first time the time- or state-dependent effect of dopamine signalling on action selection.

We have replicated the previous studies results that dopamine stimulation is rewarding and promotes repeating choices through i) the ICSS assay, and ii) optogenetically activating dopamine neurons just after the choice is made in the two-step task. While this is in agreement with the canonical view of dopamine, we showed that stimulating dopamine at outcome did not influence next choice, and neither blocking dopamine.

The lack of effect of the optogenetic manipulation of dopamine activity confronts the established notion of RPE signals in guiding policy in the RL framework (Sutton and Barto, 1998). If animals were model-free, we would expect that dopamine activation at outcome time would reinforce the action taken, as rewards would do. But that was not the case. If instead animals were model-based, and dopamine activation at outcome time increases the value of that second-step state – as we have shown rewards do –, we would expect this would influence first-step choice respecting the task's transition structure. But again, this was not the case.

However, we have seen in Chapter 3 that behaviour in our two-step task is consistent with a state inference account. Therefore, rewards on each trial not only provides state value information but constitute the *observable event*, which can be used to *infer* the current state of the task – i.e., 'is the up port currently highly rewarded or is the down port currently highly rewarded?'. Therefore, during such state inference, rewards are not used to update the *value* associated with each state or action, but instead, they are used to track the *latent* or *unobservable feature* of the task – here, the anticorrelated structure of the reward probabilities. Crucially, that means that rewards provide information on each trial to update subjects' *belief* of the state of the task.

Our data show that although RPE signals conveyed through dopamine activity update the values of states and actions, these are not conveying the state prediction error used to update the state belief (Akam and Walton, 2021). Therefore, when the model of the environment is already acquired, rewards have a dopamine-independent influence on subsequent inference-based behaviour.

This does not mean that dopamine does not have any role during state inference behaviour. Dopamine activation at outcome time reduced the latency to initiate next trial. This is in line with dopamine having a motivational role during well-known environments. Other studies using dopaminergic pharmacological manipulations, despite showing limited or no effect on learning, they showed an effect in task engagement and choice consistency (Lee et al., 2015; Verharen et al., 2019; Korn et al., 2021).

This opens up many new questions to be addressed in the future. First, how does the causal influence of dopamine on subsequent behaviour evolve from naïve to expert behaviour? One possibility is that during initial learning, when the model of the environment is not known, that dopamine has a major role in endowing state and actions with value, based on observable events in the environment. However, with time, this value learning over observable events can be transferred to learning from inferred or latent states. Therefore, from naïve to expert behaviour there might be a change from value (i.e. standard RL, continuously learning independent values) to inference learning. Another question is whether this dopamine independence is specific of inference-based behaviour or is that also true for other model-consistent behaviour, or even model-free? And if dopamine is not involved, what other circuitry is recruited for learning? One possibility could be that cortical regions learn and represent the latent structure of the environment (Hampton et al., 2006; Bradfield et al., 2015; Bartolo and Averbeck, 2020; Vertechi et al., 2020) which influence dopamine dynamics (Starkweather et al., 2018). Meta-RL has also been suggested to support reward-guided behaviour implemented in prefrontal cortex and trained by dopamine signalling (Wang et al., 2018).

5.4.2 Optogenetic manipulation limitations

Recently, it has been suggested that ICSS behaviour could not be achieved through activation of dopamine neurons at physiological frequencies (Millard et al., 2022). However, this is not the case in our experiment, as the stimulation parameters used were previously shown to produce physiological phasic dopamine activation (Coddington and Dudman, 2018b). However, as evidenced by our data, the timing of dopamine signalling is determinant on its effect on behaviour, therefore, future simultaneous optogenetic manipulation and recording

of dopamine neurons will further clarify the dopamine dynamics resulting from the optical stimulation, and whether that would explain different effects on behaviour, or whether specific projections have a different role in learning (Pan et al., 2021).

Second, although we confirmed the lack of effect on subsequent behaviour from dopaminergic stimulation by also obtaining a null effect of inhibiting dopamine at outcome, we also did not observe any change in behaviour compared to the control group after inhibiting dopamine at choice or a change in response latencies on the next trial after inhibiting at outcome. This is not simply as the transfection was ineffective as the bias switching assay demonstrated strong avoidance of the optically stimulated choice in the GtACR2 group. Nonetheless, one possibility is that the effects of inhibition on the choice task may be complicated by "rebound effects", where inhibited dopamine neurons can transiently increase their activity coincident with light stimulation offset. When we optogenetically inhibited dopamine activity after choice, the rebound excitation of dopamine activity would happen exactly at the time of second-step state. This could have masked the choice avoidance behaviour when inhibiting after choice in the two-step task. However, this was just a further positive control, and therefore, does not affect our conclusions on the dopamine-independent effect of rewards on subsequent choices during inference-based behaviour.

5.4.3 Conclusion

Although dopamine multiplexes different sets of behavioural and environmental variables, it has a limited causal role on animals' behaviour. Neither activating nor inhibiting dopamine produced any change in animals' subsequent choice. However, dopamine activation decreased the latency to initiate the next trial, consistent with a motivational role.

6 Discussion

An adaptation of the two-step task for mice has allowed us to study i) the psychological processes underlying inference-guided flexible behaviour under the RL framework, ii) the influence of the behavioural strategy together with the environment latent states and its structural complexity on dopaminergic activity across midbrain and striatal targets, and dopamine concentrations, and iii) the causal impact (or lack of it) in such an environment. In this chapter, we will summarize and discuss the main results, and suggest potential future directions.

6.1 Summary

The main goal of this thesis was to study the behavioural and dopaminergic underpinnings of flexible behaviour in a real-world-like scenario where inference guides behaviour. The external environment is complex and not fully observable, we do not know all the rules that control its dynamics. However, we can infer them using available information in it. In chapter 3 we put the example of rush hour, in which the best path depends on the time of the day. Another example is the cyclical change between weekdays and weekends. This constitutes a latent or non-observable feature of the environment which can be inferred through observable events in the surrounding world. The state of the environment, whether is weekend or not, has an impact on our behaviour.

We reproduced this type of partially observable environment in which hidden latent variables can be learned to infer the state of the world. To do so, we adapted the two-step task for mice. We examined possible RL models to explain subjects' choice behaviour. Mice's behaviour was consistent with inference RL respecting the task causal and correlation structure. We then recorded somatic and axonal calcium activity in VTA and midbrain dopaminergic projections to striatal regions (NAc and DMS), respectively, using fibre photometry. We also recorded dopamine concentrations in the same striatal regions. Both dopaminergic activity and concentration dynamics in all regions included information about the hidden latent variable and the structural complexity in the task for state and action value updating. Simultaneously, dopamine also carried motivational and movement-related signals. But despite dopamine richly reporting task variables and computations conditioned on the intrinsic structure of the task, dopamine manipulation had no effect on animals' next trial

choice. To sum up, when inference guides behaviour, dopamine has an asymmetric impact: it multiplexes a wide range of environmental- and policy-related information but has a limited influence on the actual behavioural policy.

6.2 Can inference accounts improve the understanding of flexible behaviour?

During the 20th century, associative versus cognitive accounts were the main focus of discussions to explain animals' instrumental behaviour (Thorndike, 1911; Hull, 1936; Tolman, 1948). Reconciling accounts postulated that there could be different kinds of learning, and single-theories turned into dual-system accounts (Tolman, 1949; Trapold and Overmier, 1972; Dickinson, 1985). Currently, one prominent strand of thinking is that instrumental behaviour is controlled by 2 systems, habitual and goal-directed processes. At the same time, instrumental behaviour can be controlled through a Pavlovian motivational drive (Dickinson and Pérez, 2018).

To formalise the computational principles underlying instrumental behaviour, the RL framework was used to explain reward-guided behaviour. This framework was also postulated to describe the neuronal dynamics in the basal ganglia (Doya, 1999). Model-free and model-based RL controllers were used to capture features of both habits and goal-directed behaviour, respectively, which were suggested to work jointly to control behaviour. Prefrontal regions were simulated to perform model-based computations, while the dorsolateral striatal circuitry would constitute the model-free controller (Daw et al., 2005).

To study these two processes model-free and model-based behaviour and its neuronal implementations, Daw *et al.* (2011) developed a multistep probabilistic task, the two-step task, which allowed to dissociate model-free from model-based behaviour. Therefore, this task allowed to study these two processes simultaneously. Another main advantage of this task is the large amount of trials on single sessions and over many days in which both processes could be studied in a trial-by-trial basis. This made the two-step task ideal for neurophysiological recordings.

Since then, the two-step task has been widely and extensively adapted, not just for humans (Wunderlich et al., 2012b; Otto et al., 2013b; Smittenaar et al., 2013; Sebold et al., 2014; Voon et al., 2014; Cushman and Morris, 2015; Deserno et al., 2015; Doll et al., 2015a; Morris and Cushman, 2019; Wyckmans et al., 2019), but also for rats (Miller et al., 2017; Hasz and Redish, 2018; Groman et al., 2019), and mice (Akam et al., 2021b; Korn et al., 2021). However, computational and modelling studies (Hampton et al., 2006, 2007; Behrens et al., 2007; Nassar et al., 2010, 2012; Zylberberg et al., 2018) on the two-step task and its versions have shown that the model-free – model-based distinctions might not be so clear – i.e., model-free RL strategies can generate choice behaviour that resembles model-based (Akam et al., 2015; Kool et al., 2016; Feher da Silva and Hare, 2018).

Model-free and model-based RL distinction might not be sufficient to explain behaviour in the two-step task, and more generally, might not be sufficient to describe instrumental behaviour in a natural environment. The two-step task, like in real-world situations, is not fully observable, the reward probabilities of the different options dynamically change without being explicitly told or knowing the hidden variable controlling this change. In such situations,

it has been shown, that humans (Hampton et al., 2006, 2007; Behrens et al., 2007; Nassar et al., 2010, 2012; Zylberberg et al., 2018) and animals (Gallistel et al., 2001; Gershman et al., 2010; Costa et al., 2015; Bartolo and Averbeck, 2020; Vertechi et al., 2020) can use Bayesian inference to guide their behaviour. Akam *et al.* (2015) modelling the two-step task and a reduced version of it (which we used in this thesis) exposed that inference-based RL models could also generate choice behaviour that looked like the traditional model-based. This computational study further showed that the reduced version was equally, or even better, in distinguishing the different possible RL strategies.

In this thesis, we trained animals in the reduced version of the two-step task, described by Akam *et al.* (2015). In this reduced version, the different transition probabilities from action to second-step state were maintained (although with increased contrast between common and rare transition – instead of 0.7/0.3 were 0.8/0.2 for common/rare transitions). No choice was present in the second-step state. Instead of dynamically changing through gaussian random walks, the reward probabilities at the second-step states changed in blocks in an anticorrelated fashion (Figure 6.1). The anticorrelated changes in the reward probability of the second-step state introduced a clear latent variable that could be learned to infer the state of the task – is the task in the state where *state a* (up port in our task) is good, or in the state where *state b* (down port) is good.

In chapter 3, we examined different RL strategies mice could be using to solve the two-step task. We confirmed that inference-guided RL strategies better explained mice's choice behaviour than the traditional TD RL models. These inference-guided models updated action values respecting the transition structure of the task (i.e. second-step state values were either

both updated in an anticorrelated fashion through TD error or using Bayesian state inference, and action values were updated using the updated second-step values together with the task's transition structure).

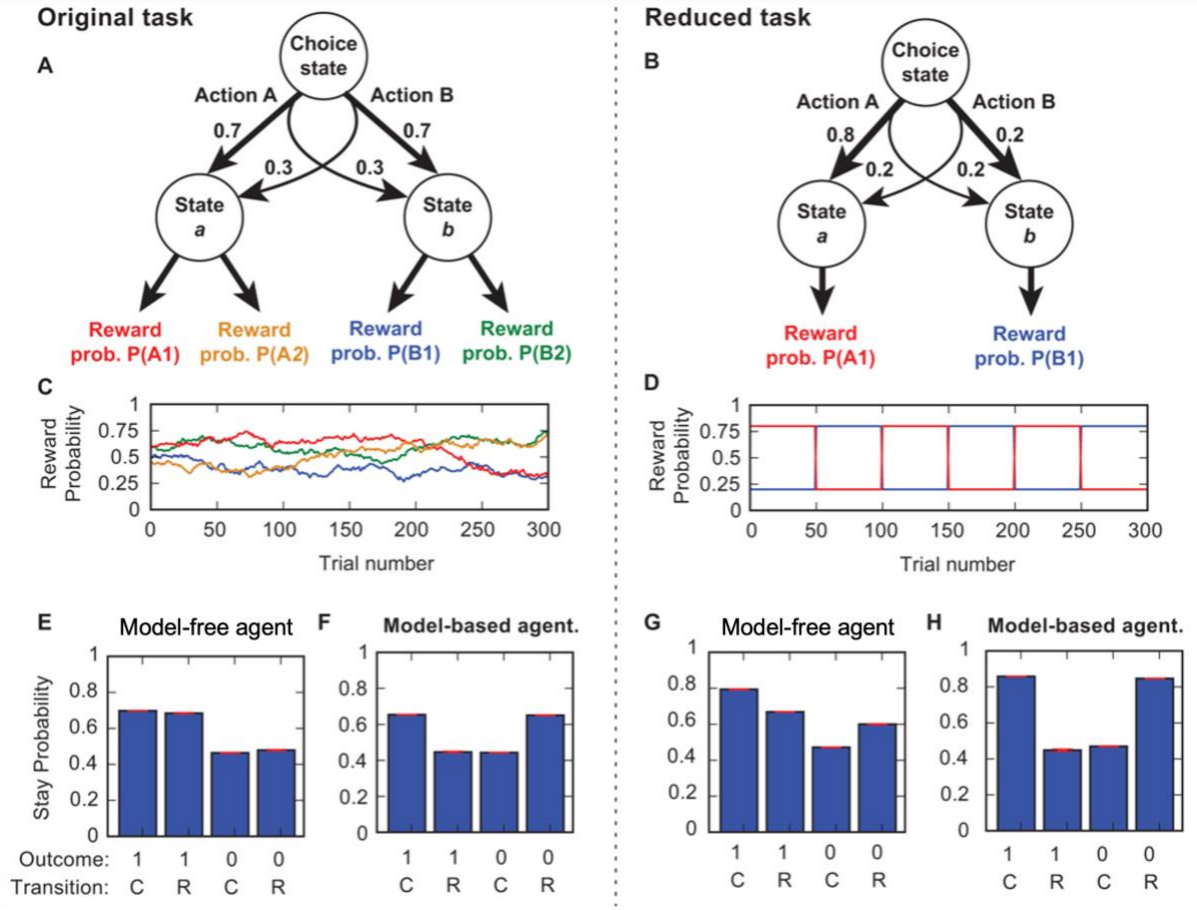


Figure 6.1. *Comparison between the original two-step task from Daw et al. (2021) and the reduced task. A, C, E, F) Original task. B, D, G, H) Reduced task. A, B) Task structure. C, D) Example of reward probability of each second-step state across a session. C) Reward probability changes following gaussian random walks. D) Reward probability changes in blocks in an anticorrelated way. E – H) Stay probability based on the transition (C, common; R, rare) and the outcome (1, rewarded; 0, unrewarded) experienced in the previous trial. E, G) Simulated behaviour from a model-free strategy. F, H) Simulated behaviour from a model-based strategy. Figure modified from Akam et al. (2015).*

Not only behaviour was conditioned on the single latent variable – the anticorrelation structure of the reward probabilities – but dopaminergic neuronal activity and concentration dynamics also reflected computations conditioned on this hidden variable. Dopamine reflected both experienced and inferred value update. Action value updates were also consistent with state inference and including an internal model of the task structure.

But our modelling also reflected that including multi-trial perseveration as an extra component improved all RL models. In our flexible model, which included model-free, model-based, and inference-based strategies, if multi-trial perseveration is included, the model-free component had a minimal influence on choice behaviour. Therefore, multi-trial perseveration captured the effects of a model-free behaviour. Dopamine dynamics also reflected the presence of these two types of signals, model-free and model-consistent inference components.

This is in line with previous reports showing that hybrid approaches are better at fitting human and animal behaviour (Daw et al., 2011; Hasz and Redish, 2018; Groman et al., 2019) and striatal activity (Daw et al., 2011).

But here, we suggest that the learning rule to update values is conditioned on the animal's *belief state*, the inferred state given the trial history. Choice behaviour and dopamine modulation in all the regions and concentrations support this conception. An inference account also can explain the null effects of optogenetic activation and inhibition on the next trial choice, if we consider that the inference about the current state of the world is not

mediated by dopamine, but dopamine is just evaluating states and actions based on the *belief state* (this will be further discussed in the next section).

Importantly, inference accounts that include belief state not only explain tasks with partial observability but can also explain behaviour and dopamine activity in fully observable tasks. Nakahara *et al.* (2004) showed that dopamine RPEs were conditioned on the 'context' of the task (or belief state) – i.e., which of the 4 'contexts' the task was in. This model could also explain behaviour in a classical conditioning task if one considers there is just a single 'context' or state. Similarly, Starkweather *et al.* (2017) further demonstrated that a TD model with belief state could explain dopamine RPEs both in a deterministic (fully observable) and in a probabilistic Pavlovian task. State inference can also explain animals' foraging behaviour (Vertechi et al., 2020).

In conclusion, our data together with other reports suggest that including state inference in RL models can potentially improve our computational description of Pavlovian and instrumental behaviour in a wide range of tasks, and therefore, a wide range of real-world behaviours, together with the neuronal implementations of such behaviours.

6.3 Rethinking the role of dopaminergic RPE in policy update

In our two-step task, rewards had a big influence on mice's next choice behaviour (chapter 3, Figure 3.2D-G) conditioned on the experienced transition (common or rare) – mice repeated actions that were rewarded after a common transition, while switched even if it was rewarded after a rare transition. Rewards (and omissions) also had a big influence on

dopamine modulation across all regions. Reward delivery (or the auditory outcome cue) drove big positive dopamine fluctuations, while negative phasic responses were produced in reward omissions. At the same time, outcome influenced dopamine responses on the next trial reflecting action and state value update conditioned on an internal model of the task structure. Dopamine dynamics in all regions, though most characteristically in NAc, were consistent with RPE signal.

Importantly, in our task, the transition structure means that rewards do not always promote repeating action. This allowed us to test if and how this effect of rewards that we see in the two-step task is mediated by dopamine. To do this, we optogenetically activated or inhibited VTA dopamine activity at outcome cue onset. Neither of the two manipulations had an influence on next trial's choice: it neither promoted nor inhibited choice. Dopamine manipulation neither had an effect alone nor interacting with transition as actual rewards do. This was despite control experiments (in different tasks) showing that our activation protocol promoted repeating actions, and inhibition promoted avoiding actions. Therefore, these results go against the canonical view in RL that RPE is used as a teaching signal to update state and action values and thus shape future behavioural policy.

It was previously shown that optogenetic activation of dopamine neurons can drive conditioned place preference (Tsai et al., 2009) and reinforce instrumental behaviour (Adamantidis et al., 2011; Hamid et al., 2016). This optogenetic phasic dopamine activations could substitute positive RPEs to promote repeating actions (Steinberg et al., 2013), while brief pauses could mimic the negative RPEs (Chang et al., 2016) and promote avoidance behaviour (Hamid et al., 2016; Parker et al., 2016).

We were able to replicate these previous reports' results in our optogenetic control experiments. Optogenetic activation of VTA dopamine neurons generated self-stimulation behaviour, consistent with dopaminergic activation being reinforcing. Furthermore, optogenetic activation after choice in the two-step task promoted repeating that same choice. Optogenetic inhibition in our control experiment instead promoted avoidance behaviour.

All this indicates that dopamine activation reinforces actions but does not have the same effects as natural rewards. Therefore, rewards' effects in updating or changing behaviour in the two-step task are not mediated by dopamine. Our dopaminergic recordings and manipulation experiments reveal a striking disconnect between value reporting and policy updating. How can we reconcile these results with RL framework and RPE?

This can be reconciled by recognising that the task is partially observable. As we have described earlier, both behaviour and dopamine modulation are consistent with state inference. In the two-step task, not all the features of the task can be directly observed but can influence observable features in the task. Rewards on each trial were observable features in the task which were controlled by the underlying structure of the reward probabilities at the second-step states which changed in blocks in an anticorrelated fashion. The anticorrelated structure of the reward probabilities constituted the latent or unobservable feature of the task, that could be used to infer the current state of the task, belief state. Crucially, this means that rewards on each trial, not only carry information about state and action value, but also provide information about the current state of the task (i.e. rewards and omissions on each trial can be used to infer whether the task is in the state where the up

second-step port is highly rewarded, or the down port instead). State information provided by task outcomes (rewards and omissions) can influence future choices by changing subjects' beliefs about the state of the task.

Consequently, our results in the optogenetic dopaminergic manipulation are consistent under the assumption that dopamine neurons carry RPEs evoked by rewards, used to update the values of states and actions, but not the state prediction errors used to update beliefs (Akam and Walton, 2021).

Previous studies' results support this interpretation. Using probabilistic reward-guided tasks, it was found that optogenetic inhibition of dopamine neurons in single trials biased animals against the chosen option in the subsequent trial (Hamid et al., 2016; Parker et al., 2016). These results were interpreted as evidence that task behaviour was dopamine-dependent, and that dopamine inhibition acted as negative RPE biasing behaviour. But in both cases, the reinforcing effect of rewards was still present even after inhibiting dopamine neurons – rewarded trials with or without inhibition promoted repeating behaviour. While this is inconsistent with dopaminergic RPEs mediating reward effects, it is consistent with the inference-based account. Similarly, pharmacological manipulation of dopamine signalling in reward-guided decision-making tasks often has limited or no effect on learning (Costa et al., 2015; Verharen et al., 2019; Korn et al., 2021). All these further suggest a dopamine-independent influence of rewards on policy.

Supporting the dopamine-independent influence of dopaminergic RPEs in policy update, Bayer & Glimcher (2005) also found a mismatch between dopaminergic RPEs and behavioural

policy with experience. They recorded from dopamine neurons in SNc while non-human primates were performing a saccade timing task. In such task, subjects needed to decide the latency to perform a saccade to get reward. The task had 2 different blocks (or states) which reversed unsignalled to the subject. In each block, there was a specific time window in which a saccade would result in reward, and in each time window, there were different time intervals with increasing reward size (Bayer and Glimcher, 2005). Similarly, to the two-step task, this task was partially observable which required subjects to infer the state of the task to guide their behaviour. They found that dopaminergic neuronal firing modulation was consistent with conveying an RPE signal. This RPE coding appeared to be correlated with saccade reaction time. However, when monkeys were fully trained on the task and had more experience in it, RPEs presented a weaker correlation with saccade timing errors, concluding that RPEs weakly influenced subjects' behavioural policy with experience (Bayer and Glimcher, 2005).

Therefore, in environments where an internal model of the task structure is learned, instead of continuously relearning values, subjects appear to update to dynamics changes by updating their belief of the current state of the world using observable features in the task, which is not mediated by dopamine.

What are dopaminergic RPEs doing then?

Our results are consistent with a model in which there are two different processes operating in parallel: state inference and evaluation. One system (not dopaminergic mediated) would be responsible for generating inferences about the current state of the world. Then in parallel,

the dopaminergic system would have access to these inferences or belief states to continuously evaluate the states and actions taken in each state.

Let's see how this model would fit our optogenetic results. Optogenetically activating dopamine neurons just after the choice was made reinforced that action. This is consistent with the reinforcing effects of dopamine. However, when dopamine was activated at outcome time, when RPEs occur naturally in the task, no reinforcing effects are seen on the next action. This is consistent with dopamine reinforcing the state that leads to high dopamine, in this case, the second-step state. Therefore, even if dopamine activation would promote repeating poking the experienced second-step, on each trial, actions are driven by the belief state. This belief state system would infer the state of the task by taking observations (e.g., rewards) and actions taken, and would predict the next observation. Differences between the predicted and actual observations would then generate state prediction errors used to update the belief state. The dopaminergic system would then combine the observed and inferred state features to estimate value and action policy. Consequently, the increased second-step state value generated by the optogenetic manipulation have no effect on the subsequent choice.

Where in the brain does the state inference happen?

Frontal cortical regions could be one of the substrates of state inference. In a probabilistic reversal task, ventromedial PFC BOLD activity was better explained by a state-based model than the standard TD RL (Hampton et al., 2006). Dorsolateral PFC neuronal population activity, also in a probabilistic reversal task, was shown to encode choice switches reflecting a change in the belief state (Bartolo and Averbeck, 2020). Inactivation of medial PFC disturbed

dopamine computations conditioned on the belief state, without disturbing the computations in a fully observable task (Starkweather et al., 2018). Similarly, OFC neuronal activity reflected unobservable task states, which affected task performance and errors (Schuck et al., 2016). Lesioning or inactivating OFC signalling impaired reversals in a probabilistic task (Walton et al., 2010), impaired performance when the best action needed to use inferred information (Bradfield et al., 2015), and shifted inference-based behaviour to stimulus-driven behavioural strategy in a foraging task (Vertechi et al., 2020). Therefore, PFC and OFC could be potential regions involved in learning and storing the predictive model of the task and inferring the current state of the task which would be used by the dopaminergic system for value computation and action generation.

Another potential substrate, although more hypothetical and targeted studies would be required to test this interaction, is the cerebellar system. The cerebellum has been presented as a location for encoding internal models of body movement (Wolpert et al., 1998), with a major role in sensorimotor control (Manto et al., 2012). But also, it has been suggested a role in cognitive aspects of behaviour and that the same principles for motor control, could also be applied to the control of mental representations (Ito, 2008). Sensory predictions errors have been shown in cerebellar activity (Schlerf et al., 2012; Popa and Ebner, 2018). Under uncertainty (Blackwood et al., 2004), and while solving structured tasks where specific rules need to be followed (Kim et al., 1994; Rao et al., 1997) lateral cerebellum and the dentate nucleus (one of the cerebellar output nuclei) presented increased BOLD activity. Recent computational models support the idea of cerebellar implications in state inference (Narain et al., 2018; Palacios et al., 2021).

Cerebellar computations were postulated to influence basal ganglia computations through a disynaptic pathway through thalamus (Bostan et al., 2010; Caligiore et al., 2017). However, recent studies have shown direct input from cerebellar nuclei into the midbrain dopaminergic system (Carta et al., 2019; Washburn et al., 2022). Furthermore, reward signals have also been shown in the cerebellar cortex (Wagner et al., 2017; Kostadinov et al., 2019).

Although evidence is very sparse and quite speculative, the cerebellar system could be presented as another potential substrate for state inference that could inform the dopaminergic system. But also, the cerebello-basal ganglia-thalamo-cortical system has been related to motor and cognitive functions (Caligiore et al., 2017). Therefore, it could also be possible that cortical and cerebellar regions could jointly contribute to state prediction, or that their implication may depend on the stage of learning.

Naïve from expert behaviour

Although our optogenetic manipulation is in line with dopaminergic RPEs not being used to update the behavioural policy, it is plausible that during initial task acquisition, dopaminergic RPEs drive value learning. Initially, when an internal model of the task is not yet built, observable events (i.e., rewards) would be used to update state and action values. With task experience, an internal model would be progressively created and updated, and hidden latent (i.e., non-observable) features in the task would be used for value estimations. Therefore, a dynamic arbitration may occur between values learnt over observable and inferred features. This would be consistent with Bayer & Glimcher (2005) results, exhibiting the uncoupling between RPEs and behaviour with experience. This arbitration may be controlled through uncertainty (Daw et al., 2005; Aitchison et al., 2021). As values conditioned on correctly

inferred latent states are stable across changes in those states, they can be learnt with higher accuracy than values over observable features, which change when latent states change. This suggests that over task acquisition, control would transfer from values learnt over observable features to those over latent features. Thus, behavioural flexibility would shift from value-learning to inference control. However, future experiments would be needed to explore this idea (below it is described a potential task that could address this question).

Another possible learning mechanism could be meta-RL, where dopaminergic RPE would be used to train another system, a recurrent neural network potentially prefrontal cortex, driving synaptic plasticity. Flexible behaviour would then be controlled through recurrent dynamics in the prefrontal system (Wang et al., 2018).

6.4 Dopamine a multiplex signal: RPE, motivation and movement

Despite the lack of effect of our optogenetic manipulation on next trial choice behaviour, this does not imply dopamine is not involved in inference-guided tasks. Our fibre photometry recordings of dopaminergic somatic (VTA) and axonal (NAc and DMS) activity and striatal concentrations report a rich set of behavioural and task-related variables. At different timescales, dopamine conveys simultaneously RPE, motivation and movement-related signals.

As discussed in detail in chapter 4, recent work has questioned the role of dopamine in multiplexing RPE and motivational signals. Instead, they exposed a functional dissociation between somatic VTA dopaminergic activity and ventral striatum dopamine concentrations.

They concluded that a motivational signal controlling action vigour was conveyed by dopaminergic striatal concentrations, but this was not present in VTA dopaminergic activity. They further suggested that striatal dopamine modulation was not consistent with RPE signalling but instead reflected reward rate. They postulated a markedly uncoupling between dopamine activity and release (Hamid et al., 2016; Mohebi et al., 2019).

Our recordings do not support this dissociation; instead we observed that dopamine activity and concentrations multiplex a wide range of signals. First, the biphasic modulation – positive when entering a new state, and negative when exiting the state – is a key signature of an RPE (Behrens et al., 2008; Rutledge et al., 2010; Niv et al., 2012). This was present in all regions, activity and release. It was most prominent in NAc, both in axonal activity and dopamine concentrations. In VTA and DMS, instead, the initial positive modulation was much larger than the subsequent negative one when new information was revealed. This difference between VTA somatic and NAc axonal activity and concentrations, can reflect local modulation in the terminals (Threlfell and Cragg, 2011; Kramer et al., 2022; Liu et al., 2022; Reynolds et al., 2022). Therefore, our data is consistent with Kim *et al.* (2020) corroborating RPE coding in dopaminergic phasic signals.

At a slower timescale that lasted the entirety of the trial, reward rate influenced dopaminergic activity and concentrations. This modulation by reward rate described separate variance from state and action values. It was reflected as a baseline DC shift in response to experienced reward rate. This parallel signal is consistent with the motivational account (Hamid et al., 2016; Berke, 2018; Mohebi et al., 2019), but present in VTA, NAc and DMS. This motivational function of VTA dopaminergic activity is further corroborated by our optogenetic activation

which speeded subsequent trial re-initiation when stimulating at outcome. This is also supported by previous optogenetic (Hamid et al., 2016) and pharmacological manipulations (Lee et al., 2015; Verharen et al., 2019; Korn et al., 2021) showing increased task engagement. Finally, in parallel to both RPE and motivational–consistent signals, we also observed other task-related modulations. Again, supporting previous reports on movement signals in the dopaminergic system (Howe and Dombeck, 2016; Parker et al., 2016; da Silva et al., 2018; Lee et al., 2019), in particular at the start and end of movement sequence (Jin and Costa, 2010; Dodson et al., 2016), we found that dopaminergic activity and concentrations reflected a lateralised action sequence signal. Contralateral movements, independent of value were present first in DMS activity and concentrations, followed by NAc axonal activity but not concentrations. VTA dopaminergic modulation, although not significant, showed similar modulation as NAc axonal activity. Therefore, this modulation did not reflect a pure movement signal, instead, across regions the different steps of the task (or actions required, a first movement from initiation to choice, and a second movement from choice to second-step) are reflected. However, other reports have also revealed a pure movement-related activity in response to movement velocity, acceleration, force, or movement kinematics (Puryear et al., 2010; Barter et al., 2015; Engelhard et al., 2019; Hughes et al., 2020; Kremer et al., 2020). Our data do not address these types of signals, and analysis of mice's actual movements at each timepoint in the task would be required.

6.5 Positive and negative coding, an intrinsic property of the dopaminergic system?

In chapter 3, we observed an asymmetric effect of rewards and omissions controlling subsequent choice behaviour. Although an ideal observer would use both rewards and omissions to guide behaviour as both outcomes are informative, this asymmetry in mice's behaviour might reflect either cognitive limitations or processes related to ethological animal behaviour. This asymmetry has also been found in humans' behaviour (Sebold et al., 2014; Wyckmans et al., 2019).

This behavioural asymmetry in response to positive and negative outcomes limits our capability to address the asymmetric response in the dopamine system by positive and negative RPE. Our dopaminergic recordings in chapter 4 reflect an asymmetric influence of rewards and omissions on action and state value update – rewards produced a bigger modulation reflecting the update than omissions. This could just reflect the behavioural asymmetry based on previous outcome.

Despite this, we observed a marked asymmetry between positive and negative modulation between regions. While NAc activity and release presented a positive – to – negative modulation consistent with RPE, VTA activity was mainly driven positively. This could be in line with studies reporting an asymmetric reporting of positive and negative RPE in midbrain activity (Schultz et al., 1993, 1997; Bayer and Glimcher, 2005; Pan et al., 2005; Bayer et al., 2007; D'Ardenne et al., 2008; Mohebi et al., 2019). It has been suggested that while dopamine would carry positive predictions, other system such as the serotonin, would carry the negative

predictions (Daw et al., 2002). At the same time, our results also support the symmetric coding in ventral striatum (D'Ardenne et al., 2008; Hart et al., 2014).

Further studies specifically focusing on this question, and potentially recording different systems simultaneously, would shed light on this current debate.

6.6 Future directions

Many different questions arise from our results, such as the mechanism for multiplexing and the interaction with other neuronal systems. Here, we will describe two other potential directions: the exploration-exploitation trade-off guiding behavioural policy and dopamine's role, and the evolution of dopaminergic signals and influence from naïve to expert behaviour. We will also discuss a new behavioural task to study the latter one.

Exploration-exploitation trade-off

During flexible behaviour in a dynamic environment, both humans and animals need to balance exploration and exploitation. While exploitation, choosing that option with a higher assigned value, can be the optimal behaviour based on experience and the internal model of the environment, exploration of new options may discover new better options in naturalistic environments.

How the exploration-exploitation trade-off is controlled behaviourally and by which neuronal circuitries is not well known. In RL, the balance between them can be achieved through ε-greedy or soft-max action-selection policy. In the ε-greedy policy, ε is the probability a random

action could be chosen. Exploration through a soft-max policy, the probability of taking each action is weighted based on their value estimate by the inverse temperature. If the inverse temperature is 1, then equals to a ε-greedy policy.

Both humans and animals have shown evidence of this kind of random exploration (Wilson et al., 2014), with dopamine potentially controlling the inverse temperature parameter (Humphries et al., 2012; Cinotti et al., 2019). Another type of exploration has also been shown based on the uncertainty (Wilson et al., 2014) involving both cortical (Daw et al., 2006; Frank et al., 2009) and dopaminergic (Naudé et al., 2016) control.

Others have also shown that exploration-exploitation trade-off can be controlled through dynamic asymmetric learning rate. The degree of the asymmetric positive and negative learning rate conditioned on background reward probability can control explorative behaviour and increase performance (Cazé and Van Der Meer, 2013; Ohta et al., 2021). Although this was not observed in a 2-armed bandit task in humans (Gershman, 2015), as discussed by Ohta *et al.* (2021), this could be the result of the difficulty in assessing exploratory behaviour when only two options are present.

Future further behavioural modelling on the two-step task can explore the presence of different asymmetric learning rates at the beginning and end of block for example, or dynamic inverse temperature based on the history of rewards. It is also possible, that the sustained dopaminergic activity and concentrations that correlated with reward rate, which we have interpreted as a motivational signal, could instead be controlling exploration-exploitation

behaviour. Further analysis of this feature together with closed-loop optogenetic manipulations would be needed to uncover this role during flexible behaviour.

The evolution of learning: from naïve to expert behaviour

As described earlier, inference-guided behaviour potentially emerges with experience, as computations over belief states are better in predicting future observations. We hypothesised that initially, traditional value-learning guides both behaviour and dopamine signalling. This value computation will progressively be taken over by inference-based ones, and cortical control.

In line with this hypothesis, Bayer & Glimcher (2005) observed an uncoupling between RPE and behaviour with learning. Coddington & Dudman (2018) recorded learning from naïve to expert behaviour during a cue-reward association task. They observed that dopamine signals reflected the learned association across time through RPE correlates which included cue- and movement-related components. Hamid *et al.* (2021) showed that striatal dopamine waves become more consistent and directional with experience, which were contingent on animals' agency in the task.

Further research tracking behavioural, and dopamine (and cortical) signalling evolution is further needed to understand the mechanisms underlying learning how to flexibly act.

Inference task: learning and building a model of the environment

Despite the advantages of the two-step task in studying different behavioural strategies and separating action from outcome, it has some limitations for studying inference-guided behaviour and understanding how this behaviour emerges.

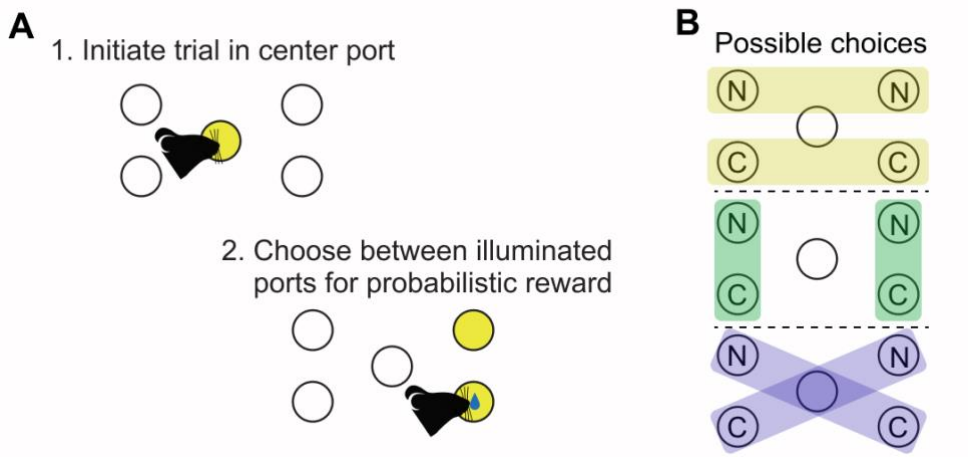


Figure 6.2. Inference task. A) Task structure. Each trial starts with the initiation state, where animals need to poke the illuminated center port. This triggers the choice state where animals need to make a choice between the 2 illuminated ports. B) Possible choices. In total there are 4 different choice ports, but they just become available 2 at a time. Neutral ports, N, are those 2 ports that are rewarded with 50% probability. Correlated ports, C, are the 2 ports that reward probability change in block in an anticorrrelated fashion. The reward probability of the correlated ports are 80% and 20%. Yellow areas, choices between the same type of port. Green areas, inference-informative choices between different port type, in which inferred behaviour can be tested. Blue areas, non-informative choices between different port type.

To achieve this, first, we need a task requiring minimum shaping of the animals, where animals can learn the rules of the task on their own through exploration. To do so, a simple task could involve 2 different states. An initiation state where animals need to poke the centre

illuminated port, and a choice state, where animals need to choose between two illuminated options which are probabilistically rewarded (Figure 6.2A).

Then, to study the evolution of inference learning, we need a hidden latent variable that would control observable features in the task. We could use the same latent variable as in the two-step task that animals can learn, the anticorrelated structure of the reward probabilities. Therefore, this task can have two choice ports with reward probabilities that change in block in an anticorrelated fashion. In each block, one would be rewarded with 80% probability, while the other with 20% probability. To avoid correlations between animals' performance and probability of reversals, this would happen independently of their behaviour (after 20-50 trials). Different to a simple 2-armed bandit task, we would include 2 other choice ports, both rewarded with 50% probability. These would be the neutral ports (Figure 6.2B). On each trial though, choice could be made between 2 illuminated ports.

This task would allow us to first understand how behaviour evolves. At the same time, simultaneous recordings of cortical regions and dopamine would allow us to track how behavioural changes are reflected at the neuronal level as a model of the task is created.

The structure of this task would also allow us to make specific predictions if inference guides behaviour by asking how an outcome in a correlated port influences subsequent choices. Figure 6.2B depicts the different choice options offered on each trial. Choices between correlated and neutral ports produce different predictions if animals use inference to guide their behaviour or if instead they are learning independent values or assign a value to each side – left or right. Vertical choices (green options in Fig, 6.2B) (on a trial only two ports

illuminate, so for vertical choices, either the two ports on the right or the left would illuminate) are inference-informative choices. For example, if animals received reward on a correlated port in the previous trial, that would decrease the value of the non-experienced correlated port; therefore, if on the current trial, the non-experienced correlated port is offered together with the neutral port on the same side, choosing the neutral port would be a sign of inference. Choices between correlated and neutral ports on opposite sides (Figure 6.2B, blue choices) would not be as informative, as if animals instead are just catching values on each side based on previous experience, this would result in the same behavioural output as an inference-based strategy. Choices between ports of the same type – neutral or correlated – would be also offered. Despite not all the possible choices being informative for inference-based behaviour, this task design would ensure that all ports are experienced the same number of times.

Using this task we could ask several questions: how does the behavioural strategy emerge? Across experience, are different strategies used? Are dopamine signals conditioned on the belief state in the correlated ports but not the neutral ones? How does this signal emerge? Is there a drastic shift from independent value learning to inference-based? Is a model of the environment represented in the cortical circuitry? How are these signals related to dopamine activity?

6.7 Conclusion

Our findings emphasise that flexible behaviour involves two processes operating in parallel: inference about the current state of the world, and evaluation of those states and actions

taken in them. The involvement of dopamine in updating values has rightly been a major focus of accounts of flexible decision-making. However, in the structured environments common both in the lab and real world, it is only half the picture, as subjects can adapt to change by updating their beliefs about the world's state. We showed here that although such belief updates strongly influence dopamine activity, this relationship is not reciprocal, as stimulating dopamine failed to reproduce natural reward effects. A pressing direction for future work is to understand the interplay between these two processes in flexibly adapting to a changing world.

REFERENCES

Adamantidis AR, Tsai H-C, Boutrel B, Zhang F, Stuber GD, Budygin EA, Touriño C, Bonci A, Deisseroth K, de Lecea L (2011) Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. J Neurosci Off J Soc Neurosci 31:10829–10835.

Adams CD (1982) Variations in the sensitivity of instrumental responding to reinforcer devaluation. Q J Exp Psychol Sect B 34.

Adams CD, Dickinson A (1981) Instrumental responding following reinforcer devaluation. Q J Exp Psychol 33:109–121.

Agnati LF, Bjelke B, Fuxe K (1992) Volume Transmission in the Brain. Am Sci 80:362–373.

Aitchison L, Jegminat J, Menendez JA, Pfister J-P, Pouget A, Latham PE (2021) Synaptic plasticity as Bayesian inference. Nat Neurosci 24:565–571.

Akam T, Costa R, Dayan P (2015) Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task Daunizeau J, ed. PLOS Comput Biol 11:e1004648.

Akam T, Lustig A, Rowland J, Kapanaiah SKT, Esteve-Agraz J, Panniello M, Marquez C, Kohl M, Kätzel D, Costa RM, Walton M (2021a) pyControl: Open source, Python based, hardware and software for controlling behavioural neuroscience experiments. bioRxiv:2021.02.22.432227.

Akam T, Rodrigues-Vaz I, Marcelo I, Zhang X, Pereira M, Oliveira RF, Dayan P, Costa RM (2021b) The Anterior Cingulate Cortex Predicts Future States to Mediate Model-Based Action Selection. Neuron 109:149-163.e7.

Akam T, Walton ME (2019) pyPhotometry: Open source Python based hardware and

software for fiber photometry data acquisition. Sci Rep 9:3521.

Akam T, Walton ME (2021) What is dopamine doing in model-based reinforcement learning? Curr Opin Behav Sci 38:74–82.

Azcorra M, Gaertner Z, Davidson C, Ramakrishnan C, Fenno L, Kim YS, Deisseroth K, Awatramani R, Dombeck DA (2022) Dopaminergic axons track somatic signaling in behaving mice. bioRxiv.

Babayan BM, Uchida N, Gershman SJ (2018) Belief state representation in the dopamine system. Nat Commun 9:1891.

Baik JH (2020) Stress and the dopaminergic reward system. Exp Mol Med 52.

Balleine BW, Dickinson A (1998) Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. In: Neuropharmacology.

Balleine BW, O'Doherty JP (2010) Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. Neuropsychopharmacology 35:48–69.

Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang 68.

Barter JW, Li S, Lu D, Bartholomew RA, Rossi MA, Shoemaker CT, Salas-Meza D, Gaidis E, Yin HH (2015) Beyond reward prediction errors: the role of dopamine in movement kinematics. Front Integr Neurosci 9:39.

Bartolo R, Averbeck BB (2020) Prefrontal Cortex Predicts State Switches during Reversal Learning. Neuron 106:1044-1054.e4.

Bartolo R, Averbeck BB (2021) Inference as a fundamental process in behavior. Curr Opin Behav Sci 38:8–13.

Baxter MG, Parker A, Lindner CC, Izquierdo AD, Murray EA (2000) Control of response

selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex. J Neurosci 20:4311–4319.

Bayer HM, Glimcher PW (2005) Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. Neuron 47:129–141.

Bayer HM, Lau B, Glimcher PW (2007) Statistics of Midbrain Dopamine Neuron Spike Trains in the Awake Primate. J Neurophysiol 98:1428–1439.

Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS (2008) Associative learning of social value. Nature 456:245–249.

Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. Nat Neurosci 10.

Beier KT, Steinberg EE, DeLoach KE, Xie S, Miyamichi K, Schwarz L, Gao XJ, Kremer EJ, Malenka RC, Luo L (2015) Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input-Output Mapping. Cell 162:622–634.

Berke JD (2018) What does dopamine mean? Nat Neurosci 21:787–793.

Berridge KC (2007) The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology (Berl) 191:391–431.

Blackburn JR, Phillips AG, Jakubovic A, Fibiger HC (1989) Dopamine and Preparatory Behavior: II. A Neurochemical Analysis. Behav Neurosci 103.

Blackwood N, ffytche D, Simmons A, Bentall R, Murray R, Howard R (2004) The cerebellum and decision making under uncertainty. Cogn Brain Res 20:46–53.

Bogacz R (2020) Dopamine role in learning and action inference. Elife 9.

Bostan AC, Dum RP, Strick PL (2010) The basal ganglia communicate with the cerebellum. Proc Natl Acad Sci 107:8452–8456.

Bradfield LA, Dezfouli A, van Holstein M, Chieng B, Balleine BW (2015) Medial Orbitofrontal

Cortex Mediates Outcome Retrieval in Partially Observable Task Situations. Neuron 88:1268–1280.

Bromberg-Martin ES, Matsumoto M, Hikosaka O (2010a) Dopamine in Motivational Control: Rewarding, Aversive, and Alerting. Neuron 68.

Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O (2010b) A pallidus-habenula-dopamine pathway signals inferred stimulus values. J Neurophysiol 104:1068–1076.

Brown HD, McCutcheon JE, Cone JJ, Ragozzino ME, Roitman MF (2011) Primary food reward and reward-predictive stimuli evoke different patterns of phasic dopamine signaling throughout the striatum. Eur J Neurosci 34:1997–2006.

Bu M, Farrer MJ, Khoshbouei H (2021) Dynamic control of the dopamine transporter in neurotransmission and homeostasis. npj Park Dis 7.

Bush RR, Mosteller F (1951) A mathematical model for simple learning. Psychol Rev 58:313–323.

Cagniard B, Balsam PD, Brunner D, Zhuang X (2006) Mice with chronically elevated dopamine exhibit enhanced motivation, but not learning, for a food reward. Neuropsychopharmacology 31.

Cai LX, Pizano K, Gundersen GW, Hayes CL, Fleming WT, Holt S, Cox JM, Witten IB (2020) Distinct signals in medial and lateral VTA dopamine neurons modulate fear extinction at different times. Elife 9.

Caligiore D, Pezzulo G, Baldassarre G, Bostan AC, Strick PL, Doya K, Helmich RC, Dirkx M, Houk J, Jörntell H, Lago-Rodriguez A, Galea JM, Miall RC, Popa T, Kishore A, Verschure PFMJ, Zucca R, Herreros I (2017) Consensus Paper: Towards a Systems-Level View of Cerebellar Function: the Interplay Between Cerebellum, Basal Ganglia, and Cortex. Cerebellum 16:203–229.

Cardinal RN, Parkinson JA, Hall J, Everitt BJ (2002) Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. Neurosci Biobehav Rev 26.

Carlsson A (1959) The occurrence, distribution and physiological role of catecholamines in the nervous system. Pharmacol Rev 11.

Carta I, Chen CH, Schott AL, Dorizan S, Khodakhah K (2019) Cerebellar modulation of the reward circuitry and social behavior. Science (80-) 363.

Castro-Rodrigues P, Akam T, Snorasson I, Camacho M, Paixão V, Maia A, Barahona-Corrêa JB, Dayan P, Simpson HB, Costa RM, Oliveira-Maia AJ (2022) Explicit knowledge of task structure is a primary determinant of human model-based action. Nat Hum Behav.

Cazé RD, Van Der Meer MAA (2013) Adaptive properties of differential learning rates for positive and negative outcomes. Biol Cybern 107.

Cazorla M, deCarvalho FD, Chohan MO, Shegda M, Chuhma N, Rayport S, Ahmari SE, Moore H, Kellendonk C (2014) Dopamine d2 receptors regulate the anatomical and functional balance of basal ganglia circuitry. Neuron 81.

Chang CY, Esber GR, Marrero-Garcia Y, Yau H-J, Bonci A, Schoenbaum G (2016) Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. Nat Neurosci 19:111–116.

Chen JS, Amseli A (1980) Recall (versus recognition) of taste and immunization against aversive taste anticipations based on illness. Science (80-) 209.

Chen T-W, Wardill TJ, Sun Y, Pulver SR, Renninger SL, Baohan A, Schreiter ER, Kerr RA, Orger MB, Jayaraman V, Looger LL, Svoboda K, Kim DS (2013) Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature 499:295–300.

Chiu PH, Lohrenz TM, Montague PR (2008) Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. Nat Neurosci 11:514–520.

Chuhma N, Mingote S, Moore H, Rayport S (2014) Dopamine neurons control striatal cholinergic neurons via regionally heterogeneous dopamine and glutamate signaling. Neuron 81.

Cieślak PE, Ahn W-Y, Bogacz R, Rodriguez Parkitna J (2018) Selective Effects of the Loss of NMDA or mGluR5 Receptors in the Reward System on Adaptive Decision-Making. eNeuro 5.

Cinotti F, Fresno V, Aklil N, Coutureau E, Girard B, Marchand AR, Khamassi M (2019) Dopamine blockade impairs the exploration-exploitation trade-off in rats. Sci Rep 9.

Coddington LT, Dudman JT (2018a) The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat Neurosci 21:1563–1573.

Coddington LT, Dudman JT (2018b) The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat Neurosci 21:1563–1573.

Coddington LT, Dudman JT (2019) Learning from Action: Reconsidering Movement Signaling in Midbrain Dopamine Neuron Activity. Neuron 104:63–77.

Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N (2012) Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482:85–88.

Collins AL, Greenfield VY, Bye JK, Linker KE, Wang AS, Wassum KM (2016) Dynamic mesolimbic dopamine signaling during action sequence learning and expectation violation. Sci Rep 6:20231.

Colwill RM, Rescorla RA (1985a) Instrumental Responding Remains Sensitive to Reinforcer Devaluation After Extensive Training. J Exp Psychol Anim Behav Process 11.

Colwill RM, Rescorla RA (1985b) Postconditioning devaluation of a reinforcer affects instrumental responding. J Exp Psychol Anim Behav Process 11:120–132.

Condon MD, Platt NJ, Zhang YF, Roberts BM, Clements MA, Vietti-Michelina S, Tseu MY,

Brimblecombe KR, Threlfell S, Mann EO, Cragg SJ (2019) Plasticity in striatal dopamine release is governed by release-independent depression and the dopamine transporter. Nat Commun 10.

Corbett D, Wise RA (1980) Intracranial self-stimulation in relation to the ascending dopaminergic systems of the midbrain: A moveable electrode mapping study. Brain Res 185.

Corbit LH, Balleine BW (2003) The role of prelimbic cortex in instrumental conditioning. Behav Brain Res 146.

Costa VD, Tran VL, Turchi J, Averbeck BB (2015) Reversal learning and dopamine: a bayesian perspective. J Neurosci 35:2407–2416.

Cushman F, Morris A (2015) Habitual control of goal selection in humans. Proc Natl Acad Sci U S A 112:13817–13822.

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD Responses Reflecting Dopaminergic Signals in the Human Ventral Tegmental Area. Science (80-) 319:1264–1267.

da Silva JA, Tecuapetla F, Paixão V, Costa RM (2018) Dopamine neuron activity before action initiation gates and invigorates future movements. Nature 554:244–248.

Dabney W, Kurth-Nelson Z, Uchida N, Starkweather CK, Hassabis D, Munos R, Botvinick M (2020) A distributional code for value in dopamine-based reinforcement learning. Nature 577:671–675.

Darwin C (1859) On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life. London: J. Murray.

Daw ND, Dayan P (2014) The algorithmic anatomy of model-based evaluation. Philos Trans R Soc B Biol Sci 369:20130478–20130478.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69:1204–1215.

Daw ND, Kakade S, Dayan P (2002) Opponent interactions between serotonin and dopamine. Neural networks Off J Int Neural Netw Soc 15:603–616.

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704–1711.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.

de Jong JW, Afjei SA, Pollak Dorocic I, Peck JR, Liu C, Kim CK, Tian L, Deisseroth K, Lammel S (2019) A Neural Circuit Mechanism for Encoding Aversive Stimuli in the Mesolimbic Dopamine System. Neuron 101:133-151.e7.

de Jong JW, Fraser KM, Lammel S (2022) Mesoaccumbal Dopamine Heterogeneity: What Do Dopamine Firing and Release Have to Do with It? Annu Rev Neurosci 45.

Decker JH, Otto AR, Daw ND, Hartley CA (2016) From Creatures of Habit to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based ReinforcementLearning. Psychol Sci 27:848.

DeLong MR (1983) The neurophysiologic basis of abnormal movements in basal ganglia disorders. Neurobehav Toxicol Teratol 5.

Denk F, Walton ME, Jennings KA, Sharp T, Rushworth MFS, Bannerman DM (2005) Differential involvement of serotonin and dopamine systems in cost-benefit decisions about delay or effort. Psychopharmacology (Berl) 179:587–596.

Deserno L, Huys QJMM, Boehme R, Buchert R, Heinze H-JJ, Grace AA, Dolan RJ, Heinz A, Schlagenhauf F (2015) Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proc Natl Acad Sci

U S A 112:1595–1600.

Dezfouli A, Balleine BW (2019) Learning the structure of the world: The adaptive nature of state-space and action representations in multi-stage decisionmaking. PLoS Comput Biol 15.

Di Chiara G (1995) The role of dopamine in drug abuse viewed from the perspective of its role in motivation. Drug Alcohol Depend 38.

Dickinson A (1985) Actions and habits: the development of behavioural autonomy. Philos Trans R Soc London B, Biol Sci 308:67–78.

Dickinson A, Balleine B (1994) Motivational control of goal-directed action. Anim Learn Behav 22.

Dickinson A, Balleine B, Watt A, Gonzalez F, Boakes RA (1995) Motivational control after extended instrumental training. Anim Learn Behav 23.

Dickinson A, Nicholas DJ (1983) Irrelevant incentive learning during instrumental conditioning: The role of the drive-reinforcer and response-reinforcer relationships. Q J Exp Psychol Sect B 35.

Dickinson A, Nicholas DJ, Adams CD (1983) The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. Q J Exp Psychol Sect B 35.

Dickinson A, Pérez OD (2018) Actions and habits: Psychological issues in dual-system theory. In: Goal-Directed Decision Making: Computations and Neural Circuits.

Dodson PD, Dreyer JK, Jennings KA, Syed ECJ, Wade-Martins R, Cragg SJ, Bolam JP, Magill PJ (2016) Representation of spontaneous movement by dopaminergic neurons is cell-type selective and disrupted in parkinsonism. Proc Natl Acad Sci U S A 113:E2180-8.

Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND (2015a) Model-based choices involve prospective neural activity. Nat Neurosci 18:767–772.

Doll BB, Shohamy D, Daw ND (2015b) Multiple memory systems as substrates for multiple decision systems. Neurobiol Learn Mem 117.

Doya K (1999) What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Networks 12:961–974.

Doya K, Samejima K, Katagiri K, Kawato M (2002) Multiple Model-Based Reinforcement Learning. Neural Comput 14:1347–1369.

Dreyer JK, Herrik KF, Berg RW, Hounsgaard JD (2010) Influence of phasic and tonic dopamine release on receptor activation. J Neurosci 30.

Ellwood IT, Patel T, Wadia V, Lee AT, Liptak AT, Bender KJ, Sohal VS (2017) Tonic or phasic stimulation of dopaminergic projections to prefrontal cortex causes mice to maintain or deviate from previously learned behavioral strategies. J Neurosci 37:1221–17.

Engelhard B, Finkelstein J, Cox J, Fleming W, Jang HJ, Ornelas S, Koay SA, Thiberge SY, Daw ND, Tank DW, Witten IB (2019) Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. Nature 570:509–513.

Eshel N, Tian J, Bukwich M, Uchida N (2016) Dopamine neurons share common response function for reward prediction error. Nat Neurosci 19:479–486.

Evenden JL, Robbins TW (1983) Increased response switching, perseveration and perseverative switching following d-amphetamine in the rat. Psychopharmacology (Berl) 80:67–73.

Fallon JH, Moore RY (1978) Catecholamine innervation of the basal forebrain. IV. Topography of the dopamine projection to the basal forebrain and neostriatum. J Comp Neurol 180:545–580.

Farassat N, Costa KM, Stojanovic S, Albert S, Kovacheva L, Shin J, Egger R, Somayaji M, Duvarci S, Schneider G, Roeper J (2019) In vivo functional diversity of midbrain

dopamine neurons within identified axonal projections. Elife 8.

Feher da Silva C, Hare TA (2018) A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability. PLoS One 13:e0195328.

Feher da Silva C, Hare TA (2020) Humans primarily use model-based inference in the two-stage task. Nat Hum Behav:1–14.

Fiorillo CD (2013) Two dimensions of value: dopamine neurons represent reward but not aversiveness. Science 341:546–549.

Fiorillo CD, Newsome WT, Schultz W (2008) The temporal precision of reward prediction in dopamine neurons. Nat Neurosci 11:966–973.

Fiorillo CD, Tobler PN, Schultz W (2003) Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. Science (80-) 299:1898–1902.

Fiorillo CD, Yun SR, Song MR (2013) Diversity and homogeneity in responses of midbrain dopamine neurons. J Neurosci Off J Soc Neurosci 33:4693–4709.

Floresco SB, Yang CR, Phillips AG, Blaha CD (1998) Basolateral amygdala stimulation evokes glutamate receptor-dependent dopamine efflux in the nucleus accumbens of the anaesthetized rat. Eur J Neurosci 10:1241–1251.

François C, Yelnik J, Tandé D, Agid Y, Hirsch EC (1999) Dopaminergic cell group A8 in the monkey: anatomical organization and projections to the striatum. J Comp Neurol 414:334–347.

Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci 12.

Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE (2007) Genetic triple

dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A 104.

Frank MJ, Seeberger LC, O'Reilly RC (2004) By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. Science (80-) 306.

Funamizu A, Ito M, Doya K, Kanzaki R, Takahashi H (2012) Uncertainty in action-value estimation affects both action choice and learning rate of the choice behaviors of rats. Eur J Neurosci 35.

Gadagkar V, Puzerey PA, Chen R, Baird-Daniel E, Farhang AR, Goldberg JH (2016) Dopamine neurons encode performance error in singing birds. Science (80-) 354:1278–1282.

Gallistel CR, Mark TA, King AP, Latham PE (2001) The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. J Exp Psychol Anim Behav Process 27:354–372.

Garris PA, Ciolkowski EL, Pastore P, Wightman RM (1994) Efflux of dopamine from the synaptic cleft in the nucleus accumbens of the rat brain. J Neurosci 14.

Gentet LJ, Williams SR (2007) Dopamine gates action potential backpropagation in midbrain dopaminergic neurons. J Neurosci 27.

German DC, Manaye KF (1993) Midbrain dopaminergic neurons (nuclei A8, A9, and A10): Three-dimensional reconstruction in the rat. J Comp Neurol 331.

Gershman SJ (2014) Dopamine ramps are a consequence of reward prediction errors. Neural Comput 26:467–471.

Gershman SJ (2015) Do learning rates adapt to the distribution of rewards? Psychon Bull Rev 22.

Gershman SJ, Blei DM, Niv Y (2010) Context, Learning, and Extinction. Psychol Rev 117.

Gershman SJ, Uchida N (2019) Believing in dopamine. Nat Rev Neurosci 20:703–714.

Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron 66:585–595.

Gottfried JA, O'Doherty J, Dolan RJ (2003) Encoding Predictive Reward Value in Human Amygdala and Orbitofrontal Cortex. Science (80-) 301:1104–1107.

Grace AA, Bunney BS (1984a) The control of firing pattern in nigral dopamine neurons: Single spike firing. J Neurosci 4.

Grace AA, Bunney BS (1984b) The control of firing pattern in nigral dopamine neurons: Burst firing. J Neurosci 4.

Graybiel AM (2008) Habits, Rituals, and the Evaluative Brain. Annu Rev Neurosci 31:359–387.

Greenstreet F, Vergara HM, Pati S, Schwarz L, Wisdom M, Marbach F, Johansson Y, Rollik L, Moskovitz T, Clopath C, Stephenson-Jones M (2022) Action prediction error: a value-free dopaminergic teaching signal that drives stable learning. bioRxiv:2022.09.12.507572.

Groman SM, Massi B, Mathias SR, Curry DW, Lee D, Taylor JR (2019) Neurochemical and Behavioral Dissections of Decision-Making in a Rodent Multistage Task. J Neurosci 39:295–306.

Guru A, Seo C, Post RJ, Kullakanda DS, Schaffer JA, Warden MR (2020) Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map. bioRxiv:2020.05.21.108886.

Hall H, Sedvall G, Magnusson O, Kopp J, Halldin C, Farde L (1994) Distribution of D1- and D2-dopamine receptors, and dopamine and its metabolites in the human brain. Neuropsychopharmacology 11.

Hamid AA, Frank MJ, Moore CI (2021) Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. Cell 184:2733-2749.e16.

Hamid AA, Pettibone JR, Mabrouk OS, Hetrick VL, Schmidt R, Vander Weele CM, Kennedy RT, Aragona BJ, Berke JD (2016) Mesolimbic dopamine signals the value of work. Nat Neurosci 19:117–126.

Hampton AN, Adolphs R, Tyszka MJ, O'Doherty JP (2007) Contributions of the Amygdala to Reward Expectancy and Choice Signals in Human Prefrontal Cortex. Neuron 55.

Hampton AN, Bossaerts P, O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci 26:8360–8367.

Hart AS, Rutledge RB, Glimcher PW, Phillips PEM (2014) Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. J Neurosci Off J Soc Neurosci 34:698–704.

Hasz BM, Redish AD (2018) Deliberation and Procedural Automation on a Two-Step Task for Rats. Front Integr Neurosci 12:30.

Hikosaka O, Nakamura K, Nakahara H (2006) Basal ganglia orient eyes to reward. J Neurophysiol 95:567–584.

Holland PC (2004) Relations Between Pavlovian-Instrumental Transfer and Reinforcer Devaluation. J Exp Psychol Anim Behav Process 30.

Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. Nat Neurosci 1:304–309.

Houk JC, Adams JL, Barto AG (1995) A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement. In: Computational neuroscience. Models of information processing in the basal ganglia. Cambridge, MA, US: The MIT Press., pp 249–270.

Howe MW, Dombeck DA (2016) Rapid signalling in distinct dopaminergic axons during

locomotion and reward. Nature 535:505–510.

Howe MW, Tierney PL, Sandberg SG, Phillips PEM, Graybiel AM (2013) Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature 500:575–579.

Hughes RN, Bakhurin KI, Petter EA, Watson GDR, Kim N, Friedman AD, Yin HH (2020) Ventral Tegmental Dopamine Neurons Control the Impulse Vector during Motivated Behavior. Curr Biol 30:2681-2694.e5.

Hull CL (1936) The conflicting psychologies of learning—a way out. Psychol Rev 42:491.

Humphries MD, Khamassi M, Gurney K (2012) Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. Front Neurosci.

Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP (2012) Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol 8.

Ikemoto S, Panksepp J (1999) The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. Brain Res Brain Res Rev 31:6–41.

Ito M (2008) Control of mental activities by internal models in the cerebellum. Nat Rev Neurosci 9:304–313.

Ito M, Doya K (2009) Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci Off J Soc Neurosci 29:9861–9874.

Jang AI, Costa VD, Rudebeck PH, Chudasama Y, Murray EA, Averbeck BB (2015) The role of frontal cortical and medial-temporal lobe brain areas in learning a Bayesian prior belief on reversals. J Neurosci 35.

Jeffreys H (1961) Theory of Probability, 3rd ed. Clarendon Press, Oxford.

Jin X, Costa RM (2010) Start/stop signals emerge in nigrostriatal circuits during sequence learning. Nature 466:457–462.

Jones JL, Day JJ, Aragona BJ, Wheeler RA, Wightman RM, Carelli RM (2010) Basolateral Amygdala Modulates Terminal Dopamine Release in the Nucleus Accumbens and Conditioned Responding. Biol Psychiatry 67.

Jones S, Paul ES, Dayan P, Robinson ESJ, Mendl M (2017) Pavlovian influences on learning differ between rats and mice in a counter-balanced Go/NoGo judgement bias task. Behav Brain Res 331.

Kamin LJ (1968) "Attention-like" processes in classical conditioning. In: Symposium on the Prediction of Behaviour.

Kawagoe R, Takikawa Y, Hikosaka O (1998) Expectation of reward modulates cognitive signals in the basal ganglia. Nat Neurosci 1:411–416.

Keiflin R, Pribut HJ, Shah NB, Janak PH (2019) Ventral Tegmental Dopamine Neurons Participate in Reward Identity Predictions. Curr Biol 29:93-103.e3.

Killcross S, Coutureau E (2003) Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. Cereb Cortex 13:400–408.

Kim HR, Malik AN, Mikhael JG, Bech P, Tsutsui-Kimura I, Sun F, Zhang Y, Li Y, Watabe-Uchida M, Gershman SJ, Uchida N (2019) A unified framework for dopamine signals across timescales. bioRxiv:803437.

Kim HR, Malik AN, Mikhael JG, Bech P, Tsutsui-Kimura I, Sun F, Zhang Y, Li Y, Watabe-Uchida M, Gershman SJ, Uchida N (2020) A Unified Framework for Dopamine Signals across Timescales. Cell 183:1600-1616.e25.

Kim S-G, Uğurbil K, Strick PL (1994) Activation of a Cerebellar Output Nucleus During Cognitive Processing. Science (80-) 265:949–951.

Kishida KT, Saez I, Lohrenz T, Witcher MR, Laxton AW, Tatter SB, White JP, Ellis TL, Phillips PEM, Montague PR (2016) Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. Proc Natl Acad Sci 113:200–205.

Kiyatkin EA (1994) Behavioral significance of phasic changes in mesolimbic dopamine-dependent electrochemical signal associated with heroin self-injections. J Neural Transm 96.

Koob GF (1992) Dopamine, addiction and reward. Semin Neurosci 4.

Kool W, Cushman FA, Gershman SJ (2016) When Does Model-Based Control Pay Off? PLOS Comput Biol 12:e1005090.

Kool W, Gershman SJ, Cushman FA (2017) Planning complexity registers as a cost in metacontrol. J Cogn Neurosci 30.

Korn C, Akam T, Jensen KH, Vagnoni C, Huber A, Tunbridge EM, Walton ME (2019) Distinct roles for DAT and COMT in regulating dopamine transients and reward-guided decision making. bioRxiv:823401.

Korn C, Akam T, Jensen KHR, Vagnoni C, Huber A, Tunbridge EM, Walton ME (2021) Distinct roles for dopamine clearance mechanisms in regulating behavioral flexibility. Mol Psychiatry 26.

Kostadinov D, Beau M, Pozo MB, Häusser M (2019) Predictive and reactive reward signals conveyed by climbing fiber inputs to cerebellar Purkinje cells. Nat Neurosci.

Kramer PF, Brill-Weil SG, Cummins AC, Zhang R, Camacho-Hernandez GA, Newman AH, Eldridge MAG, Averbeck BB, Khaliq ZM (2022) Synaptic-like axo-axonal transmission from striatal cholinergic interneurons onto dopaminergic fibers. Neuron.

Kravitz A V, Tye LD, Kreitzer AC (2012) Distinct roles for direct and indirect pathway striatal

neurons in reinforcement. Nat Neurosci 15:816–818.

Kremer Y, Flakowski J, Rohner C, Lüscher C (2020) Context-Dependent Multiplexing by Individual VTA Dopamine Neurons. J Neurosci 40:7489 LP – 7509.

Krieckhaus EE, George W (1968) Acquisition Of Sodium By Rats: Interaction Of Innate Mechanisms And Latent Learning. J Comp Physiol Psychol 65.

Krugel LK, Biele G, Mohr PNC, Li SC, Heekeren HR (2009) Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. Proc Natl Acad Sci U S A 106.

Kupchik YM, Kalivas PW (2017) The Direct and Indirect Pathways of the Nucleus Accumbens are not What You Think. Neuropsychopharmacology 42.

Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A (2017) Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision Article Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. Curr Biol 27:821–832.

Lak A, Okun M, Moss MM, Gurnani H, Farrell K, Wells MJ, Reddy CB, Kepecs A, Harris KD, Carandini M (2020) Dopaminergic and Prefrontal Basis of Learning from Sensory Confidence and Reward Value. Neuron 105.

Lammel S, Hetzel A, Häckel O, Jones I, Liss B, Roeper J (2008) Unique Properties of Mesoprefrontal Neurons within a Dual Mesocorticolimbic Dopamine System. Neuron 57.

Lammel S, Ion DI, Roeper J, Malenka RC (2011) Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. Neuron 70:855–862.

Lammel S, Lim BK, Ran C, Huang KW, Betley MJ, Tye KM, Deisseroth K, Malenka RC (2012) Input-specific control of reward and aversion in the ventral tegmental area. Nature 491:212–217.

Lee E, Seo M, Dal Monte O, Averbeck BB (2015) Injection of a dopamine type 2 receptor antagonist into the dorsal striatum disrupts choices driven by previous outcomes, but not perceptual inference. J Neurosci 35:6298–6306.

Lee RS, Mattar MG, Parker NF, Witten IB, Daw ND (2019) Reward prediction error does not explain movement selectivity in DMS-projecting dopamine neurons. Elife 8.

Lee SJ, Lodder B, Chen Y, Patriarchi T, Tian L, Sabatini BL (2021) Cell-type-specific asynchronous modulation of PKA by dopamine in learning. Nature 590.

Lees AJ, Hardy J, Revesz T (2009) Parkinson's disease. Lancet (London, England) 373:2055–2066.

Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S (2017) Behavioural and neural characterization of optimistic reinforcement learning. Nat Hum Behav 1.

Lerner TN, Shilyansky C, Davidson TJ, Evans KE, Beier KT, Zalocusky KA, Crow AK, Malenka RC, Luo L, Tomer R, Deisseroth K (2015) Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. Cell 162:635–647.

Liu C, Cai X, Ritzau-Jost A, Kramer PF, Li Y, Khaliq ZM, Hallermann S, Kaeser PS (2022) An action potential initiation mechanism in distal axons for the control of dopamine release. Science (80-) 375.

Liu C, Goel P, Kaeser PS (2021) Spatial and temporal scales of dopamine transmission. Nat Rev Neurosci 22.

Liu C, Kaeser PS (2019) Mechanisms and regulation of dopamine release. Curr Opin Neurobiol 57.

Liu C, Kershberg L, Wang J, Schneeberger S, Kaeser PS (2018) Dopamine Secretion Is Mediated by Sparse Active Zone-like Release Sites. Cell 172.

Ljungberg T, Apicella P, Schultz W (1992) Responses of monkey dopamine neurons during

learning of behavioral reactions. J Neurophysiol 67.

Ljungberg T, Enquist M (1987) Disruptive effects of low doses of d-amphetamine on the ability of rats to organize behaviour into functional sequences. Psychopharmacology (Berl) 93:146–151.

Lohani S, Martig AK, Underhill SM, DeFrancesco A, Roberts MJ, Rinaman L, Amara S, Moghaddam B (2018) Burst activation of dopamine neurons produces prolonged post-burst availability of actively released dopamine. Neuropsychopharmacology 43.

Lohrenz T, McCabe K, Camerer CF, Montague PR (2007) Neural signature of fictive learning signals in a sequential investment task. Proc Natl Acad Sci U S A 104.

Lopes G, Bonacchi N, Frazão J, Neto JP, Atallah B V., Soares S, Moreira L, Matias S, Itskov PM, Correia PA, Medina RE, Calcaterra L, Dreosti E, Paton JJ, Kampff AR (2015) Bonsai: an event-based framework for processing and controlling data streams. Front Neuroinform 9:7.

Lubbock J (1889) On the senses, instincts, and intelligence of animals: with special reference to insects. K. Paul, Trench & Company.

Lubow RE, Moore AU (1959) Latent inhibition: the effect of nonreinforced pre-exposure to the conditional stimulus. J Comp Physiol Psychol 52:415.

Lynd-Balta E, Haber SN (1994) The organization of midbrain projections to the ventral striatum in the primate. Neuroscience 59:609–623.

Mackintosh NJ (1965) Selective attention in animal discrimination learning. Psychol Bull 64.

Manto M, Bower JM, Conforto AB, Delgado-García JM, Da Guarda SNF, Gerwig M, Habas C, Hagura N, Ivry RB, Marien P, Molinari M, Naito E, Nowak DA, Ben Taib NO, Pelisson D, Tesche CD, Tilikete C, Timmann D (2012) Consensus paper: Roles of the cerebellum in motor control-the diversity of ideas on cerebellar involvement in movement. In:

Cerebellum.

Marcott PF, Mamaligas AA, Ford CP (2014) Phasic dopamine release drives rapid activation of striatal D2-receptors. Neuron 84.

Martel JC, Gatti McArthur S (2020) Dopamine Receptor Subtypes, Physiology and Pharmacology: New Ligands and Concepts in Schizophrenia. Front Pharmacol 11.

Matsumoto M, Hikosaka O (2009) Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature 459:837–841.

Matuschek H, Kliegl R, Vasishth S, Baayen H, Bates D (2017) Balancing Type I error and power in linear mixed models. J Mem Lang 94.

McClure SM, Daw ND, Read Montague P (2003) A computational substrate for incentive salience. Trends Neurosci 26:423–428.

Menegas W, Bergan JF, Ogawa SK, Isogai Y, Venkataraju KU, Osten P, Uchida N, Watabe-Uchida M (2015) Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. Elife 4.

Millard SJ, Hoang IB, Greer Z, O'Connor SL, Wassum KM, James MH, Barker DJ, Sharpe MJ (2022) The cognitive basis of intracranial self-stimulation of midbrain dopamine neurons. bioRxiv:2022.08.11.503670.

Miller KJ, Botvinick MM, Brody CD (2017) Dorsal hippocampus contributes to model-based planning. Nat Neurosci 20:1269–1276.

Miller KJ, Shenhav A, Ludvig EA (2019) Habits without values. Psychol Rev 126:292–311.

Miller NE (1935) A reply to "Sign-Gestalt or Conditioned Reflex?" Psychol Rev 42.

Missale C, Russel Nash S, Robinson SW, Jaber M, Caron MG (1998) Dopamine receptors: From structure to function. Physiol Rev 78.

Mohebi A, Pettibone JR, Hamid AA, Wong J-MT, Vinson LT, Patriarchi T, Tian L, Kennedy RT,

Berke JD (2019) Dissociable dopamine dynamics for learning and motivation. Nature 570:65–70.

Molinoff PB, Axelrod J (1971) Biochemistry of catecholamines. Annu Rev Biochem 40.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.

Montague PR, King-Casas B, Cohen JD (2006) Imaging valuation models in human choice. Annu Rev Neurosci 29.

Morgan CL (1886) II.—ON THE STUDY OF ANIMAL INTELLIGENCE. Mind os-XI.

Morgan CL 1852-1936 (1894) An introduction to comparative psychology. England: W. Scott, limited.

Morita K, Kato A (2014) Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. Front Neural Circuits 8:36.

Morris A, Cushman F (2019) Model-Free RL or Action Sequences? Front Psychol 10:2892.

Moss MM, Zatka-Haas P, Harris KD, Carandini M, Lak1 A (2021) Dopamine axons in dorsal striatum encode contralateral visual stimuli and choices. J Neurosci 41.

Nakahara H, Itoh H, Kawagoe R, Takikawa Y, Hikosaka O (2004) Dopamine neurons can represent context-dependent prediction error. Neuron 41:269–280.

Nakanishi S, Hikida T, Yawata S (2014) Distinct dopaminergic control of the direct and indirect pathways in reward-based and avoidance learning behaviors. Neuroscience 282.

Narain D, Remington ED, Zeeuw CI De, Jazayeri M (2018) A cerebellar mechanism for learning prior distributions of time intervals. Nat Commun 9:469.

Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI (2012) Rational regulation of learning dynamics by pupil-linked arousal systems. Nat Neurosci 15.

Nassar MR, Wilson RC, Heasly B, Gold JI (2010) An approximately Bayesian delta-rule model

explains the dynamics of belief updating in a changing environment. J Neurosci 30.

Naudé J, Tolu S, Dongelmans M, Torquet N, Valverde S, Rodriguez G, Pons S, Maskos U, Mourot A, Marti F, Faure P (2016) Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking. Nat Neurosci 19:471–478.

Nirenberg MJ, Chan J, Pohorille A, Vaughan RA, Uhl GR, Kuhar MJ, Pickel VM (1997) The dopamine transporter: comparative ultrastructure of dopaminergic axons in limbic and motor compartments of the nucleus accumbens. J Neurosci 17:6899–6907.

Niv Y, Daw ND, Joel D, Dayan P (2007) Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl) 191:507–520.

Niv Y, Edlund JA, Dayan P, O'Doherty JP (2012) Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. J Neurosci 32:551–562.

Niv Y, Joel D, Dayan P (2006) A normative perspective on motivation. Trends Cogn Sci 10:375–381.

Oades RD (1985) The role of noradrenaline in tuning and dopamine in switching between signals in the CNS. Neurosci Biobehav Rev 9:261–282.

Obeso JA, Rodriguez-Oroz MC, Rodriguez M, Lanciego JL, Artieda J, Gonzalo N, Olanow CW (2000) Pathophysiology of the basal ganglia in Parkinson's disease. Trends Neurosci 23:S8–S19.

Oh Y, Heien ML, Park C, Kang YM, Kim J, Boschen SL, Shin H, Cho HU, Blaha CD, Bennet KE, Lee HK, Jung SJ, Kim IY, Lee KH, Jang DP (2018) Tracking tonic dopamine levels in vivo using multiple cyclic square wave voltammetry. Biosens Bioelectron 121.

Ohta H, Satori K, Takarada Y, Arake M, Ishizuka T, Morimoto Y, Takahashi T (2021) The asymmetric learning rates of murine exploratory behavior in sparse reward environments. Neural Networks 143.

Ostlund SB, Balleine BW (2005) Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. J Neurosci 25.

Otomo K, Perkins J, Kulkarni A, Stojanovic S, Roeper J, Paladini CA (2020) In vivo patch-clamp recordings reveal distinct subthreshold signatures and threshold dynamics of midbrain dopamine neurons. Nat Commun 11.

Otto AR, Gershman SJ, Markman AB, Daw ND (2013a) The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. Psychol Sci 24.

Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013b) Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci U S A 110.

Otto AR, Skatova A, Madlon-Kay S, Daw ND (2015) Cognitive Control Predicts Use of Model-Based Reinforcement-Learning. J Cogn Neurosci 27:319.

Owesson-White CA, Roitman MF, Sombers LA, Belle AM, Keithley RB, Peele JL, Carelli RM, Wightman RM (2012) Sources contributing to the average extracellular concentration of dopamine in the nucleus accumbens. J Neurochem 121:252–262.

Palacios ER, Houghton C, Chadderton P (2021) Accounting for uncertainty: inhibition for neural inference in the cerebellum. Proceedings Biol Sci 288:20210276.

Paladini CA, Roeper J (2014) Generating bursts (and pauses) in the dopamine midbrain neurons. Neuroscience 282.

Pan W-X, Schmidt R, Wickens JR, Hyland BI (2005) Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J Neurosci Off J Soc Neurosci 25:6235–6242.

Pan WX, Coddington LT, Dudman JT (2021) Dissociable contributions of phasic dopamine activity to reward and prediction. Cell Rep 36.

Parker NF, Cameron CM, Taliaferro JP, Lee J, Choi JY, Davidson TJ, Daw ND, Witten IB (2016) Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nat Neurosci 19:845–854.

Patriarchi T, Cho JR, Merten K, Howe MW, Marley A, Xiong W-H, Folk RW, Broussard GJ, Liang R, Jang MJ, Zhong H, Dombeck D, von Zastrow M, Nimmerjahn A, Gradinaru V, Williams JT, Tian L (2018) Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. Science (80-) 360:eaat4422.

Pavlov IP (1927) Conditioned reflexes. An investigation of the physiological activity of the cerebral cortex. Translated and edited by E. V. ANREP. Cond reflexes An Investig Physiol Act Cereb cortex Transl Ed by E V ANREP.

Phillips AG, Carter DA, Fibiger HC (1976) Dopaminergic substrates of intracranial self-stimulation in the caudate-putamen. Brain Res 104.

Phillips PEM, Stuber GD, Heien MLA V., Wightman RM, Carelli RM (2003) Subsecond dopamine release promotes cocaine seeking. Nature 422:614–618.

Popa LS, Ebner TJ (2018) Cerebellum, Predictions and Errors. Front Cell Neurosci 12:524.

Poulin J-F, Caronia G, Hofer C, Cui Q, Helm B, Ramakrishnan C, Chan CS, Dombeck DA, Deisseroth K, Awatramani R (2018) Mapping projections of molecularly defined dopamine neuron subtypes using intersectional genetic approaches. Nat Neurosci:1.

Puryear CB, Kim MJ, Mizumori SJY (2010) Conjunctive encoding of movement and reward by ventral tegmental area neurons in the freely navigating rodent. Behav Neurosci 124:234–247.

Rao SM, Bobholz JA, Hammeke TA, Rosen AC, Woodley SJ, Cunningham JM, Cox RW, Stein EA, Binder JR (1997) Functional MRI evidence for subcortical participation in conceptual reasoning skills. Neuroreport 8.

Rescorla RA (1969) Pavlovian conditioned inhibition. Psychol Bull 72:77–94.

Rescorla RA (1970) Reduction in the effectiveness of reinforcement after prior excitatory conditioning. Learn Motiv 1.

Rescorla RA, Solomom RL (1967) Two-Process Learning Theory: Relationships between Pavlovian Conditioning and Instrumental Learning. Psychol Rev 74.

Rescorla RA, Wagner A (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical Conditioning II: Current Research and Theory (Black A, Prokasy W, eds), pp 64–99. New York, NY: Appleton-Century-Crofts.

Reynolds JNJ, Avvisati R, Dodson PD, Fisher SD, Oswald MJ, Wickens JR, Zhang YF (2022) Coincidence of cholinergic pauses, dopaminergic activation and depolarisation of spiny projection neurons drives synaptic plasticity in the striatum. Nat Commun 13.

Rice ME (2000) Distinct regional differences in dopamine-mediated volume transmission. In: Progress in Brain Research.

Rice ME, Cragg SJ (2004) Nicotine amplifies reward-related dopamine signals in striatum. Nat Neurosci 7.

Rice ME, Cragg SJ (2008) Dopamine spillover after quantal release: Rethinking dopamine transmission in the nigrostriatal pathway. Brain Res Rev 58.

Richfield EK, Penney JB, Young AB (1989) Anatomical and affinity state comparisons between dopamine D1 and D2 receptors in the rat central nervous system. Neuroscience 30:767–777.

Robbins TW, Everitt BJ (1982) Functional studies of the central catecholamines. Int Rev Neurobiol 23:303–365.

Roeper J (2013) Dissecting the diversity of midbrain dopamine neurons. Trends Neurosci

36:336–342.

Roitman MF, Stuber GD, Phillips PEM, Wightman RM, Carelli RM (2004) Dopamine operates as a subsecond modulator of food seeking. J Neurosci 24:1265–1271.

Romanes GJ (1886) Animal intelligence / by George J. Romanes. | Wellcome Collection, 4th ed. London: K. Paul, Trench & co.

Romo R, Schultz W (1990) Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. J Neurophysiol 63.

Rothenhoefe KM, Costa VD, Bartolo R, Vicario-Feliciano R, Murray EA, Averbeck BB (2017) Effects of ventral striatum lesions on stimulus-based versus action-based reinforcement learning. J Neurosci 37.

Rozeboom WW (1957) Secondary extinction of lever-pressing behavior in the albino rat. J Exp Psychol 54.

Rutledge RB, Dean M, Caplin A, Glimcher PW (2010) Testing the reward prediction error hypothesis with an axiomatic model. J Neurosci 30:13525–13536.

Sadacca BF, Jones JL, Schoenbaum G (2016) Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. Elife 5.

Salamone J, Correa M, Mingote S, Weber S (2005) Beyond the reward hypothesis: alternative functions of nucleus accumbens dopamine. Curr Opin Pharmacol 5:34–41.

Salamone JD, Correa M (2002) Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. Behav Brain Res 137:3–25.

Salamone JD, Correa M (2012) The mysterious motivational functions of mesolimbic dopamine. Neuron 76:470–485.

Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of Action-Specific Reward

Values in the Striatum. Science (80-) 310:1337–1340.

Santo-Angles A, Fuentes-Claramonte P, Argila-Plaza I, Guardiola-Ripoll M, Almodóvar-Payá C, Munuera J, McKenna PJ, Pomarol-Clotet E, Radua J (2021) Reward and fictive prediction error signals in ventral striatum: asymmetry between factual and counterfactual processing. Brain Struct Funct 226.

Sarafyazd M, Jazayeri M (2019) Hierarchical reasoning by neural circuits in the frontal cortex. Science (80-) 364.

Saunders A, Oldenburg IA, Berezovskii VK, Johnson CA, Kingery ND, Elliott HL, Xie T, Gerfen CR, Sabatini BL (2015) A direct GABAergic output from the basal ganglia to frontal cortex. Nature 521.

Schlerf J, Ivry RB, Diedrichsen J (2012) Encoding of Sensory Prediction Errors in the Human Cerebellum. J Neurosci 32:4913 LP – 4922.

Schuck NW, Cai MB, Wilson RC, Niv Y (2016) Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. Neuron 91:1402–1412.

Schultz W (1986) Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. J Neurophysiol 56.

Schultz W (1998) Predictive Reward Signal of Dopamine Neurons. J Neurophysiol 80:1–27.

Schultz W (2002) Getting formal with dopamine and reward. Neuron 36:241–263.

Schultz W, Apicella P, Ljungberg T (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neurosci 13.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.

Schultz W, Dayan P, Montague RR (2003) A Neural Substrate of Prediction and Reward.

Science (80-) 275:1593–1599.

Schultz W, Romo R (1990) Dopamine neurons of the monkey midbrain: Contingencies of responses to stimuli eliciting immediate behavioral reactions. J Neurophysiol 63.

Schultz W, Ruffieux A, Aebischer P (1983) The activity of pars compacta neurons of the monkey substantia nigra in relation to motor activation. Exp Brain Res 51:377–387.

Sebold M, Deserno L, Nebe S, Schad DJ, Garbusow M, Hägele C, Keller J, Jünger E, Kathmann N, Smolka M, Rapp MA, Schlagenhauf F, Heinz A, Huys QJM (2014) Model-based and model-free decisions in alcohol dependence. Neuropsychobiology 70:122–131.

Sharpe MJ, Batchelor HM, Mueller LE, Yun Chang C, Maes EJP, Niv Y, Schoenbaum G (2020) Dopamine transients do not act as model-free prediction errors during associative learning. Nat Commun 11:106.

Sharpe MJ, Chang CY, Liu MA, Batchelor HM, Mueller LE, Jones JL, Niv Y, Schoenbaum G (2017) Dopamine transients are sufficient and necessary for acquisition of model-based associations. Nat Neurosci 20:735–742.

Simon DA, Daw ND (2011) Neural correlates of forward planning in a spatial decision task in humans. J Neurosci 31.

Singmann H, Bolker B, Westfall J, Aust F (2018) afex': analysis of factorial experiments. R package.

Smittenaar P, FitzGerald THB, Romei V, Wright ND, Dolan RJ (2013) Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. Neuron 80:914.

Soares-Cunha C, Coimbra B, Sousa N, Rodrigues AJ (2016) Reappraising striatal D1- and D2-neurons in reward and aversion. Neurosci Biobehav Rev 68.

Soares S, Atallah B V., Paton JJ (2016) Midbrain dopamine neurons control judgment of time.

Science (80-) 354:1273–1277.

Starkweather CK, Babayan BM, Uchida N, Gershman SJ (2017) Dopamine reward prediction errors reflect hidden-state inference across time. Nat Neurosci 20:581–589.

Starkweather CK, Gershman SJ, Uchida N (2018) The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. Neuron 98:616-629.e6.

Stauffer WR, Lak A, Yang A, Borel M, Paulsen O, Boyden ES, Schultz W (2016) Dopamine Neuron-Specific Optogenetic Stimulation in Rhesus Macaques. Cell 166:1564-1571.e6.

Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH (2013) A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci 16:966–973.

Stopper CM, Tse MTL, Montes DR, Wiedman CR, Floresco SB (2014) Overriding phasic dopamine signals redirects action selection during risk/reward decision making. Neuron 84:177–189.

Sulzer D, Cragg SJ, Rice ME (2016) Striatal dopamine neurotransmission: Regulation of release and uptake. Basal Ganglia 6.

Sutton RS, Barto AG (1998) Reinforcement learning : an introduction. MIT Press.

Sutton RS, Barto AG (2018) Reinforcement learning: An introduction. MIT press.

Syed ECJ, Grima LL, Magill PJ, Bogacz R, Brown P, Walton ME (2016) Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat Neurosci 19:34–36.

Takahashi YK, Batchelor HM, Liu B, Khanna A, Morales M, Schoenbaum G (2017) Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. Neuron 95:1395-1405.e3.

Takahashi YK, Langdon AJ, Niv Y, Schoenbaum G (2016) Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum. Neuron 91:182–193.

Taylor JR, Robbins TW (1986) 6-Hydroxydopamine lesions of the nucleus accumbens, but not of the caudate nucleus, attenuate enhanced responding with reward-related stimuli produced by intra-accumbens d-amphetamine. Psychopharmacology (Berl) 90:390–397.

Thorndike EL (1898) Animal intelligence: An experimental study of the associative processes in animals. Psychol Rev 2:i–109.

Thorndike EL (1911) Animal intelligence. New York.

Threlfell S, Cragg SJ (2011) Dopamine signaling in dorsal versus ventral striatum: the dynamic role of cholinergic interneurons. Front Syst Neurosci 5:11.

Threlfell S, Lalic T, Platt NJ, Jennings KA, Deisseroth K, Cragg SJ (2012) Striatal Dopamine Release Is Triggered by Synchronized Activity in Cholinergic Interneurons. Neuron 75:58–64.

Tian J, Huang R, Cohen JY, Osakada F, Kobak D, Machens CK, Callaway EM, Uchida N, Watabe-Uchida M (2016) Distributed and Mixed Information in Monosynaptic Inputs to Dopamine Neurons. Neuron 91:1374–1389.

Tobia MJ, Guo R, Schwarze U, Boehmer W, Gläscher J, Finckh B, Marschner A, Büchel C, Obermayer K, Sommer T (2014) Neural systems for choice and valuation with counterfactual learning signals. Neuroimage 89.

Tolman EC (1933) Sign-Gestalt or conditioned reflex. Psychol Rev 40.

Tolman EC (1938) The determiners of behavior at a choice point. Psychol Rev 45:41.

Tolman EC (1948) Cognitive maps in rats and men. Psychol Rev 55:189–208.

Tolman EC (1949) There is more than one kind of learning. Psychol Rev 56:144–155.

Tolman EC, Gleitman H (1949) Studies in learning and motivation: I. Equal reinforcements in both end-boxes, followed by shock in one end-box. J Exp Psychol 39.

Trapold MA, Overmier JB (1972) The second learning process in instrumental learning. Curr

Res theory 2:427–452.

Tricomi E, Balleine BW, O'Doherty JP (2009) A specific role for posterior dorsolateral striatum in human habit learning. Eur J Neurosci 29:2225–2232.

Tritsch NX, Ding JB, Sabatini BL (2012) Dopaminergic neurons inhibit striatal output through non-canonical release of GABA. Nature 490.

Tsai H-C, Zhang F, Adamantidis A, Stuber GD, Bonci A, de Lecea L, Deisseroth K (2009) Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science 324:1080–1084.

Tsutsui-Kimura I, Matsumoto H, Akiti K, Yamada MM, Uchida N, Watabe-Uchida M (2020) Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. Elife 9.

Ungless MA, Magill PJ, Bolam JP (2004) Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. Science 303:2040–2042.

Van Hoeck N, Watson PD, Barbey AK (2015) Cognitive neuroscience of human counterfactual reasoning. Front Hum Neurosci 9.

Verharen JPH, Adan RAH, Vanderschuren LJMJ (2019) Differential contributions of striatal dopamine D1 and D2 receptors to component processes of value-based decision making. Neuropsychopharmacology 44:2195–2204.

Verharen JPH, Zhu Y, Lammel S (2020) Aversion hot spots in the dopamine system. Curr Opin Neurobiol 64:46–52.

Vertechi P, Lottem E, Sarra D, Godinho B, Treves I, Quendera T, Oude Lohuis MN, Mainen ZF (2020) Inference-Based Decisions in a Hidden State Foraging Task: Differential Contributions of Prefrontal Cortical Areas. Neuron 106:166-176.e6.

Voon V, Derbyshire K, Rück C, Irvine MA, Worbe Y, Enander J, Schreiber LRN, Gillan C,

Fineberg NA, Sahakian BJ, Robbins TW, Harrison NA, Wood J, Daw ND, Dayan P, Grant JE, Bullmore ET (2014) Disorders of compulsivity: a common bias towards learning habits. Mol Psychiatry 2015 203 20:345–352.

Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. Nature 412:43–48.

Wagner AR (1969) Stimulus validity and stimulus selection in associative learning. Fundam issues Assoc Learn:90–122.

Wagner MJ, Kim TH, Savall J, Schnitzer MJ, Luo L (2017) Cerebellar granule cells encode the expectation of reward. Nature 544:96–100.

Walton ME, Behrens TEJ, Buckley MJ, Rudebeck PH, Rushworth MFS (2010) Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron 65:927–939.

Walton ME, Kennerley SW, Bannerman DM, Phillips PEM, Rushworth MFS (2006) Weighing up the benefits of work: Behavioral and neural analyses of effort-related decision making. Neural Networks 19:1302–1314.

Wang JX, Kurth-Nelson Z, Kumaran D, Tirumala D, Soyer H, Leibo JZ, Hassabis D, Botvinick M (2018) Prefrontal cortex as a meta-reinforcement learning system. Nat Neurosci 21:860–868.

Warden CJ (1927) The historical development of comparative psychology. Psychol Rev 34.

Warden CJ (1930) A Note on the Early History of Experimental Methods in Comparative Psychology. Pedagog Semin J Genet Psychol 38:466.

Washburn S, Oñate M, Yoshida J, Vera J, K. B. R, Khatami L, Nadim F, Khodakhah K (2022) Cerebellum Directly Modulates the Substantia Nigra Dopaminergic Activity. bioRxiv:2022.05.20.492532.

Wassum K, Ostlund S, Maidment N (2012) Phasic mesolimbic dopamine signaling precedes and predicts performance of a self-initiated action sequence task. Biol Psychiatry 71.

Watabe-Uchida M, Eshel N, Uchida N (2017) Neural Circuitry of Reward Prediction Error. Annu Rev Neurosci 40:373–394.

Watabe-Uchida M, Zhu L, Ogawa SK, Vamanrao A, Uchida N (2012) Whole-Brain Mapping of Direct Inputs to Midbrain Dopamine Neurons. Neuron 74.

Watkins CJCH (1989) Learning from delayed rewards.

Weiss F, Markou A, Lorang MT, Koob GF (1992) Basal extracellular dopamine levels in the nucleus accumbens are decreased during cocaine withdrawal after unlimited-access self-administration. Brain Res 593.

Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD (2014) Humans use directed and random exploration to solve the explore-exploit dilemma. J Exp Psychol Gen 143:2074–2081.

Wimmer GE, Daw ND, Shohamy D (2012) Generalization of value in reinforcement learning by humans. Eur J Neurosci 35:1092–1104.

Witten IB, Steinberg EE, Lee SY, Davidson TJ, Zalocusky KA, Brodsky M, Yizhar O, Cho SL, Gong S, Ramakrishnan C, Stuber GD, Tye KM, Janak PH, Deisseroth K (2011) Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron 72:721–733.

Wittmann BC, Daw ND, Seymour B, Dolan RJ (2008) Striatal activity underlies novelty-based choice in humans. Neuron 58:967–973.

Wolpert DM, Kawato M (1998) Multiple paired forward and inverse models for motor control. Neural Networks 11.

Wolpert DM, Miall RC, Kawato M (1998) Internal models in the cerebellum. Trends Cogn Sci 2.

Wunderlich K, Dayan P, Dolan RJ (2012a) Mapping value based planning and extensively trained choice in the human brain. Nat Neurosci 15.

Wunderlich K, Smittenaar P, Dolan RJ (2012b) Dopamine enhances model-based over model-free choice behavior. Neuron 75:418–424.

Wyckmans F, Otto AR, Sebold M, Daw N, Bechara A, Saeremans M, Kornreich C, Chatard A, Jaafari N, Noël X (2019) Reduced model-based decision-making in gambling disorder. Sci Rep 9:19625.

Yapo C, Nair AG, Clement L, Castro LR, Hellgren Kotaleski J, Vincent P (2017) Detection of phasic dopamine by D1 and D2 striatal medium spiny neurons. J Physiol 595.

Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci 19.

Yin HH, Knowlton BJ, Balleine BW (2006) Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. Behav Brain Res 166.

Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci 22.

Yuen J, Goyal A, Rusheen AE, Kouzani AZ, Berk M, Kim JH, Tye SJ, Blaha CD, Bennet KE, Jang DP, Lee KH, Shin H, Oh Y (2021) Cocaine-Induced Changes in Tonic Dopamine Concentrations Measured Using Multiple-Cyclic Square Wave Voltammetry in vivo. Front Pharmacol 12.

Yves K, Jérôme F, Clément R, Christian L (2018) VTA dopamine neurons multiplex external with internal representations of goal-directed action. bioRxiv:408062.

Zhang H, Sulzer D (2004) Frequency-dependent modulation of dopamine release by nicotine. Nat Neurosci 7.

Zweifel LS, Parker JG, Lobb CJ, Rainwater A, Wall V, Fadok JP, Darvas M, Kim MJ, Mizumori SJY, Paladini CA, Phillips PEM, Palmiter RD (2009) Disruption of NMDAR-dependent burst firing by dopamine neurons provides selective assessment of phasic dopamine-dependent behavior. Proc Natl Acad Sci U S A 106.

Zylberberg A, Wolpert DM, Shadlen MN (2018) Counterfactual Reasoning Underlies the Learning of Priors in Decision Making. Neuron 99.