Learning about the distribution of rewards through dopamine and the basal ganglia pathways

Moritz Möller

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Acknowledgements

Personal acknowledgements

This thesis would not exist without the relentless support and nurturing of my supervisors, Prof. Rafal Bogacz and Prof. Sanjay Manohar. I was tremendously lucky to have their expert guidance on my first poster, my first paper, my first peer-review and my first conference. Thank you for your encouragement, your knowledge and your mentorship. Thanks also go to Prof. Michael Browning for several inspiring discussions—it was he who suggested to work out the prediction I describe in chapter 3, subsection 3.3.3.

I had the great pleasure to work with a group of exceptional colleagues. I want to thank them for the camaraderie, the advice and support and the inspiring discussions we had in the lab, over lunch and overseas at conferences. Special thanks go to Dr Benoît Duchet (my academic older brother) and to Jan Grohn (my favourite collaborator). In particular, I want to thank Benoît for showing me how to do things right at Oxford (teaching, research and beyond), and Jan for the energy and the creativity he contributed to our project, as well as for collecting the data that chapter 5 of this thesis is based on. I also want to thank my friend Mary Sanford for helping me proof-read this thesis, and to acknowledge Dr John McManigle who's thesis template I found very useful.

Moving from academics to institutions, I'd first like to express my deep gratitude to the Medical Research Council for the studentship that allowed me to conduct the research for this thesis. Further, I'd like to thank all the members of the Brain Network Dynamics Unit and the Nuffield Department of Clinical Neuroscience, in particular Vicky Anderton, Savita Anderson and Dr Vivienne Collins, who made the administrative aspects of my degree seem like a breeze. Thanks also go to St Cross College of which I am a proud member.

To conclude, I'd like to thank my family and friends in Oxford and abroad: I could not have done it without you! Thanks for your support, friendship and advice in good and challenging times. Because of you I will always most fondly remember my time as a doctoral student at the University of Oxford. And finally, thank you to Magdalena Hasmüller, for all her support.

Published materials

I have published several of the results I report in this thesis. Most results reported in chapter 3 (in particular those in section 3.2 and in subsections 3.3.1 and 3.3.2), as well as those in appendix A, were published by Möller and Bogacz (2019). The results reported in chapter 5 and appendix D are under review at the time of writing; a preprint version was made available by Möller, Grohn, et al. (2021).

I used materials (text and figures) from these publications in the indicated sections of this thesis. I hereby confirm that I wrote/produced all the materials that I use there. I further confirm that the other authors have given me written permission to include the materials in my thesis.

Abstract

The basal ganglia pathways and their dopaminergic innervation are essential for reinforcement learning and are thought to play an important role in value-based decision making. Yet, it is not fully understood how these components interact and which computations they perform. In this thesis, we aim to improve this understanding, building on the Actor learning Uncertainty (AU) hypothesis of Mikhael and Bogacz (2016). According to this hypothesis, the basal ganglia might learn statistics (averages and uncertainties) of reward distributions and use them to inform choices, with dopamine acting as a teaching signal as well as a modulator.

First, we reapply the AU model, shifting the focus from reward uncertainty to tasks in which subjects need to work for rewards. We test whether the model can learn the payoffs and costs of actions and whether it can reproduce the effects of dopaminergic medication on the willingness to work for food. Second, we develop a new variant of the AU model—the scaled prediction error model—to understand how tracking uncertainty can improve learning. We investigate the new model's performance, biological plausibility and consistency with behavioural data. Third, we test the AU model empirically, using behaviour in a learning task. The model predicts that there should be an association between unexpected rewards and subsequent risk-seeking. We test that prediction, controlling for alternative explanations using trial-by-trial modelling.

We find that the AU model can learn the payoffs and costs of actions and reproduce related behavioural effects. We further find that tracking uncertainty can improve learning if the estimate of uncertainty is used to scale reward prediction errors. This also makes the model consistent with the empirical phenomenon of dopaminergic range adaptation. Finally, we confirm the AU model's prediction: unexpected rewards are associated with risk-seeking. Our results consolidate and extend our theoretical understanding of the role that dopamine and the basal ganglia play in learning about the distribution of rewards.

Contents

List of Abbreviations

AC . . . . . Actor-Critic

AI . . . . . Artificial Intelligence

AU . . . . . Actor learning Uncertainty

BG . . . . . Basal Ganglia

BMI . . . . . Brain-Machine Interface

CP . . . . . Caudate-Putamen

DDM . . . . . Drift-Diffusion Model

EEG . . . . . Electro-Encephalogram

EPSP . . . . . Excitatory PostSynaptic Potential

FMRI . . . . . Functional Magnetic Resonance Imaging

GBP . . . . . British pound sterling

GPe . . . . . Globus Pallidus, external segment

GPi . . . . . Globus Pallidus, internal segment

ITI . . . . . Inter-Trial Interval

NAc . . . . . Nucleus Accumbens

OEIRS . . . . . Outcome Prediction Errors Induce Risk Seeking

PEIRS . . . . . Prediction Errors Induce Risk Seeking

PIRS . . . . . Predictions Induce Risk Seeking

PD . . . . . Parkinson's Disease

PE . . . . . Prediction Error

RL . . . . . Reinforcement Learning

RPE . . . . . Reward Prediction Error

SE . . . . . Standard Error

SNc . . . . . Substantia Nigra pars compacta

SNr . . . . . Substantia Nigra pars reticulata

SPN . . . . . Spiny Projection Neuron.

STDP . . . . . Spike-Timing Dependent Plasticity

STN . . . . . SubThalamic Nucleus

TD . . . . . Temporal Difference

VTA . . . . . Ventral Tegmental Area

1 Introduction

Contents

1.1 How I became a neuroscientist

I entered the field of computational neuroscience in 2017. My background was in physics, but I had also studied reinforcement learning algorithms and their mathematical foundations. What fascinated me about reinforcement learning was how general these algorithms were, and how little domain knowledge one needed to inject into them. Feed a naive agent with images, hand over control, and before long (and without any other help!) it will play Pac-Man better than you.

The way programmers like Mnih, Kavukcuoglu, et al. (2015) use reward functions to guide their reinforcement learning agents reminded me of the way Ferster and Skinner (1957) used food reinforcements to shape the behaviour of their pigeons. It seemed logical that some process in the brain should resemble a

reinforcement learning algorithm. So when I finally learned about the firing patterns of dopamine neurons and their similarity to the patterns of reward prediction errors in the temporal-difference learning algorithm, I had a feeling of deep satisfaction: human reasoning and evolution had come to the same conclusion. There was a direct correspondence between a variable in an algorithm and a neural signal measured in the brain of a behaving animal.

Such uncanny correspondences between mathematical models and the natural world were familiar to me as a physicist, but working out the details of, say, the cosmic microwave background did not seem nearly as exciting as working out the links between artificial and biological intelligence. Consequently, when the time came to choose a PhD topic I decided to leave physics and go to work in computational neuroscience. In this thesis, I report the results of this work.

1.2 The goals and contributions of this thesis

We focus on the neuromodulator dopamine and the direct and indirect pathways in the basal ganglia circuit. Our central goal is to better understand how these structures interact and which computations they perform together. In particular, we want to understand their role in learning and decision making: how are the neural structures shaped by experience, and how do they affect choices?

Our starting point is a recent hypothesis, the Actor learning Uncertainty (AU) model. According to that model, the major basal ganglia pathways (direct and indirect) might learn and store statistics of reward distributions associated with actions and stimuli, through dopamine-dependent plasticity in the striatum (Mikhael and Bogacz 2016). These statistics—which encode reward uncertainties as well as reward averages—might then be used flexibly to inform choices between actions, with dopamine balancing the relative impact of different statistics. Could this be the computational role of the basal ganglia? In this thesis, we address that question in three related research projects.

The first project (chapter 3) revolves around a new interpretation of the reward statistics that the basal ganglia might learn, first proposed by Bogacz (2017b). Instead of focusing on average reward and reward uncertainty as Mikhael and Bogacz (2016), we will focus on payoffs and costs, i.e., the positive and negative consequences of actions. For example, in a typical experimental setting a rat might press a lever to obtain a food pellet. The costs of that behaviour might be physical exertion, the payoff might derive from the calories in the food. Could these two values—the payoff and the cost—be learned from experience, and stored in the basal ganglia pathways? And if so, how would they influence decisions, and what role does dopamine play in weighing them against each other? In this project, we look at an existing model in a new way, applying it to tasks it was not originally designed for, and test what predictions it makes in that new domain. This includes a mathematical analysis as well as simulations and fits to empirical data.

In the second project (chapter 4), we try to understand how tracking reward uncertainty can improve reinforcement learning performance. The AU model tells us how uncertainty can be learned, but not how it should be used or why it is needed. In chapter 4, we show that learned estimates of uncertainty can be used to scale prediction errors. We demonstrate how such scaling follows from Bayesian principles and that it improves learning performance, especially in very unpredictable environments. We further show that scaled prediction errors might explain empirical findings concerning the scaling of dopaminergic prediction errors with reward magnitude (Tobler et al. 2005). The main result of this chapter is the scaled prediction error model, which can be seen as an updated version of the AU model.

The third project (chapter 5) is dedicated to testing a prediction of the AU model: unexpected rewards should induce risk-seeking. This follows from how uncertainty is represented in the basal ganglia pathways and from the fact that dopamine acts as a modulator as well as a teaching signal. Even though

these premises concern the neural level, we argue that their consequences should be observable in behaviour. To test this prediction, we will use trial-by-trial modelling to analyse a behavioural dataset from a reinforcement learning task.

Ultimately, all these projects contribute to further develop a computational model of the basal ganglia circuit—the AU model of Mikhael and Bogacz (2016). We explore new applications of the model in chapter 3, develop a new version of it in chapter 4 and test its predictions empirically in chapter 5. The more theoretical chapters (chapters 3 and 4) come first, the more empirical chapter (chapter 5) comes last.

All chapters are self-contained: they include detailed descriptions of the relevant models and procedures, as well as discussions of the results. They are preceded by a review of the relevant literature in chapter 2 and followed by a general discussion in chapter 6. In the literature review, we provide an overview of what is already known in our area of research. That way we establish the framework in which we work. In the general discussion, we conclude the thesis by challenging this framework and exploring alternative perspectives.

2

Literature review

Contents

In this chapter, we review what is known about the topic of this thesis, which

is the role of dopamine and the basal ganglia in learning and choice. We group the relevant literature into four main themes—reward prediction errors and dopamine in section 2.1, other effects of dopamine in section 2.2, the structure and function of the basal ganglia in section 2.3 and the mechanisms of basal ganglia learning in section 2.4. For each theme, we aim to present the key established facts and theories, as well as relevant recent research, with a particular focus on modelling work.

2.1 Reward prediction errors and dopamine

Our first theme revolves around a concept that is central to reinforcement learning: the reward prediction error (RPE). We will see that RPEs appear in models of animal learning (subsection 2.1.1), in machine learning (subsection 2.1.2) and in neuroscience (subsection 2.1.3). We will present the reward prediction error hypothesis of dopamine which unifies these different perspectives and review research inspired by that hypothesis in subsection 2.1.4. Many reviews discuss these topics; here, we followed Watabe-Uchida et al. (2017).

2.1.1 Reward prediction errors in animal learning

The prediction error concept first appeared as an explanation of certain puzzling effects in classical conditioning experiments. At the time, it was thought that conditioning could be simply summarised as "One stimulus evokes the original response to another because of their pairing" (Rescorla 1988).

However, some experimental designs (such as the blocking design) indicated that mere pairing is not enough to change the associations between stimuli. It became clear that associations between stimuli change only if *unexpected* events happen. For example, unexpected delivery of food after the sound of a bell may increase the association between the bell's sound and food. If the delivery of food could be predicted (perhaps using another cue), it loses its ability to change associations, and the pairing of food and bell is ineffective.

This theory is formalised in the immensely influential Rescorla Wagner model (RW, Rescorla 1972), in which the difference between the expected and experienced stimulus intensity—the so-called prediction error—drives learning. For example, let r be the intensity of a rewarding stimulus, and let v be the intensity that is expected. The RW model then suggests that presenting the stimulus should change v by

$$\Delta v = \alpha\delta, \tag{2.1}$$

with α the learning rate and $\delta = r - v$ the prediction error (in this case, it is a reward prediction error). If the reward intensity is as expected, then δ is zero and v does not change—no learning happens.

This idea applies to Pavlovian conditioning experiments in which the outcome does not depend on the animal's response; it also applies to instrumental conditioning where the outcome is contingent on the animal's response.

2.1.2 Reward prediction errors in artificial intelligence

While psychologists studied animal learning, computer scientists developed reinforcement learning (RL), which is now one of the main subfields of artificial intelligence (Sutton and Barto 2018). According to the RL paradigm, a programmer should be able to solve a problem merely by defining goals, which means specifying a reward signal. The reinforcement learning algorithm would then learn how to best reach those goals through interaction with its environment, using a trial-and-error approach.

RL algorithms often use so-called state values, which measure how good a certain state of the environment is for reaching the given goal. State values must be learned from experience, in a process called value learning. The temporal difference (TD) learning algorithm is an example of a value learning method.

TD learning is particularly interesting here because it uses reward prediction errors to compute updates of the value prediction system: if at any point the predicted value is inconsistent with experienced rewards, the inconsistency is quantified and used to compute an update for the prediction system. If the state value is predicted correctly, the prediction error is zero and no learning takes place. The absence of learning in the absence of prediction errors is a feature that TD learning shares with the RW model we described above.

Reinforcement learning algorithms received a lot of attention recently: powered by a combination of improved RL algorithms and fast modern hardware, computers reached superhuman performance in complex strategy games such as Go (Mnih, Kavukcuoglu, et al. 2015; Silver et al. 2016). It is now clear that comparatively simple learning methods can lead to sophisticated behavioural strategies, given sufficient amounts of data and processing power.

2.1.3 The neuroscience of conditioning

While psychologists studied animal learning and computer scientists developed reinforcement learning, neuroscientists established fundamental facts about the brain circuitry involved in conditioning. They found that the neurotransmitter dopamine (DA) played a key role in "stamping in" associations and behaviours.

This insight was gained using the intracranial self-stimulation (ICS) paradigm (Wise and Rompre 1989): self-administrated electrical stimulation at precisely localised sites in rat brains was used to generate a map of brain sites that would support sustained self-stimulation. In the resulting map, effective ICS sites often coincided with the locations of dopamine neurons or their axons. This suggested that it was dopamine that reinforced the behaviours that led to stimulation. Later, voltammetry measurements confirmed that dopamine release is necessary for the efficacy of ICS (Garris et al. 1999).

In addition, it was established that blocking dopamine receptors through pharmacological interventions makes self-stimulation ineffective (Wise and Rompre

1989; Wise 2004). Based on all these discoveries, the dopamine system emerged as what was called the "common path for all rewards" (Wise and Rompre 1989). ICS was viewed as a way to directly tap into this reward pathway without a detour over rewarding stimuli such as food. This view is discredited now—more nuanced theories have taken its place. We will review them in the next sections, starting with the reward prediction error hypothesis.

2.1.4 The reward prediction error hypothesis of dopamine

The results from those different fields—animal learning, artificial intelligence and neuroscience—converged in the 1990s. The breakthrough came in the form of electrophysiology in conditioned monkeys. It was discovered that monkeys' dopamine neurons respond to unexpected rewards, but not to rewards predicted by a cue (Mirenowicz and Schultz 1994).

The model that was proposed to explain this effect suggested that dopamine firing corresponds to a reward prediction error signal, very similar to that which occurs in the TD algorithm. According to that model, the brain uses those dopaminergic prediction errors to update a reward prediction system, which is itself the basis of the behaviour observed in conditioning experiments (Montague, Dayan, et al. 1996; Schultz et al. 1997). This synthesis, commonly known as the *reward prediction error hypothesis of dopamine*, is one of the most influential developments in neuroscience in the last decades.

Many studies have investigated and refined aspects of the reward prediction error hypothesis of dopamine since it was first proposed. One line of research aimed at determining the details of the dopamine signal and its precise relationship to rewards. Using electrophysiological recordings from the substantia nigra pars compacta (SNc) of monkeys during a saccade task, it could be confirmed that dopamine cell firing depended on previously experienced rewards just as predicted by the TD error theory (Bayer and Glimcher 2005). We know today that the computation of RPEs involves GABA neurons in the ventral

tegmental area (VTA, Eshel, Bukwich, et al. 2015) and that VTA dopamine neurons homogeneously encode RPEs during Pavlovian conditioning (Eshel, Tian, et al. 2016). It was also shown that dopamine activity predicted choices as expected: using electrophysiological recordings from the SNc of monkeys interacting with a 2-armed bandit, it was shown that predictions of choice outcomes based on dopamine responses were more precise than predictions based on previous rewards. This is consistent with the idea that dopamine might mediate the effects of reward on behaviour (Morris, Nevet, et al. 2006).

Another line of research sought to find out whether the RPE signals observed in animals could also be found in humans. Using a combination of dopaminergic drugs and FMRI in a learning task, RPE signals were found in the human striatum and it was shown that those signals were associated with learning (Pessiglione et al. 2006). Later, direct observations of RPEs in the human VTA were achieved via FMRI (D'Ardenne et al. 2008). More direct evidence was gained using microelectrodes to measure neural signals in the SNc of patients with Parkinson's disease: single neurons could be shown to respond to unexpected rewards (Zaghloul et al. 2009). All this supports dopaminergic reward prediction errors in humans.

More recently, new technologies made it possible to test the causal connection between dopamine and learning: optogenetic stimulation of VTA neurons is sufficient for creating place preferences (Tsai et al. 2009), for overcoming blocking in Pavlovian conditioning (Steinberg et al. 2013), and for creating stimulus-reward associations (Saunders et al. 2018).

By and large, the bulk of the collected evidence suggests that dopamine plays a key role in learning and that the reward prediction error theory makes correct predictions in many situations. However, reward-related dopamine signals do not always behave exactly like TD errors: recordings in monkeys showed that dopamine bursts, unlike the TD errors, do not generally scale with reward magnitude (Tobler et al. 2005). This could mean that dopamine bursts encode

relative or *scaled* prediction errors instead of absolute prediction errors. How such scaling might work and how learning with scaled prediction errors differs from learning from absolute prediction errors is the subject of chapter 4.

2.2 The direct behavioural effects of dopamine

In the previous section, we presented the reward prediction error hypothesis of dopamine. That hypothesis paints the neurotransmitter as a teaching or feedback signal with indirect effects on behaviour—dopamine affects behaviour through changing the neural circuitry as a response to actions and events.

Though being very influential and appealing, this hypothesis is only a part of the dopamine story. Dopamine also has direct, immediate effects on behaviour which are not captured by the reward prediction theory. In this section, we will focus on these effects. We will first look at the role of dopamine in motivation (subsection 2.2.1). Then, we will review the relation between dopamine and risk-taking (subsection 2.2.2). Finally, we will discuss how dopamine may influence behaviour directly and indirectly at the same time (subsection 2.2.3).

2.2.1 Dopamine and motivation

The direct effects of dopamine on behaviour are often linked to the concept of *motivation*. Motivation is a broad and vague notion; we thus start this discussion with a few definitions. Following Salamone and Correa (2012), we define motivated behaviour as behaviour that is purposeful or goal-directed. In particular, this includes movements towards or away from certain stimuli.

Motivated behaviour can often be decomposed into two phases: the approach phase and the consummatory phase. The approach phase (also referred to as *seeking*, *appetitive*, *anticipatory* or *instrumental*) comes first and is characterised by *direction* (towards or away from a stimulus) and *activation* (the speed or vigour of the movement). It is followed by the consummatory phase (also referred to as *taking* or *completion*).

Now, which of these aspects is linked to dopamine? Many studies have contributed to answering this question, often using dopamine depletion in the nucleus accumbens or antagonistic drugs in low doses. The general finding is that dopamine affects the approach phase, in particular by modulating activation (Salamone and Correa 2012). This fits in well with the finding that electrical stimulation of dopamine cells makes animals run quicker and act more vigorously, hence contributing to what is called pre-reward arousal (Wise 2004). It also fits with the finding that dopamine affects wanting, but not liking (Berridge and Robinson 1998).

Together, these results suggest that dopamine does more than the reinforcing of associations we discussed in section 2.1: it also seems to energise approach behaviour directly. We refer to this as the motivation hypothesis of dopamine. A huge literature is attached to this hypothesis (Ikemoto and Panksepp 1999; Salamone, Correa, et al. 2009), influential studies were conducted to work out its details (e.g. Nicola 2010), and it still inspires new research (e.g. Hughes et al. 2020).

Although the motivation hypothesis of dopamine has not been formulated as concisely, parsimoniously and elegantly as the reward prediction error hypothesis, there are still some theoretical frameworks attached to it. For example, one popular view holds that behaviour follows economic principles, such as price and demand (Hursh et al. 1988). Within this framework, it was proposed that dopamine is required to overcome the costs associated with obtaining rewards (Salamone, Correa, et al. 2009). This idea of costs also features prominently in modelling work which attempted to integrate the reward prediction error hypothesis with the motivation hypothesis (Niv, Daw, and Dayan 2006; Niv 2007; Niv, Daw, Joel, et al. 2007). Although the model has been challenged by recent evidence (Zénon et al. 2016), it has shown that introducing effort-dependent costs alongside rewards is a promising way to capture the motivational effects of

dopamine. We will explore this direction in chapter 3, where we investigate how the payoffs and costs of actions might be learned and used for decision-making.

2.2.2 Dopamine and risk-taking

Another direct effect of dopamine on behaviour can be observed when animals are taking risks: dopamine seems necessary to overcome risk-aversion, just as it seems necessary to overcome the aversion to costs.

In rats, it was first shown that dopamine-enhancing medication increases risk-seeking and that dopamine antagonists decrease risk-seeking (St Onge and Floresco 2009). More recently, risk preferences in rats were modulated on the level of single trials: using optogenetic stimulation that mimicked the effect of a dopamine pause just before a choice could bias rats towards safer choices (Zalocusky et al. 2016). Taken together, these results suggest that increasing dopamine levels might cause risk-seeking while decreasing them might cause risk aversion.

In humans, it is known that dopamine-enhancing medication drives excessive gambling in patients with Parkinson's disease (Voon et al. 2006; Gallagher et al. 2007; Weintraub et al. 2010). It has also been demonstrated that phasic responses in dopaminergic brain areas modulate moment-by-moment risk-preference in humans: the tendency to take risks correlated positively with the magnitude of task-related dopamine release in a decision making task (Chew et al. 2019). The effect of dopamine on risk preferences is captured in the Actor learning Uncertainty (AU) model of the basal ganglia which we discuss in detail in section 2.4.

2.2.3 The different roles of dopamine

We have seen that dopamine has direct, immediate effects on behaviour. Among these are the modulation of response vigour and the modulation of risk preferences. On the other hand, in section 2.1 we have reviewed a line of evidence that

suggests that dopamine broadcasts a prediction error signal, which presumably drives neural plasticity without affecting behaviour directly. Dopamine thus seems to play more than one role.

If this is so, do the target structures have a way to distinguish between the different types of dopamine signals, in order to react in the appropriate way? Several theories have been proposed to explain how the direct and indirect functions of dopamine are separated from each other.

One very prominent idea is that the two signals use different timescales: the tonic level of dopamine carries information about motivation, which changes slowly over time, while phasic activity carries learning signals (Niv, Daw, Joel, et al. 2007). This idea was recently challenged by the observation of fast motivational signals (Hamid et al. 2016). Another hypothesis suggests that cholinergic interneurons in the striatum might indicate how dopamine signals should be interpreted (Berke 2018). This idea is supported by the finding that the activity of those interneurons seems to be synchronised with prediction error type activity in dopamine neurons (Morris, Arkadir, et al. 2004).

None of these hypotheses has been confirmed yet. More generally, it is not known whether a separating mechanism between the different functions of dopamine exists at all—interference between the two signals and their behavioural correlates has never been studied systematically, as far as we are aware. In chapter 5, we argue that such interference should be observable on the level of behaviour and present empirical evidence to support that claim.

2.3 The basal ganglia system

We have now reviewed the relevant parts of what is known about dopamine, focusing on its direct and indirect effects on behaviour. But dopamine is only one link in the neural pathways which control learning and action selection. The next link in those pathways—a major target of dopamine signalling—is the basal ganglia system. The basal ganglia are a group of subcortical nuclei which can

be found in all vertebrates. The earliest known organism with basal ganglia is the lamprey; the system has not changed much through evolution (Grillner and B. Robertson 2015).

The signals in the basal ganglia are much more complex than the comparatively homogenous and interpretable dopamine signal. The system has been studied in great detail for many decades; much research has been carried out to understand its structure and function. Hence, our exposition must necessarily be superficial. We recommend the review of Klaus, Alves da Silva, et al. (2019) as a source of further details.

To improve the readability of this section, we structure it into two parts. We start with a short synopsis (subsection 2.3.1) in which we present a simplified model of the basal ganglia. There, we describe only some key elements of the system—the bare minimum required to understand the models in this thesis. Our discussion will be on a high level, and we will focus on brevity instead of completeness. We then move on to provide a more detailed and accurate overview over the basal ganglia component parts in subsections 2.3.2 - 2.3.6.

2.3.1 A simplified model

The models we study in this thesis often assume a simplified, abstracted model of the basal ganglia, with only the most essential features represented. Here, we want to sketch out such a simplified model to serve as a first-order approximation to the system, following Möller and Bogacz (2019). The second-order corrections to that approximation are provided in the subsequent subsections.

For our simplified model of the basal ganglia, we consider only three brain regions: the cortex, the striatum, and the thalamus. We think of those as three layers, arranged along the vertical axis in figure 2.1.

The middle layer—the striatum—is divided into two populations, the D1 and the D2 population (D1 and D2 are the types of dopamine receptors that the corresponding neurons express). This division gives rise to two parallel descending

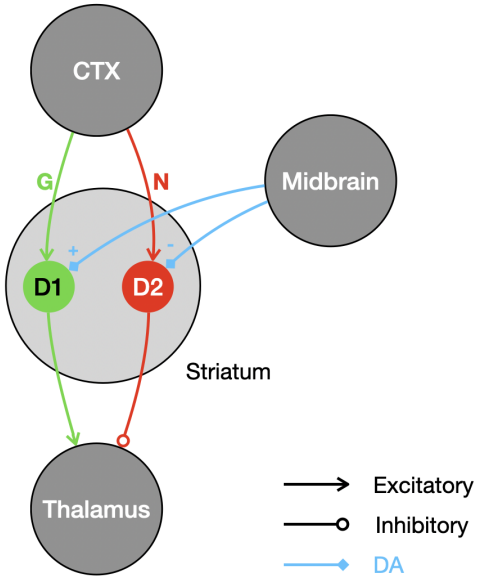

Figure 2.1: The simplified basal ganglia circuit. Selected nuclei and connections are shown as circles and arrows. CTX is short for cortex. Green connections correspond to the direct pathway, red connections correspond to the indirect pathway. Dopamine projections are shown in blue.

pathways called the direct and the indirect pathway, shown in green and red respectively in figure 2.1.

The effects of the direct and the indirect pathway on the thalamus are differential: the direct pathway effectively excites the thalamus; the indirect pathway effectively inhibits it. Note that the projections from the striatum to the thalamus in figure 2.1 are abstractions—in the brain, there are several intermediate nuclei between the striatum and the thalamus (see subsection 2.3.2 and figure 2.2). However, the convergent connectivity and parallel organisation of the basal ganglia makes it possible to think of those projections as two main pathways with opposite effects, at least as a first-order approximation.

The final elements of the simplified basal ganglia model are the dopamine projections from midbrain regions, shown in blue in figure 2.1. We assume dopamine to be a homogenous, global signal that targets the striatal D1 and D2 populations. The effect of dopamine on these populations is differential: it increases excitability and potentiates synapses in the D1 population while it

decreases excitability and depresses synapses in the D2 population (we provide more details on striatal plasticity in section 2.4). This is due to differential receptor expression, which—along with the differential connectivity—defines the two striatal populations.

What is encoded in those three layers? One popular view often used in modelling (Collins and Frank 2014; Mikhael and Bogacz 2016) is that the cortex supplies contextual information, i.e., cues, stimuli, sensory data or information on the state of the environment. The other populations (D1, D2 and Thalamus) encode actions. Each action is represented by a distinct subpopulation of each nucleus, and the connectivity between the nuclei is action specific. For example, assume there is a subpopulation in D1 associated with pressing a lever. A corresponding subpopulation could be found in D2 as well as in the thalamus, which is known to relay motor commands to the relevant cortical areas (Sommer 2003). The two striatal subpopulations associated with the lever press would then project exclusively to the lever-press subpopulation in the thalamus, together forming what is often called an action channel (Redgrave et al. 1999).

We may thus view the two pathways as two distinct state-action mappings—one from cortex to D1, one from cortex to D2. Action specific projections then descend from the striatum and converge at the level of the thalamus. Due to the differential effects of the two pathways on their target, the activity in the thalamus is often considered to be the result of a competition between the direct and the indirect pathway. By modulating the excitability of the striatal populations, dopamine can bias that competition, tilting the balance towards the direct pathway by rising above its baseline level and towards the indirect pathway by falling below.

In summary, our simplified model (which hides most of the internal structure of the basal ganglia system) features two parallel descending pathways that converge on the level of the thalamus, where they have opposite effects. Those pathways are differentially modulated by dopamine and structured into distinct action channels.

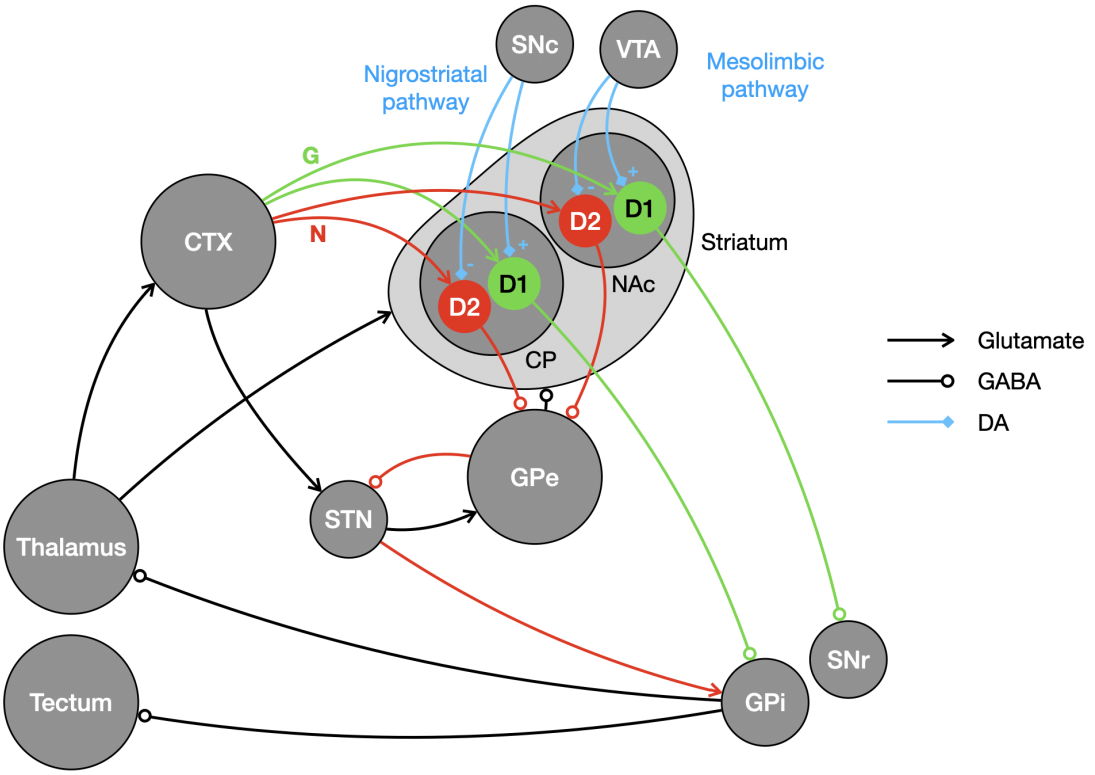

Figure 2.2: The basal ganglia circuit. This diagram shows selected nuclei and the connections between them. The arrows coloured in green correspond to the direct pathway, the red arrows represent the indirect pathway. Dopamine projections are shown in blue.

2.3.2 The functional anatomy

After the high-level synpopsis, we now provide a more detailed overview and unpack the internal structure of the basal ganglia system (see figure 2.2 for a visual representation). Our overview follows Redgrave et al. (1999), who in turn refer to the more extensive treatments of Gerfen and C. J. Wilson (1996) and Mink (1996).

The input nuclei of the basal ganglia are the striatum (already mentioned in our synopsis) and the subthalamic nucleus (STN). Both the striatum and the STN receive extensive excitatory glutamatergic input from the cortex and the thalamus, as well as from the amygdala and the hippocampus[1].

[1]In many descriptions of the basal ganglia, the thalamic input to the striatum is neglected. However, Wolff et al. (2019) recently showed that this input is essential for the execution of learned behaviours (in contrast to cortical inputs, which seem to be essential for learning but not execution). In this thesis, we will assume that like cortex, thalamus provides contextual and

The striatum consists mainly of medium spiny projection neurons (SPNs) which inhibit their targets GABAergically. The STN is much smaller than the striatum (there are about 200 striatal neurons for each STN neuron, Oorschot 1996), and extends excitatory glutamatergic projections.

The output nuclei of the basal ganglia are the internal segment of the globus pallidus (GPi) and substantia nigra pars reticulata (SNr). They project to the thalamus (as pointed out in our synopsis), but also to premotor areas such as the tectum. Those targets in turn project to motor centres further downstream, as well as back to the striatum and the cortex. The basal ganglia are thus embedded in pathways that form loops, as well as in descending pathways that allow the basal ganglia to contribute to motor control directly. The cells in the basal ganglia output nuclei are tonically active and inhibit their targets through GABAergic projections, a phenomenon referred to as *tonic inhibition*.

Finally, the basal ganglia feature an internal nucleus, the external segment of the globus pallidus (GPe). The GPe is at the centre of the basal ganglia and inhibits its targets GABAergically, as do most other parts of the basal ganglia. The internal connectivity of the basal ganglia is highly structured: the striatum and the STN both project to the GPe and directly to the output nuclei. The GPe sends projections back to the striatum and the STN, reciprocating their projections.

One often thinks of the basal ganglia as organised into parallel loops—circuits that are similar in structure but differ in how they are connected to the rest of the brain (Middleton and Strick 2000). On the coarsest level, one distinguishes between the *limbic*, the *associative* and the *motor* loop. Those loops can be thought of as copies of the same circuit that are connected to different inputs and outputs and bundled up in the same structure. One accordingly divides the striatum into the ventral striatum (also called nucleus accumbens, NAc), the dorsolateral striatum and the dorsomedial striatum[2]. The ventral striatum is part of the limbic

sensory information to the striatum. This view is consistent with its role in motor execution.

[2]Taken together, the dorsomedial and dorsolateral striatum form the dorsal striatum, which is also referred to as caudate-putamen (CP).

loop, the dorsomedial striatum is part of the associative loop and the dorsolateral striatum is part of the motor loop.

The loops are called such because they start and end in the same cortical region (McHaffie et al. 2005). For example, a loop might begin with cortical motor areas projecting to the striatum. From there, projections descend through the basal ganglia to thalamic populations, which finally project back to the same cortical motor areas from which the neural pathway started. In their review, Middleton and Strick (2000) further describe that neurons within a certain loop encode loop-specific information, irrespective of which part of the brain they belong to. For example, the neurons in the motor loop might encode information related to movements, whether they are in the cortex, the striatum or the thalamus. More recently, Nambu (2008) noted that there is overlap between the loops, and that a strict separation might thus be unlikely. Overall, it appears as if the basal ganglia, the cortex and the thalamus formed various open and closed loops, though it is not fully understood how these circuits are organised.

Beyond the loop structure, there are also descending pathways that allow the basal ganglia to take part in motor control. This might involve direct connections between the basal ganglia and downstream pre-motor areas, as well as indirect connections between the basal ganglia and motor circuits via the thalamus (Redgrave et al. 1999). In this thesis, we will focus on these pathways, viewing the basal ganglia as a feed-forward network rather than a recurrent network.

2.3.3 The dopaminergic innervation

We have gained an overview of the basal ganglia anatomy. Now, how are dopamine neurons connected to this circuit? A detailed account of the dopamine system is given by Björklund and Dunnett (2007); for us, the key fact is that the striatum is a major target of diffuse dopaminergic projections. These originate in the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc).

The SNc mainly projects to the CP (this is called the nigrostriatal pathway), while the VTA mainly targets the NAc (this is referred to as the mesolimbic pathway). Recent evidence suggests that these two distinct anatomical pathways might reflect distinct functions (Howe and Dombeck 2016), with RPE and motivation signals restricted to the mesolimbic pathway. Beyond this distinction, dopamine signals seem to be targeted not at single cells, but at broad populations. One often speaks of *broadcasting* and a *global* signal.

Dopamine's effect on the striatal SPNs are two-fold: it modulates their activity and controls synaptic plasticity. We discuss both effects, starting with modulation in the next subsection. Plasticity will be the topic of section 2.4 below.

2.3.4 The major pathways

In our short synopsis in subsection 2.3.1, we already introduced the idea that the basal ganglia circuit is split into two major pathways defined by connectivity and neurochemical properties (Gerfen and Surmeier 2011; Smith et al. 1998; Surmeier et al. 2007). This split begins in the striatum, where one finds two SPN populations of about equal size.

The neurons in one of these populations express D1 type dopamine receptors and project to the GPi and the SNr directly. We will refer to these neurons as D1-SPNs; their inhibitory projections to the basal ganglia output nuclei define the direct pathway (the entire direct pathway is marked in green in figure 2.2). By inhibiting the tonically active GPi cells, D1-SPNs can release the tonic inhibition of the thalamus and the tectum.

The cells in the other striatal population express D2 type dopamine receptors and project to the GPe. We will refer to these cells as D2-SPNs. D2-SPNs can increase activity in the output nuclei indirectly via the GPe and the STN, hence increasing the tonic inhibition of the thalamus and the tectum. This multi-synaptic pathway is called the indirect pathway and is marked in red in figure 2.2.

The overall picture suggests that direct and indirect pathway have differential effects on their ultimate target structures (one releases inhibition, the other increases it). This view is generally accepted; however, recent findings suggest that D1-SPNs can actually cause inhibition *and* excitation in SNr populations (Freeze et al. 2013). The same holds for D2-SPNs. This might be due to lateral inhibition on the level of the striatum and emphasises the complexity of the neural circuitry.

Owing to the different dopamine receptors they express, D1-SPNs and D2-SPNs are differentially affected by dopaminergic modulation: dopamine makes D1-SPNs more excitable and D2-SPNs less excitable (Gerfen, Engber, et al. 1990). Supposedly this could be the mechanism that mediates the direct consequences of dopamine signalling discussed in section 2.2. In figure 2.3 we visualise this dopaminergic modulation by reproducing the results of two different studies: the study of Thurley et al. (2008) and the study of Hernández-López et al. (2000). The figure shows how the slopes of the fI curves[3] of D1- and D2-expressing neurons depend on the presence or absence of dopamine.

To summarise this subsection, the basal ganglia features two major pathways—direct and indirect—which have differential effects on the basal ganglia output nuclei and are differentially modulated by dopamine release.

2.3.5 The function of the direct and indirect pathways

How do the direct and the indirect pathway affect to behaviour? The connectivity structure suggests that the converging pathways should generally have opposite effects on behaviour. The direct pathway can release the inhibition of motor centres and is thought to be prokinetic. In contrast, the indirect pathway increases the inhibition of motor centres and is thought to be antikinetic. For these reasons, the direct pathway is sometimes referred to as the *Go-pathway*, while the indirect pathways is called the *NoGo-pathway* (see e.g. Frank et al. 2004).

[3] A neuron's fI curve shows the neuron's firing rate depends on the input current.

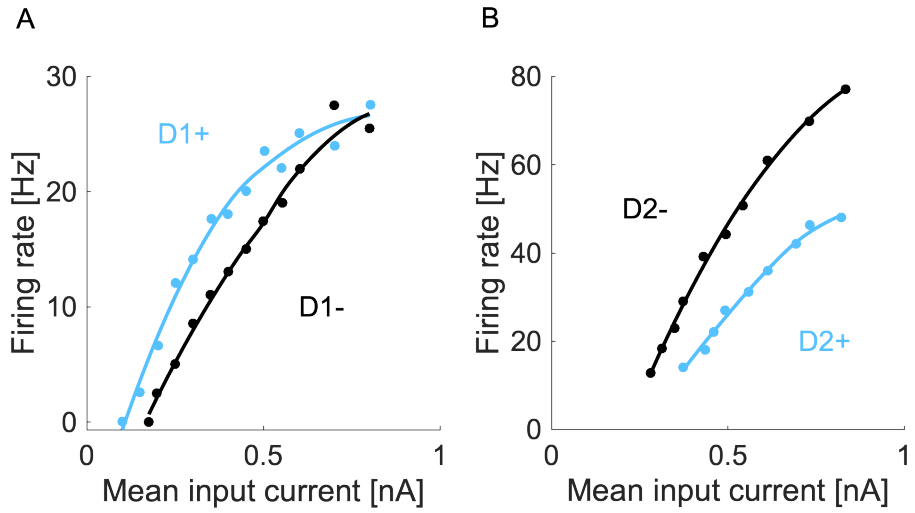

Figure 2.3: The fI curves of D1- and D2-expressing neurons at different levels of receptor activation A The fI curves of D1-expressing neurons, replotted from Thurley et al. (2008). The blue points are recorded from a neuron at a higher level of D1 receptor activation (e.g. with dopamine present), the black points are recorded at a lower level of receptor activation (e.g. without dopamine). Smooth curves were obtained by smoothing the data using the LOESS method with a span of 20, followed by spline interpolation. Those smooth curves serve as visual guides (black and blue lines). **B** The fI curves of D2-expressing neurons, replotted from Hernández-López et al. (2000). The blue points are recorded from a neuron at a higher level of D2 receptor activation (e.g. in the presence of the D2 agonist quinpirole), the black points are recorded from a neuron in the control group at a lower level of D2 activation (e.g. in the absence of the agonist). As in panel A, smoothed curves (black and blue lines) have been added as visual guides. This figure was reproduced from Möller and Bogacz (2019) and adapted for this thesis.

This view is supported by the finding that STN lesions (which represent a disruption of the indirect pathway) tend to cause premature responding (Baunez et al. 1995). Further, optogenetic stimulation of direct and indirect dorsomedial SPNs in freely moving mice leads to increased movement and decreased movement, respectively (Kravitz et al. 2010).

On a more specific level, one influential hypothesis states that the basal ganglia facilitate movement by removing inhibition from one motor circuit via the direct pathway, while inhibiting all other competing motor circuits via the indirect pathway (Mink 1996). Recent experimental results provide new evidence for this idea: it was shown that SPNs in both pathways are (and need to be) active during normal movement, and less active during rest (Cui et al. 2013; Tecuapetla et al. 2014).

2.3.6 Action selection and action channels.

We have provided a detailed discussion of the anatomy of the basal ganglia and its dopaminergic innervation. We have also discussed the two pathways in terms of their physiological properties and their function. In the final part of our review of the basal ganglia we discuss another important concept: action channels. We already briefly introduced the concept in our synopsis in subsection 2.3.1—here, we dive deeper into the details and the supporting evidence.

Many modellers share the view that the basal ganglia is (at least partly) a circuit for action selection (Redgrave et al. 1999; Gurney et al. 2001a; Gurney et al. 2001b; Frank et al. 2004). According to this view, different actions are represented by parallel channels called *action channels* that descend through the basal ganglia. Several action channels might be active at the input stage, but only one action at a time should be active at the output stage. Anatomically, action channels are generally more closely associated with the dorsal striatum—the ventral striatum is thought to reflect value estimates (O'Doherty et al. 2004; Samejima et al. 2005).

Evidence that corroborates the existence of action channels comes from studies that decode movements from striatal activity (Klaus, Martins, et al. 2017; Markowitz et al. 2018). These results support the view of the striatum as a map of action space, i.e., as a collection of distinct ensembles of neurons that correspond to distinct movements.

We have now gained an overview of the basal ganglia system, both from a high level and in detail. For the intents and purposes of most of this thesis, we might view it as an action selection system that features two antagonistic sets of action channels (corresponding to the direct and indirect pathway), the balance of which is differentially modulated by dopamine.

2.4 Learning in the basal ganglia

We have already discussed a mechanism that gives dopamine direct control over behaviour: it modulates the activity of striatal neurons in a pathway-specific manner. Presumably, that mechanism underlies dopamine's direct effects on behaviour. We reviewed those above in section 2.2.

But what about the indirect effects of dopamine? In section 2.1, we have seen that dopamine is well described as a prediction error signal that drives learning. In this section, we will focus on the mechanism behind such learning: dopamine-dependent synaptic plasticity in the striatum.

We will first review what is known about this plasticity (subsection 2.4.1). Then, we will look at two models of learning in the basal ganglia (the OpAL model in subsection 2.4.2 and the AU model in subsection 2.4.3). In addition to modelling plasticity, these models also cover many other aspects we have encountered, such as pathway competition and dopamine as a modulator. This will conclude our literature review, as it combines and summarises the key aspects of what is known about the topic of this thesis.

2.4.1 The plasticity of cortico-striatal synapses

The reward prediction error hypothesis of dopamine predicts that there should be dopamine-dependent plasticity in the targets of dopamine signalling, such as the striatum. The predicted plasticity was indeed found in synapses between cortical projections and striatal SPNs.

Initially, this was achieved using a self-stimulation paradigm, with stimulation targeting dopamine neurons in the rat's SNc. The study reported dopamine-dependent potentiation in cortico-striatal synapses (Reynolds et al. 2001). Later, a more detailed investigation revealed that dopamine has differential effects

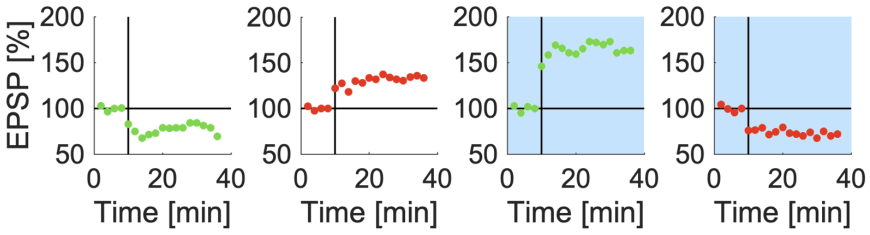

Figure 2.4: Dopamine dependent plasticity in the basal ganglia pathways. Excitatory postsynaptic potential (EPSP) amplitudes are plotted before and after spike-timing-dependent plasticity (STDP) was induced. The data is taken from Shen et al. (2008), figures 3B, 3E, 3F and 1H. The black vertical line marks the time of STDP, the horizontal line marks the average EPSP amplitude before STDP. The panels with green colour show measurements from D1-SPNs, the panels with red colour show measurements from D2-SPNs. A white background indicates that the measurement was taken in mice whose dopamine neurons were lesioned with 6-OHDA. A blue background indicates the presence of dopamine receptor agonists (SKF81297 for D1 receptors and quinpirole for D2 receptors). This figure is a reproduction of figure 3C of Bogacz (2017b).

on synaptic plasticity in D1 and D2 cells (Shen et al. 2008): it causes long-term potentiation in synapses involving D1-SPNs, and long-term depression in synapses involving D2-SPNs (see figure 2.4 for a reproduction of these results).

Are these the mechanisms behind adaptive changes in behaviour? Is the interface between cortex and striatum the (or at least one) site of reinforcement learning in the brain? It appears that this might indeed be the case: in a recent study in mice, the effect of dopamine bursts or pauses on striatal SPNs was mimicked with weak optogenetic stimulation (Yttri and Dudman 2016). Applying that stimulation in a closed-loop procedure was sufficient to induce lasting changes in behaviour. The direction of the changes depended on the pathway that was targeted and was consistent with the differential effects of dopamine on striatal plasticity.

Overall, the evidence suggests that cortico-striatal synapses are subject to dopamine- and spike-timing-dependent plasticity, with the effect of dopamine depending on the post-synaptic cell type. It further appears as if this plasticity is related to behavioural adaptation. These findings, along with others (such as the reward prediction error hypothesis and the motivational effects of dopamine) have been

integrated and summarised in compact models of learning and action selection in the basal ganglia. We will review two of these models in the next subsections.

2.4.2 The Opponent Actor Learning model

The Opponent Actor Learning model (OpAL, Collins and Frank 2014) might be considered the canonical model of learning in the basal ganglia. It is the first model to represent the motivational *and* the reinforcing properties of dopamine in a common framework.

OpAL is an actor-critic model—it consists of two components, called actor and critic. The critic is in charge of reward prediction, the actor is responsible for action selection. The critic is conventional: the value of cues is predicted, deviations δ between predicted and experienced rewards are signalled through phasic dopamine and drive value updates according to an RW-type rule (see equation 2.1).

The actor, however, is different from typical actor models which usually feature one activation for each action. This activation reflects a preference for that action and is used to decide which action to take. The OpAL model features not one but two activations, G_a and N_a, for each action a. The weighed difference between these activations reflects the preference:

$$T_a = \beta_G G_a - \beta_N N_a \tag{2.2}$$

with T_a the preference for action a[4]. The two activations correspond to the direct and the indirect pathway contributions to each action channel. G represents the direct (Go) pathway, N represents the indirect (NoGo) pathway. The weights β_G and β_N represent pathway-specific modulation which is attributed to tonic dopamine.

[4]Here, T stands for thalamic activity—the output of the basal ganglia. Collins and Frank (2014) use a different notation, Act_a instead of T_a. We use T_a to increase consistency across the thesis.

The activations associated with an action (and hence preference for that action) may change with experience, through dopamine-dependent synaptic plasticity at the interface between cortex and striatum. Such updates are modelled by the OpAL learning rules

$$\Delta G_a = \alpha_G G_a \delta \tag{2.3}$$

$$\Delta N_a = \alpha_N N_a (-\delta). \tag{2.4}$$

$$\tag{2.5}$$

These rules are consistent with many known features of dopamine-dependent plasticity in the striatum, such as those found by Reynolds et al. (2001) and by Fisher et al. (2017). In particular, they capture the fact that phasic dopamine leads to long-term potentiation in D1-SPNs and depression in D2-SPNs.

The factors G_a and N_a on the right-hand side of the update equations are meant to reflect Hebbian plasticity, and make the learning rules nonlinear. Without them, N would just be a negative mirror-image of G, and the model would be equivalent to an ordinary actor-critic model. This nonlinearity is therefore an essential feature of the OpAL model.

The internal circuitry of the basal ganglia is assumed to achieve action selection at the level of the GPi. Formally, the probability $P(a)$ that action a is selected is given as a softmax distribution

$$P(a) = \frac{\exp(T_a)}{\sum_{a'} \exp(T_a)}$$

In the OpAL model, dopamine affects action selection in two different ways: indirectly through δ, which changes G_a and N_a in the process of learning, and directly through β_G and β_N, which control the contribution of G_a and N_a to the Activation T_a of actions, and hence the likelihood that a is selected.

The OpAL model can be applied to a wide range of different phenomena: the effects of pathway-specific optogenetic stimulation, the impact of motivation in probabilistic selection tasks, the effect of D2 blocking on the willingness to work, the acquisition of motor skills and the acquisition of cognitive skills.

OpAL explains some of these effects more elegantly than others. For instance, to reproduce the effects of blocking D2 receptors on the willingness to work for food, OpAL must first be trained for two actions separately ("work" with a negative reward and "eat food" with a positive reward). If one then tests the model on a combination of both actions, it displays preferences similar to those measured in the experiment. This complicated setup is necessary because OpAL will present an action either as generally negative (using N) or generally positive (using G). It cannot learn the positive *and* the negative consequence of the same action at the same time, as would be needed in a situation where an action requires effort (negative reward) but leads to food (positive reward). To circumvent this, two action values must be learned separately and then somehow combined.

Another complication arises from the nonlinearity of the OpAL learning rules: the Hebbian factors cause an inevitable decay to zero of all actor weights if the model tries to learn alternating deterministic rewards (Bogacz 2017b). OpAL thus predicts that if an agent is exposed to alternating rewards for long enough, it will eventually lose all preferences and act randomly. Here, OpAL's predictions are quite unrealistic. Taken together, we may summarise that the OpAL model has issues with the convergence of actor weights and needs rather ad hoc constructions to produce the correct results in certain conditions.

Those complications notwithstanding, OpAL is among the most important and influential models of the basal ganglia because it shows that the dual actor architecture with dopamine-controlled balance provides enough flexibility to accommodate a plethora of data connected to dopamine, learning, motivation and reward.

Other models have adopted the general architecture of OpAL to explain how two antagonistic networks can acquire independent evaluations of states and actions, which—as OpAL has first shown—are essential to explain the effects of dopamine. Among them is the AU model (Mikhael and Bogacz 2016) which we will review in the next subsection. In chapter 3, we will show that AU can overcome the technical problems of OpAL: AU can be shown to converge to sensible values for alternative reward sequences and other schedules, and it can explain the effect of dopamine on the willingness to work.

2.4.3 The Actor learning Uncertainty model

Similar to OpAL, the AU model proposed by Mikhael and Bogacz (2016) aims to capture the role of the basal ganglia pathways in learning and action selection. Also similar to OpAL, AU features two parallel action selection networks, G and N, which correspond to the two basal ganglia pathways and have differential effects on action selection.

However, OpAL and AU differ in *what* they learn, and hence in their learning rules. The learning rules of the AU model are[5]

$$\delta = r - \frac{1}{2}\left(G_a - N_a\right) \tag{2.6}$$

$$\Delta G_a = \alpha f_\epsilon\left(\delta\right) - \lambda G_a \tag{2.7}$$

$$\Delta N_a = \alpha f_\epsilon\left(-\delta\right) - \lambda N_a \tag{2.8}$$

with $f_\epsilon(x) = x$ for $x > 0$ and $f_\epsilon(x) = \epsilon x$ for $x < 0$. We call α the learning rate, ϵ the slope parameter and λ the unlearning rate. The nonlinear transformation $f_\epsilon(\delta)$ of the prediction error is visualised in figure 2.5A. The updates 2.7 and 2.8 only apply if the resulting weights are still positive - if an update would render

[5]The rules given in equation 2.6 - equation 2.8 are not exactly equal to the rules given by Mikhael and Bogacz (2016)—instead, we use the formulation of Möller and Bogacz (2019). The difference between the formulations is the factor $1/2$ in equation 2.6. This factor does not change the model qualitatively, it could be completely absorbed in a rescaling of G and N. However, it is slightly more convenient for our calculations in chapter 3.

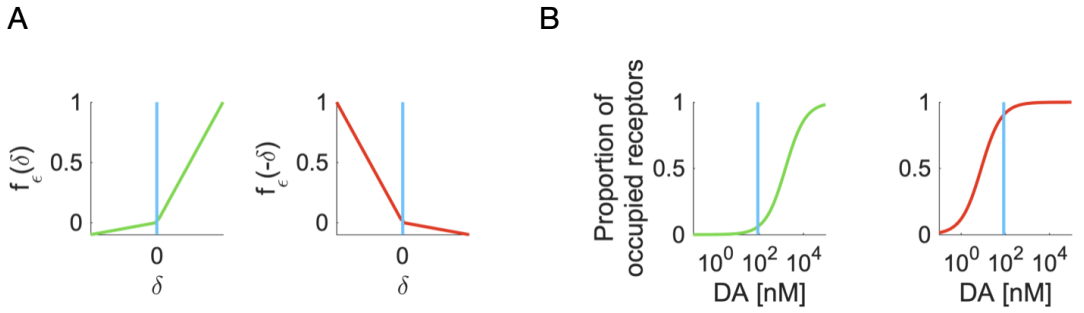

Figure 2.5: Pathway-specific nonlinear transformations of prediction errors and receptor occupancies. A We show the nonlinear functions $f_\epsilon(\delta)$ that feature in the AU learning rules equation 2.7 and equation 2.8. The green line shows the transformed reward prediction error in equation 2.7, the red line shows the transformed reward prediction error in equation 2.8. The blue line marks the dopamine baseline level (i.e. $\delta = 0$). For the plots, we used $\epsilon = 0.1$. **B** We plot the proportion of occupied receptors in the striatum as a function of dopamine concentration. The curves are based on the results of Dreyer et al. (2010). The blue vertical lines indicate the baseline dopamine concentration in the ventral striatum, based on the results of Dodson et al. (2016). The green curve corresponds to the occupancy of D1 receptors, the red curve corresponds to the occupancy of D2 receptors. Panel A of this figure was reproduced from figure 3E of Möller and Bogacz (2019), Panel B is a partial reproduction of figure 3D of Bogacz (2017b).

a weight negative, that weight is set to zero instead. The AU learning rules are designed to reflect the key features of striatal plasticity, as we will discuss in subsection 2.4.4 below.

In OpAL, G and N encode preferences for actions with positive and negative outcomes, respectively. In AU, G and N encode information about the means and spreads[6] of the reward distributions associated with individual actions. Mikhael and Bogacz (2016) show that after learning, the difference $1/2 \times (G_a - N_a)$ represents the mean reward Q_a associated with action a. The sum $1/2 \times (G_a + N_a)$, on the other hand, represents the spread S_a of rewards associated with a.

In AU, both the mean reward Q_a and the reward spread S_a might enter action selection. Like in OpAL, AU assumes that the preference for an action a depends

[6]In this thesis, the spread s of a distribution is defined as a statistic similar to the standard deviation, but based on the modulus of distances rather than their square: $s := \mathbb{E}\,|r - \mathbb{E}(r)|$.

on the corresponding activation T_a, which is a weighted difference of the two network outputs G_a and N_a[7]:

$$T_a = DG_a - (1 - D)N_a = Q_a + (2D - 1)S_a \tag{2.9}$$

Here, $D \in [0, 1]$ is the tonic dopamine level in the striatum. The differential weighting of dopamine expressed in this equation reflects the properties of neuromodulation in the striatum—see figure 2.3.

The baseline level of dopamine is $D = 1/2$. If D is at that level, both pathways are perfectly balanced and the weighted sum collapses into an estimate of the mean reward: $T_a = Q_a$. If dopamine deviates from the baseline level, the spread S_a of the reward distribution can enter the final evaluation of the action, biasing the agent either towards or away from risk (i.e. towards or away from actions with high reward spreads).

In AU, the function of dopaminergic modulation thus becomes very transparent: if dopamine rises above baseline, risk bonuses are added to action values. This interpretation makes it possible to link the model to empirical findings concerning the effect of dopamine blockers on risk preferences in animals (e.g. St Onge and Floresco 2009, see section 2.2 for more examples). The AU model can explain those effects, as it is able to capture different types of risk preferences—a feature that previous reinforcement models lack.

In summary, the AU model has learning rules that track the mean and spread of the reward distribution associated with single actions, and it can describe phenomena related to risk preferences. In this thesis, we build on these results and develop the model further in three distinct directions.

[7]Here, we use the AU formulation given by Bogacz (2017b) instead of the formulation of Mikhael and Bogacz (2016) because it makes the role of tonic dopamine more transparent. Note that the activation T_a proposed by Bogacz (2017b) is mathematically equivalent up to an overall factor to that proposed by Collins and Frank (2014) given above in equation 2.2, by way of the substitution $D = 1/2\,(1 + (\beta_G - \beta_N)\,/\,(\beta_G + \beta_N))$.

First, we show that the AU model can do more than just capture risk preferences. In chapter 3, we develop the idea of Bogacz (2017b): that the AU model can learn the payoffs and costs of actions. We then apply it to decisions that involve different payoffs and costs, and show that it can explain the effects of dopamine depletion on the willingness to work (Salamone, Steinpreis, et al. 1991).

Second, we use the AU model as a starting point and inspiration to explore new models. In chapter 4, we develop the scaled prediction error model. Like AU, the scaled prediction error model is a model of the basal ganglia, and like AU it can learn statistics of reward distributions.

Third, we use the AU model to test whether the different roles of dopamine (teaching signal and motivation signal) interfere on the level of behaviour. Here, the AU model allows us to derive detailed behavioural predictions, which we then test empirically. That is the subject of chapter 5.

We have now reviewed almost all prior work we need to move on to those chapters. The last missing part is an overview over how pathway-specific striatal plasticity maps onto the AU learning rules, which will become important in chapter 4.

2.4.4 The biological implementation of uncertainty learning

The AU learning rules are given in equation 2.7 and equation 2.8. They are based on three premises. We will discuss them in turn.

First, the efficacy of direct pathway synapses is assumed to increase as a result of positive reward prediction errors (i.e. $\delta > 0$), and decrease as a result of negative reward prediction errors (i.e. $\delta < 0$). The opposite is assumed to hold for indirect pathway synapses: their efficacy should decrease with positive prediction errors and increase with negative prediction errors. This premise corresponds to the overall slope of the nonlinearities in figure 2.5A, and equivalently to the sign of the prediction error in equations 2.7 and 2.8. Bogacz (2017b) points out that this premise is consistent with data we discussed above in figure 2.4.

Second, it is assumed that for D1-SPNs, positive prediction errors have a stronger effect on plasticity than negative prediction errors. This is encoded by the slope parameter ϵ, which takes values between 0 and 1 and must be less than 1 for the model to work. For D2-SPNs, AU assumes the opposite: negative prediction errors should have a stronger plasticity effect than positive prediction errors. Mikhael and Bogacz (2016) argue that this premise is realistic, based on the different affinities of D1 and D2 receptors: while D1 receptors are mostly unoccupied at baseline dopamine levels, D2 receptors are almost saturated—this is visualised in figure 2.5B. Due to this baseline setting, additional dopamine should lead to a large difference in the occupation of D1 receptors, but only a small change in the occupancy of D2 receptors. A decrease in dopamine, on the other hand, is strongly felt in D2 receptor occupancy but does not change D1 receptor occupancy much.

Third, an activity-dependent decay (or "unlearning") is assumed to occur in the synaptic weights whenever they are activated in the absence of prediction errors. This is reflected in the unlearning rate λ, which must be larger than zero for the model to work. On the neural level, that premise translates into mild long-term depression after co-activation of the pre- and postsynaptic cells at baseline dopamine levels. Recently, this effect has been observed at cortico-striatal synapses in vivo (Fisher et al. 2017): in anaesthetised rats, presynaptic activity followed by postsynaptic activity caused LTD when dopamine was absent.

In summary, we discussed the three premises of the AU learning rules—the different overall effects of dopamine on plasticity in each pathway, the nonlinear effects of dopaminergic prediction errors and synaptic unlearning. We saw that all three premises are supported by the physiological properties of D1-SPNs and D2-SPNs.

2.5 Summary of the literature review

The goal of this chapter was to review what is known about the neural processes behind reinforcement learning and choice. We first focused on dopamine and reviewed its putative role as a reward prediction error signal in section 2.1. In section 2.2 we then saw that dopamine is also important for taking actions despite associated costs and risks.

In sections 2.3 and 2.4, we turned to the basal ganglia system, which is a major target of dopamine projections. The effects of dopamine on the basal ganglia are twofold: dopamine modulates the activity of SPNs in the striatum, but also gates plasticity of cortico-striatal synapses. The basal ganglia are organised into two antagonistic pathways; dopamine exerts differential effects on these pathways, which in turn exert opposing effects on behaviour.

In addition to the most important empirical findings, we have reviewed the most influential theories in the field (the reward prediction error hypothesis and the motivation hypothesis), as well as the most relevant recent models (OpAL and AU). Those models map learning and decision making onto the physiological mechanisms and the structure of the basal ganglia and their dopaminergic innervation.

3

Learning the payoffs and costs of actions

Contents

This chapter is about the payoffs-costs hypothesis of the basal ganglia. The central

idea here is that actions often have negative as well as positive consequences—writing a PhD thesis is hard, solitary and seemingly endless work, but it will eventually lead to a degree and a sense of great achievement.

Several parts of this chapter were published by Möller and Bogacz (2019). The text and the figures of section 3.2 and subsection 3.3.1 were taken from from that publication and adapted slightly for this thesis. The results and the figure of subsection 3.3.2 also stem from that publication, but have been rewritten for this thesis. All materials were produced by the author of this thesis. We use them here with the full consent of all co-authors (written permission was obtained).

3.1 The payoffs-costs hypothesis

The payoffs-costs hypothesis states that positive consequences of actions are encoded in the direct pathway of the basal ganglia, while negative consequences are encoded in the indirect pathway. We will start this chapter by tracing the origin of this idea. We will then summarise the status quo of the theory, and point out the open questions.

3.1.1 The origins of the hypothesis

Above in section 2.4.3, we have reviewed the AU model of learning in the basal ganglia in some detail. Originally this model was designed to capture dopamine's effect on risk preferences. It was soon noticed, though, that the AU model might apply to phenomena beyond risk: Bogacz (2017b) pointed out that—if applied to a situation in which actions have positive as well as negative consequences—the AU model could learn the magnitudes of these payoffs and costs. Payoffs would be encoded in the direct pathway, while costs would be encoded in the indirect pathway. This would explain why dopaminergic medication affects the willingness to work for reward, as reported for example by Salamone, Steinpreis, et al. (1991).

Bogacz (2017b) argues that if, for example, an animal would need to press a lever to obtain a food pellet, it could learn the cost of pressing the lever and store this information in the indirect pathway channel associated with that action. The payoff associated with the food pellet would be stored in the corresponding direct pathway channel. After learning, the differential dopaminergic modulation of the basal ganglia pathways could enable a flexible weighing of payoffs against costs, tuned to internal and external factors (such as hunger or scarcity of food in the environment). For example, hunger could raise the tonic level of dopamine. This would emphasise the payoff of lever-pressing in the decision making process, biasing the animal to ignore the costs of pressing the lever in order to obtain the food pellet.

The idea that information about positive outcomes is stored in the direct pathway and information about negative outcomes is stored in the indirect pathway was already present in the OpAL model developed by Collins and Frank (2014)—we reviewed this work in detail above in subsection 2.4.2. However, according to OpAL, an action could either have a payoff *or* a cost. Bogacz (2017b) was the first to show how the payoff *and* cost of a single action could be learned.

3.1.2 The status quo and open questions

What is the current status of the payoffs-costs hypothesis? The main ideas of the hypothesis were outlined by Bogacz (2017b), as described above. In that work, it was suggested that the AU model might be able to learn the payoffs and costs of an action if they are registered at different moments in time.

Bogacz (2017b) trained the AU model on an alternating sequence of rewards (positive rewards p alternated with negative rewards $-n$). It was shown that the AU weights G and N will converge to values approximately proportional to p and n, given that the parameters of the model are tuned appropriately. This showed that the AU model can learn about the payoffs and costs of an an action in the basic but important case of alternating deterministic rewards.

Bogacz (2017b) further applied the AU model to an experiment that investigated the effect of a dopamine antagonist on the willingness to work for food (Salamone, Steinpreis, et al. 1991). For simplicity, Bogacz (2017b) assumed that the antagonist acts on the tonic level of dopamine D, rather than modelling a pathway-specific effect.

Given these results, what are the open problems and challenges? We will point out three questions. First, can the AU model learn payoffs and costs of actions in general? Bogacz (2017b) has shown that it is possible in a special case, but does this generalise to other reward scenarios? In particular, do the results of Bogacz (2017b) generalise to stochastic reward schedules? Answering those questions requires a more general definition of payoffs and costs, ideally one that can be applied to stochastic as well as to deterministic reward schedules.

Second, how must the parameters of the AU model be tuned to enable the learning of payoffs and costs? Bogacz (2017b) derives a condition for a deterministic reward schedule, but it is unclear whether that condition will be enough for more general cases. Further, the condition only ensures convergence to values proportional to the payoffs and costs. It is not known which requirements must be satisfied to learn exact (or at least approximately exact) values.

Third, how does AU relate to data from experiments that involve payoffs and costs? Bogacz (2017b) offered an analysis of the experiment of Salamone, Steinpreis, et al. (1991), but omitted the pathway specificity of the pharmacological intervention. Do these results still hold if the details of the intervention are considered? Furthermore, can we make predictions for other experiments, to potentially test the payoffs-costs-hypothesis empirically?

In this chapter, we address those questions in turn. Our contributions are grouped into a mathematical analysis (section 3.2) and simulations (section 3.3). These contributions are new results; however, some of them (in particular those in sections 3.2.2, 3.2.3 and 3.3.1) are closely related to existing work of Mikhael and Bogacz (2016) and of Bogacz (2017b). In section 3.4.2, we point out the differences

between our results and previous work, to emphasize the unique contributions of this thesis.

For those who want to gain a more intuitive understanding of the structure of the AU rules in the context of the learning of payoffs and costs, we provide a constructive explanation of the rules in appendix A. There, we show how one might arrive at the AU learning rules if one tried to construct a set of learning rules that learn payoffs and costs from scratch. Although rather didactic in nature, this construction supports the results we present in the main text.

3.2 Mathematical analysis

In our mathematical analysis, we first extend the notion of payoffs and costs to stochastic reward schedules, offering a definition of payoffs and costs in terms of statistics of reward distributions (section 3.2.1). This link then allows us to show that the AU model can learn payoffs and costs when training on stochastic reward schedules. We further obtain new conditions for the parameters of the model in section 3.2.2. We extend our analysis to deterministic rewards in section 3.2.3, obtaining another condition for the parameters. We finish the section with an investigation of robustness (section 3.2.4): what is the impact of parameter detuning on the accuracy of the estimates of payoffs and costs, and on the ultimate evaluation of actions?

3.2.1 Payoffs, costs and reward distributions

In chapter 2 we mentioned that the rules in equations 2.7 and 2.8 were originally meant to describe the learning of reward statistics: Mikhael and Bogacz (2016) showed that after learning, particular combinations of G and N encode the mean $\mathbb{E}r$ and the mean spread $\mathbb{E}\left|r - \mathbb{E}r\right|$ of the received rewards. For further reference, we define

$$q := \mathbb{E}r \tag{3.1}$$

$$s := \mathbb{E}\,|r - q| \tag{3.2}$$

How are the mean and the mean spread of the reward distribution related to payoff and cost? Consider the reward distribution of an action that reliably requires effort (corresponding to a negative reward of about $-n$, $n > 0$) to produce a payoff (which corresponds to positive reward of about p, $p > 0$). Repeat that action multiple times, and record all received rewards, the costs as well as the payoffs. Finally, consider how all these received rewards are distributed. The reward distribution will turn out bimodal, as schematically shown in figure 3.1.

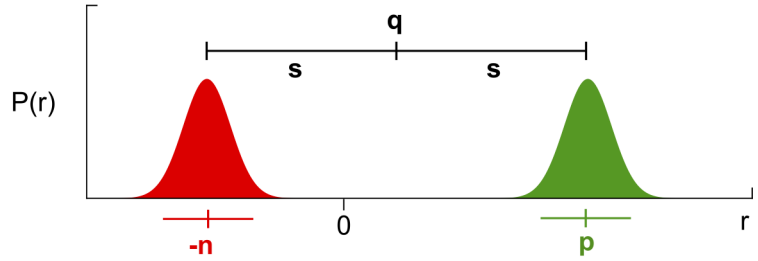

Figure 3.1: Payoff and cost in relation to the statistics of the reward distribution. We show a representative reward distribution for an action with payoffs and costs. The negative rewards (costs) are coloured red, while the positive rewards (payoffs) are coloured green. The mean q and the mean spread s are indicated above the distribution, the mean cost $-n$ and the mean payoff p are indicated below the distribution. This figure was taken from Möller and Bogacz (2019).

The distribution has two peaks, one centred around the payoff p and one centred around the cost $-n$. If we neglect noise, we may describe this distribution mathematically as

$$P(r = p) = 1/2 \tag{3.3}$$

$$P(r = -n) = 1/2 \tag{3.4}$$

Using equations 3.1 and 3.2, we can then show that $q = 1/2 \times (p - n)$ and $s = 1/2 \times (p + n)$. Those statistics are also shown in figure 3.1.

We find that payoffs and costs are both exactly one spread s away from the centre q of the distribution—the payoff above, and the cost below. This implies that there is, at least in this representative case, a strong connection between payoffs and costs and the statistics of the reward distribution:

$$p = q + s \tag{3.5}$$

$$-n = q - s \tag{3.6}$$

We conclude that if G and N are to represent payoff and cost, they must approach $q + s$ and $-q + s$ respectively. Equivalently, we can ask for $1/2\,(G - N)$ and $1/2\,(G + N)$ to approach q and s in the course of learning.

3.2.2 Stochastic reward schedules

After defining payoffs and costs in terms of reward statistics, we can now test whether and how the AU model can learn those values if rewards are stochastic. To do this, we first determine the values that G and N converge to when trained on a stochastic reward schedule, i.e., on rewards sampled at random from a fixed distribution. We then set those values equal to the payoffs and costs of the reward distribution, requiring $1/2 \times (G - N)$ to approximate q and $1/2 \times (G + N)$ to approximate s after training is finished. From these conditions, we will be able to derive conditions for the parameters.

Working through these steps is simpler after changing variables from G and N to $Q := 1/2 \times (G - N)$ and $S := 1/2 \times (G + N)$ right away. To determine how Q and S change in response to a reward prediction error δ, we simply add and subtract the update rules in equations 2.7 and 2.8. The convenient properties of the nonlinear functions f_ϵ help to further simplify the resulting

equations: $f_\epsilon(x) - f_\epsilon(-x) = (1+\epsilon)\,x$ and $f_\epsilon(x) + f_\epsilon(-x) = (1-\epsilon)\,|x|$. Using these properties, we obtain

$$\Delta Q = \alpha_Q \delta - \lambda Q \tag{3.7}$$

$$\Delta S = \alpha_S |\delta| - \lambda S. \tag{3.8}$$

Here, for brevity of notation, we introduced the effective learning rates $\alpha_Q = \alpha(1+\epsilon)/2$ and $\alpha_S = \alpha(1-\epsilon)/2$.

Now, let us determine the strengths of the weights G and N, or equivalently of the variables Q and S, after extensive training. When training on a new reward schedule, Q and S typically change a lot during the first trials. These changes then get smaller and smaller as more experience is integrated—the learning curve plateaus. After enough trials, Q and S stop changing systematically, and start to merely fluctuate about some constant values: the equilibrium points Q^* and S^*.

How do we find these equilibrium points? The equilibrium is reached when Q and S can be expected to remain unchanged by another trial, i.e. when $\mathbb{E}(\Delta Q) = 0$ and $\mathbb{E}(\Delta S) = 0$. Using the learning rules equation 2.7 and equation 2.8, we find

$$0 = \mathbb{E}\,\Delta Q = \mathbb{E}\left[\alpha_Q (r - Q^*) - \lambda Q^*\right] = \alpha_Q (q - Q^*) - \lambda Q^* \tag{3.9}$$

$$0 = \mathbb{E}\,\Delta S = \mathbb{E}\left[\alpha_S |r - Q^*| - \lambda S^*\right] = \alpha_S \mathbb{E}\,|r - Q^*| - \lambda S^*. \tag{3.10}$$

Solving these equations for Q^* and S^* yields

$$Q^* = c_Q\, q \tag{3.11}$$

$$S^* = c_S\, \mathbb{E}\left|r - c_Q q\right|, \tag{3.12}$$

with

$$c_Q = \alpha_Q / \left(\alpha_Q + \lambda\right) \tag{3.13}$$

$$c_S = \alpha_S / \lambda \tag{3.14}$$

Next, we need to implement the conditions 3.5 and 3.6 we defined in the previous subsection. Thanks to our choice of variables, this simply amounts to requiring Q to converge to the mean reward q, and S to the mean spread s, i.e. requiring $Q^* = q$ and $S^* = s$. Inserting the values from equation 3.11 and equation 3.12, we obtain

$$c_Q \, q = q \tag{3.15}$$

$$c_S \, \mathbb{E} \left| r - c_Q q \right| = s \tag{3.16}$$

These equations are fundamental for this section. Their left-hand side represents the result of learning according to the AU rules, their right-hand side specifies what needs to be learned if G and N are to represented payoffs and costs. The equations determine how the AU parameters must be chosen for payoffs and costs to be learned: for equation 3.15 and equation 3.16 to hold, α, λ and ϵ must take values such that

$$c_Q = 1 \tag{3.17}$$

$$c_S = 1 \tag{3.18}$$

Is it at all possible to satisfy the derived conditions? What do they mean with respect to the parameters α, λ and ϵ? And finally, is there a practical way to determine sets of parameters α, λ and ϵ which—at least approximately—satisfy the conditions? We discuss each of these questions in the following paragraphs.

Firstly, is it possible to satisfy $c_Q = 1$ and $c_S = 1$ exactly? Let us start with $c_Q = 1$. Examining the definition $c_Q = \alpha_Q / \left(\alpha_Q + \lambda\right)$ quickly reveals that $c_Q \to 1$ would amount to $\lambda \to 0$. To see why this is the case, consider that $c_Q \to 1$ amounts to $\lambda / \alpha_Q \to 0$. This could be done by $\alpha_Q \to \infty$ or $\lambda \to 0$. However, α_Q is an effective learning rate, and so must take values smaller than one. Thus, we really need $\lambda \to 0$.

The other requirement is $c_S = 1$. The definition of c_S is $c_S = \alpha_S / \lambda$. If $\lambda \to 0$ and $c_S = 1$ both hold, we would need $\alpha_S \to 0$ as well. This, though, would pose a real problem: α_S is the effective learning rate for S—having it vanish would imply stopping learning for S completely.

We must conclude that strict satisfaction of the constraints $c_Q = 1$ and $c_S = 1$ is not possible. Specifically, $c_Q = 1$ can only ever hold approximately if the spread s is to be learned in finite time.

However, it is possible to tune the parameters to $c_Q \approx 1$ and $c_S = 1$. What do these conditions mean in terms of the parameters α, λ and ϵ ? In the previous paragraph, we saw that $c_Q \approx 1$ is equivalent to $\lambda / \alpha_Q \approx 0$. Since both λ (the unlearning rate) and α_Q (an effective learning rate) are inherently positive, we may rewrite this as $\lambda / \alpha_Q \ll 1$. Inserting the definition $\alpha_Q = \alpha \left(1 + \epsilon\right) / 2$ immediately yields

$$2\lambda \ll \alpha \left(1 + \epsilon\right) \tag{3.19}$$

The other condition, $c_S = 1$, is easily translated analogously. We need only use the definitions $c_S = \alpha_S / \lambda$ and $\alpha_S = \alpha \left(1 - \epsilon\right) / 2$ to obtain

$$2\lambda = \alpha \left(1 - \epsilon\right). \tag{3.20}$$

Equations 3.19 and 3.20 provide the exact relations between the parameters α, λ and ϵ that need to hold for payoffs and costs to be estimated accurately. They

cannot be further simplified, but we may use them to gain some more insight into the required magnitudes of the individual parameters: by substituting 2λ according to equation 3.20 on the right-hand side of equation 3.19, one obtains a condition of the form $1 - \epsilon \ll 1 + \epsilon$. Now, given that the intended range for ϵ is $[0, 1]$, we reach the conclusion $\epsilon \approx 1$. Reinserting this into equation 3.19 yields $\lambda \ll \alpha$.

In conclusion, we found that it is necessary (though not strictly sufficient) for accurate learning of stochastic payoffs and costs to maintain a small but non-vanishing nonlinearity ϵ in the effect of the prediction error signal, as well as a non-vanishing unlearning rate λ which is much smaller than the learning rate α. This is the first key results of this chapter.

We finish this part with a practical question: how does one determine parameters α, λ and ϵ for the learning of payoffs and costs, e.g. in simulations? To implement the conditions $c_Q \approx 1$ and $c_S = 1$, one can express λ and ϵ in terms of α, c_Q and c_S. It is tedious but without conceptual difficulty to invert the definitions of c_Q and c_S in order to yield $\epsilon = (1 - c_S(1/c_Q - 1))/(1 + c_S(1/c_Q - 1))$ and $\lambda = \alpha(1 - \epsilon)/(2c_S)$. Then, one chooses α freely at one's convenience, and c_Q and c_S close (or, in case of c_S, equal) to one. Importantly, c_Q must be chosen smaller then one to result in a positive λ. From these choices, one obtains values of ϵ and λ that work with the chosen α.

Our simulations suggest that the values $c_Q = 0.6$ and $c_S = 0.95$, in combination with a learning rate $\alpha = 0.3$, are close enough to one to allow for reasonably accurate estimations of payoff and cost. Equivalently, this means $\epsilon = 0.22449$ and $\lambda = 0.122449$. The performance under this parametrisation can be seen in figure A.1: the simulations shown there used those exact settings.

3.2.3 Deterministic reward schedules

We know now how to parametrise the AU model to make it learn stochastic payoffs and costs. But what happens if rewards follow a deterministic pattern?

Assume, for instance, that an action reliably yields a fixed cost $-n$ followed by a fixed payoff p (the case that Bogacz (2017b) focused on). Do G and N then still reflect the magnitudes of payoff and cost after learning, if tuned with the conditions derived above? We will investigate this next.

Our approach is similar to the one we took above. We first determine the points to which G and N converge if exposed to the deterministic reward schedule. We then equate those points to the values of payoff and cost. Finally, we check if any new conditions arise from these equations. Throughout, we assume that the conditions derived above hold—we are interested in additional conditions here.

So let us determine the fixed points of the learning rules for the deterministic reward schedule. The fixed points are simply those values of G and N (or equivalently of the alternative variables Q and S) that are invariant under the updates caused by further exposure to the reward schedule[1]. As in the previous section, we denote the fixed points by G^* and N^*, or Q^* and S^*.

First, we focus on determining the fixed point of Q. Note that each encounter with the action yields two updates of Q: one due to the cost and one due to the payoff. Formally this means

$$Q_{\text{after action}} = Q_{\text{before action}} + (\Delta Q)_{\text{cost}} + (\Delta Q)_{\text{payoff}}. \tag{3.21}$$

To find Q^*, we require that an update does not change the value of Q, i.e., $Q_{\text{after action}} = Q_{\text{before action}}$ or

$$(\Delta Q)_{\text{cost}} + (\Delta Q)_{\text{payoff}} = 0. \tag{3.22}$$

Using the update rules equation 2.7 and equation 2.8, we find that

[1] In the deterministic case considered in this subsection, the invariance is exact. In the stochastic case we explored in subsection 3.2.2, the fixed points were invariant on average.

$$\begin{aligned}(\Delta Q)_{\text{cost}} &= \alpha_Q \left(r_{\text{cost}} - Q_{\text{before action}}\right) - \lambda Q_{\text{before action}} \\ &= \alpha_Q \left(-n - Q_{\text{before action}}\right) - \lambda Q_{\text{before action}}.\end{aligned}$$

Similarly, using $Q_{\text{after cost}} = Q_{\text{before action}} + (\Delta Q)_{\text{cost}}$ we find that

$$\begin{aligned}(\Delta Q)_{\text{payoff}} &= \alpha_Q \left(r_{\text{payoff}} - Q_{\text{after cost}}\right) - \lambda Q_{\text{after cost}} \\ &= \alpha_Q \left(p - Q_{\text{after cost}}\right) - \lambda Q_{\text{after cost}} \\ &= \alpha_Q \left(p - \left(Q_{\text{before action}} + (\Delta Q)_{\text{cost}}\right)\right) - \lambda \left(Q_{\text{before action}} + (\Delta Q)_{\text{cost}}\right).\end{aligned}$$

Finally, we substitute $(\Delta Q)_{\text{cost}}$ from above into this expression, and then insert $(\Delta Q)_{\text{cost}}$ and $(\Delta Q)_{\text{payoff}}$ into equation 3.22. Solving the equation for $Q_{\text{before action}}$ yields

$$Q^* = \frac{1}{2 - \alpha_Q - \lambda} \left(n(\alpha_Q + \lambda - 1) + p\right), \tag{3.23}$$

where $\alpha_Q = \alpha\,(1 + \epsilon)\,/2$. Now, recall that the definition of Q in terms of G and N is $Q = 1/2 \times (G - N)$, and that the true payoffs and costs of this reward schedule are p and n. With $G^* = p$ and $N^* = n$, equation 3.23 reads

$$\frac{1}{2 - \alpha_Q - \lambda} \left(n(\alpha_Q + \lambda - 1) + p\right) = \frac{1}{2}\,(p - n)\,. \tag{3.24}$$

Just as equation 3.15 and equation 3.16, this equation represents a meeting point of the AU learning rules on the left-hand side and the payoffs-costs hypothesis on the right-hand side. For the equality to hold, we must only have

$$\alpha_Q + \lambda = 0. \tag{3.25}$$

This is a novel condition, different from those that we found above. The definition of α_Q and the previously derived conditions in equation 3.19 and equation 3.20 may be used to transform this novel condition into the simpler form $\alpha \ll 1$. We find that learning good approximations of the payoffs and costs in the deterministic alternating reward schedule requires a small learning rate.

Next, we repeat the same analysis for S. Since we search for additional conditions on the parameters, we are free to use the original conditions in equation 3.19 and equation 3.20 to simplify our calculations. The only complication we encounter is the appearance of Q in the update rules of S, which we resolve by substituting Q with Q^*, acknowledging that the fixed points of S and Q depend on each other. We arrive at

$$S^* = \frac{1}{2}(p+n)\,. \tag{3.26}$$

Again, we use the definition $S = 1/2 \times (G+N)$ to compare the result of learning with the value required to represent payoffs and costs. We find that now further conditions arise here. Thus, equation 3.25 is the only additional condition for successful learning of payoff and cost from rewards that follow an alternating pattern.

From the results presented in this section, we conclude that the learning rules in equations 2.7 and 2.8 facilitate learning of the magnitudes of fixed payoffs and costs that occur reliably one after the other. However, we also saw that for this to work, the learning rate α must be small, in addition to the conditions that we derived in the previous section.

3.2.4 Robustness

So far, we saw that the AU model can learn payoffs and costs only approximately: we derived conditions for the parameters of the AU learning rules, and saw that these conditions cannot be satisfied exactly. In our simulations below we have

to use parameter settings that violate the conditions. Any biological mechanism that implements the learning rules will also have to violate the conditions.

It is therefore important to ask how robust the model is with respect to parameter detuning. How much violation of the conditions can the rules take without breaking? Here, we first describe the effect of parameter detuning on the values to which G and N converge. Then, we will show that the algorithm will produce useful results even under substantial detuning of the parameters.

We are interested in the encoding of payoffs and costs after learning, and should therefore investigate the equilibrium values G^* and N^*. Those equilibrium values may be obtained via combination of the equilibrium values of Q^* and S^* given in equation 3.9 and equation 3.12:

$$G^* = Q^* + S^* = c_Q q + c_s \mathbb{E}\left|r - c_q q\right| \approx c_Q q + c_s s \tag{3.27}$$

$$-N^* = Q^* - S^* = c_Q q - c_s \mathbb{E}\left|r - c_q q\right| \approx c_Q q - c_s s. \tag{3.28}$$

Here, we assumed that the average spread around $c_Q q$ is approximately equal to the average spread around q, which is a good approximation if only a tiny fraction of all rewards fall between $c_Q q$ and q. This will hold for most cases of interest, especially for the bimodal reward distributions we discussed above.

Next, we can use the relation of payoffs p and costs n to the statistics q and s of the reward distribution they generate. These relations are given in equation 3.5 and equation 3.6; inverting and inserting those yields

$$G^* \approx \frac{1}{2}\left(c_Q + c_S\right) p - \frac{1}{2}\left(c_Q - c_S\right) n \tag{3.29}$$

$$N^* \approx -\frac{1}{2}\left(c_Q - c_S\right) p + \frac{1}{2}\left(c_Q + c_S\right) n. \tag{3.30}$$

We find that as long as $c_Q = c_S$, the Go and No-Go weights converge to values proportional to the payoffs and costs. Thus, as long as $c_Q = c_S$, the payoffs and costs are encoded separately in the two pathways.

Expressed in terms of the elementary parameters α, λ and ϵ, and solved[2] for ϵ, this condition becomes

$$\epsilon = \sqrt{(2\lambda/\alpha)^2 + 1} - 2\lambda/\alpha. \tag{3.31}$$

The condition in equation 3.31 is equivalent to the condition in equation 25 of Bogacz (2017b) for the convergence of G and N to values proportional to p and n. Bogacz (2017b) arrived here from deterministic reward sequences and a proportionality requirement, while our route started at stochastic rewards and the exact representation of payoffs and costs. It is reassuring that both routes, albeit different, arrive at the same result.

According to equation 3.31, if λ/α is very small (i.e. if unlearning is weak relative to learning), then ϵ approaches one, rendering the learning rules approximately linear. If, on the other hand, λ/α is very large (i.e. unlearning is very strong compared to learning), then ϵ approaches zero, rendering the learning rules maximally nonlinear.

This relationship between ϵ and λ is not surprising; in fact, the intuitive approach to the AU learning rules we offer in the appendix A suggests that unlearning is necessary to balance the unconstrained strengthening of the weights that results from introducing the nonlinearity (compare figure A.1B and figure A.1C). Equation 3.31 makes this explicit: the stronger the nonlinearity (i.e. the closer ϵ gets to zero), the stronger the required unlearning.

In summary of the last few paragraphs, we can now say that a detuning of the parameters may leave the pathway-specific encoding of payoffs and costs intact,

[2]Note that a second solution with $\epsilon < 0$ exists; we ignore that second solution because it is not biologically plausible.

given that detuning in one parameter (say λ) is compensated by changing another parameter (say ϵ) in the opposite direction. However, every such detuning will introduce an overall scaling.

Now, after investigating the effect of detuning on G^* and N^*, let us explore the effect of detuning on the overall evaluation of actions. We saw in chapter 2 that in the AU model the ultimate evaluation of an action is given by the corresponding thalamic activity T, defined as a function of G, N and D in equation 2.9. Substituting the equations 3.29 and 3.30 into the definition of T we obtain

$$T = p\left(\frac{1}{2}c_Q - \frac{1}{2}c_S + Dc_S\right) - n\left(\frac{1}{2}c_Q + \frac{1}{2}c_S - Dc_S\right). \tag{3.32}$$

We find that when $c_Q = c_S \neq 1$, the activation T becomes scaled by a constant c_S, but as this scaling is the same for all actions, the network can still select actions on the basis of payoffs and costs modulated by motivation signal D, in the same way as it normally would. Importantly, the effect of dopamine—to emphasize the payoffs when increased, and emphasize the costs when decreased—is still present as long as $c_S > 0$ even if $c_Q \neq c_S$. These signature effects of the proposed mechanism are thus robust even under significant detuning.

However, we find that when $c_Q \neq c_S$, the dopaminergic motivation signal D would have a relatively smaller effect on changing the weighting between payoffs and costs; for example, the payoffs or costs could no longer be ignored by setting D to its extreme values of 0 or 1. From this analysis, we may conclude that while action selection is quite robust under violation of the derived conditions, dopaminergic regulation works most effectively if $c_Q = c_S$.

3.3 Simulations

How do the AU learning rules work in practice? Which predictions does the AU model make in realistic experimental settings? Can it explain empirical

findings? In this section, we will use simulations and model fitting to answer these questions.

We first test our theoretical results under various circumstances (section 3.3.1). Next, we simulate the experiment of Salamone, Steinpreis, et al. (1991) to show that the payoffs-costs hypothesis explains the effects of dopaminergic medication on the willingness to work for food rewards (section 3.3.2). Finally, we simulate a well-known learning task for humans, obtaining a testable prediction for the effect of a D2 agonist on learning performance (section 3.3.3).

3.3.1 Learning from different reward schedules

The previous section revealed what to expect from training the AU learning rules on certain reward schedules. In particular, we looked at totally predictable or totally random reward schedules. Here, we aim to confirm and extend the mathematical results of the previous section with numerical simulations.

Figure 3.2 shows the results of simulating AU learning for four different reward schedules. In all those simulations, G and N change according to the learning rules in equations 2.7 and 2.8. The parameters we used here (see figure 3.2B) roughly fulfil the conditions in equation 3.15 and equation 3.16 for learning the correct magnitudes of payoffs and costs, but are also chosen to facilitate quick convergence.

For the simulation in figure 3.2A, we used a reward schedule in which a cost $-n$ is reliably followed by a payoff p. This is exactly the schedule we analysed in subsection 3.2.3. We find that both weights constantly oscillate due to the alternation of payoffs and costs. This oscillating behaviour is superimposed with learning curves that take the weights from their initial values towards the correct levels. After 30 trials, G and N provide good approximations of p and n.

Figure 3.2C is similar to figure 3.2A, with a slight variation: just as in figure 3.2A, payoffs and costs alternate reliably. But while the cost is held constant at $-n$, this time the payoff is sampled from a fixed distribution (a normal distribution

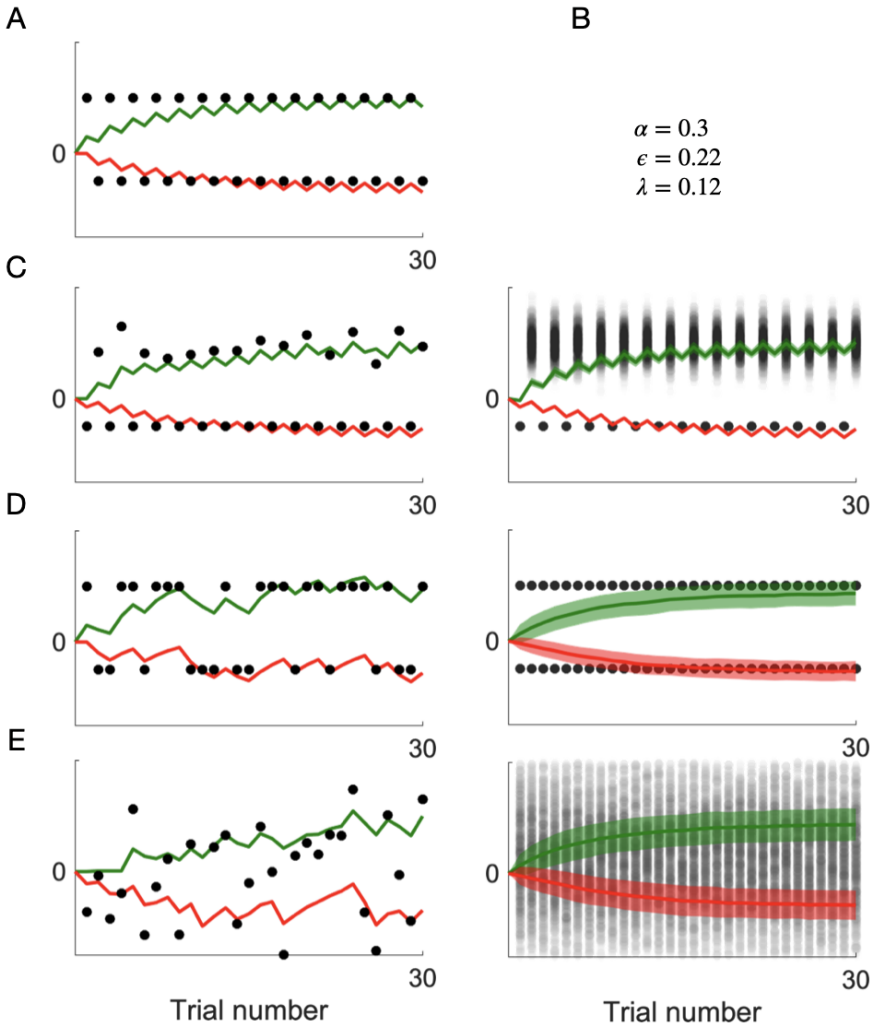

Figure 3.2: Simulations of learning the payoffs and costs of different reward schedules. In all graphs, G is depicted in green, while $-N$ is depicted in red. Rewards are indicated by solid black dots in the panels on the left, and by transparent dots on the right. Each simulation shows how G and N change over 30 trials. **A** Learning based on predictable, alternating rewards. **B** The parameter settings used in the simulations. **C–E** Single and averaged simulations of learning stochastic reward schedules. On the left, we show a single sequence of learning, with rewards sampled from different distributions. On the right, we show averages over many such sequences of learning. The mean weights are represented as green and red lines. The shaded green and red areas around these lines of G and N in the right column indicate one standard deviation. For all panels, the reward distributions were parametrised such that $p = 20$ and $n = 10$. This figure is a reproduction of figure 5 of Möller and Bogacz (2019).

with mean p and a standard deviation of $p/4$) in each trial. Thus, the schedule includes both stochastic and deterministic components: a fixed cost, followed by an uncertain reward. We find that under such conditions, N eventually represents

the cost n, while G converges towards the mean payoff p.

Finally, figure 3.2D and figure 3.2E contain simulations of rewards drawn completely at random from fixed distributions. In figure 3.2D, the rewards are either p or $-n$, with probabilities $1/2$ each[3]. In figure 3.2E, rewards are sampled from a normal distribution with mean $\mu_r = 1/2 \times (p - n)$ and standard deviation of $\sigma_r = 1/2\sqrt{\pi/2}\,(p + n)$[4].

The stochastic nature of the reward schedules in figure 3.2C - 3.2E causes the evolution of the weights G and N to be different each time the simulation is run. For those schedules, we show a single example run on the left and an average over 1000 runs on the right. We find that G and N approximate payoffs and costs as expected from the analytic treatment above. We may conclude that the appropriately tuned AU rules are capable of learning the payoffs and costs for various reward schedules, both in theory and in practice.

3.3.2 The effect of D2 blocking on the willingness to work for food

So far, we focused on how different reward schedules change the synaptic weights G and N. In this section, we introduce actions and choices between them, based on learned payoffs and costs. We do so to model a classic experiment, reported by Salamone, Steinpreis, et al. (1991). This experiment has already been modelled by Bogacz (2017b), who also used the AU rules to explain the effects of D2 blocking on the willingness to work for food. Our account here is similar, but features a more realistic rendering of the pathway-specific effects of the pharmacological intervention. We discuss the differences between the two approaches in detail in section 3.4.2.

The experiment had two conditions. In one condition ("press for pellet"), rats were given the opportunity to obtain food pellets through lever-pressing. They

[3]This is the distribution defined in equation 3.3 that we used to define payoff and cost in terms of statistics of the reward distribution.

[4]We chose σ such that that the distribution would have a spread $s = 1/2(p + n)$. For this, we use the relation $s = \sqrt{2/\pi}\sigma$ which holds for the normal distribution (Mikhael and Bogacz 2016).

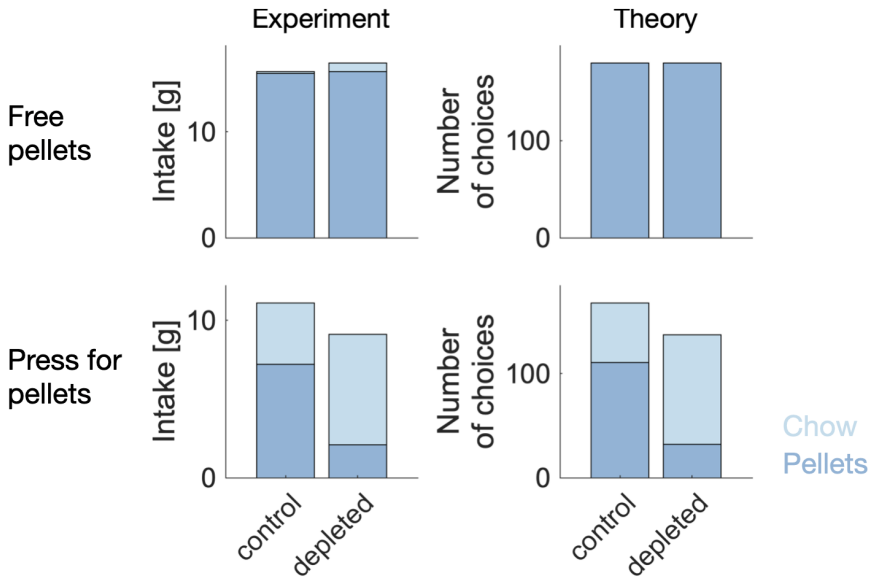

Figure 3.3: The effect of haloperidol on the willingness to work for food. We show the results of the experiment reported by Salamone, Steinpreis, et al. (1991) (left column), and the corresponding predictions of the AU model (right column). In all panels, the bars indicate the amount of food consumed. The dark blue bars correspond to consumed pellets, the light blue bars correspond to consumed lab chow. The height of the two bars combined indicates the overall consumption. The first row shows the results of the condition in which both pellets and lab chow were freely available. The second row corresponds to the condition in which the rats had to press a lever to obtain pellets. This figure was taken from Möller and Bogacz (2019) and adapted for this thesis.

were also given free access lab chow at the same time. In another condition ("free pellet"), both pellets and lab chow were freely available. In both conditions, the amounts of consumed chow and pellets were measured for a control group of rats and for a group that had been treated with the D2 antagonist haloperidol. The results of the experiment are shown in figure 3.3, left column.

We first observe that normal rats seem to prefer food pellets over lab chow, even if work is required to obtain them. We further see that dopamine depletion strongly reduced the number of pellets consumed in the condition that required work, but not in the condition where pellets were freely available. One interpretation of this is that the D2 antagonist disinhibited the indirect pathway. Because of this the learned costs of lever-pressing weighed stronger and the rats' willingness to work for food was reduced. In short, costs amplified by the D2 blocker might

have biased the rats away from the pellets. To test whether this interpretation can explain the results quantitatively, we model the experiment with the AU learning rules.

To do this, we need to connect the action evaluation T_a in equation 2.9 to choice probabilities for the actions "get pellet" and "get chow". The usual way to do this would be a softmax function, used for example by Collins and Frank (2014). However, this is not the best way here: looking at the bottom left panel of figure 3.3, we find that the group treated with the D2 antagonist differed from the control group not only in their willingness to work for food but also in their overall food consumption—the rats with D2 antagonist consumed less food in total. In our framework, this might correspond to choosing neither of the two possible actions, which would not be possible with a softmax rule.

We follow Bogacz (2017b) in using a slightly different choice mechanism instead: first, we add random noise to all T_a, to account for choice stochasticity. Then, the action a with the highest associated T_a is chosen. If $T_a < 0$ for all actions, no action is performed. It is thus possible for the modelled rats to consume neither pellet nor chow.

How can we model the effect of the drug haloperidol? Haloperidol is a D2 antagonist; it blocks the D2 receptors on the medial spiny neurons of the No-Go pathway. This blocking reduces the (inhibiting) impact of dopamine on the activity of that pathway. To account for this in our model, we introduce the factor $\kappa_N \in [0, 1]$ into our expression for the thalamic activity:

$$T_a = DG_a - (1 - \kappa_N D)\, N_a. \tag{3.33}$$

The parameter κ_N controls how much dopamine affects the activity of the indirect pathway channels N_a: $\kappa_N = 1$ (no blocking) recovers the normal thalamic activity given in equation 2.9, while $\kappa_N = 0$ (total blocking) fully removes the impact of dopamine on the indirect pathway, leading to completely uninhibited activity

of N_a. In the control group of the experiment, κ_N is set to 1 (no medication is administered, no blocking happens). In the group that received the medication, κ_N is a free parameter that must be fitted to the data. This way of modelling the effect of haloperidol is different to the way that was used by Bogacz (2017b), who did not introduce the paramter κ_N, but instead used D as a free parameter for the effect group. Our approach here is more realistic, since only one type of receptor is targeted by the drug.

We fit the thus modified AU model to the experimental results of Salamone, Steinpreis, et al. (1991) (see subsection 3.5.2 for details of the procedure). We show the results of this fit in figure 3.3. Comparing the left and the right column in that figure, we find a strong correspondence between the simulated behaviour and the experimental results: for both conditions, the model captures the ratio between pellet consumption and chow consumption, as well as the overall amount of consumed food. In particular, it captures the difference between the control group and the effect group in the "press for pellet" condition.

This suggests that the payoffs-costs hypothesis can explain the impact of a D2 antagonist on the willingness to work for food. We find $\kappa_N = 0.7507$ as the best fitting blocking parameter. This corresponds to a blocking of roughly 25 % of the D2 receptors.

Let us take a closer look at the underlying mechanisms. In figure 3.4B and C, we show the values of G and N for the two options after learning. From the left column (labelled 'pellet') we can read off that for the best fitting model, pressing the lever to obtain a pellet has a cost of 13.0 (lever-pressing) and a payoff of 15.1 (pellet-eating). At baseline dopamine level (figure 3.4B), this results in a thalamic activity of $+1.0$ for this action. For the D2-blocked group, however, the dopaminergic inhibition of the indirect pathway is decreased by 25%. With that taken into account, the thalamic activity is -0.6. The intervention turns the pellet option from generally desirable into generally undesirable.

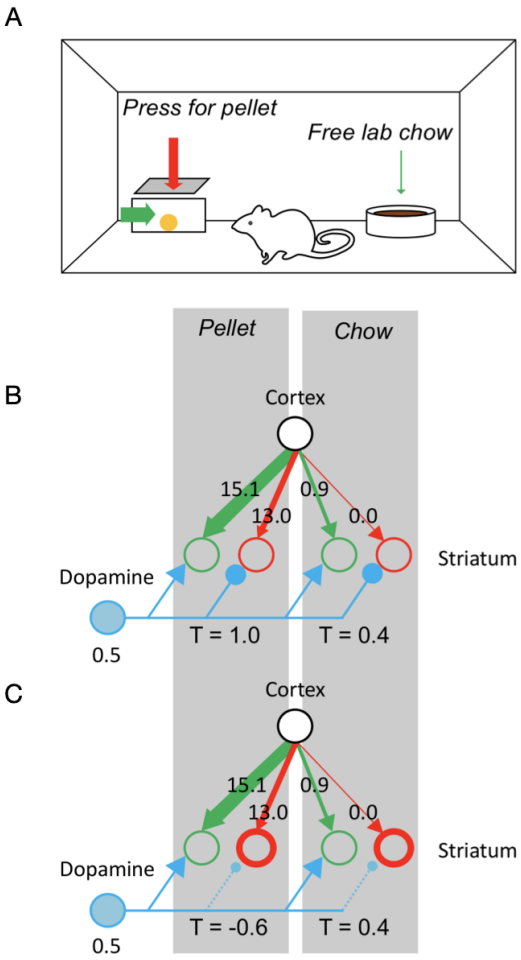

Figure 3.4: Payoffs and costs after learning. A A representation of the task of Salamone, Steinpreis, et al. (1991). Rats can press a lever to obtain food pellets or feed on freely available lab chow. **B** Weights G and N for the two options after learning, in the control group. The green arrows correspond to the direct pathway weights, the red arrows correspond to the indirect pathway weights. The width of the arrow indicates the strength of the weight, which is also annotated. The green and red circles represent action-specific striatal populations of direct and indirect pathway SPNs respectively. The dopaminergic modulation is shown in blue. The weights that correspond to the pellet option are shown on the left, the weights that correspond to the chow option are shown on the right. The resulting thalamic activity T is annotated for each action. The weights and the thalamic activity were extracted from a simulation of the experiment. Maximum likelihood paramters were used. **C** Analogous to B, but for the group treated with haloperidol. The D2-blocking effect is shown as reduced dopaminergic inhibition on the indirect pathway populations (reduced size of blue inhibitory dopamine projections, increased size of indirect pathway populations). This figure was taken from Möller and Bogacz (2019) and adapted for this thesis.

No such effect occurs to the chow option: as we can see in the right column (labelled 'chow'), eating chow has a payoff of 0.9 (a pellet is more than 15 times

as attractive!), but no cost at all. Since cost (and hence the indirect pathway) do not play a role in evaluating the chow option, D2 blockers do not affect it. The thalamic activity is 0.4 in the control group as well as in the D2-blocked group. Hence, eating chow is generally desirable for both groups (but less desirable than eating pellets in the control group).

Overall, we may conclude that the payoffs-costs hypothesis can explain the effect of haloperidol on the willingness to work for food, both qualitatively and quantitatively. From the model fit we could estimate the efficacy of D2 blocking, as well as the effective payoffs and costs of actions in this experiment. Using the thus calibrated model, one might now predict the effects of other drugs as well. This could allow for quantitative, hypothesis-driven characterisation of dopaminergic drugs.

3.3.3 The effect of a D2 agonist on learning from wins and losses

In the previous subsection, we used our model to explain the results of an experiment that has already been done. Next, we present a prediction for an experiment the results of which have not yet been published at the time of writing (as far as we are aware). We will see that this experiment might be an important test of the payoffs-costs hypothesis.

The experiment we consider in this section assesses the effect of a D2 agonist on performance in a learning task in humans. The task (which is designed to contrast learning from positive outcomes with learning from negative outcomes) is taken from Pessiglione et al. (2006) and adjusted only slightly.

The task consists of a sequence of trials. A trial can either be a *win* trial or a *loss* trial. In either condition, participants have to make a choice between two options (say between A and B in the win condition and C and D in the loss condition). In the win condition, the possible outcomes of the choice are +1 GBP or nothing. Both options can lead to both outcomes, but not with the same likelihood: one of the options (say A) has a probability of 0.7 to lead to +1 GBP, the other (say B)

leads to +1 GPB with a probability of only 0.3. In the loss condition, the outcomes are -1 GBP or nothing. Again, one option (say C) has a probability of 0.7 to lead to -1 GBP; the other option (say D) leads to that outcome with a probability of 0.3. The two conditions occur equally often but are randomly mixed during 120 trials.

Participants who want to maximize their financial gain in this task have to learn and exploit the dependencies between options and outcomes. For the win condition, they must determine that option A is more likely to result in the win +1 GBP. To do this, they must try out both options and learn from the wins they observe. For the loss condition, it is key to determine that option D is less likely to result in the loss -1 GBP. Here, participants must learn from losses. Comparing participants' performance on the two conditions then allows assessing how well they learn from wins relative to how well they learn from losses.

How would dopaminergic medication alter performance in this task? To answer this question based on the payoffs-costs hypothesis, we simulate behaviour using the AU model. For each of the options, we define an action channel with one G and one N weight. We go through the trials, using the choice rule outlined in section 3.3.2 to make decisions and the learning rules in equation 2.7 and 2.8 to adjust the weights depending on the outcome.

How do we model effect of the D2 agonist? As in the previous section, we model the thalamic activity of an action channel with equation 3.33. A D2 agonist inhibits activity in the indirect pathway. Therefore the factor κ_N should be larger than 1 in the effect group while staying at 1 in the control group.

How do we model choice? The choice rule we defined above allows "doing nothing" as an outcome. This outcome occurs when the thalamic activity of all options is below zero. Of course, doing nothing is not an acceptable response in a two-alternative forced-choice paradigm like the one we are modelling here. We solve this issue by repeated sampling: after the thalamic activity for all options is computed from G and N, we add Gaussian noise to each channel. If all activities

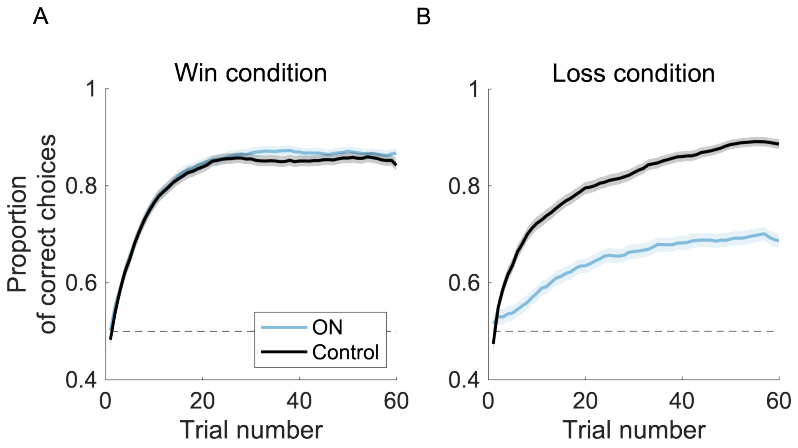

Figure 3.5: The effect of a D2 agonist on learning from gains and losses. Trial-by-trail learning curves were computed per condition (win and loss) and group (control and D2 agonist group—ON), from $N = 1000$ simulated participants per group. We show the proportion of correct choices as a function of the trial number. A moving average with a window size of 10 was used to smooth the curves. The blue and black lines correspond to the ON and control group, respectively. The shaded areas indicate the standard error of the mean across simulated participants. The dashed line indicates the chance level. **A** Learning in the win condition. **B** Learning in the loss condition.

are below zero, we sample the noise again. Otherwise, we select the action with the highest activity. These steps are repeated until an action is selected.

We have now fully specified the model, and can use it to make a prediction. If the model captures the essential relevant mechanisms, what behavioural effects will a D2 agonist have in this task? To assess this, we tuned the AU model such that it predicts learning curves that appear realistic when compared with human performance. We then simulate $N = 1000$ participants for both the control and the effect group and compute the average learning curves for each condition in each group. Those are shown in figure 3.5. For details of the simulation procedure, see subsection 3.5.3.

We find a small but significant difference between the control group and the D2-agonist group for the win condition ($t_{59} = 4.85$, $p < 0.001$, two-tailed paired t-test), and a clear difference for the loss condition ($t_{59} = 28.93$, $p < 0.001$, two-tailed paired t-test). The D2 agonist group is slightly better in detecting the option that is more likely to win, but much worse in avoiding the option that is more likely to lose. From this, we conclude: our model predicts that a D2 agonist

should mainly decrease asymptotic performance in the loss condition. The effect is strong, and should be visible even if $N \ll 1000$. The effect in the win condition is very small and might not be visible in studies with lower power.

Why does the model predicts this? The model will interpret losses as costs and store information about their magnitude in N. Wins will be considered payoffs, stored in G. Therefore, the two conditions of the task recruit different pathways for decision making. The D2 agonist acts by inhibiting D2-SPNs, hence inhibiting activity in the indirect pathway. In the model, this corresponds to decreasing the impact of N—and hence of costs—on choices. This leads to decreased loss aversion and ultimately worse performance in the loss condition. The direct pathway is not affected. The impact of payoffs on decisions is thus not modulated, so there is no big effect in the win condition.

We have seen that our model makes a very clear prediction for the effect of a D2-agonist on learning from losses versus learning from gains: treatment with a D2 agonist should decrease performance in the loss condition, but not alter performance in the win condition substantially.

3.4 Discussion

In this chapter, we have shown that the AU model can learn payoffs and costs in various different scenarios. We have determined how parameters must be tuned for this to work. Further, we have demonstrated that the model can explain the effect of a D2 antagonist on the willingness to work for rewards. Finally, we have made a prediction for the effect of a D2 agonist on learning from wins versus learning from losses.

In the remainder of this chapter, we discuss two aspects of our results. First, we unpack an implicit assumption of our theory. Payoffs and costs can only be learned as described above if dopamine reacts to positive *and* negative results, i.e., to rewards as well as to costs. While it is well known that dopamine responds to rewards, its response to negative outcomes is less clear.

Second, we will have a closer look at how our results here relate to previous results by Bogacz (2017b) and Mikhael and Bogacz (2016). Since we build on their work, some of our analyses are closely related to what they have done. In 3.4.2, we therefore point out the unique contributions of the work presented above.

3.4.1 Dopamine responses to negative events

The AU model can only learn about payoffs and costs if it receives them as an input. The relevant input channel is the reward signal r. For our model to work, we need that signal to provide an evaluation of all events, both positive and negative. The AU model can then extract the average payoffs and costs from that signal.

The reward signal is assumed to affect striatal dopamine release, such that dopamine broadcasts reward prediction errors. Therefore, it is implicit in our theory that dopamine should respond to positive as well as to negative events, providing an evaluation adjusted by learned expectations. Is this a realistic assumption?

Plenty of evidence suggests that dopamine responds to positive events (i.e. rewards, such as food, water, cocaine, ...) and reward-predicting stimuli—we reviewed this in chapter 2. Its association with negative events is more controversial, but several pieces of evidence suggest that negative outcomes are represented by dopamine pauses: Ungless et al. (2004) show this for pain, Matsumoto and Hikosaka (2009) show it for air puffs (which are adverse for monkeys) and Zaghloul et al. (2009) show it for monetary losses. In addition, effort-related cost signals have recently been found in the human striatum (Suzuki et al. 2020). The assumption that dopamine might provide evaluations of negative as well as positive outcomes seems to be supported by empirical findings.

3.4.2 Relation to previous work

The work that we presented in subsections 3.2.2, 3.2.3 and 3.3.2 is closely related to the results of Mikhael and Bogacz (2016) and Bogacz (2017b). Here, we point out the differences between those results and ours.

In subsection 3.2.2, we performed a stochastic fixed-point analysis of the AU rules to determine how parameters must be chosen to ensure that G and N converge to payoff and cost. A similar analysis was conducted by Mikhael and Bogacz (2016). There, the aim was to find conditions on the parameters that would ensure that $G - N$ and $G + N$ converge to values proportional to q and s respectively. The condition that was found was $\lambda \ll \alpha(1 + \epsilon)$. Our aim in subsection 3.2.2 was different: we aimed to find conditions that would ensure that G and N converge to p and n *exactly*, which is equivalent to $G - N$ and $G + N$ converging to q and s *exactly*. This requirement is stronger than the requirement of Mikhael and Bogacz (2016). Accordingly, we find more conditions than Mikhael and Bogacz (2016), namely $2\lambda \ll \alpha(1 + \epsilon)$ and $2\lambda = \alpha(1 - \epsilon)$. The first of these conditions mirrors the condition of Mikhael and Bogacz (2016), with the factor 2 stemming from our modification of equation 2.6. The second condition is a new result of our analysis.

Furthermore, we introduced a new parametrisation of the rules (c_Q an c_S instead of ϵ and λ, see equation 3.13 and 3.14), and determined exactly how one parametrisation can be translated into the other one. The parametrisation in terms of c_Q and c_S allows to precisely specify the desired precision of the estimates of q and s, and is thus a convenient way to determine parameters ϵ and λ for simulations.

In subsection 3.2.3, we looked for conditions that would ensure that the learning rules converge to the correct values also in the case of deterministic reward schedules. A similar analysis was reported by Bogacz (2017b). There are two differences between that analysis and the one presented here.

First, Bogacz (2017b) used an approximation to simplify the calculations. Considering a cost followed by a payoff, updates were computed assuming that the weights would only change after both had been received. The change after the cost but before the payoff was neglected. Here, we do not neglect this effect. Taking it into account leads to a new condition: the learning rate α must be small. While this seems intuitive, it has not been formally shown previously.

Second, the condition of Bogacz (2017b) was derived from the requirement that G should converge to a value *proportional* to p, while N should converge to a value *proportional* to n. Again, our requirement—that G and N should converge to p and n exactly—is a special case of the requirement of Bogacz (2017b), and again we obtain two conditions where Bogacz (2017b) obtains only one.

Finally, in subsection 3.3.2, we used the AU model in combination with a choice rule to explain behaviour in an experiment conducted by Salamone, Steinpreis, et al. (1991). A similar analysis was reported by Bogacz (2017b). However, Bogacz (2017b) assumed that the effect of haloperidol would affect the level of tonic dopamine D for both pathways. Here, we modelled the effect of haloperidol differently, by introducing a pathway-specific blocking factor κ_N (see equation 3.33). This affects both the quality of the fit and the values of the fitted parameters.

Overall, while there a parallels between the work presented in this chapter and the work of Mikhael and Bogacz (2016) and Bogacz (2017b), there are also clear differences. The work presented here—most of which published by Möller and Bogacz (2019)—represents the most recent version of the payoffs-costs hypothesis at the time of writing.

3.5 Methods

3.5.1 Simulations of learning payoffs and costs

To simulate the learning of payoffs and costs from different reward distributions, we use the AU learning rules 2.6, 2.7 and 2.8 with parameters fixed at $\alpha = 0.3$, $\epsilon = 0.22$ and $\lambda = 0.12$. For stochastic rewards (figure 3.2B, 3.2C, 3.2D) we simulate 1000 runs each. We set $p = 20$ and $n = 10$ for all simulations.

3.5.2 Simulations of the task of Salamone et al. (1991)

To model the task, we define the two possible actions: $a \in \{\text{pellet}, \text{chow}\}$. Selection of an action a is always followed by two rewards, n_a and p_a. Each reward is followed by an update of the corresponding weights, which means that there are two updates in each trial. This reflects the sequence of events in the task: the effort of pressing the lever is registered before the rewarding pellet.

Learning in this task is modelled with the AU rules (equation 2.6, 2.7 and 2.8), decision-making is set up as follows: first, the thalamic activity of each action is computed as

$$T_a = DG_a - ((1 - \kappa_N D)\, N_a + E, \tag{3.34}$$

with $E \sim \mathcal{N}(0, \sigma^2)$. If $T_a < 0$ for all actions, no action is selected and we jump to the next trial without any updates. Otherwise, the action with largest T_a is selected and rewards and updates commence before the next trial starts.

The task has a 2-by-2 design: there is a control group and a group treated with a D2 antagonist; both groups complete a condition with free pellets and a condition in which pellets must be earned by lever presses.

This design is reflected in our parameter setting: the parameter κ_N is free in the antagonist group and fixed to $\kappa_N = 0$ in the control group. Further, the cost n_{pellet} is a free parameter in the "press for pellets" condition and fixed at $n_{\text{pellet}} = 0$ in

the "free pellets" condition. The parameters σ and p_{pellet} are free without any restrictions, and all other parameters are fixed: $p_{\text{chow}} = 1$ and $n_{\text{chow}} = 0$ define the rewards following the action *chow*, $\alpha = 0.1$, $\epsilon = 0.632653$ and $\lambda = 0.0204082$ parametrise the learning rules, $D = 0.5$ parametrises the choice rule and $G_0 = 0.1$, $N_0 = 0.1$ define the initial values of the weights.

To simulate the experiment, we first go through a sequence of forced choices (180 trials for each action) to ensure that payoffs and costs are learned properly. We then proceed to 180 trials of free choices, with learning is still taking place.

The free parameters are determined by minimising a loss function. To compute the loss, we simulate the entire experiment repeatedly ($N = 100$) and aggregate choices per group and condition—let n_{agc} denote the number of times action a was selected in the group g and condition c. The loss function is defined as

$$L = \sum_c \sum_g \sum_a \left(\frac{n_{agc}}{n_{gc}} - \frac{f_{agc}}{f_{gc}} \right)^2,$$

where f_{agc} is the amount of food a consumed by group g in condition c as reported by Salamone, Steinpreis, et al. (1991). The normalisation factors $n_{gc} = \sum_a n_{agc}$ and $f_{gc} = \sum_a f_{agc}$ make the experimental data comparable with the choice frequencies extracted from the simulations.

The loss function was minimised using a combination of grid search and Matlabs's *fminsearch* function. The parameters that minimized the loss function were $p_{\text{pellet}} = 15.511751$, $n_{\text{pellet}} = 14.510517$, $\kappa_N = 0.7507$ and $\sigma = 1.066246$.

After minimising the loss function, the data in figure 3.3 were generated by simulating the experiment with N=6, which is the sample size of Salamone, Steinpreis, et al. (1991).

3.5.3 Simulation of the task of Pessiglione et al. (2006)

We modelled the task with two trial types, win and loss. On a win trial, the learner had to choose between $a_{\text{high-win}}$ and $a_{\text{low-win}}$. Choosing $a_{\text{high-win}}$ produced

a reward of $r = 1$ with probability $p = 0.7$ and a reward of $r = 0$ with probability $p = 0.3$. The other action $a_{\text{low-win}}$ produced a reward of $r = 1$ with probability $p = 0.3$ and a reward of $r = 0$ with probability $p = 0.7$. Probabilities of winning were thus inversely related for the two actions.

On a loss trial, the learner had to choose between $a_{\text{high-loss}}$ and $a_{\text{low-loss}}$. Choosing $a_{\text{high-loss}}$ produced a reward of $r = -1$ with probability $p = 0.7$ and a reward of $r = 0$ with probability $p = 0.3$. Here, the other action $a_{\text{low-loss}}$ produced a reward of $r = -1$ with probability $p = 0.3$ and a reward of $r = 0$ with probability $p = 0.7$. Again, probabilities were inversely related for the two actions, but this time the non-zero outcome was a loss.

The two trial types were randomly intermixed. In each run we included 60 trials of each trial type, yielding a sequence of 120 trials per run. We simulated $N = 1000$ runs for both the control group and the effect group.

To model learning and choice, we used the AU learning rules (equation 2.6; 2.7 and 2.8) and a choice rule rule based on equation 3.34 as above in subsection 3.5.2. However, in contrast to the procedure described in subsection 3.5.2, we did not accept no action as a possible outcome. If T_a was smaller than 0 for all actions, we sampled equation 3.34 again, until T_a was larger than zero for at least one action a. Then, the action with the largest T_a was chosen.

We parametrised the model with $\alpha = 0.4$, $\sigma = 0.2$, $\epsilon = 0.44$, $\lambda = 0.1$, $D = 0.5$, $G_0 = 0$ and $N_0 = 0$. We used $\kappa_N = 1$ for the control group and $\kappa_N = 1.8$ for the effect group. Parameters were chosen such that the learning curves resembled the empirical curves reported by Pessiglione et al. (2006).

4

The scaled prediction error model

Contents

In the previous chapter, we provided detailed analysis of the AU learning rules. Using analytical methods and simulations, we showed that those rules can be used to learn payoffs and costs (or, equivalently, the mean and spread of a reward signal), at least approximately. We also saw that the dopaminergic modulation of learned payoffs and costs provided a mechanistic explanation for the effect of dopaminergic drugs on the willingness to work for food.

However, we have not yet explained why it is *beneficial* for an organism to track payoffs and costs separately—or, equivalently, why it is beneficial to track reward uncertainty in addition to the average reward. In this chapter, we explore one reason to do this. We introduce a new version of the AU model—the scaled prediction error model—which can leverage its estimate of reward uncertainty to improve its own learning and action selection performance. To understand intuitively how and why this works, let us take a step back and consider a fundamental problem that all organisms face.

For any organism, better decisions mean better chances of survival. Reward prediction is an important aspect of this—for example, if an organism can predict the size of a food reward associated with some behaviour, it can decide whether it is worth to engage in that behaviour or not. Reward predictions are typically based on values learned from previous reward observations; we discussed various reward learning models in chapters 2 and 3.

Now, Piray and Daw (2020a) argue that when trying to predict rewards, the organism faces two challenges. The first challenge is the dynamic nature of the environment: reward sizes and contingencies might change over time, in ways that cannot be predicted. Such genuine changes in the environment can be quantified by the typical rate of change, which is called *process noise*. The second challenge is *observation noise*: even if the environment is stable, rewards will vary

from experience to experience. This could be due to the random nature of the environment, but also to variability in the organism's own behaviour, or to noise in the organism's perception and evaluation systems[1].

What is the best reward prediction method an organism could use when facing process noise and observation noise? Similar problems occur in engineering, for example in the context of navigation. There, a very versatile solution has been found. That solution, called the *Kalman filter* (Simon 2006), is very widely used—it even played a role in the moon landing (Grewal and Andrews 2010). The Kalman filter describes how estimates of a variable must be updated when new noisy observations of that variable become available. For certain types of signals, it can be shown that the Kalman filter is indeed the best method for prediction in the presence of noise. The method has proven useful not only in engineering, but also to model neural and behavioural processes (Gershman 2014)—we will discuss this in more detail below.

However, if one wants to use a Kalman filter to predict rewards, one runs into a problem: the Kalman filter requires estimates of the magnitudes of process noise and observation noise as parameters. Where to take these values from? An organism might either estimate them somehow, or use fixed (perhaps genetically determined) values. The latter option bears a risk: if the world changes, the quality of the organism's predictions might decline strongly.

Solutions for this have been proposed. For example, Piray and Daw (2020b) present a model that tracks process noise and observation noise alongside reward, allowing for *adaptive* Kalman filtering. However, their model is biologically implausible and computationally expensive, especially with regards to tracking observation noise. It is hence not suitable to describe biological learning on the mechanistic or the algorithmic level.

[1]The stock market provides a nice example of this: consider the changes of a stock prize as a reward signal. Most of the variability of that signal will be due to random fluctuations—this can be classified as observation noise. However, a part of the signal's variability will reflect genuine lasting changes in the stock prize. This part should be classified as process noise.

This leads us to the central question of this chapter: how might organisms track observation noise in a biologically plausible, computationally simple way, and use it for adaptive reward prediction? We have seen in previous chapters that there are simple, biologically plausible models (the AU model, for example) that can track quantities related to observation noise. Might this be part of the answer?

In the following sections, we derive and analyse a new model—the *scaled prediction error model*—that tracks observation noise and uses it for adaptive reward prediction. The model is computationally simple, biologically plausible and can be derived from Bayesian principles. Its simplicity is achieved through several approximations, which we make explicit in the derivation of the model (section 4.1). We then show that the model outperforms models that do not adapt to observation noise, both in reward prediction (subsection 4.2.3) and in action selection (subsection 4.2.4). Finally, we analyse the model's biological plausibility, in particular concerning the basal ganglia system (subsection 4.3.1), and test whether evidence of scaled prediction error learning can be found in behavioural data from a recent study (subsection 4.3.2).

4.1 Derivation

One way to derive the scaled prediction error learning rules is the Bayesian mode-matching method, which is a novel method to approximate Bayesian learning[2]. We introduce this method in subsection 4.1.1. We then apply the method to the problem of tracking the mean and standard deviation of a signal in subsection 4.1.2 and find a new set of learning rules.

Those rules can be shown to learn unbiased estimates of mean and standard deviation (see subsection 4.1.3), and they feature a scaled prediction error signal —hence the name.

[2]We did not find any previous accounts of mode-matching, but we cannot rule out that they exist (for example in the statistics literature).

4.1.1 The mode-matching method

The mode-matching method is based on Bayesian principles. Let us consider the problem of learning the mean and standard deviation of a signal. A fully Bayesian learner would maintain a belief about the values of the mean and standard deviation at all times, encoded as a probability distribution over all possible pairs of values. It would also maintain a generative model of the signal.

When the learner is provided with new information (say another sample of the signal), it applies Bayes law to combine its current belief (now the prior) and the likelihood of the observation (computed using the generative model) into a posterior distribution, which encodes its belief after observing the sample. This process is then repeated ad infinitum, with the posterior after one sample turning into the prior for the next.

Now, consider a learner that cannot encode arbitrary belief distributions. Instead, it can only adapt a few of the parameters of a belief distribution with otherwise fixed shape. For example, it might encode a belief using a normal distribution with fixed width, and update it by adapting the mean. How might such a learner—let us call it a fixed-shape learner—approximate a fully Bayesian learner best?

Here, we propose the mode-matching method: after observing a new sample, the fixed-shape learner should change the parameters of its belief distribution such that the maximum of the distribution (its mode) is aligned with the maximum of the true posterior. We show this process schematically in figure 4.1.

After the update, the fixed-shape learner's belief is still different from the true posterior. This is because the shape of the true posterior is generally not the same as the fixed belief shape that the learner uses. Hence, mode-matching is only an approximation of Bayesian learning, and some features are lost in this approximation.

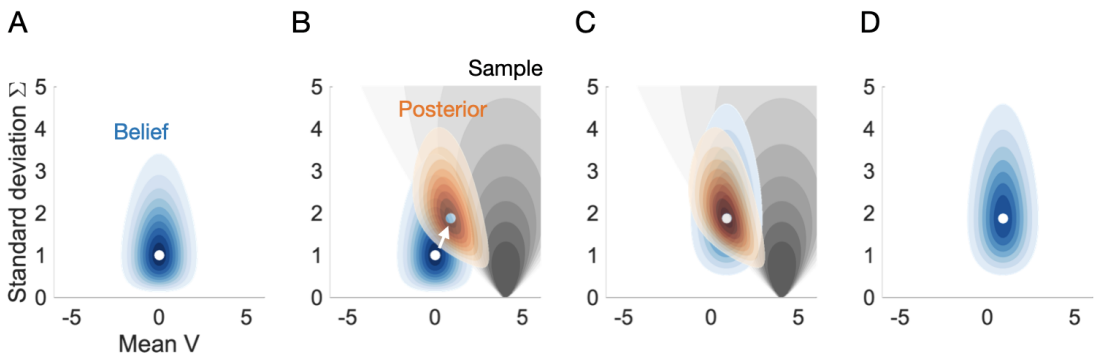

Figure 4.1: The mode-matching method. The fixed-shape belief distribution is represented by a blue shading; darker shades of blue indicate higher probabilities. Similarly, the true posterior distribution is represented as an orange shading, and the likelihood of the sample is represented by a grey shading. The axes represent the mean and standard deviation of a signal. Units are arbitrary. **A** The fixed-shape belief distribution encodes the learner's knowledge before a new observation is made. **B** A new observation is made. The likelihood of the observed value is indicated by the grey shading. Using the fixed-shape belief as a prior, a posterior can be computed. The posterior's mode is different from the mode of the fixed-shape belief; therefore, an update (indicated by a white arrow) is required. **C** The fixed-shape belief has been modified such that its mode aligns with the mode of the true posterior. **D** The modified fixed-shape belief represents the learner's knowledge after the new observation has been taken into account. The distributions were computed using the densities given in equation 4.6–4.8

Mode-matching is formally related to the variational Bayes scheme (Dayan et al. 1995; Kingma and Welling 2013; Bogacz 2017a; Buckley et al. 2017), which works by minimizing the Kullback-Leibler divergence between the true posterior and a fixed belief shape (usually a multivariate normal distribution). However, mode-matching does not minimise the Kullback-Leibler divergence; instead, it minimises the distance between the modes of the distributions.

The learning rules that can be derived with the mode-matching method are not as precise as those derived from variational Bayes, let alone fully Bayesian learning. What makes mode-matching interesting is that it can be used to derive relatively simple, tractable learning rules, as we shall see in the next section.

4.1.2 New learning rules via mode-matching

Let us consider a situation in which an organism tracks the size of a reward associated with some behaviour. By engaging in that behaviour it samples the

reward size r. Using these samples, it attempts to estimate the mean reward V that can be expected from performing the behaviour at any given time.

To derive the learning rules for this situation, we start by specifying a model for the reward-generating process, and the learner's fixed-shape belief distributions over the process variables. We model rewards as normally distributed around a mean V, with a standard deviation S:

$$r \sim \mathcal{N}\left(V, S^2\right). \tag{4.1}$$

Note that S quantifies trial-by-trial fluctuations, and therefore observation noise. The distribution in equation 4.1 is stationary—this means that the environment is modelled as stable.

We further assume that the learner maintains beliefs about V and S in form of a normal distribution over possible values of V and a gamma distribution over possible values of S:

$$V \sim \mathcal{N}\left(\mu_V, \sigma_V\right) \tag{4.2}$$

$$S \sim \Gamma\left(a, b\right) \tag{4.3}$$

The learner can change its beliefs by adapting the mean (and hence mode) μ_V of the normal distribution, and the mode $\kappa_S := (a-1)/b$ of the gamma distribution. The standard deviation σ_V and the rate parameter b stay fixed.

How should we interpret this belief encoding? Allowing μ_V and κ_S to vary means that the learner considers both the mean reward V and the observation noise S as unknown—it can adapt its beliefs about these variables. Fixing σ_V and b means that the learner's uncertainty about the mean and the standard deviation of the signal are kept constant—it cannot adapt those. The learner will thus not become more certain about either the mean or the standard deviation as it gathers

more and more data. Fixing σ_V and b keeps the resulting learning rules simple. An additional advantage of this design arises when the environment fails to be stationary—then, high certainty about the tracked variables would prevent the learner from adapting to new situations. We investigate such non-stationary environments in more detail in subsection 4.2.2.

The model of the reward generating process and the learner's belief system form our central assumptions—the rest follows. The learner we are about to derive will interpret all rewards it sees as being sampled from a normal distribution with fixed mean and variance, and it will make its inferences accordingly.

Now, let us use mode-matching to derive learning rules from our assumptions. To find out how the learner should update μ_V and κ_S after sampling a reward r, we must first find the mode of the true posterior distribution. For this, we can use a well-known way to simplify calculations.

Bayes' theorem states that

$$P(\theta|x) = P(x|\theta)P(\theta)/P(x)$$

with θ the parameters that are to be inferred and x the data that is observed. Now we notice that

$$\log\, P(\theta|x) = \log\, P(x|\theta) + \log\, P(\theta) - \log\, P(x),$$

with the last term independent of the parameters θ. We can define the function

$$E := \log\, P(x|\theta) + \log\, P(\theta),$$

and it is easy to see that the parameters $\theta_{\max}$ that maximise the function E also maximise the posterior distribution $P(\theta|x)$ (this is true because the logarithm is strictly monotonic). The function E, often called *energy* in analogy to statistical

physics, is related to the famous of free energy function which plays a key role in many contemporary theories of brain function (K. Friston 2010; Bogacz 2020; Gershman 2019).

In the case at hand, the function E is given as

$$E := \log\left(p\left(r|V,S\right)\right) + \log\left(p\left(V|\mu_V,\sigma_V\right)p\left(S|a,b\right)\right) \tag{4.4}$$

$$= \log\left(S^{-1}\exp\left(-\frac{1}{2}\frac{\left(r-V\right)^2}{S^2}\right)\exp\left(-\frac{1}{2}\frac{\left(V-\mu_V\right)^2}{\sigma_V^2}\right)S^{a-1}\exp\left(-Sb\right)\right) + C \tag{4.5}$$

with C a term that does not depend on V or S, and

$$p\left(r|V,S\right) = \left(2\pi S^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{\left(r-V\right)^2}{S^2}\right) \tag{4.6}$$

$$p\left(V|\mu_V,\sigma_V\right) = \left(2\pi\sigma_V^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{\left(V-\mu_V\right)^2}{\sigma_V^2}\right) \tag{4.7}$$

$$p\left(S|a,b\right) = \frac{b^a}{\Gamma(a)}S^{a-1}\exp\left(-Sb\right) \tag{4.8}$$

$$\tag{4.9}$$

the probability density functions associated with the distributions 4.1, 4.2 and 4.3. To find the maximum of E with respect to V and S, and hence the mode of the posterior, we can investigate the gradient $(\partial E/\partial V, \partial E/\partial S)$ of E, which vanishes at the maximum. Evaluation the conditions $\partial E/\partial V = 0$ and $\partial E/\partial S = 0$, we find

$$V - \mu_V = \frac{\sigma_V^2}{S^2}\left(r - V\right) \tag{4.10}$$

$$S - \kappa_S = \frac{1}{b}\left(\left(\frac{r-V}{S}\right)^2 - 1\right) \tag{4.11}$$

To interpret these equations, note that the left-hand side yields the distance of S and V from the mode of their respective prior distributions. The right-hand side

quantifies the mismatch between what was expected based on V and S and what actually happened: based on V and S, the reward r was expected to be close to V, and $(r - V)^2$ was expected to be close to S^2. The mismatches are weighted with a measure of prior narrowness, σ_V^2 / S^2 in equation 4.10 and $1/b$ in equation 4.11.

We now have to solve these equations for V and S to find the mode of the true posterior. We could try and find the exact solutions, but considering that the equations are nonlinear, we would have to expect complicated expressions. Here we will not choose that route.

Instead, we shall restrict ourselves to approximate solutions. We assume that the priors of both V and S are very narrow (which corresponds to assuming that $\sigma_V^2 \ll 1$ and $1/b \ll 1$). This allows us to linearise and solve equation 4.10 and 4.11 easily. We find that the mode of the posterior is approximately at

$$V_{\max} = \mu_V + \frac{\sigma_V^2}{\kappa_S^2} (r - \mu_V) + \mathcal{O}(2) \tag{4.12}$$

$$S_{\max} = \kappa_S + \frac{1}{b} \left(\left(\frac{r - \mu_V}{\kappa_S} \right)^2 - 1 \right) + \mathcal{O}(2), \tag{4.13}$$

where $\mathcal{O}(2)$ reminds us that we have neglected terms of second or higher-order in $1/b$ and σ_V^2. The mode of the posterior is now found—at least approximately.

The final step of the mode-matching method consist in updating the mode of the fixed-shape belief distribution—which is (μ_V, κ_S))—by aligning it with the maximum of the true posterior, which is (approximately) given by $(V_{\max}, S_{\max})$ in equation 4.12 and 4.13.

If we were just looking for computationally lightweight learning rules that approximate Bayesian learning, we could stop here. However, we are ultimately interested in modelling learning in biological systems, in particular the basal ganglia system. We must hence consider that changes in synaptic strength can only depend on local information (such as pre- and postsynaptic potentials)

and low-dimensional global feedback signals (such as dopamine release in the striatum). We can achieve this here by applying yet another set of approximations. First, we identify certain factors as learning rates: σ_V^2/κ_S is replaced by α_v, and $1/b$ by α_s. Then, we simplify the equations by making the learning rates constant; we hence omit the κ_S-dependence of α_v. Finally, switching to an easier notation with $s = \kappa_S$ and $v = \mu_V$, we arrive at a set of very simple learning rules:

$$\delta = \frac{r - v}{s} \tag{4.14}$$

$$\Delta v = \alpha_v \delta \tag{4.15}$$

$$\Delta s = \alpha_s \left(\delta^2 - 1\right), \tag{4.16}$$

with a global feedback signal δ. These rules track the mean reward v as well as the observation noise s. Both v and s are fed back into the learning system as they enter what we will call the *scaled* prediction error δ.

4.1.3 Do the new rules work?

Do the newly derived rules do what they are meant to do? In this subsection, we use a stochastic fixed-point analysis to show that in theory, V and S should converge to the mean and the standard deviation of the reward signal. We confirm this with a first simulation.

We have already performed a stochastic fixed-point analysis in section 3.2.2 in the previous chapter; the very same approach will help us here. Let us assume that rewards are indeed generated by sampling from a distribution with mean μ and standard deviation σ (this could be a normal distribution or any other distribution with well defined mean and standard deviation).

We consider a situation in which the learner has already found the correct values of the variables it maintains, i.e., $(v, s) = (\mu, \sigma)$. From there, what are the *expected updates*? A straightforward calculation yields

$$\mathbb{E}\Delta s = \alpha_s \left(\frac{\mathbb{E}(r-v)^2}{s^2} - 1 \right) = \alpha_s \left(\frac{\mathbb{E}(r-\mu)^2}{\sigma^2} - 1 \right) = \alpha_s \left(\frac{\sigma^2}{\sigma^2} - 1 \right) = 0 \quad (4.17)$$

$$\mathbb{E}\Delta v = \alpha_v \frac{\mathbb{E}(r-v)}{s} = \alpha_v \frac{\mathbb{E}r - \mu}{\sigma} = \alpha_v \frac{\mu - \mu}{\sigma} = 0, \quad (4.18)$$

with $\mathbb{E}r = \mu$ and $\mathbb{E}(r-\mu)^2 = \sigma^2$ by definition. We find that the expected change away from $(v, s) = (\mu, \sigma)$ is zero, which makes (μ, σ) a stochastic fixed-point. We may conclude that in equilibrium, the rules given in equations 4.14, 4.15 and 4.16 should give us unbiased estimates of the reward mean and standard deviation.

After this theoretical analysis, we conduct the first practical test of the new rules. To do so, we generate a signal by sampling from a normal distribution with slowly varying mean and standard deviation. We feed the thus produced rewards to our new learning rules and then compare the estimates that the rules generate with the ground truth. The result of this test is shown in figure 4.2.

We find that the rules track the mean and standard deviation of the reward signal faithfully. This is apparent both from the time series in panel A and the correlations between ground truth and estimate in panel B.

All in all, both our mathematical analysis and our first simulations suggest that the scaled prediction error learning rules fulfil their purpose—they seem to provide good estimates of a signal's mean and standard deviation.

4.2 Performance tests

So far we do not know whether using the scaled prediction error learning rules is at all beneficial. They are certainly more complicated than learning rules such as the Rescorla-Wagner rule (equation 2.1), but are they also *better* in any way? Why should an organism use them? The Kalman filter we mentioned above can be shown to be optimal for a certain class of signals; at this point, we have no such guarantees for the scaled prediction error model.

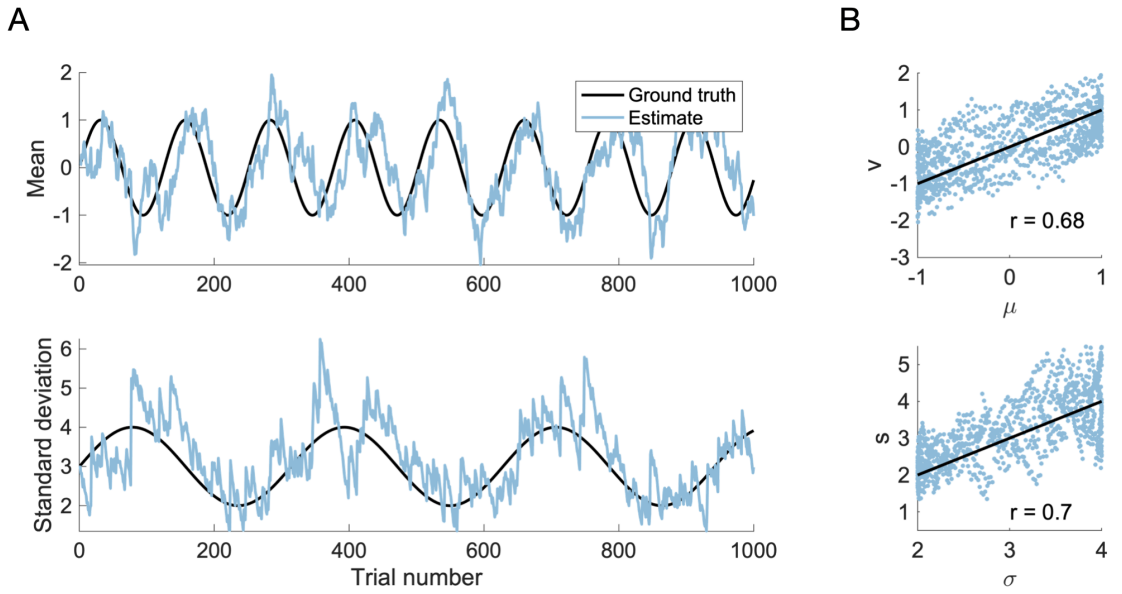

Figure 4.2: Tracking a signal with slowly varying mean and standard deviation with scaled prediction error learning. Rewards were generated by sampling from a normal distribution with mean $\mu(t) = \sin(t/20)$ and standard deviation $\sigma(t) = 3 + \sin(t/50)$, with t the trial number. The scaled prediction error model (equations 4.14–4.16) with learning rates $\alpha_v = 0.2$ and $\alpha_s = 0.2$ was used to track the statistics of the reward distribution. **A** Time series. Estimates from the model (blue) and ground truth variables (black) are plotted as functions of trial number. **B** Correlations. We show the correlation between estimate and ground truth for the reward mean in the upper panel and the reward standard deviation in the lower panel. The black lines indicate the identity function, r is Pearson's correlation coefficient.

We also do not know yet how reliable the new rules are. Do they always work well, or are there certain conditions to be met? Which types of signals can be learned? Or, formulated differently, what are the types of environments in which scaled prediction error learning would pay off for an organism?

In this section, we investigate the stability and the performance of the scaled prediction error model. In subsections 4.2.1 and 4.2.2, we focus on stability, i.e., we determine systematically under which conditions the rules work well and as expected. Then, in subsections 4.2.3 and 4.2.4, we compare the new rules with existing rules with respect to performance, in the context of reward prediction as well as action selection.

We will find that the new learning rules work well for slowly varying signals with

high observation noise, but not for signals with very low levels of observation noise or signals that change rapidly—an organism can rely on scaled prediction error learning if it aims to predict relatively variable rewards in relatively stable environments. We will further find that the scaled prediction error model outperforms standard models under these conditions, both in reward prediction tasks and in action selection tasks. In summary, whenever an organism faces unknown or variable levels of reward observation noise, it pays off for it to use scaled-prediction error learning.

4.2.1 Instabilities at low noise levels

In subsection 4.1.3, we have shown that the scaled prediction error learning rules have a stochastic fixed point at the mean and standard deviation of the underlying signal. This is a necessary condition for the rules to provide us with stable, unbiased estimates of these quantities, but it is not sufficient: we do not know whether the model will always reach that fixed point without problems.

In figure 4.2, the rules track the statistics of interest fairly robustly. However, the same simulation with a slightly modified target signal shows that this is not true for all target signals: in figure 4.3, we find that large jumps of s start to occur whenever the standard deviation of the target signal drops to very low values.

This simulation suggests that signals with low noise levels might be problematic, and indicates that a more systematic investigation is required. Which parameters might be relevant? Apart from the standard deviation σ of the target signal, it seems plausible that the learning rate α_s for the standard deviation might play an important role: it controls the size of the updates and hence determines the size of the large jumps that we observe at low noise levels.

Having selected the parameters of interest (α_s and σ), we can now vary them systematically and evaluate their effect on performance using simulations. The result of this approach is shown in form of a performance map in figure 4.4. The map provides several insights: we find that there is indeed a domain (a region in

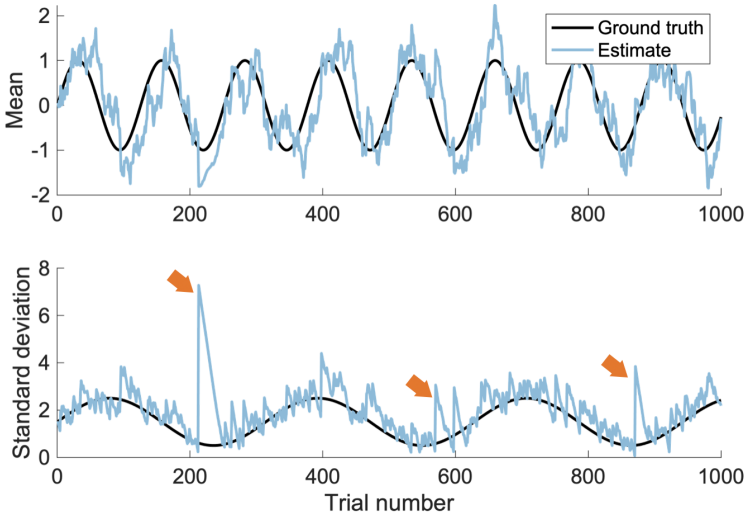

Figure 4.3: Instabilities occur at low variances. The simulation shown here is similar to that shown in figure 4.2, but with a standard deviation of $\sigma(t) = 3/2 + \sin(t/50)$ instead of $\sigma(t) = 3 + \sin(t/50)$. Instabilities (large, unwarranted jumps) are marked with orange arrows in the lower panel, which shows the ground truth (black) and estimate (blue) of the signals' standard deviation.

the space spanned by α_s and σ) in which the performance degrades very strongly; it is the "orange" region in figure 4.4, which we will call the domain of instability. We also find that stability seems to depend merely on the ratio σ/α_s, but not on the absolute values of those parameters. We can use the results of the simulations to derive a rule of thumb: to avoid the domain of instability, choose a learning rate that is at least 7 times smaller than the lowest level of noise that you wish to track.

Can we understand why there is a domain of instability? While a mathematically rigorous analysis of the phenomenon is beyond the scope of this chapter, we will formulate a hypothesis about the underlying mechanism.

We propose that the instabilities might occur due to a sequence of events. First, let us assume that learning has converged, i.e., that $s \approx \sigma$. Now, assume that the learner samples a reward r that is very close to the prediction, i.e., $r \approx v$. What would the resulting update for s be? Using the learning rule given in equation 4.16, we find

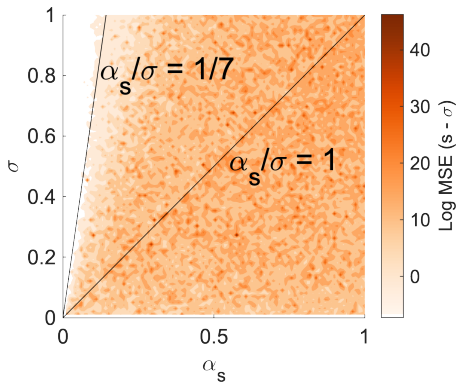

Figure 4.4: Instabilities at low variances. The mean squared difference between estimated and actual standard deviation as a function of learning rate α_s and noise level σ. The error is represented as a heat map, with darker shades corresponding to larger errors. Black lines indicate certain ratios between α_s and σ, see annotations.

$$\Delta s = \alpha_S \left(\left(\frac{r - v}{s} \right)^2 - 1 \right) \approx \alpha_S \left(\left(\frac{0}{\sigma} \right)^2 - 1 \right) = -\alpha_S. \tag{4.19}$$

The updated estimate $s + \Delta s$ is then $s + \Delta s \approx \sigma - \alpha_S$. Now, if $\alpha_S \to \sigma$, we would have $s + \Delta s \to 0$ and hence $|\delta| \to \infty$ in the next update (remember that s is the denominator of δ). Uncontrollably large jumps must therefore be expected when $\alpha_s \approx \sigma$, which is exactly what we observed in our systematic analysis shown in figure 4.4.

More generally speaking, the magnitude of jumps away from equilibrium appears to depend on the typical penetration depth of s in the area below $s = \sigma$, where the update magnitude rises steeply as s gets smaller. This typical penetration depth is controlled by the learning rate α_s, which explains why choosing low learning rates allows one to learn signals with low noise levels. Our analyses suggest that one should choose a learning rate $\alpha_s \ll \sigma$.

Are these instabilities also to be expected in a biological system that implements the scaled prediction error learning rules? Large updates are driven by large scaled prediction errors, which—as we will discuss below—are supposedly broadcast through dopamine release. This release cannot become arbitrarily large, due to a limited number of dopamine neurons, and a limited firing range

for each neuron. Hence, we should expect that biological prediction errors cannot grow beyond some maximal size. This should prevent instabilities from becoming too problematic in biological systems and might be an effective way to ensure that the scaled prediction error model is not derailed even if the noise levels of the signal drop unexpectedly.

Another approach to tackle the instabilities is to change the model such that the denominator of the scaled prediction error cannot become too small. We will see one way to implement this below in section 4.3.1, where we discuss how scaled prediction errors might be computed in the basal ganglia system. Simply adding internal noise to the reward signal might also have the desired effect.

Finally, one might also be able to solve this problem by dynamically adapting the learning rate α_s to the current estimate s. For example, one might avoid instabilities by setting $\alpha_s = s/10$. Would this work, or would it bring about other problems? How might this be implemented neurally? The answers to these questions are beyond the scope of this thesis, but should certainly be interesting directions for future research.

4.2.2 Instabilities in dynamic environments

We know now that the learning rules are unreliable when observation noise is too low, i.e. when rewards are too predictable. Next, we will investigate another source of instability: changes of the variable itself, called process noise. In our derivation (section 4.1) we have modelled the environment as stable. What if this assumption is violated? What if an organism uses scaled prediction error learning in an unstable environment? We have already seen that the rules are capable of tracking a slowly changing signal in figure 4.2, but what happens if changes become more rapid? To test this, we add a drift to the reward process: after each observation, the mean reward changes by a random amount sampled from a normal distribution:

$$r_t \sim \mathcal{N}\left(\mu_t, \sigma^2\right) \tag{4.20}$$

$$\mu_{t+1} \sim \mathcal{N}\left(\mu_t, \nu^2\right). \tag{4.21}$$

Here, ν is the process noise.

High process noise will require learners to update their estimate v rapidly. The speed of such updating is controlled by the learning rate α_v. It thus seems that α_v will be a relevant variable. To systematically assess the model's performance at different levels of process noise, we therefore simulate the model for different combinations of α_v and ν, like we did in the previous subsection with α_s and σ. The details of those simulations are provided in subsection 4.5.2, the results are shown in figure 4.5.

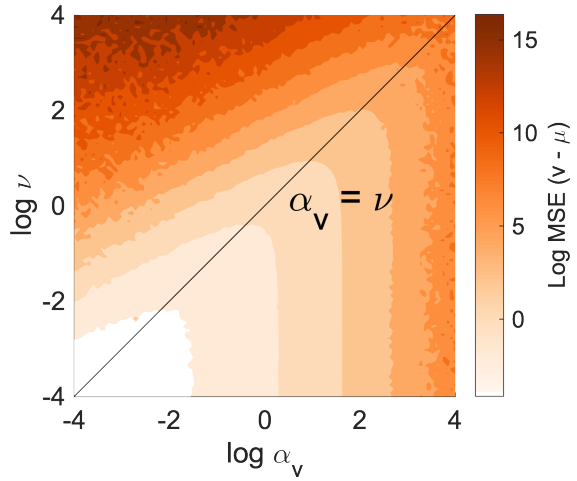

Figure 4.5: Performance and process noise. The mean squared difference between v and μ is shown as a function of the process noise ν and the value learning rate α_v. The error is represented as a heatmap, with darker shades corresponding to larger errors. The black diagonal indicates the equality of ν and α_v. The reward noise was fixed at $\sigma = 1$.

We find that the model's performance generally drops with higher ν. We further find that optimal performance is reached for $\alpha_v \approx \nu$. Indeed, one can prove this mathematically—we show the proof in appendix C, and discuss it below. For now, we can use this insight when parametrising our model.

To summarise the last two subsections, let us reiterate our findings: we found that the scaled prediction error learning rules work well when tracking slowly varying signals with high levels of observation noise. Performance deteriorates strongly if the observation noise level σ drops to a level comparable to the learning rate α_s. We further found that our model works better at lower levels of the process noise ν, and that performance is best at $\alpha_v = \nu$. These findings define a domain of applicability: if $\alpha_s \ll \sigma$ and $\alpha_v = \nu$, we may use the rules with confidence. In the next sections, we will restrict ourselves to this domain of applicability. Overall our results indicate that scaled prediction error learning is suitable for a relatively stable environment in which rewards are very unpredictable.

4.2.3 Improvements in reward prediction

After focusing on stability in the last two subsections, we now turn to performance: how do the scaled prediction error rules compare to established rules such as the Rescorla-Wagner model (equation 2.1) with respect to accurate reward predictions or policy optimisation? We first focus on reward prediction. We will find that the scaled prediction error model generally outperforms the conventional RW model.

To compare the performances of scaled prediction error learning and RW learning, we use the same reward process as above in section 4.2.2: noisy samples are taken around the mean reward; the mean drifts slowly. This process is formally defined in equation 4.20 and 4.21. Both learners track the reward signal and provide predictions of the value (i.e., the current mean of the reward distribution), at every trial. The learners' performance is judged by measuring the average precision of their predictions.

We compare the models for different levels of observation noise σ, while keeping the process noise ν constant at $\nu = 1$. The results of those comparisons are presented in figure 4.6 (see subsection 4.5.3 for details of the implementation).

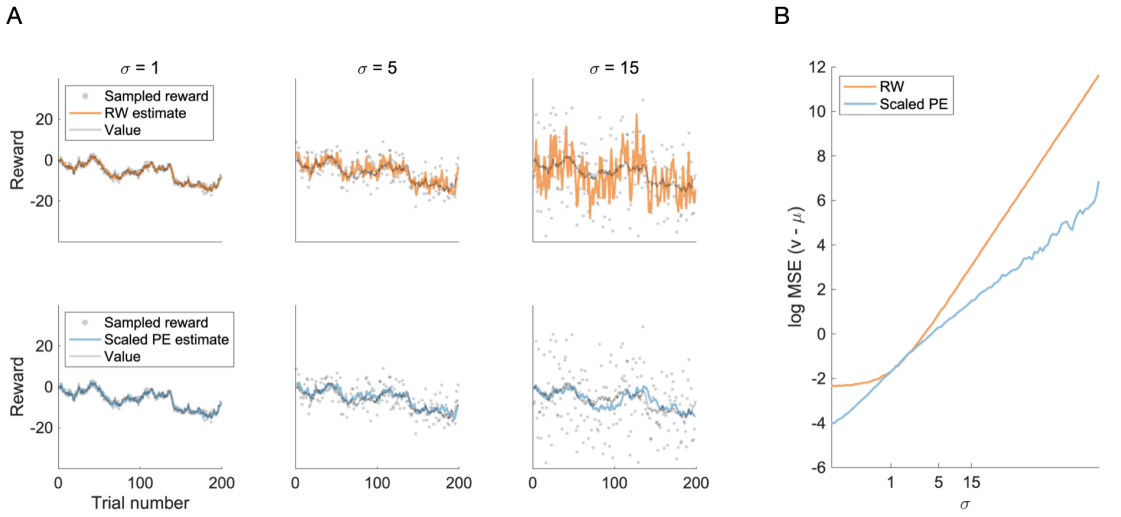

Figure 4.6: Reward prediction performance of the Rescorla-Wagner model and scaled prediction error model. **A** The first 200 trials of reward prediction for the Rescorla-Wagner learner (upper row, orange colour) and the scaled prediction error learner (lower row, blue colour). The true value (grey line), the observed rewards (grey dots) and the learner's estimate (coloured line) are shown as a function of trial number. Columns correspond to selected levels of observation noise ($\sigma = 1, 5, 15$). **B** Learning performance averaged over trials. We show the logarithm of the average squared error of the learners' predictions as a function of observation noise. The blue line corresponds to the scaled prediction error learner, the orange line corresponds to the RW learner.

Looking first at the time series in figure 4.6A, we find that there is a qualitative difference between the Rescorla-Wagner learner in the top row and the scaled prediction error learner in the bottom row: as the noise level σ increases, the Rescorla-Wagner learner's predictions fluctuate stronger and stronger. This is not so for the scaled prediction error learner, who's predictions fluctuate as much for low noise levels as they do for high noise levels.

This effect is also visible in the aggregated performance measure, shown in figure 4.6B: the mean squared errors of the learners' predictions grow with observation noise for both learners, but they grow stronger for the Rescorla-Wagner. While both learners show similar performance around $\sigma = 1$, the RW learner is orders of magnitude worse than the scaled prediction error learner for both higher and lower levels of noise.

It thus appears as if the scaled prediction error model has an advantage over

RW-type learning, at least in the configuration we examined in figure 4.6. Does this hold generally? The scaled prediction error learner was parametrised with $\alpha_v = \nu$ based on our insights from the sections above. RW-type learning could potentially also be improved by choosing a different learning rate. To draw a general conclusion, we performed the above experiment for a range of different parametrisations for the RW learner. The results of this are shown in figure 4.7, the procedures are described in subsection 4.5.3.

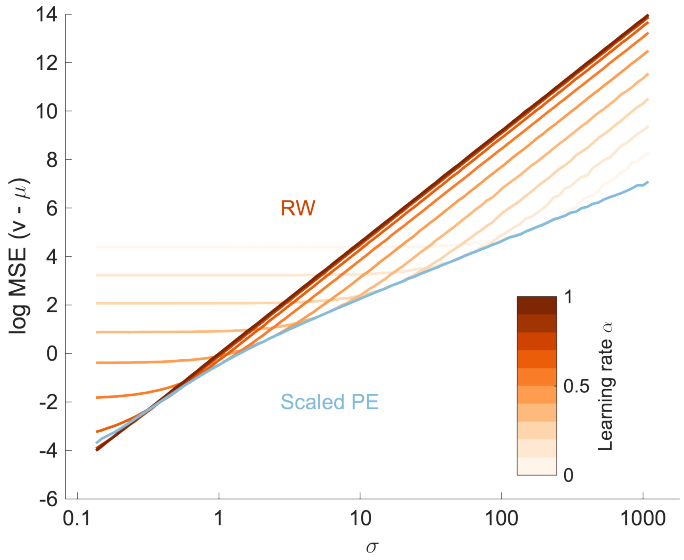

Figure 4.7: Reward prediction performance as a function of observation noise for different learning rates. We show the logarithm of the mean squared difference between the mean of the reward distribution and the learner's prediction thereof, as a function of observation noise. Orange lines correspond to RW learners, the blue line corresponds to a scaled prediction error learner parametrised with $\alpha_v = 1$ and $\alpha_s = 0.01$. The different shades of orange correspond to different learning rates, as indicated by the colour bar. The process noise was fixed at $\nu = 1$.

We find a very stereotyped effect for the average performance of RW learners: as the level of noise increases, prediction accuracy does not change much up to a certain point (call this the plateau) and grows steadily after that point (call this the slope). This is the case irrespective of the learning rate. Smaller learning rates have a plateau that extends to higher noise levels but also provides a lower accuracy. The steepness of the slope is invariant across learning rates. This is to be expected: the estimates of the RW learners fluctuate proportionally to the fluctuations of the signal in the limit of large σ. Errors will thus grow linearly

with σ. The axes in figure 4.7 are logarithmic, so the graphs of those linear relationships will be lines with slope one.

Overall, the behaviour of the RW learners is such that for each given level of observation noise there is an optimal learning rate: if one selects any one position on the x-axis of figure 4.7, there is always a single orange curve with the lowest y-coordinate (and hence the smallest average error) at that position. In general, we find: the higher the observation noise, the lower the optimal learning rate. This appears consistent with intuition—if observation noise is high, there is less useful information in any single observation and an organism should therefore update its estimate more carefully.

The scaled prediction error learner shows different behaviour. There is also a slope (prediction accuracy steadily decreases with increasing observation noise), but no plateau. The steepness of the slope changes at $\sigma = 1$. For higher levels of observation noise, the slope of the scaled prediction error learner is shallower than those of the RW learners. We find that for any given level of noise σ larger than one, the performance of the scaled prediction error model is about as good as the performance of the best RW model. This suggests that in the regime of high observation noise we might view the scaled prediction error model as an RW learner that reaches optimal performance by fine-tuning itself to the estimated level of observation noise.

Can one do better than this? In fact, one can show that the scaled prediction error model (parametrised with $\alpha_v = 1$) is approximately optimal in the situation investigated here: for high levels of observation noise, scaled prediction error learning approximates the steady-state Kalman filter (we show this in appendix C), which is approximately optimal for the types of signals we use here.

As mentioned above, to use a Kalman filter, one needs to provide it with the correct values of σ and ν. This is also true for the steady-state version of the Kalman filter, but it is not the case for the scaled prediction error model: here one only needs to provide α_v—which corresponds to ν, as shown above—but not σ,

which the model can track by itself. We can thus think of scaled prediction error learning as *adaptive* steady-state Kalman filtering.

However, to work optimally, the scaled prediction error model still needs to be provided with the correct value for α_v. To make the model more autonomous, one might extend it with a mechanism to track ν alongside σ, for example the mechanism proposed by Piray and Daw (2020a). This is an interesting direction for further research, but goes beyond the scope of this chapter.

Let us collect the results thus far. We used simulations as well as mathematical analysis, and learned that the learning rate for the observation noise should be substantially smaller then the observation noise itself, $\alpha_s \ll \sigma$. We also learned that the learning rate for the reward should be similar to the process noise, $\alpha_v = \nu$. If these conditions are fulfilled, the scaled prediction error model is approximately optimal for signals with $\nu < \sigma$. In particular, it will be at least as good as any RW learner, and about as good as a steady-state Kalman filter.

Scaled prediction error learners thus appear particularly well suited to track signals with unknown or changing levels of observation noise, as they can adapt themselves to whatever level of noise they experience. In contrast, an RW learner would either have to be fine-tuned based on prior knowledge, or it would perform suboptimally due to under- or overfitting.

4.2.4 Improvements in actor learning

However important reward prediction may be, ultimately it is only a tool for action selection. Reward prediction and action selection might be largely separated in some methods, but they can also be directly interconnected in others. One interconnected method is the the actor-critic model which we briefly discussed in chapter 2. The actor-critic is popular in neuroscience (Bennett et al. 2021) as well as in AI (Mnih, Badia, et al. 2016). In this subsection, we will investigate how the scaled prediction error model can become part of an actor-critic setup, and test whether this can improve action selection.

As the name implies, the actor-critic model has two components: an actor and a critic component. The actor can be thought of as a state-action map[3]. The learner uses this state-action map to select the appropriate action for a given state.

The critic can be thought of as a state-value map. The learner may use this map to evaluate how good states are with respect to its goals, or to judge whether the outcome of an action was good or bad. Values are learned rewards; the critic is thus a reward prediction system.

The state-action map and the state-value map are plastic, and the learner constantly improves both, increasing the quality of reward predictions and decisions simultaneously. This improvement usually requires reward prediction errors to be computed and broadcast. For example, it has been found that the reward prediction errors of the TD algorithm work well as a teaching signal for both the critic *and* the actor (Sutton, Barto, et al. 1998).

It is easy to see how the scaled prediction error model might fit into an actor-critic framework: instead of using a critic that relies on absolute prediction errors for its reward predictions, one could deploy the scaled prediction error learner as a critic. One could then use the scaled prediction errors it produces as a feedback signal for both actor and critic.

In which type of task would we likely see a difference between a scaled actor-critic and a standard actor-critic? Our results above indicate that to benefit from scaling, the learner would have to experience different levels of reward observation noise during the task. In the following sections, we present two such tasks, purposefully designed to challenge learners with varying levels of reward noise. One of these tasks has a discrete action space, the other one has a continuous action space.

[3]Often, this is referred to as policy in the reinforcement learning literature, for example by Sutton, Barto, et al. (1998).

Scaled prediction error learning in discrete action spaces: the distracted bandit task

First, let us consider a task with a discrete action space, i.e., a task in which the learner must select one action out of a few possible options. Bandit tasks fall into this category, as do many other typical tasks in cognitive neuroscience. The particular task we study here shall be called the *distracted bandit* task (we are not aware that this task was mentioned or named elsewhere). The distracted bandit is a 3-armed bandit with a Gaussian reward distribution associated with each arm. The variance of those reward distributions is the same for all three arms, but the means vary: two arms are associated with high means that differ slightly (say an average of 9.5 points versus an average of 10 points). The third arm, called the *distractor*, is associated with a low mean (say 0 points on average). The reward distributions are sketched schematically in figure 4.8A.

Decision-making tasks with distractor options have previously been studied by neuroscientists such as Chau et al. (2014) and behavioural economists such as Huber et al. (1982). The difference between those tasks and ours is that in our task the values of the options must be learned from experience (hence the name distracted *bandit*). The tasks in the literature provide to participants with the values explicitly, learning is not required.

The distracted bandit yields two challenges for a learner, an easy one and a hard one. The easy challenge consists in learning that the non-distractor options should be preferred over the distractor option. This is easy because the reward contrast between the distractor and the other options is large (relative to the reward variance within each option). The hard challenge is to learn which of the two non-distractor options is the better choice. This is harder because the contrast between them is much smaller.

We test two models in the distracted-bandit task: a standard actor-critic and a scaled actor-critic (see subsection 4.5.4 for the definitions of the models). The results of the tests are shown in figure 4.8B. We find that both models show quite

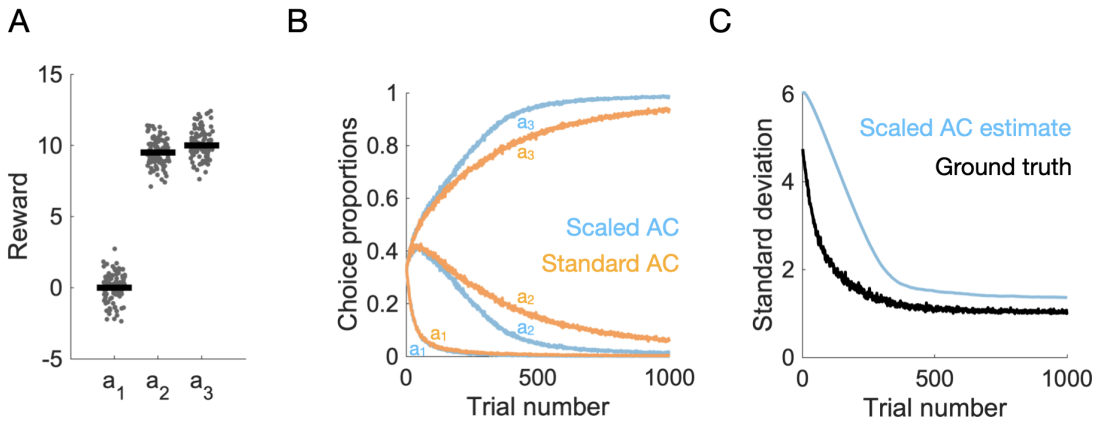

Figure 4.8: Performance in the distracted bandit task. **A** The reward structure of the distracted bandit task. Options a_2 and a_3 yield high average rewards, while option a_1 yields low average rewards and serves as a distractor. Rewards are sampled at random, from normal distributions with means 0, 9.5, and 10, and a standard deviation of 1. Horizontal bars indicate the means of the reward distributions, dots indicate individual rewards. **B** Choice probabilities in the distracted bandit task, for a standard actor-critic and a scaled actor-critic. The probabilities of choosing options a_1, a_2 and a_3 are shown as functions of trial number. The probabilities are obtained by averaging over multiple repetitions of the experiment. The orange lines represent the choice frequencies of the standard actor-critic, the blue lines represent the scaled actor-critic. **C** Estimated and empirical standard deviation. The back curve shows the standard deviation of the experienced reward distribution as a function of trial number. The blue curve is the corresponding averaged estimate of the scaled prediction error learner. AC stands for Actor-Critic.

similar behaviour at the beginning of the task. The choice probabilities then diverge after about 200 trials: the probability of choosing option a_3 (the best option) increases more steeply for the scaled actor-critic, with a corresponding steeper decrease in the probability of option a_2. The choice probabilities of the two models diverge just as the probability of choosing the distractor option (a_1 in fig 4.8B) drops to zero. From that point on, the scaled actor-critic clearly outperforms the standard actor-critic.

The reason for this is quite intuitive: after solving the simple challenge and discarding option a_1, the scaled actor-critic uses its scaling mechanism to normalise its reward perception, hence "zooming in" on the reward distributions of the remaining options. This is illustrated in figure 4.8C, which shows the scaled learners' estimate of the reward observation noise, and the empirical reward

standard deviation for comparison. As performance increases, the empirical as well as the estimated reward standard deviation drop. This then magnifies the reward prediction errors, and thus also increases the actor updates (see equations 4.32 and 4.33). By increasing the size of these updates, the scaled actor-critic model can increase its learning speed when entering the hard challenge.The standard actor-critic learner does not use such techniques. Its actor updates do not get a boost during the hard challenge, and it thus learns slower than the scaled actor-critic learner.

From what we have seen so far, it appears as if a scaled critic is beneficial for learning in the distracted bandit task. We have, however, only tested the models with one set of parameters. To test whether the result holds more generally, we must evaluate performances systematically for various sets of parameters, varying both the learning rate α_ϕ of the actor and the learning rate α_v of the critic. The results of these experiments are shown in figure 4.9.

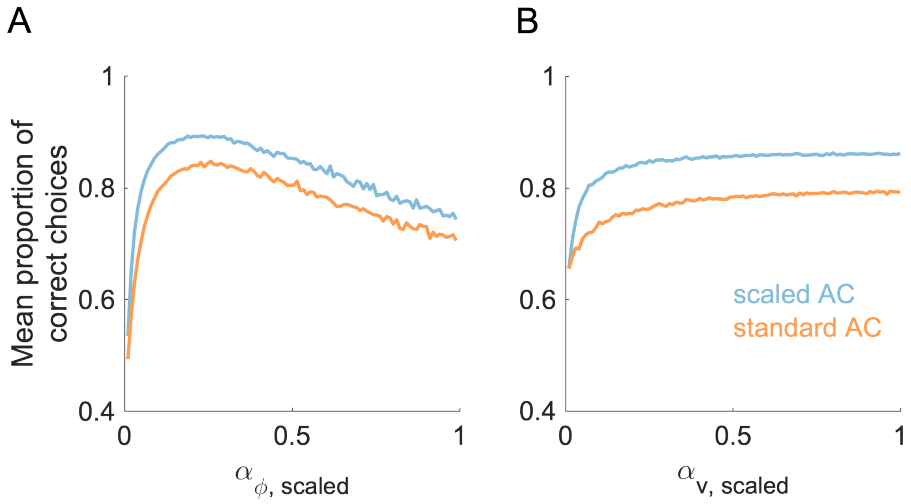

Figure 4.9: The impact of learning rates on performance in the distracted bandit task. **A** Performances of a scaled actor-critic model and a standard actor-critic model in the distracted bandit task, as a function of the actor learning rate $\alpha_{\phi,\ \text{scaled}}$ of the scaled actor-critic model. The actor learning rate $\alpha_{\phi,\ \text{standard}}$ of the standard actor-critic model was proportional to $\alpha_{\phi,\ \text{scaled}}$, see subsection 4.5.4 and equation 4.39 for details. Performance is defined as the average probability of choosing the highest-value option, aggregated over all trials. The scaled actor-critic performance is shown in blue, the standard actor-critic performance is shown in orange. **B** Performances as a function of critic learning rate $\alpha_{v,\ \text{scaled}}$ of the scaled actor-critic model. Again, the critic learning rate $\alpha_{v,\ \text{standard}}$ of the standard model was proportional to $\alpha_{v,\ \text{scaled}}$. The results are represented as in A.

We find that the performance of the standard actor-critic and the performance of the scaled actor-critic both depend on the learning rates α_ϕ and α_v in very similar ways. Importantly, we find that the scaled actor-critic model reaches a higher peak performance than the standard actor-critic, suggesting that it is the overall stronger model.

Scaled prediction error learning in continuous action spaces: the diminishing returns task

Next, we consider a task with a continuous action space. Here, the learner must select an action from a continuous range, for example a real number from the interval $[0, +\infty)$. In neuroscience, continuous action spaces occur in the context of motor learning, where subjects or agents need to optimise their movements within the continuous space of possible trajectories (Merel et al. 2019). Note that for this type of learning one requires an actor-critic, or a related model—RW-type models alone are not suitable.

Here, we consider a task in which the learner is asked to choose a number a out of the interval $[0, +\infty)$. Some a will result in higher rewards than others; the learner must try to find the a that yields the largest rewards on average. It achieves this by exploring the actions around the best action ϕ it has found so far—let us call ϕ the learner's *best guess*. Using the reward data it generates in this exploration, the learner updates its best guess as appropriate. In the case at hand, the learner uses a stochastic exploration policy in form of a Gaussian distribution of fixed width around the best guess ϕ.

Unbeknownst to the learner, the rewards are computed by simply taking the logarithm of a. Hence, larger a are always better, but the increase in reward brought about by moving to higher values of a becomes ever smaller the larger a is. For this reason, we refer to the task as the *diminishing returns task*. The reward function and the learner's policy are sketched in figure 4.10A.

As with the distracted bandit task, we test two models with the diminishing returns task: a standard actor-critic model and a scaled actor-critic model, adapted

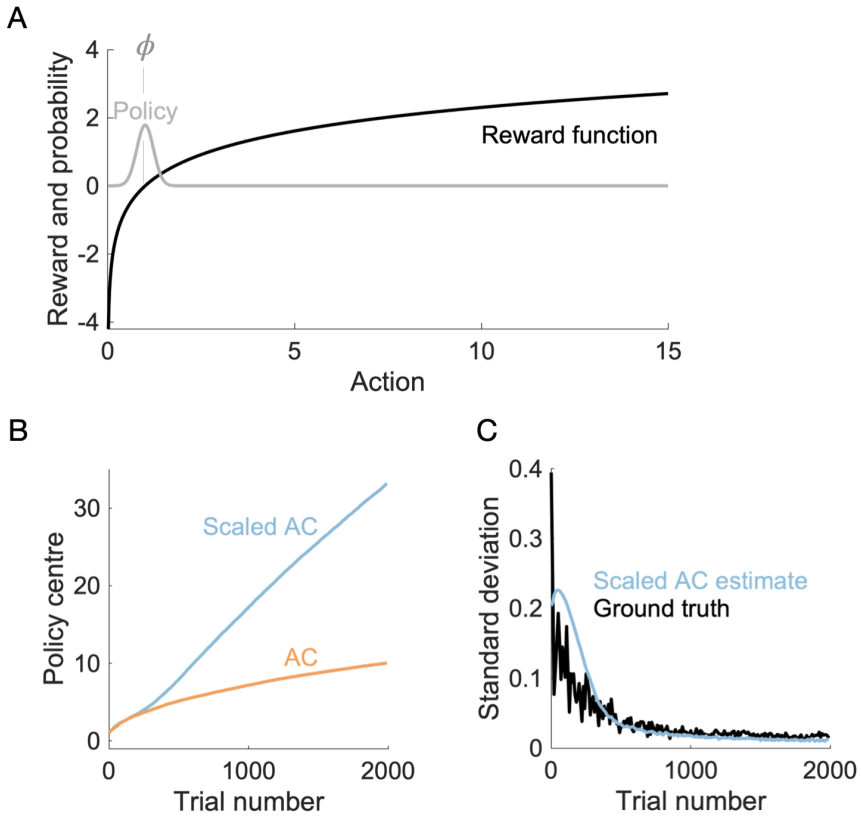

Figure 4.10: Performance in the diminishing returns task. A Schematic representation of the task. The x-axis corresponds to the action space of the task. The black line indicates the reward function $r(a)$ which grows logarithmically with a, $r(a) = \log(a)$. The grey line represents the learner's policy, a normal distribution centred around the learner's best guess. The learner selects actions by sampling from that distribution. **B** Learning curves in the diminishing returns task. The learner's best guess is plotted as a function of the trial number. The orange curve represents the standard actor-critic, the blue curve represents the scaled actor-critic. AC stands for actor-critic. **C** Estimated and empirical standard deviation. The back curve shows the standard deviation of the experienced reward distribution as a function of trial number. The blue curve is the corresponding averaged estimate of the scaled prediction error learner. AC stands for Actor-Critic.

to the continuous action space setup (see 4.5.5 for details). The measure of performance in this task is simply the learners' best guess ϕ—higher ϕ mean higher rewards, thus higher performance. The learners' performance is shown in figure 4.10B.

We find that both models perform equally well at the beginning of the task. After several hundred trials, however, the scaled actor-critic starts to outperform the standard actor-critic. This is similar to the results we obtained for the distracted

bandit—compare figure 4.8B and figure 4.10B.

The mechanism behind the superior performance of the scaled learner is the same as above–the learner continuously increases the size of the actor updates by continuously decreasing s (this is shown in figure 4.10C). This way, it can compensate for the decreasing magnitude of typical prediction errors $r - v$ (the reward differences between the actions that the learner explores are getting smaller as the learner becomes better). These two effects balance each other out; the learner maintains a constant learning speed. This is not so for the standard learner—its learning speed drops as it becomes better.

As with the distracted bandit task, we must investigate performances more systematically to validate this result. We again perform simulations for several different parametrisations, and compare performances between the scaled actor-critic and a standard actor-critic. The results of these tests are shown in figure 4.11.

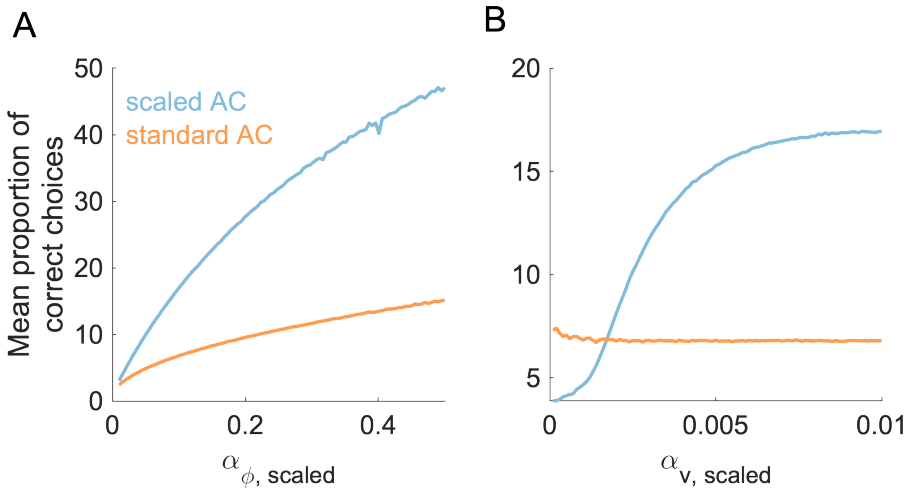

Figure 4.11: The impact of learning rates on performance in the diminishing returns task. A Performances of a scaled actor-critic and a standard actor-critic in the diminishing returns task as a function of the actor learning rate $\alpha_{\phi,\,\text{scaled}}$ of the scaled actor-critic model. The actor learning rate $\alpha_{\phi,\,\text{standard}}$ of the standard actor-critic model was proportional to $\alpha_{\phi,\,\text{scaled}}$, see subsection 4.5.5 and equation 4.53 for details. Here, performance is defined as the average best guess ϕ aggregated over all trials. The scaled actor-critic performance is shown in blue, the standard actor-critic performance is shown in orange. **B** Performances as a function of critic learning rate α_v of the scaled actor-critic model. Again, the critic learning rate $\alpha_{v,\,\text{standard}}$ of the standard model was proportional to $\alpha_{v,\,\text{scaled}}$. The representation of the results is analogous to A.

As for the distracted bandit, we find that the scaled actor-critic outperforms the standard actor-critic. The scaled learner reaches a much higher peak performance, which suggests that it is the overall stronger model (even more so here than in the distracted bandit task).

In conclusion, we can say that the scaled actor-critic model outperforms the standard actor-critic model in the two tasks we have discussed. The mechanism behind this is the amplification of actor updates as a consequence of adaptive prediction error scaling. This boosts the learning speed whenever the reward variability shrinks.

4.2.5 Summary of the performance tests

We conclude this section on performance with a short review of our findings: we have compared the performance of the scaled prediction error model with the performance of standard models—the RW model and the AC model—in both reward prediction and action selection tasks. We have found consistently that the scaled prediction error learning can outperform conventional learning if the learner faces unknown or varying levels of reward observation noise.

Overall, it appears to be advantageous for an organism to implement scaled prediction error learning unless rewards are very predictable (i.e., unless reward observation noise is very low). This is because the organism will likely experience different levels of reward observation noise or different reward gradients in its life, and scaled prediction error learning allows it to fine-tune its learning algorithm accordingly. On normative grounds, the scaled prediction error model has a case.

4.3 Empirical tests

In the previous sections, we derived a new set of learning rules—the scaled prediction error model—by approximating Bayesian learning. We then discussed the robustness of the learning rules and discovered a domain of stability within

which they work well. We have compared the performance of the scaled prediction error model to standard learning rules and found that scaled prediction error learners have an advantage over standard learners when confronted with unknown or varying levels of reward observation noise.

Taken together, these findings suggest that animals might benefit from applying scaled prediction error learning. So, can we find evidence for scaled prediction error learning in animal behaviour, or in the animal brain? In this section, we compare the predictions of our learning rules to empirical findings.

First, we discuss the biological plausibility of our theory in subsection 4.3.1. We will analyse whether dopamine signals could be scaled prediction errors, and whether the basal ganglia circuit could implement scaled prediction error learning.

Second, we look for evidence of scaled prediction error learning on the level of behaviour. In particular, we will test whether our model might account for recently observed effects of reward magnitude on learning speed (subsection 4.3.2). We will find that our model accounts for the effect to some degree, but that other explanations fit better. We then propose other ways to test the scaled prediction error model using behaviour (subsection 4.3.3).

4.3.1 Biological plausibility

Could the scaled prediction error learning rules (equations 4.14, 4.15 and 4.16) be implemented in the dopamine system and the basal ganglia pathways? We start our analysis by mapping the components of the new learning rules onto components of the basal ganglia system, just we did in section 2.4.4, where we discussed the biological plausibility of the AU learning rules. Many of the ideas of that section can be applied again. As in the AU model, we consider a distributed encoding of the two reward statistics in the two main basal ganglia pathways. We propose that the mean of the reward signal is encoded in the difference between direct and indirect pathway activity: $v = 1/2 \times (G - N)$. The standard deviation

of the signal is encoded in the sum of the activities: $s = 1/2 \times (G + N)$. We further suggest that striatal dopamine release broadcasts scaled prediction errors, $\delta = (r - v)/s$, and that the update rules given in equations 4.15 and 4.16 are implemented by dopamine-dependent plasticity in the striatum.

In the next sections, we will analyse the plausibility of these suggestions. We first discuss the relationship between dopamine responses and scaled prediction errors. We then propose a mechanism that might implement the scaling. Finally, we discuss how the scaled prediction error learning rules can be mapped on striatal plasticity rules.

Scaled prediction errors are consistent with dopamine activity

In a seminal study, Tobler et al. (2005) investigated how the responses of dopamine neurons to unpredictable rewards depended on reward magnitude, using electrophysiology in monkeys. Three different visual stimuli were paired with three different reward magnitudes (0.05 ml, 0.15 ml and 0.5 ml of juice). After being shown one of the stimuli, the monkeys received the corresponding reward with a probability of 50%. Seeing the stimulus allowed the monkey to predict the magnitude of the reward that could occur, but not whether it would occur in a given trial. Reward delivery thus came as a surprise and evoked a dopamine response. Interestingly, these responses did not scale with the magnitude of the received rewards. The measured dopamine responses are shown in figure 4.12A.

This result was unexpected—standard RW learning would predict that the residual prediction errors in rewarded trials should grow linearly with reward magnitude (the prediction of the standard theory is shown in figure 4.12B). Our new scaled prediction error rules, on the other hand, predict exactly what has been observed (see 4.12C). Dopamine activity corresponding to scaled reward prediction errors is thus plausible; in fact, with respect to the results of Tobler et al. (2005), it is more plausible than dopamine activity corresponding to unscaled reward prediction errors.

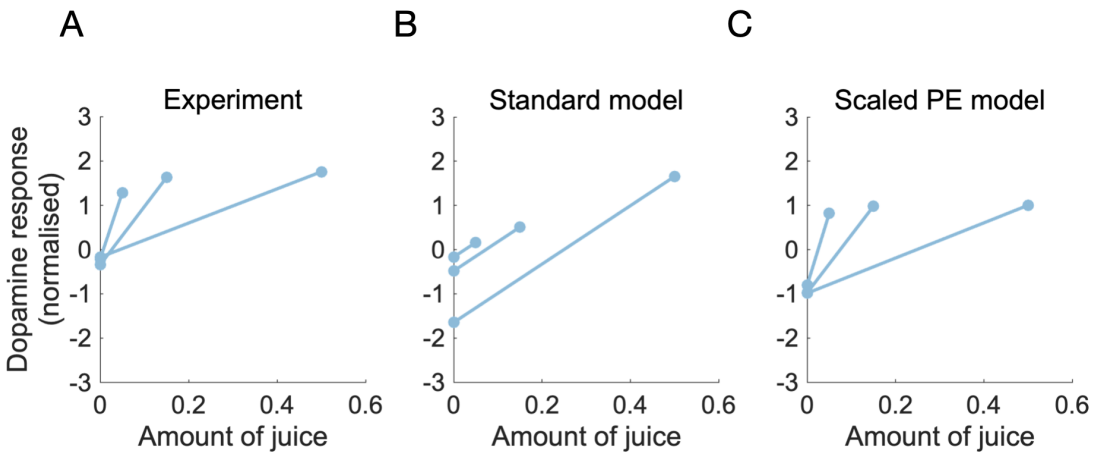

Figure 4.12: Dopamine responses to unpredictable rewards—experimental data and simulations. Normalised dopamine responses are shown as a function of reward magnitude, for three different conditions. The representation of data is similar to that in figure 4C of Tobler et al. (2005). **A** Experimental data, extracted from figure 4C (animal A) of Tobler et al. (2005). **B** Simulated data, using a standard RW model. **C** Simulated data, using the scaled prediction error model. See subsection 4.5.6 for details of the simulation procedure.

At this point, one may object that the results of Tobler et al. (2005) might also be explained by scaling with respect to the reward range—reward range and reward standard deviation cannot be dissociated in that experiment. While that is true, another recent experiment can dissociate them: Rothenhoefer et al. (2021) used two reward distributions with the same reward range but different reward standard deviations in a Pavlovian conditioning task (see figure 4.13A).

After exhaustive training, single unit recordings were performed to measure dopamine responses to rewards that deviated from the expected value. It was found that the same deviation from the expected value caused stronger dopamine responses for the distribution with the smaller standard deviation (figure 4.13B, first panel). This is consistent with scaling by reward standard deviation, but not with scaling by reward range—both distributions had the same range, so scaling by range should yield similar responses for both conditions.

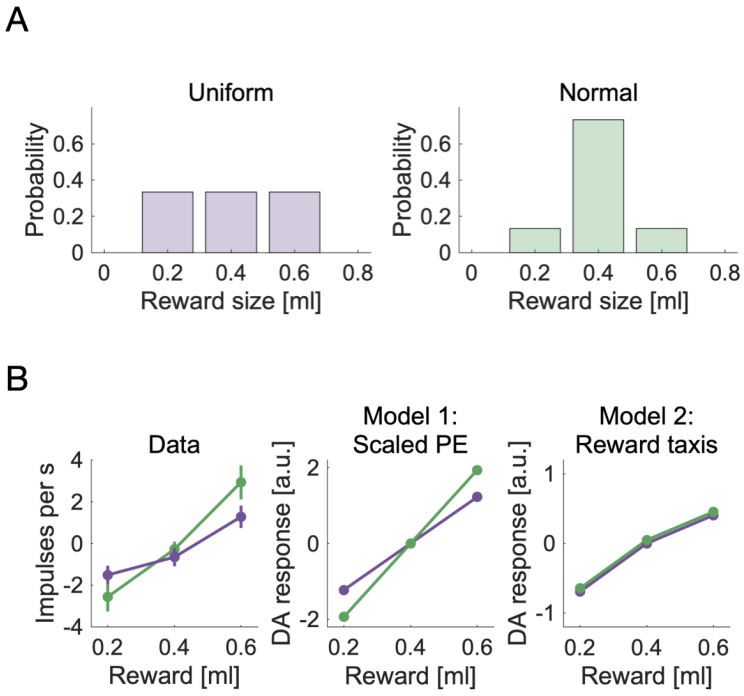

Figure 4.13: Modelling dopamine responses in the experiment of Rothenhoefer et al. (2021). **A** The reward distributions used by Rothenhoefer et al. (2021). The panel is reproduced from Rothenhoefer et al. (2021), figure 1A. **B** Dopamine responses to rewards sampled from the distributions in A. We show the empirical values, reproduced from Rothenhoefer et al. (2021), figure 2E, and the responses according to the scaled prediction error model and the reward taxis model. Purple lines correspond to the uniform reward distribution, green lines correspond to the normal reward distribution. We computed the model responses as $\delta = log(r/\mu)$ for the reward taxis model, and as $\delta = (r - \mu)/\sigma$ for the scaled prediction error model. Here, r is the reward size, μ is the mean and σ the standard deviation of the reward distribution.

Scaled prediction errors can be computed through a dopaminergic feedback loop

Next, we discuss how the scaled prediction error $\delta = (r - v)/s$ might be computed in the basal ganglia system. Expressed in terms of G and N, the scaled prediction error is given as

$$\delta = \frac{r - \frac{1}{2}(G - N)}{\frac{1}{2}(G + N)}. \tag{4.22}$$

This seems to be a fairly complicated combination of terms, and it is difficult to see how a simple network might compute it. Surprisingly, there is a fairly simple

approximate implementation based on a feedback loop. In this subsection, we will describe that mechanism, using a minimal dynamical model of the basal ganglia network.

First, where is the feedback loop? Let us assume that the prediction error δ is computed by subtracting the thalamic activity T—the output of the basal ganglia—from the reward signal r. Formally, we assume $\delta = r - T$. The thalamic activity might be captured by $T = DG - (1 - D)N$ (see equation 2.9). In this equation, D is the tonic level of dopamine, which has a value of $D = 0.5$ at baseline. If we now admit the prediction error δ to contribute to the tonic level of dopamine, we obtain a feedback loop: dopamine release modulates the thalamic activity, which itself inhibits dopamine release. Concretely, we write $D = 1/2 + \delta/2$.

To see how this feedback loop computes, let us view the basal ganglia system as a dynamical system. In chapter 3 our description of the basal ganglia has been trial-by-trial: we described the activations G and N of the striatal populations, the activations T of the thalamic populations and the striatal dopamine release δ by single numbers for each trial. These numbers changed over trials (due to learning or changing inputs, for example), but not within a trial.

To examine the computation of scaled prediction errors, we want to explore the dynamics in the basal ganglia system on a faster timescale, i.e., within a single trial. To do this, we model the relevant populations' activities as leaky integrators with effective connectivity as sketched in figure 4.14A, using differential equations in continuous time (this is a common approach in neuroscience, see H. R. Wilson and Cowan 1972 for a classic example). The dynamical model complements our trial-by-trial models and helps us to understand how variables like the scaled prediction error are computed in real-time. Modelling a system on two timescales simultaneously is not uncommon; for example, such descriptions are used in predictive coding models (Bogacz 2020), where one differentiates between inference (which happens fast, in continuous time) and updating (which happens slowly, between trials).

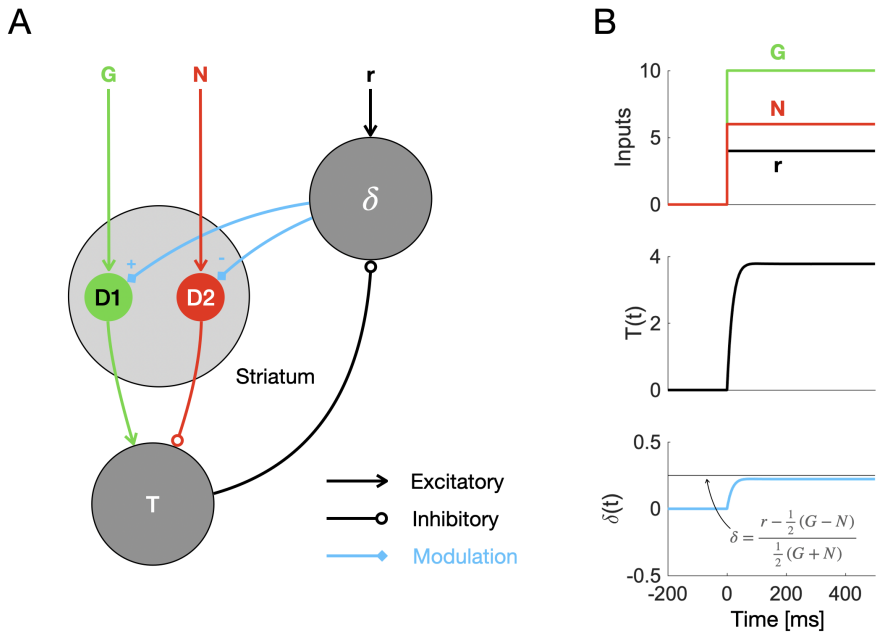

Figure 4.14: The computation of scaled prediction errors through a dopaminergic feedback loop. **A** The connectivity underlying a dynamical model of the simplified basal ganglia circuit. Circles correspond to neural populations; arrows between them indicate connections. **B** The computation of a scaled prediction error in continuous time, according to a dynamical model of the basal ganglia. We show how the relevant variables, T and δ, evolve as a function of time, assuming a step-function activation for the input nodes G, N and r. The black line in the lowest panel indicates the level of dopamine required for exact scaled prediction error learning.

The dynamical system sketched in figure 4.14A corresponds to a set of differential equations,

$$\tau_\delta \dot{\delta} = -\delta + (r - T) \tag{4.23}$$

$$\tau_T \dot{T} = -T + \left(\frac{1+\delta}{2} G - \frac{1-\delta}{2} N \right). \tag{4.24}$$

Here, τ_δ and τ_T are the characteristic timescales of the striatal dopamine release and thalamic activation. The system is set up such that its equilibrium point is consistent with our trial-wise description ($\delta = r - T$ and $T = G(1+\delta)/2 - N(1-\delta)/2$ at $\dot{\delta} = \dot{T} = 0$). This asserts that the two levels of description are consistent with each other. Using these equilibrium equations, we can determine

the equilibrium value of δ (by inserting one equation into the other and solving for δ). We find

$$\delta = \frac{r - \frac{1}{2}(G - N)}{1 + \frac{1}{2}(G + N)} \tag{4.25}$$

For $1/2 \times (G + N) \gg 1$, this approximates the scaled prediction error in equation 4.22. This suggests that the circuit can compute an approximation to the scaled prediction error. The approximation will be accurate as long as $G + N$ is sufficiently large. Although the additional 1 in the denominator prevents perfect scaling, it might in fact be beneficial: it could prevent the catastrophically large prediction errors that cause the instabilities we discussed in section 4.2.1, which occur when the denominator becomes very small. So far, it looks as though the circuit has an equilibrium point at approximately the right value. However, it is not yet clear whether and how this equilibrium is reached. To learn more about these aspects, we need to simulate the system.

To simulate the computation of the prediction error, we assume G, N and r to be set externally, for example through cortical inputs. G and N then represent precisely timed reward predictions, while r represents the reward signal itself. We model G, N and r as step-functions that jump from zero to their respective values at the same time, as illustrated by the first panel of figure 4.14B. The time constants τ_δ and τ_T are set to realistic values taken from the literature (see subsection 4.5.7 for details). A simulation of the system is shown in figure 4.14B. We find that δ settles to its equilibrium value quite quickly (after tens of milliseconds) and without oscillations. This is likely due to the difference in time constants—the thalamic activity changes much faster than the striatal dopamine concentration. Our results suggests that even a simple system as the one in figure 4.14A can compute scaled prediction errors through a feedback loop.

The mechanism outlined here might resolve a puzzling feature of the AU model: in the AU model, tonic dopamine D is assumed to modulate the thalamic

activity T and hence decisions. However, the prediction error δ is assumed to be unaffected by modulation through tonic dopamine. For these assumptions to hold, one would either require very precisely timed pauses in dopaminergic modulation (during the computation of the prediction error) or a value learning substrate other than the basal ganglia that computes reward prediction errors independently. The scaled prediction error model does not require these extra assumptions—as we have seen, here the dopaminergic modulation of prediction errors is not only acceptable but in fact beneficial, as it allows for recursive control of dopamine release.

It is important to note that the analysis we present here has some substantial limitations. One important issue arises from the fact that the dopaminergic modulation D must only take values between 0 and 1 for the model to be biologically plausible (we explain this in more detail below in chapter 5, section 5.1.2). In the model defined by equations 4.23 and 4.24, we do not enforce this. The variables will hence assume implausible values for some inputs. We could enforce a biologically plausible range for our variables, for example by using saturation functions (we will take this approach below in section 5.1.2).

However, doing so would make it difficult to solve the equations for their equilibrium point—numerical methods would be needed to evaluate how well δ would approximate the scaled prediction error. Another issue lies in our assumption of perfectly timed, step-function like inputs. Does the computation still work if delays and gradual onsets are taken into account?

While these questions might be interesting directions for future research, they go beyond the scope of this thesis. We take the results in this subsection to be the first proof of principle for a possible neural mechanism that implements the computation of scaled prediction errors, instead of a detailed explanation. [4]

[4]The idea that prediction errors might contribute to dopaminergic modulation of the basal ganglia pathways will occur again later in this thesis, in chapter 5. There, we will show that such feedback has implications on the level of behaviour as well, in particular in the context of decision making under uncertainty. We will also present behavioural evidence consistent with such feedback.

The scaled prediction error learning rules are consistent with striatal plasticity

After establishing that dopaminergic scaled prediction errors are plausible, we now move on to discuss how the update rules given in equations 4.15 and 4.16 could be implemented in the basal ganglia circuit. As mentioned above, we assume that $v = \frac{1}{2}(G - N)$ and $s = \frac{1}{2}(G + N)$. These assumptions can be used to rewrite the learning rules given in equations 4.15 and 4.16 in terms of G and N. We get

$$\Delta G = \alpha_v f_\beta(\delta) - \alpha_s \tag{4.26}$$

$$\Delta N = \alpha_v f_\beta(-\delta) - \alpha_s \tag{4.27}$$

with $f_\beta(\delta) = \beta\delta^2 + \delta$ and $\beta = \alpha_s/\alpha_v$. When comparing these rules with the AU rules specified in equation 2.7 and equation 2.8, we find several similarities: both sets of rules have a decay term (linear decay in the scaled prediction error rules, exponential decay in the AU rules), and both sets of rules feature a nonlinear transformation of the dopamine signal δ. We plot the transformation of the scaled prediction error rules in figure 4.15A. The curves in that figure bear remarkable similarity to corresponding curves of the AU rules (see figure 4.15B).

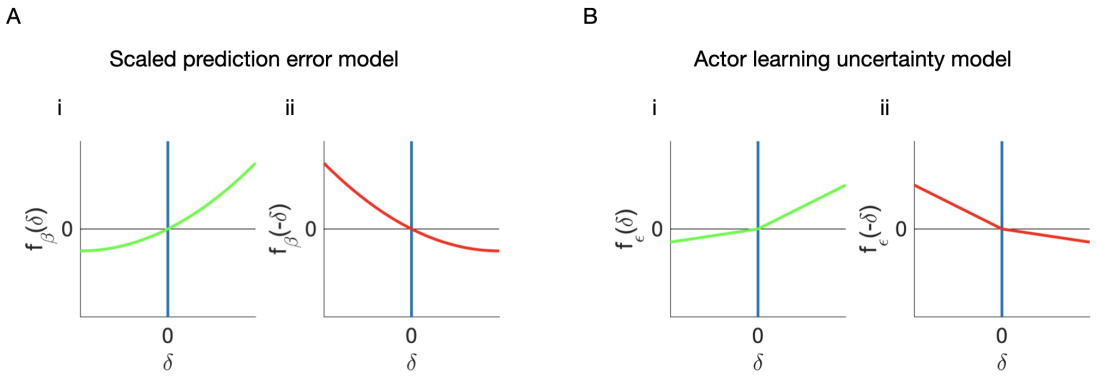

Figure 4.15: Nonlinear transformations of dopaminergic prediction errors in the scaled prediction error model and the AU model. A The nonlinear transformation of dopaminergic prediction errors in the scaled prediction error model. The transformation in the direct pathway (i) and the transformation in the undirect pathway (ii) are mirror images of each other. **B** Same as A, but for the AU model (see also figure 2.5A).

This implies that the implementation of the scaled prediction error rules might look very similar to the implementation of the AU rules. The same arguments apply in both cases: in short, the known properties of baseline D1 and D2 receptor occupancy and dopamine-dependent, pathway-specific three-factor plasticity in the striatum are consistent with the curvatures and overall slopes of the curves in figure 4.15—for a detailed discussion, see subsection 2.4.4 above. We may conclude that striatal plasticity might implement the learning rules of the scaled prediction error model, just as it might approximate the AU learning rules of chapter 3.

The scaled prediction error model appears biologically plausible

All in all, we have found plausible neural substrates for the scaled prediction error signal and the learning rules, and have provided an idea of how the scaled prediction error signal may be computed. We may conclude that the scaled prediction error model is at least as plausible as the AU model in chapter 3, and hence a promising candidate to model learning in the basal ganglia. After this analysis on the neural level, we move on to discuss behavioural evidence for scaled prediction error learning.

4.3.2 Behavioural plausibility

We look for evidence of scaled prediction error learning in a recent experiment done with monkeys (Ferrucci et al. 2019). The experiment featured two tasks, one carried out well after the other . The two tasks were similar in all but one aspect. They both consisted of a set of two-alternative forced choices or "problems". Each problem was presented six times in total. The outcomes of the choices were deterministic: one alternative was rewarded whenever it was chosen, the other alternative was never rewarded. To maximise their reward, monkeys had to determine the rewarded alternative through trial-and-error.

The difference between the tasks was in the sizes of the rewards: in the first task, all correct choices were rewarded with a medium amount of juice (0.3 ml). In

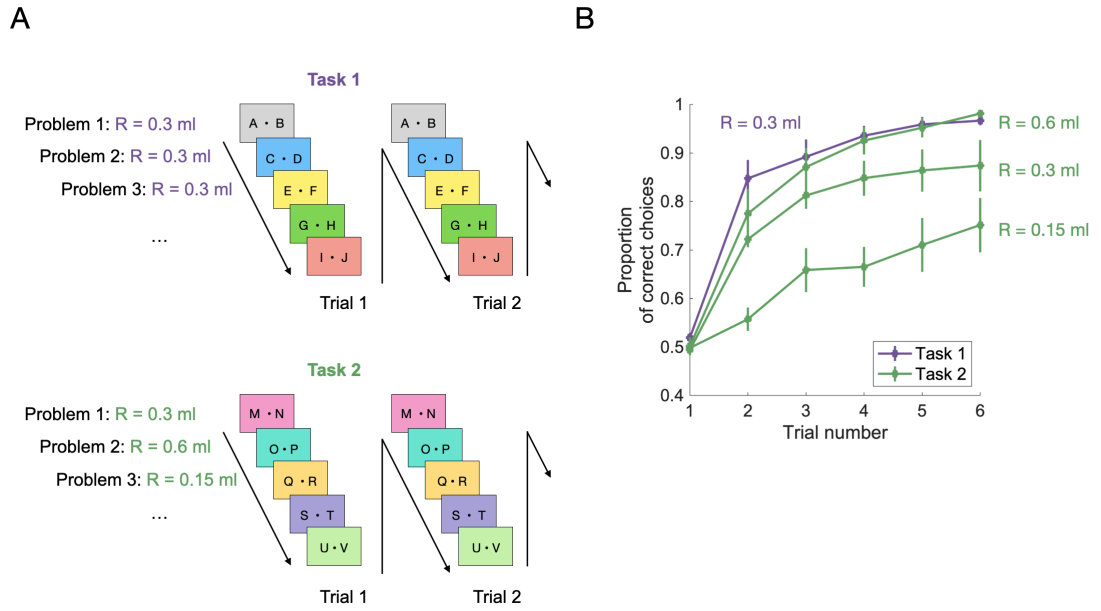

Figure 4.16: Trial structure and results of Ferrucci et al. (2019). A The structure of the two tasks. In task one, only one reward size is used. In task two, three different reward sizes occur. The reward size is constant within a problem but differs across problems. All problems are presented before moving to the next trial. Different sets of problems were used for the two tasks. **B** Learning curves. Learning curves are shown for each reward size in each task. The curve from the first task is purple, the curves from the second task are green. Errorbars indicate the standard error of the mean across subjects (N= 3). The learning curves are reproduced from the data of Ferrucci et al. (2019).

the second task, correct choices were rewarded with either a low (0.15 ml), a medium (0.3 ml) or a high (0.6 ml) amount of juice. Rewards were consistent within a problem, but differed across problems. The structure of the two tasks is illustrated schematically in figure 4.16A.

Consider the second task. How would we expected the reward magnitude to affect learning performance? Standard RL theory (for example a typical actor-critic model) suggests that higher rewards make for faster learning, and hence that performance should increase with reward size. This is exactly what was observed here: in figure 4.16B, we show the learning curves from the second task (in green), split by reward size. The curves are ordered just as the associated rewards predict.

Now, consider just the subset of problems in task two that yielded medium

rewards (i.e., 0.3 ml). This subset is completely equivalent to the entire set of problems in task one; hence, standard theory predicts that the learning performance associated with medium rewards in task two should be similar to the performance in task one. This, however, was not the case. Instead, the learning performance in task one was similar to the performance associated with high rewards in task two, as can be seen in figure 4.16B.

One way to interpret this is between-task adaptation. The two tasks feature different reward distributions; a learner that normalises reward distributions in some way might explain the adaptation. The phenomenon reported by Ferrucci et al. (2019) might therefore be a signature of scaled prediction error learning.

To test whether scaled prediction error learning could explain the between-task adaptation reported by Ferrucci et al. (2019), we fitted a standard actor-critic and the scaled actor-critic model to the behavioural data. In addition to scaled prediction error learning, we considered two other alternative explanations.

The first alternative explanation is reward range adaptation with respect to recent experiences (Rangel and Clithero 2012)[5]. That theory is based on the hypothesis that the brain represents rewards relative to the range of rewards that were experienced in the current context. Experiences are believed to be mapped to a fixed reward spectrum, ranging from "worst experience in the current context" to "best experience in the current context". The corresponding neural value signals then reflect the relative reward within a context, not the absolute reward of an experience. Value signals linked to the same stimulus may thus change as the context changes; such adaptations are thought to happen slowly, over tens of trials.

The reward range adaptation hypothesis is supported by solid evidence (Padoa-Schioppa 2009), and differs from our scaled prediction error model in two aspects:

[5]Note that in this and in similar references, the word *value* is used where this thesis would use the word *reward*. This is because in our framework, "value" refers to a quantity to be learned within the system of interest (the basal ganglia)—a stimulus will acquire value if it predicts reward. "Reward", on the other hand, is considered an input to the learning system.

the first difference is that in the scaled prediction error model, adaptation happens at the stage of the reward prediction error, but not before. The rewards that the learning rules take as input are considered to be on an absolute scale. Reward range adaptation, on the other hand, would suggest that those rewards are already encoded relative to the range of other rewards.

The second difference is in the nature of scaling: the scaled prediction error model suggests that prediction errors are scaled by the standard deviation of the reward distribution. Reward range adaptation, on the other hand, proposes that rewards (and hence prediction errors) are scaled with respect to the range of recent rewards. Such range adaptation might be achieved using buffered memories of recent experiences, similar to mechanisms such as decision by sampling (Stewart et al. 2006). In situations where the worst reward is no reward at all, reward range adaptation might also be achieved by scaling with the recent maximum reward.

These two differences—*where* the scaling happens, and *how*—should be observable on the neural level as well as on the level of behaviour, in particular if the standard deviation of the reward distribution is different from its range.

The second alternative explanation we consider is normalisation at choice time, often referred to as divisive normalisation (Louie et al. 2011). That theory suggests that when a decision must be made, the learned values of the options are retrieved and then normalised relative to each other. The choice is then based on the normalised values of the options. In contrast to reward range adaptation, which is assumed to occur gradually over several trials, normalisation at choice time occurs very quickly and flexibly before each decision. Normalisation at choice time differs from the scaled prediction error model and reward range adaptation in that it suggests scaling of decision variables at decision time, as opposed to scaling of feedback signals at outcome time (Rangel and Clithero 2012). It also differs in the scaling factor that is applied to the decision variables. Here, we

consider normalisation by the sum of the values of the available options—Louie et al. (2011) call this divisive normalisation.

In total, the models we considered are a standard actor-critic, a scaled actor-critic, an actor-critic with reward range normalisation and a RW model with divisive normalisation at choice time. We fit all models to the empirical learning curves obtained from both tasks. The results of these fits are shown in figure 4.17. Details of the models and the procedure can be found in subsection 4.5.8.

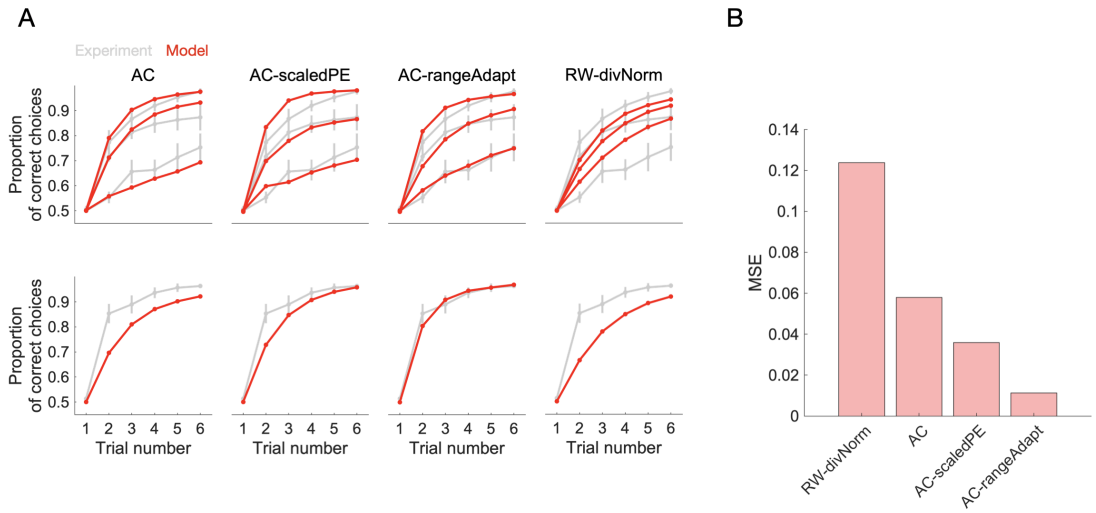

Figure 4.17: Modelling learning in the experiment of Ferrucci et al. (2019). AC stands for Actor-Critic. **A** Learning curves from experiment and fitted models, superimposed. Choice accuracy (i.e. relative frequency of choosing the rewarded option) is plotted as a function of trial number. Error bars indicate the standard error of the mean across participating monkeys. The grey curves show the empirical results and are identical across columns (the same data is also shown in figure 4.16B). The red curves show the predictions of the fitted models. The first row corresponds to task 2 of Ferrucci et al. (2019), the second row corresponds to task 1. Note that the errorbars on the red curves are too small to be visible. **B** Mean squared errors (MSE) after fitting. The bars indicate the value of the fitting loss function at the minimum (see 4.5.8 for details), and thus the goodness of fit.

We find that the scaled actor-critic fits the empirical data better than the standard actor-critic and the divisive normalisation model. However, we also find that reward range adaptation explains the data better than the scaled actor-critic (see figure 4.17B). The reward range adaptation model uses reward ranges to normalise value signals; we may thus hypothesise that the adaptation phenomenon

in this study is more likely to be related to reward ranges than to reward standard deviations. This is consistent with intuition: figure 4.16A shows that the learning curves for the largest available rewards—the upper end of the reward range—within each task coincide. It seems plausible that an animal might focus its learning resources on the largest available reward, thus achieving comparable performances.

Overall, it seems that this study does not provide strong evidence for the scaled prediction error model: though scaled prediction error learning describes the empirical data better than the standard actor-critic and the divisive normalisation model, our model still systematically predicts lower learning performance than observed in task 1 (see figure 4.17A, second column, lower panel). Furthermore, there exists an alternative explanation that is well backed by another line of evidence and explains the data better than the scaled prediction error model.

The absence of clear evidence of scaled prediction error learning in the experiment of Ferrucci et al. (2019) of course does not mean that scaled prediction error learning is not used by animals. For example, it is possible that the scaling is problem-specific (in contrast to the scaling in the scaled actor-critic model, which applied across problems in the task above). In other words, monkeys might learn a scaling term s for each problem, and not just a single s for all problems. The study of Tobler et al. (2005) supports this view.

But would problem- or stimulus-specific normalisation not imply the same speed of learning in all conditions in the experiment of Ferrucci et al. (2019)? Not necessarily: stimulus-specific normalisation might be too slow to show in the experiment of Ferrucci et al. (2019). The details of the procedure of Tobler et al. (2005) suggest that normalisation within the context of a stimulus might take time: before the recordings were performed, conditioning had been going on for weeks, with about 1500 occurrences of each stimulus. In contrast, the monkeys of Ferrucci et al. (2019) encountered each problem only six times. It seems plausible that stimulus-specific normalisation might take more experience than this.

In summary, the study of Ferrucci et al. (2019) seems to provide evidence for reward range adaptation, but not for scaled prediction error learning. Based on this, we hypothesise that the scaling of scaled prediction error learning might not apply across stimuli, but might instead be stimulus-specific and slow. It appears that we should focus on prediction tasks like those that we studied in subsection 4.2.3, more than on the action-selection tasks of subsection 4.2.4.

4.3.3 Promising future experiments

We saw that the study of Ferrucci et al. (2019) did not provide evidence of scaled prediction error learning. This helped us narrow our hypothesis down: scaled prediction error learning might be stimulus-specific, and it might require tens or even hundreds of trials to appear. What experiments might be suitable to test this? Here, we want to suggest two possible designs.

First, scaled prediction error learning could be probed with a task that combines elements of the experiments of Ferrucci et al. (2019) and Tobler et al. (2005): consider a task with a small number of stimuli (say three), paired with uncertain rewards of different sizes. For example, one could use 0.15, 0.3 and 0.6 ml of fruit juice, as in the experiment of Ferrucci et al. (2019), given only in 50 % of the trials. The stimuli are presented in mixed, random order. So far, this would be very similar to the experiment of Tobler et al. (2005).

In addition, we suggest that after stimulus presentation the subject should have to press one of two buttons (say left and right) to move to the reward stage. However, at the beginning of the task, reward delivery would be independent of button choice. For example, a subject might see a stimulus that is rewarded with 0.3 ml of juice in 50 % of all trials. It would then press one button (say the left one), and receive a reward still with a probability of 50 %. Overall, the first phase of this design is indeed very similar to the experiment of Tobler et al. (2005) with the main difference that the subject must press a button to move from trial to trial.

In the second phase of the task, a contingency between buttons and reward would be introduced. This might start after 200 trials or so: from trial 201 on, one of the buttons leads to reward delivery with certainty, the other button is never rewarded. This is the case for the rest of the task (say another 100 trials). We may expect to observe random pressing in the first phase of the task followed by a gradually developing preference for the correct button in the second phase. The second phase of our design would be very similar to the experiment of Ferrucci et al. (2019), but with fewer stimuli and more trials.

Standard learning theory (and also reward range adaptation) would predict different learning curves for different reward sizes in the second phase of the task. Based on these theories, one would expect the same outcome as in task 2 of Ferrucci et al. (2019). Scaled prediction error learning theory makes a different prediction: if the participants learn stimulus-specific reward means and standard deviations, they should normalise prediction errors by the end of phase 1. This should then lead to identical learning curves across different reward sizes in phase 2.

This design should be able to differentiate between scaled prediction error learning and the other forms of learning mentioned above on the level of behaviour. Further, it would be useful to collect neural recordings of dopamine signals (using single unit or FMRI methods) as well, to determine which types of prediction error signals occur in this task, and how they are linked with behaviour.

Second, the scaled prediction error model could be probed in a reward prediction task like the one we analysed in section 4.2.3: participants might be instructed to track a signal and predict the next sample over an extended amount of trials. Unbeknownst to the participants, the signal's mean and standard deviation might drift slowly, for example like in figure 4.2.

If participants used scaled prediction error learning to track the signal, we should observe adaptation to increased or decreased standard deviations. The updates of the participants' predictions should remain similar in magnitude

throughout. If participants used unscaled learning, one should observe increased update magnitudes in times of high standard deviation and decreased update magnitudes in times of low standard deviation, with no adaptation taking place.

Trial-by-trial modelling (similar to what we use in section 5.3 of chapter 5) could be used to test the scaled prediction error model against other value estimation models on the level of behaviour. Methods such as FMRI could be used to measure the underlying prediction error signals and potentially determine the neural correlates of scaled prediction error learning, in a way similar to Behrens et al. (2007).

4.4 Discussion

We presented a new model of error-driven learning: the scaled prediction error model. We showed how it can be derived from Bayesian principles, we tested it in simulations and analysed its empirical plausibility using both behavioural and neural data. Now, we will discuss the new model more broadly. First, we will summarise our key findings. Then, we discuss how the scaled prediction error model relates to other models from neuroscience and from artificial intelligence.

4.4.1 Summary

This chapter was about the scaled prediction error model, which describes how an organism might adapt its learning mechanism to changing levels of reward observation noise σ.

First, we used the mode-matching method to find learning rules that approximate Bayesian learning. This resulted in the scaled prediction error learning rules, which can track the mean and standard deviation of a reward signal. We then tested the performance of the new rules. A thorough analysis of robustness revealed that there are certain conditions to be met for the rules to function optimally: the standard deviation learning rate α_s should be much smaller than σ, and the value learning rate α_v should be equal to the process noise ν and smaller

than σ as well. Comparing scaled prediction error learning with RW learning, we found that the new learning rules can improve performance when a learner faces unknown or varying levels of reward observation noise. This holds true for reward prediction tasks as well as action selection tasks.

Finally, we reviewed empirical evidence relating to scaled prediction error learning. On the neural level, we found that scaled prediction error learning describes dopamine responses better than conventional models in several studies. We further showed how scaled prediction errors could be computed in a dopaminergic feedback loop, and how the basal ganglia pathways might implement the learning rules of the scaled prediction error model.

We finally tested whether the rules could explain learning speed effects in a task with different reward magnitudes. We found that scaled prediction error learning could fit the data better than the standard model, but that there were other explanations of the phenomenon that fitted even better. We concluded that scaling might not apply across many stimuli, and proposed that instead it be stimulus-specific. We closed by suggesting experiments that might detect such stimulus-specific scaled prediction error learning.

4.4.2 Relation to models in neuroscience

The AU model

The scaled prediction error model is closely related to the AU model—both models describe how the basal ganglia pathways track reward uncertainty; they also share the distributed encoding of reward statistics. It is thus not surprising that the learning rules of the two models have similarities (at least qualitatively, see subsection 4.3.1 and figure 4.15). However, the scaled prediction error model differs from the AU model in several important aspects.

Of course, the scaled prediction error itself is the key new feature that drives most of the interesting effects we investigated in this chapter. It is through the scaling of the prediction error that our new model puts its estimate of the reward

observation noise to good use. The AU model tracks reward noise as well, but does not use its estimate to improve learning performance (or for anything else). In contrast, the scaled prediction error model explains not only *how* to to track σ, but also *why*.

Further, there is a small difference in the statistics that the models track: the AU model tracks mean and spread (the average absolute distance from the mean), while the scaled prediction error model tracks mean and standard deviation (the square root of the average squared distance from the mean). Finally, note that the scaled prediction error model can learn unbiased estimates of mean and standard deviation (i.e., the stochastic fixed point of the learning rules is at the exact target values). This is not true for the AU model, which can only learn values proportional to the statistics it tracks.

The Kalman-TD model

Scaled prediction error learning is not the only model that addresses the scaling of dopamine responses. One recent theory—Kalman-TD—explained those responses, as well as other phenomena such as preconditioning, as a consequence of volatility tracking (Gershman 2017). Kalman-TD applies the Kalman filter method to the computational problem of TD learning: reward prediction in the time domain. The resulting model features vector-valued learning rates that constantly adapt to observations and outcomes. It elegantly describes how covariances between cues and cue-specific uncertainties might modulate learning, and can be shown to explain several empirical phenomena. However, the Kalman-TD theory does not address the tracking of observation noise (the theory focuses on process noise). It also does not discuss how prediction error scaling might be implemented. We may thus view it as a complement rather than a competitor to the theory presented above.

The reward taxis model

Another model was recently proposed to explain the effects reported by Tobler et al. (2005) and other phenomena. The model is called *reward taxis* (Karin and Alon 2021), and explains the dopaminergic range adaptation using a logarithm: if both rewards and reward expectations were transformed by a logarithmic function, prediction errors would be given by $\delta = \log(r) - \log(v) = \log(r/v)$. In the experiment of Tobler et al. (2005) rewards were given in 50% of the trials. For a reward of size r, the expected reward would then be $v = r/2$, and the prediction error would be $\delta = \log(r/(r/2)) = \log(2)$, i.e., independent of reward size. Reward taxis can hence explain the results of Tobler et al. (2005) quite elegantly.

However, that explanation breaks down as we look at other experiments. We have already mentioned the experiment by Rothenhoefer et al. (2021), which featured two reward distributions with equal means and ranges but different standard deviations. We show those distributions in figure 4.13A. Rothenhoefer et al. (2021) first used Pavlovian conditioning in a way similar way to Tobler et al. (2005), pairing the two reward distributions with two different cues. They then recorded the dopamine responses at reward delivery, for all reward sizes of each distribution. We reproduce their data in figure 4.13B (first panel). The responses to the middle reward are similar for both distributions, but the responses to the extreme rewards differ: they seem scaled up for the normal distribution.

What would the reward taxis theory predict for the responses in this experiment? Both distributions have the same mean; reward taxis hence predicts similiar responses for both distributions. We show this in figure 4.13B (last panel). The experimental data thus falsifies the reward taxis model in this experiment. In contrast, the scaled prediction error model predicts different responses for the two distributions—we show this in figure 4.13B (middle panel). Overall, it appears as if dopamine responses to reward distributions with variable width are better captured by the scaled prediction error model than the reward taxis model.

Free energy models

Finally, we want to discuss the relation of our model to free-energy models: the scaled reward prediction errors in this chapter are formally related to the precision weighted prediction errors of the free-energy approach, especially when the recognition density (the learner's model of the world) is taken to be Gaussian (K. J. Friston et al. 2008; Buckley et al. 2017). In that case, the prediction errors that drive inference and learning in free energy models are often weighted by precisions, i.e., inverse variances. The connection to scaled reward prediction errors becomes very close when the free energy approach is applied to reward prediction, dopamine and the basal ganglia system, as has been done in the DopAct framework (Bogacz 2020). This framework integrates several theoretical ideas (free energy, reinforcement learning, habits without values and active inference), and suggests that dopaminergic prediction errors drive both learning and action planning. Precision weighted prediction errors encoded by dopamine transients feature in one variant of that model, but they are not the focus of the theory, and possible implementations or empirical consequences of these weighted prediction errors have not been investigated so far. Furthermore, it is important to note that precision, or inverse variance, scales differently to standard deviation, and might hence not explain classical observations such as those reported by Tobler et al. (2005).

4.4.3 Relation to models in artificial intelligence

Scaled reward prediction errors have been explored outside of neuroscience as well: in the field of AI-type reinforcement learning, it was noticed that normalising reward prediction errors can enable an agent to learn effectively across several different tasks (Hessel et al. 2019). This is consistent with our conclusions: different tasks come with different levels of reward observation noise, and adaptive scaling can normalise performance across tasks without requiring the need for fine-tuning. However, the rules for scaling prediction errors in AI are different from the scaled prediction error learning rules and have

not been designed with the intention to model learning in biological systems. Further, Hessel et al. (2019) have focused on typical benchmark tasks of AI-type reinforcement learning (i.e. Atari games and others), while we have explored the types of tasks that are used in neuroscience and psychology.

Prediction error scaling also occurs at a more basic level of AI, inside the optimisation algorithms that are used to improve the parameters of neural networks. A very prominent example is the Adam optimiser (Kingma and Ba 2014), which implements a variant of gradient descent in which all updates are normalised using an estimate of the second moment of the gradient distribution. By making gradient descent effective across different gradient magnitudes, adaptive optimisers such as Adam contribute to the spectacular successes of deep learning. This supports the main idea of this chapter—that scaling prediction errors can be beneficial for learning. However, here we only looked at the scaling of *reward* prediction errors. Adam-style optimisation suggests that there might be similar mechanisms for other neural error signals as well. This is an intriguing hypothesis, though we are not aware of any empirical evidence that supports it—the mechanisms and evidence presented in this chapter focus on reward prediction errors and the basal ganglia system, and generalising them to a more universal mechanism would require more work.

4.5 Methods

4.5.1 Stability for low-variance signals

To generate the data shown in figure 4.4, we simulated scaled prediction error learning for different combinations of the learning rate α_S and reward observation noise σ.

We used the scaled prediction error learning rules given in equations 4.14–4.16. The learning rate for v was set equal to the learning rate for s, $\alpha_v = \alpha_s$. The rewards were sampled from a normal distribution with mean 0 and standard deviation σ.

For both α_S and σ, we used 100 evenly spaced values between zero and one. Simulations were run for all possible combinations. For each combination, we simulated 10 runs with 100 trials per run. The variables v and s were initialised at their target values (i.e., $v_0 = 0$ and $s_0 = \sigma$). We then computed the average square difference between s and σ across all trials and runs. For each combination (σ, α_s), we thus obtained one error magnitude.

4.5.2 Stability for non-stationary signals

We generated the data shown in figure 4.5 by simulating scaled prediction error learning for different combinations of the learning rate α_v and the reward process noise ν.

We used the learning rules given in equations 4.14–4.16. The learning rate for s was kept at $\alpha_s = 0.1$. Rewards were sampled from a non-stationary normal distribution with standard deviation $\sigma = 1$ and a drifting mean. The mean performed a Gaussian random walk, $\mu_{t+1} \sim \mathcal{N}\left(\mu_t, \nu^2\right)$, initialised at $\mu_0 = 0$.

For ν and α_v, we used 100 and 99 different values respectively. Those values ranged from 0.018 to 55 and were evenly spaced on a logarithmic scale[6]. For each combination (α_v, ν), the signal was tracked for 10^4 trials. For each such run, the variables v and s were initialised at their target values (i.e., $v_0 = \mu_0$, $s_0 = \sigma$). We then computed the average squared difference between the models' predictions v and the true mean μ across all trials. For each combination (α_v, ν), we thus obtained one error magnitude.

4.5.3 Reward prediction performance

In figure 4.6, we compare the performance of the RW model with the performance of the scaled prediction error model. As above, the scaled prediction error model was defined by the learning rules given in equations 4.14–4.16. The RW model was defined by the learning rule in equation 2.1.

[6]One might object that it does not make sense to test learning rates greater than one. Here, this does not apply: the effective learning rate is α_v / s, so α_v can be larger than one as long as $s > \alpha_v$.

The scaled prediction error learning rules were parametrised with $\alpha_v = 1$ and $\alpha_s = 0.1$. The RW model was parametrised with $\alpha = 0.5$. Rewards were sampled from a normal distribution with drifting mean, just as in subsection 4.5.2. The process noise was fixed at $\nu = 1$.

For panel 4.6A, we used three different observation noise levels (1, 5 and 15). For panel 4.6B, we used 101 observation noise levels, evenly distributed on a logarithmic scale from 0.1353 to 1096.6. For each σ and model we simulated 2×10^4 trials and computed the average squared difference between the model predictions v and the true mean μ across all trials.

For figure 4.7, we used 10 different learning rates for the RW model (ranging from 0.007 to 0.993), the same parameters as above for the scaled prediction error model, and 100 different levels of observation noise, evenly distributed on a logarithmic scale from 0.1353 to 1096.6. The process noise was fixed at $\nu = 1$ as above. For each combination, we simulated 10^5 trials and computed the average squared difference between the model predictions v and the true mean μ across all trials.

s

4.5.4 The distracted bandit task

The distracted bandit task consisted of three actions, a_1, a_2 and a_3. The rewards associated to those actions were normally distributed, with means $\mu_1 = 0$, $\mu_2 = 9.5$ and $\mu_3 = 10$ and standard deviation $\sigma = 1$.

We tested two models in the distracted bandit task: a standard actor-critic model and an scaled actor-critic model. In both models, actions are generated by sampling from a softmax distribution over action weights:

$$P(a) = \exp \phi_a / \sum_b \exp \phi_b \tag{4.28}$$

with $P(a)$ the probability that action a is chosen and ϕ_a the action weight of action a. The standard actor-critic rules were

$$\delta = r - v \tag{4.29}$$

$$\Delta\phi_a = \alpha_\phi \delta \left(c_a - P(a)\right) \tag{4.30}$$

$$\Delta v = \alpha_v \delta \tag{4.31}$$

with $c_a = 1$ if action a was chosen and 0 otherwise. The actor update in equation 4.30 was derived using the policy gradient method (Sutton, Barto, et al. 1998), see appendix B.

The scaled actor-critic rules were

$$\delta = \frac{r - v}{s} \tag{4.32}$$

$$\Delta\phi_a = \alpha_\phi \delta \left(c_a - P(a)\right) \tag{4.33}$$

$$\Delta v = \alpha_v \delta \tag{4.34}$$

$$\Delta s = \alpha_s \left(\delta^2 - 1\right). \tag{4.35}$$

For both models, v and ϕ_a were initialised at zero. For the scaled actor-critic model, s was initialised at $s_0 = 6$. The learning rates we used were

$$\alpha_{\phi,\,\text{scaled}} = 0.1 \tag{4.36}$$

$$\alpha_{v,\,\text{scaled}} = 1 \tag{4.37}$$

$$\alpha_s = 0.02 \tag{4.38}$$

for the scaled actor-critic and

$$\alpha_{\phi,\,\text{standard}} = \alpha_{\phi,\,\text{scaled}} / s_0 \tag{4.39}$$

$$\alpha_{v,\,\text{standard}} = \alpha_{v,\,\text{scaled}} / s_0 \tag{4.40}$$

for the standard actor-critic. The learning rates were chosen such that the *effective* learning rates of both models were equal at the beginning of the task. For the scaled actor-critic model, this includes the effect of s_0. This parametrisation ensures that both models behave similarly at the beginning of the task, before diverging due to the adaptation mechanisms of the scaled actor-critic model.

For each candidate model, we simulated 1000 runs with 1000 trials per run, hence 10^6 trials in total. We estimated the probability of choosing an action in a given trial by computing the proportion of choices of that action in that trial, across runs.

The empirical standard deviation of the reward distribution (black line in figure 4.8C) was computed by taking the standard deviation across all rewards obtained in a certain trial. Since there were 1000 independent simulations of the task, each standard deviation was computed from 1000 samples. To obtain the average estimate of the standard deviation (blue line in figure 4.8), we averaged the estimate s across all runs of the scaled actor-critic model.

For figure 4.9, we used the same models as above. For each model and parametrisation, we simulated 5000 runs with 1000 trials per run. Performances were computed by averaging the probability of choosing the highest value action across all trials and runs. We systematically varied $\alpha_{\phi,\,\text{scaled}}$ across the range $[0, 1]$ for 4.9A and $\alpha_{v,\,\text{scaled}}$ across the range $[0, 1]$ for 4.9B, while keeping the rest of the parametrisation as described above. The learning rates of the standard actor-critic model moved with the learning rates of the scaled actor-critic model according to equations 4.39 and 4.40. We tested 100 values for each learning rate.

4.5.5 The diminishing returns task

The diminishing returns task has a continuous action space, with actions a taking values in the positive real numbers: $a \in [0, \infty)$. Actions are mapped to rewards deterministically; in particular, the learner receives a reward $r = \log a$ for performing the action a. Higher a always led to higher rewards, but the rate of reward increase slows the higher a becomes.

We tested two models in the diminishing returns task: a standard actor-critic model and a scaled actor-critic model. The standard actor-critic rules were

$$\delta = r - v \tag{4.41}$$

$$\Delta\phi = \alpha_\phi \, \delta \, (a - \phi) \tag{4.42}$$

$$\Delta v = \alpha_v \, \delta \tag{4.43}$$

$$\tag{4.44}$$

with ϕ the learner's current estimate of the best action and a the action that was sampled before the update. The scaled actor-critic rules were

$$\delta = \frac{r - v}{s} \tag{4.45}$$

$$\Delta\phi = \alpha_\phi \, \delta \, (a - \phi) \tag{4.46}$$

$$\Delta v = \alpha_v \, \delta \tag{4.47}$$

$$\Delta s = \alpha_s \left(\delta^2 - 1 \right). \tag{4.48}$$

Again, the actor update in equation 4.46 and equation 4.42 were derived using the policy gradient method (Sutton, Barto, et al. 1998), see appendix B. In both models, actions are generated by sampling from a normal distribution,

$$a \sim \mathcal{N}\left(\phi, \sigma^2\right), \tag{4.49}$$

with $\sigma^2 = 0.05$.

The initial values were $\phi_0 = 1$ and $v_0 = 0$ for both models and $s_0 = 0.2$ for the scaled actor-critic. The learning rates we used were

$$\alpha_{\phi,\,\text{scaled}} = 0.1 \tag{4.50}$$

$$\alpha_{v,\,\text{scaled}} = 0.01 \tag{4.51}$$

$$\alpha_s = 0.001 \tag{4.52}$$

for the scaled actor-critic and

$$\alpha_{\phi,\,\text{standard}} = \alpha_{\phi,\,\text{scaled}} / s_0 \tag{4.53}$$

$$\alpha_{v,\,\text{standard}} = \alpha_{v,\,\text{scaled}} / s_0 \tag{4.54}$$

for the standard actor-critic. Again, the learning rates were chosen such that the effective learning rates of both models were equal at the beginning of the task.

For each candidate model, we simulated 10 runs with 2000 trials per run. Learning curves were obtained by computing the average centre ϕ of the policy for each trial, across trials.

The empirical standard deviation and the scaled actor-critic model's estimate thereof were computed as described in subsection 4.5.4.

To produce figure 4.11, we proceeded analogously to figure 4.9: we simulated 100 runs with 2000 trials per run for each model and parametrisation, and computed performances as the average ϕ across all trials and runs. We systematically varied $\alpha_{\phi,\,\text{scaled}}$ across the range $[0, 0.5]$ for 4.11A and $\alpha_{v,\,\text{scaled}}$ across the range $[0, 0.01]$ for 4.11B, while keeping the rest of the parametrisation as described above. The learning rates of the standard actor-critic moved with the learning rates of the scaled model according to equations 4.53 and 4.54. We tested 100 values for each learning rate.

4.5.6 Simulations of the task of Tobler et al. (2005)

To simulate the relevant parts of the experiment reported by Tobler et al. (2005), we modelled Pavlovian conditioning with three different stimuli, which were associated with three different reward magnitudes ($r = 0.05$, $r = 0.15$ and $r = 0.5$). The stimuli were followed by the associated reward in one half of the trials and by no reward in the other half. The rewarded trials were selected pseudorandomly, such that there were two rewarded and two non-rewarded trials in every four successive trials.

We simulated 2000 trials per stimulus, and extracted prediction errors from the last 1500. Discarding the first 500 trials accounts for the substantial pretraining of Tobler et al. (2005).

We used two models: an RW model and a scaled prediction error model. The learning rule of the RW model is given in equation 2.1. The rule was used with $\alpha_v = 0.0067$ and $v_0 = 0$. The learning rules of the scaled prediction error model are given in equations 4.14, 4.15 and 4.16. These rules were used with $\alpha_v = \alpha_s = 0.0067$ and $v_0 = 0$, $s_0 = 1$.

To compare our simulations to the experimental data from Tobler et al. (2005), we extracted the prediction errors δ from the simulations and averaged them for each model, outcome and condition separately. There were three conditions (corresponding to the three reward sizes) with two outcomes (reward or no nothing) each, resulting in a total of six combinations per model. Finally, we normalised the six averaged prediction errors by their standard deviation for each model.

4.5.7 A dynamical model of the basal ganglia

The differential equations 4.23 and 4.24 were solved using Matlab's *ode15s*, from $t = -200$ ms until $t = 500$ ms. As inputs, we used step functions

$$r(t) = \theta(t) r_{\text{step}} \tag{4.55}$$

$$G(t) = \theta(t) G_{\text{step}} \tag{4.56}$$

$$N(t) = \theta(t) N_{\text{step}} \tag{4.57}$$

with $\theta(t) = 1$ for $t > 0$ and $\theta(t) = 0$ for $t < 0$, and $G_{\text{step}} = 10$, $N_{\text{step}} = 6$ and $r_{\text{step}} = 4$. These inputs correspond to a learned mean $v = 2$ and a learned standard deviation $s = 8$.

The time constant τ_T of the thalamic population was set to 10 ms, based on the measurement of the membrane time constant of thalamic neurons reported by Paz et al. (2007). The time constant τ_δ for striatal dopamine was set to 300 ms, based on figure 2C of Montague, McClure, et al. (2004): the dopamine transient in that figure decays to e^{-1} of its peak value in about 300 ms.

4.5.8 Simulations of the task of Ferrucci et al. (2019)

For each model, we simulated 300 participants that played two tasks with 180 sessions each. In task 1, one session consisted of one problem (i.e., one two-alternative forced choice) that was presented six times. One option was always rewarded ($r = 1$), the other was always unrewarded ($r = 0$).

In task 2 one session consisted of five problems, each presented six times. The simulated subjects made a decision for each problem before advancing to the next trial. Different problems yielded rewards of different sizes ($r = 0.5$, $r = 1$ or $r = 2$). Reward sizes were selected such that every reward size was associated with at least one and not more than two problems within a session. Further, reward sizes were balanced such that all rewards occurred equally often within three sessions.

We fitted four models: a standard actor-critic (AC), a scaled actor-critic (AC-scaledPE), an actor-critic with range adaptation (AC-rangeAdapt) and an RW model with divisive normalisation at choice time (RW-divNorm).

AC The standard actor-critic (AC) model is defined in equation 4.29 and equation 4.30, with equation 4.28 as the choice rule. It featured the free parameters α and α_v, and the variables ϕ_a and v. As above, ϕ_a encoded preference for action a, while v encoded the context value. At the start of each task we initialised v at zero. For each new session, we set $\phi_a = 0$ for all a. The baseline v was not problem-specific, but was used for all problems in a session.

AC-scaledPE The scaled standard actor-critic (AC-scaledPE) model is defined in equation 4.32 - equation 4.35, with equation 4.28 as the choice rule. It had the same free parameters and variables as AC. Additionally, it had a fixed parameter $\alpha_s = 0.01$ and a variable s, which encoded an estimate of the reward observation noise across the context. We initialised s at one at the start of each task. We fixed α_s because it could not be estimated from the learning curves. The baseline v and the scale s were not problem-specific, but were used for all problems in a session.

AC-rangeAdapt The actor-critic with range adaptation (AC-rangeAdapt) is identical to the standard actor-critic, with the single difference that the reward r that enters the prediction error is scaled by the maximum of all previous rewards[7]:

$$\delta = \frac{r}{\max\{r_{t-1}, r_{t-2}, ...\}} - v \tag{4.58}$$

$$\tag{4.59}$$

The set of previous rewards is initialised as $\{1\}$ for each participant at the beginning of each new task. Intuitively, this model is a standard actor-critic that operates on a reward signal that is continuously range-adapted by another brain module. Again, the baseline v was not problem-specific, but was used for all problems in a session.

[7] If the lowest experienced reward is zero, then the maximum reward also marks the magnitude of the reward range. Scaling by the maximum reward and scaling by the reward range are equivalent in this case.

RW-divNorm The RW model with divisive normalisation at choice time (RW-divNorm) had free parameters α and β, and variable values v_a for each action a. All v_a were initialised near zero.

The choice rule was a softmax rule over normalised values:

$$\phi_a = \beta v_a / \sum_b v_b \tag{4.60}$$

$$P(a) = \exp \phi_a / \sum_b \exp \phi_b \tag{4.61}$$

with $P(a)$ the probability that option a is chosen. The sums range over all available options. The normalisation at choice time was divisive and based on the sum of the values of the available options. The learning rule of the RW model was equation 2.1.

We fitted the models by minimising the loss function

$$L = \sum_t \left(p(1)_t^{\text{exp}} - p(1)_t^{\text{mod}} \right)^2 + \sum_c \sum_t \left(p(2)_{t,c}^{\text{exp}} - p(2)_{t,c}^{\text{mod}} \right)^2$$

The two sums correspond to the two tasks. In the first sum, $p(1)_t$ is the proportion of correct choices in trial t across all sessions and participants in task one. In the second sum, $p(2)_{t,c}$ is the proportion of correct choices in trial t and reward condition c across all sessions and participants in task two. The superscript *exp* refers to experimental data, the superscript *mod* refers to data obtained by simulating a model. Intuitively, the loss function measures how much the simulated learning curves differ from the empirical learning curves.

The loss function was minimised using Matlab's *patternsearch* function. The minimised loss was used to quantify the goodness of the model fit in figure 4.17B.

5

Prediction errors and risk seeking

Contents

We saw in chapter 2 that dopamine has direct and indirect effects on behaviour. Accordingly, the models we studied in the previous chapters (especially the AU model in chapter 3) featured two "types" of dopamine: the prediction-error-type dopamine δ and the modulation-type dopamine D. The first type, δ, occurred in the update rules, driving changes of synaptic weights. The second type, D, regulated striatal activity and hence determined the relative impact of direct and indirect pathways on the basal ganglia output.

In chapter 3, we followed Mikhael and Bogacz (2016) in treating these two types of dopamine as independent—as if they were two different neurotransmitters. In chapter 4, we allowed prediction errors to modulate the pathway balance, and showed that this interaction might explain how scaled prediction errors are computed.

Which of these two approaches is correct? Are the two functions of dopamine separated by some mechanism, or can dopaminergic reward prediction errors also affect pathway balance? This question is actively debated in the field—we reviewed the debate above in chapter 2. At the time of writing, the answer is unknown.

In this chapter, we want to address these questions using human behaviour. Our experiment was inspired by the question "Do dopaminergic reward prediction errors modulate the pathway balance?", but it was designed around behavioural correlates of these neural entities. Unexpected rewards were used to provoke dopaminergic reward prediction errors, and risk preferences were measured as a proxy for dopaminergic pathway modulation (dopamine can induce risk-seeking, as reviewed in subsection 2.2.2). The central question then becomes: "Are unexpected rewards associated with risk-seeking?"

The chapter is structured as follows: first, we introduce a novel learning task in section 5.1. We then analyse that task using the AU model and derive two concrete behavioural predictions. In section 5.2 we analyse the behaviour that was recorded in the task and test whether the predicted effects did occur. Finally, we use modelling to consolidate our findings and rule out alternative explanations in section 5.3.

A paper reporting the results in this chapter is under review at the time of writing; an interim version of that paper is available (Möller, Grohn, et al. 2021). The text and the figures in this chapter were taken from that paper and adapted for this thesis. This was done with the full consent of all co-authors (written permission was obtained).

The data we show below was collected by one of the coauthors (JG), but text and figures were composed by the author of this thesis, who also conducted the corresponding analyses.

5.1 Task and theory

5.1.1 The task

Our task consists of a sequence of 120 two-alternative forced-choice trials. On each trial, after an inter-trial interval (ITI) of 1s, two stimuli (fractal images, figure 5.1A) are drawn from a set of four stimuli and shown to the participant, who has to choose one. The stimuli shown on any given trial are selected pseudo-randomly, such that all ordered stimulus combinations (12 combinations) occur equally often (10 times each).

The choice is followed by a delay of 1s. Then a numerical reward between 1 and 99 is displayed under the chosen stimulus for 1.5s. Then, the next trial begins. Participants are instructed to maximise the total number of reward points throughout the experiment.

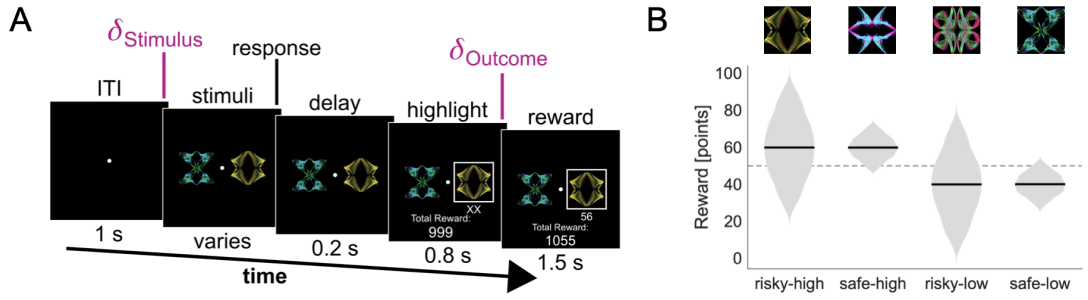

Figure 5.1: Trial structure and reward distributions. A Task structure. Each trial begins with a blank screen. After 1s, two out of four possible stimuli are selected pseudo-randomly and shown to the participant. The participant chooses one of them, which is then highlighted after a delay of 0.2s. After another delay of 0.8s a point reward is displayed underneath the chosen stimulus. During each trial, two separate reward prediction errors occur: at stimulus presentation, the participant discovers whether the two displayed stimuli are better (positive prediction error) or worse (negative prediction error) than the expected average stimulus. At outcome presentation, the participant discovers whether the obtained reward is higher or lower than the expected average reward of the selected stimulus. **B** Reward distributions. Each stimulus (top) is associated with a different reward distribution (bottom). The distributions differ in mean (60 points versus 40 points) and standard deviation (20 points versus 5 points). The reward distributions are unknown to participants. The dashed line in the background indicates the middle of the reward range, which is at 50 points. This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

The reward on each trial depends on the participant's choice: each stimulus is associated with a specific reward distribution from which rewards are sampled. The four reward distributions associated with the four stimuli are approximately Gaussian and follow a two-by-two design: the mean of the Gaussian can be either high or low (60 or 40), and the standard deviation can be either large or small (20 or 5), resulting in four reward different distributions (risky-high, risky-low, safe-high and safe-low, figure 5.1B). The names derive from the idea that it is "risky" to pick a stimulus associated with a broad reward distribution—for such stimuli outcomes might deviate a lot from the expected outcome. Correspondingly, it is "safe" to pick a stimulus with a narrow distribution—here, outcomes will mostly be as expected.

We organise the trials into three conditions: 1) *different*: trials in which the shown stimuli have different average rewards (for example risky-high and safe-low, which have average rewards of 60 and 40 respectively), 2) *both-high*: trials in

which both stimuli have a high average reward (risky-high and safe-high, both have an average reward of 60 points) and 3) *both-low*: trials in which both stimuli have a low average reward (risky-low and safe-low, both have an average reward of 40 points).

5.1.2 Learning and decision making in theory

After introducing the task, we now proceed to sketch a mechanistic theory of learning and decision-making in this task. This theory is then used to derive behavioural predictions and a computational model.

We use the AU model of Mikhael and Bogacz (2016) as the basis of our analysis. As in chapter 3, we assume that there is an action channel for each option i in our task; that the probability of choosing option i depends on the total activation T_i of that action channel; and that this activation is influenced by the direct pathway weights G_i, the indirect pathway weights N_i and dopamine D (see also equation 2.9):

$$T_i = DG_i - (1 - D)N_i. \tag{5.1}$$

As above, G_i and N_i are subject to dopamine dependent plasticity, and hence reward-driven learning. Here, we will not use the full AU learning rules (see chapter 3, equations 2.7 and 2.8, for those rules and a detailed analysis). Instead we use a simplified version. We consider the difference $Q_i = 1/2 \times (G_i - N_i)$ and the sum $S_i = 1/2 \times (G_i + N_i)$ of the variables G_i and N_i. Those variables have simple interpretations: Q_i converges to the mean of the rewards associated with option i, S_i converges to their spread (this was shown above in subsection 3.2.2). We use the learning rules

$$\delta = r - Q_i \tag{5.2}$$

$$\Delta Q_i = \alpha_Q \delta \tag{5.3}$$

$$\Delta S_i = \alpha_S(|\delta| - S_i), \tag{5.4}$$

with $\alpha_Q > \alpha_S$. These learning rules are idealised versions of the proper AU rules (compare them to equation 3.7 and equation 3.8). The idealised rules are less complex and can thus be fitted to behaviour more easily.

Expressed in the variables Q and S, the action channel activation defined in equation 5.1 becomes

$$T_i = Q_i + (2D - 1)S_i. \tag{5.5}$$

We see that the impact of reward spread on action activation is gated by the level of dopamine.

For our model to remain biologically plausible, D must take values between 0 and 1. To ensure this, we introduce a reparameterisation: let $\delta \in (-\infty, +\infty)$ quantify dopamine activity relative to the baseline, i.e., $\delta = 0$ corresponds to baseline dopamine activity, $\delta < 0$ means that dopamine activity is suppressed, and $\delta > 0$ means that dopamine activity is enhanced. We can then write

$$D = \sigma(\tilde{\omega}\delta) \tag{5.6}$$

with $\sigma(x) := (1 + e^{-x})^{-1}$ the sigmoid function and $\tilde{\omega}$ a proportionality constant. Using this notation, we can be sure that equation 5.1 will produce biologically plausible results for any value δ may take. Inserting the new parametrisation into equation 5.5 yields

$$T_i = Q_i + \tanh(\omega\delta) \times S_i, \tag{5.7}$$

with $\omega = \tilde{\omega}/2$ a rescaled proportionality constant. This is the equation that we will use in our models.

In our task, the Q_i and S_i of the four action channels should converge to the means and spreads of the four reward distributions. From equation 5.5, we can read off that with dopamine at baseline (i.e., $D = 0.5$ or $\delta = 0$) the channel activation is proportional to Q_i, and hence to the mean reward. Choices should then be biased towards options with high mean rewards. If dopamine levels are increased (i.e., $\delta > 0$), the learned spread S_i contributes positively to action activation, biasing choices towards risky options. If, on the other hand, dopamine levels are below baseline (i.e., $\delta < 0$), the learned spread S_i reduces action activation, biasing choices towards safe options.

5.1.3 Reward prediction errors in theory

Next, we explore how the participant's reward prediction might change over the course of a trial. Our analysis is based on TD learning (Sutton and Barto 2018), which has been applied to describe dopaminergic responses to rewards (we reviewed this in section 2.1). The basic assumption of that theory is that participants maintain a prediction of upcoming rewards at all times. In the context of this task, the prediction would be based on learned estimates Q_i of the average rewards associated with the four stimuli.

At the beginning of a trial (before the stimuli appear), participants do not have any specific information to base their reward prediction on. Given that our task design contains a fixed ITI and trials have the same structure throughout, we may assume that participants anticipate the appearance of stimuli at a certain time after the initial fixation. The reward prediction at that time should be an average

over all possibilities, i.e., an average over the learned values of all options that might occur, $1/4 \times \sum_{\text{all } i} Q_i$.

After the options appear, participants should adjust their reward prediction based on the learned values of the displayed options. We take the participants' updated prediction to be the average learned value over the presented options, $1/2 \times \sum_{\text{shown } i} Q_i$. The updated prediction should differ across conditions: if the participants learned accurate estimates of the values, their prediction would be $1/2 \times \sum_{\text{shown } i} Q_i = 60$ points in the *both-high* condition and $1/2 \times \sum_{\text{shown } i} Q_i = 40$ points in the *both-low* condition.

This change in participants' reward expectation through the appearance of reward-predicting stimuli constitutes a reward prediction error that we call the *stimulus prediction error* δ_{stim}[1]. The timing of the stimulus prediction error is shown in figure 5.1. This prediction error, previously described by A. I. Jang et al. (2019), should cause phasic dopamine activity according to the reward prediction error hypothesis of dopamine. The magnitude of the stimulus prediction error is given by

$$\delta_{\text{stim}} = 1/2 \times \sum_{\text{shown } i} Q_i - 1/4 \times \sum_{\text{all } i} Q_i. \tag{5.8}$$

If the values of the stimuli are learned with reasonable accuracy, this reward prediction error should have a magnitude of about $60 - 50 = 10$ in the *both-high* condition and about $40 - 50 = -10$ in the *both-low* condition. After successful learning, we hence expect a positive stimulus prediction error in the *both-high* condition and a negative stimulus prediction error in the *both-low* condition.

Next, participants will make a choice. Now, their reward expectation is the value Q_{selected} of the option they chose. Finally, a reward r is displayed, forcing

[1] It is important to note that the stimulus prediction error is a reward prediction error that occurs at the time of stimulus onset, not an error in stimulus prediction. The identity of the stimulus is relevant only insofar as it is related to reward expectations.

participants to update their reward prediction again. We call this second reward prediction error—the difference between the learned value of the option and the actual reward received—the outcome prediction error δ_{out}. Its magnitude is given by

$$\delta_{\text{out}} = r - Q_{\text{selected}}. \tag{5.9}$$

It is the outcome prediction error that drives learning about the stimuli (see equation 5.3 and equation 5.4 above). The outcome prediction error is a reward prediction error at the time of outcome presentation. Stimulus prediction errors and outcome prediction errors thus exist within the same signal—they are both reward prediction errors—they just happen at different times.

In summary, our analysis suggests that two prediction errors (and two corresponding dopamine responses) should occur in each trial: first the stimulus prediction error after the presentation of the options, and second the outcome prediction error after the presentation of the reward.

To check whether those prediction errors occur, we combined trial-by-trial estimates of their magnitude[2] with measurements of the participants' pupil diameters. We found that pupils responded to both prediction errors, reflecting the corresponding surprise. These analyses are reported in appendix D. While this does not directly confirm the occurrence of the dopamine responses we describe here, it does suggest that both stimulus onset and outcome presentation trigger cognitive processes related to reward expectation.

5.1.4 The PEIRS model

In the previous subsection, we have analysed reward prediction errors at different times in the trial. In the subsection before that, we sketched the theory of learning

[2]The trial-by-trial estimates of prediction error magnitude were extracted as part of our modelling analysis. We explain this in detail in section 5.3.

and decision-making we will use in this chapter. Now, we want to combine these two theories. Together, they will form the *Prediction Errors Induce Risk Seeking* (PEIRS) model. The key novel idea of PEIRS is to connect two dopamine-related phenomena—the dopamine responses to changes in reward expectation on the one hand, and the dopaminergic modulation of risk preferences on the other hand.

The PEIRS model is central to this chapter: it compactly summarises the hypothesis we test and the theory we suggest. Because of its importance, we now give its full mathematical description. The model is based on the learning rules in equations 5.3 and 5.4 from above:

$$\Delta Q_i = \alpha_Q (r - Q_i) \tag{5.10}$$

$$\Delta S_i = \alpha_S \left(|r - Q_i| - S_i \right). \tag{5.11}$$

We add a choice rule based on the activation T_i of action channels as modelled in equation 5.7,

$$T_i(\delta) = Q_i + \tanh(\omega\delta) \times S_i. \tag{5.12}$$

By writing T_i as a function of δ, we stress that the activation of the different options is a function of dopamine activity in the striatum at any given moment. The activations are turned into choice probabilities using a conventional softmax choice model (Daw et al. 2011),

$$P(i) = \frac{\exp(\beta T_i(\delta))}{\sum_{j\,\text{shown}} \exp(\beta T_j(\delta))}, \tag{5.13}$$

with $P(i)$ the probability that option i is selected. Crucially, we propose that the relevant modulatory dopamine activity δ is the activity that corresponds to the stimulus prediction error in equation 5.8, i.e.

$$\delta = \delta_{\text{stim}}, \tag{5.14}$$

In other words, we propose that the stimulus prediction error (a reward prediction error that occurs at the time of stimulus presentation) might cause an increase or a pause in dopamine transmission, which then affects the risk preferences in the choice by modulating the basal ganglia pathways.

Using equation 5.7, equation 5.13 and equation 5.14, we arrive at

$$P(i) = \frac{\exp(\beta(Q_i + \tanh(\omega\delta_{\text{stim}}) \times S_i))}{\sum_{j\,\text{shown}} \exp(\beta(Q_j + \tanh(\omega\delta_{\text{stim}}) \times S_j))}, \tag{5.15}$$

for the probability of choosing option i. We can see that this probability is a function of the stimulus prediction error δ_{stim} interacting with the estimate S_i of reward risk . This model, completely defined by equations 5.8, 5.10, 5.11 and 5.15, encodes our hypothesis: that prediction errors might affect risk preferences by modulating the balance of the basal ganglia pathways.

5.1.5 Behavioural Predictions

In addition to developing a mathematical model of learning and decision making in our task, we can also use our theory to derive task-specific behavioural predictions. See figure 5.2 for a schematic representation of the following logic. We have seen that the presentation of the options should cause a positive prediction error in the *both-high* condition, and a negative prediction error in the *both-low* condition. Those prediction errors should lead to transient changes in the striatal dopamine levels during the choice period (increases in the *both-high* condition, and decreases in the *both-low* condition, see figure 5.2, mechanistic level). We have also seen that dopamine levels affect choices through modulation of the BG pathways: increased dopamine makes people risk-seeking, decreased dopamine makes them risk-averse (see equation 5.7). If the average reward is similar for two

options (as it is in the *both-high* and the *both-low* condition), these risk preferences should be the decisive factor in decisions (figure 5.2, task level and past rewards level).

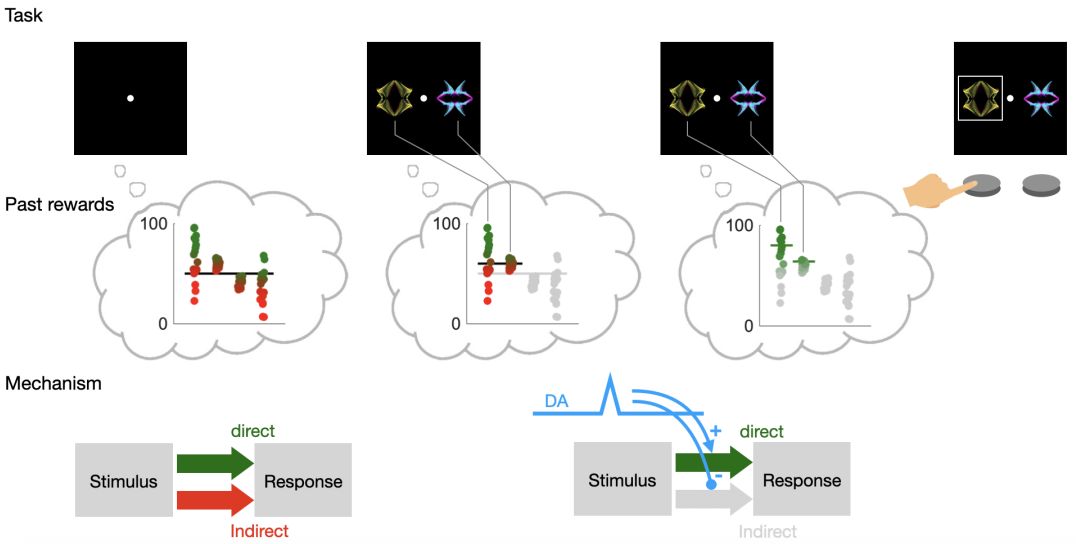

Figure 5.2: Events during the first half of a trial (ITI to response) according to our theory. As the blank screen appears the participant's reward expectation (black horizontal line in the first thought bubble) is based on all past rewards (distributions in the first thought bubble). Above-average past rewards are encoded in the direct pathway (green dots) and below-average past rewards in the indirect pathway (red dots). They are weighed equally since the direct and the indirect pathways are in balance (first diagram in bottom row). As the stimuli appear, only the relevant past rewards are considered (irrelevant rewards greyed out in the middle thought-bubble), and the reward expectation changes (black horizontal line in the middle thought-bubble is higher than the black horizontal line in the first thought-bubble). The upward change in reward expectation constitutes a TD reward prediction error which is signalled by increased dopamine transmission in the striatum (blue transient in the bottom row). This suppresses the indirect pathway and enhances the direct pathway (second diagram in bottom row). As a result, below-average past rewards are ignored, and the focus shifts to above-average past rewards (below-average rewards greyed out in the last thought bubble). The stimulus with the larger spread is now valued higher (green lines in the last thought bubble) and is therefore chosen. This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

Taken together, these premises suggest that the risk preferences in the *both-high* condition should be different to those in the *both-low* condition—risk-seeking should be stronger in *both-high* than in *both-low* because the stimulus prediction error is higher in *both-high* than in *both-low*. This should be the case even if other,

condition-independent risk preferences (such as a general risk-aversion) occur in addition to the effect we propose.

Assuming there are no other substantial risk effects, we can make another even more specific prediction: we should see a preference for the risky stimulus in the *both-high* condition (depicted in figure 5.2) and a preference for the safe stimulus in the *both-low* condition. These effects should appear gradually, since they require that both the mean and the spread of the reward distributions are learned. Specifically, risk preferences in the *both-high* and *both-low* conditions should appear slower than value preferences in the *different* condition. This follows from the underlying plasticity rules: above, in section 3.2, we show that the learning rate for spread must always be lower than the learning rate for value. In addition to this, a reasonably accurate value estimate is required for the spread estimate to converge; this also contributes to a higher learning speed for value compared to spread.

In summary, we have formulated two predictions: 1) We should see a difference in risk preference between conditions, and 2) we should see gradually emerging risk-seeking in the *both-high* condition and gradually emerging risk aversion in the *both-low* condition. These predictions are based on hypothetical neural mechanisms. However, the predictions are purely on the level of behaviour. From here on, we hence focus on behaviour, conscious that this will only provide indirect evidence of the neural mechanisms behind the effects we will describe.

5.2 Behavioural analysis

So far we introduced a learning task and performed a theoretical analysis to derive behavioural predictions. Now we will present the results of testing these predictions using data collected from that task.

5.2.1 Performance

A cohort of participants ($N = 30$, 3 excluded, see 5.5.1) engaged in the task described above. Each participant performed four blocks of 120 trials. Each block used a new set of four stimuli, mapped to the same four distributions. Participants made their choices on average 0.97 s (standard deviation 0.51 s) after stimulus onset.

First, we investigate whether participants' performance improved during the task. We expect choice accuracy (i.e., the likelihood of choosing the option with the higher average reward) to increase gradually over trials. To test this, we focus on choices in the *different* condition, where participants had to choose between stimuli with different average rewards. We find that the probability of choosing the stimulus with the higher average reward increases gradually over trials across the population. Average performance differs from chance level with high significance (figure 5.3A, t-test, $t(27) = 31.9$, $p < 0.001$) and approaches its asymptote in the second half of the block. This suggests that the participants learned to distinguish between high-value and low-value stimuli.

5.2.2 Risk preferences

Next, we test our first prediction—that there would be a difference in risk preferences between the *both-high* and the *both-low* condition. Specifically, we predicted higher risk-seeking in *both-high* than in *both-low*. To test this, we compute the average difference in risk preference between conditions for each participant across all trials. We find that most of the participants were more risk-seeking in the *both-high* condition than in the *both-low* condition (figure 5.3B; two-tailed paired t-test: $t(27) = 3.58$, $p = 0.0016$), which confirms our first prediction.

Next, we test our second prediction—that participants would be risk-seeking in the *both-high* condition and risk-averse in the *both-low* condition, and that these preferences would emerge gradually and more slowly than preferences

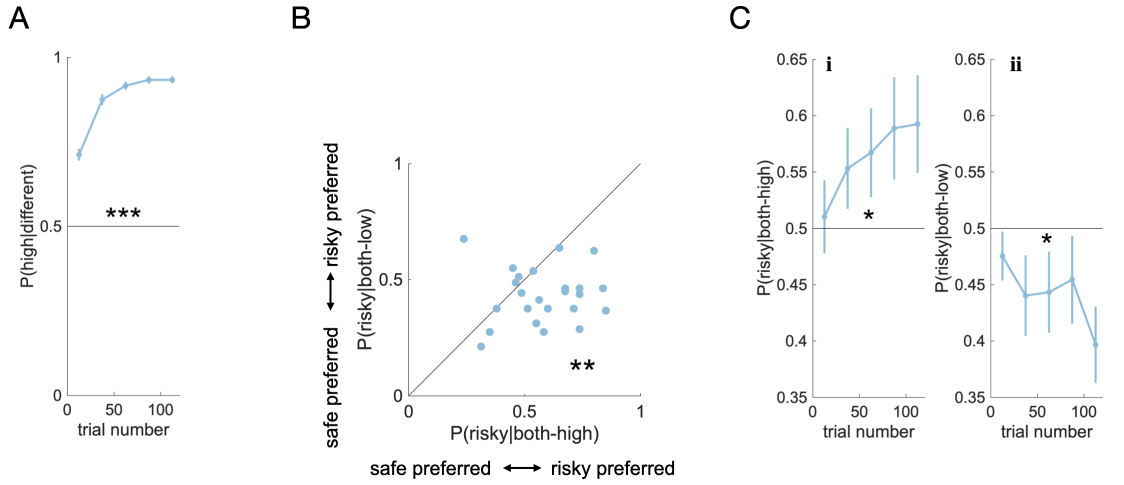

Figure 5.3: The gradual emergence of condition-specific risk preferences. A Probability of choosing a stimulus with high average reward over a stimulus with low average reward, as a function of trial number. Choices are binned according to trial number. For each bin, we show the mean (dot) and SE (bar) across subject means. The stars (*: p < 0.05, **: p < 0.01, ***: p < 0.001) indicate that the population mean across all trials is significantly different from chance (0.5). **B** Correlation between aggregate risk preferences in the *both-high* and the *both-low* condition. Each point represents one participant. The preference for the risky stimulus if mean rewards are high (x-axis) is plotted against the preference for the risky stimulus if mean rewards are low (y-axis). If a point falls below the diagonal, the participant is more risk-seeking for high-mean stimuli than for low-mean stimuli. The stars indicate that the population mean is significantly below the diagonal. **C** Probability of choosing the risky stimulus over the safe stimulus, i) if mean rewards are both high, and ii) if mean rewards are both low, as a function of trial number. The data is represented as in A. This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

for high-value stimuli. For this, we analyse choices in the *both-high* and *both-low* condition. As predicted, we find significant risk-seeking in the *both-high* condition (figure 5.3Ci; two-tailed t-test: $p = 0.0343$, $t(26) = 2.23$) and significant risk aversion in the *both-low* condition (figure 5.3Cii; two-tailed t-test: $p = 0.0317$, $t(26) = -2.27$). These preferences emerge gradually as a function of trial number, at a slower rate than the preferences for the high- over the low-value stimuli. To see this, compare figure 5.3A against figure 5.3Ci and figure 5.3Cii. The slopes of those curves differ significantly (mixed-effects model: $p = 0.001$, see section 5.5). The data thus also confirms all aspects of our second prediction.

5.2.3 Reaction times

In addition to choices, we also analyse reaction times. In particular, we explore how reaction times depend on trial number, and on reward variability—do participants deliberate longer when choosing between risky stimuli than when choosing between safe stimuli? The results of these analyses are shown in figure 5.4.

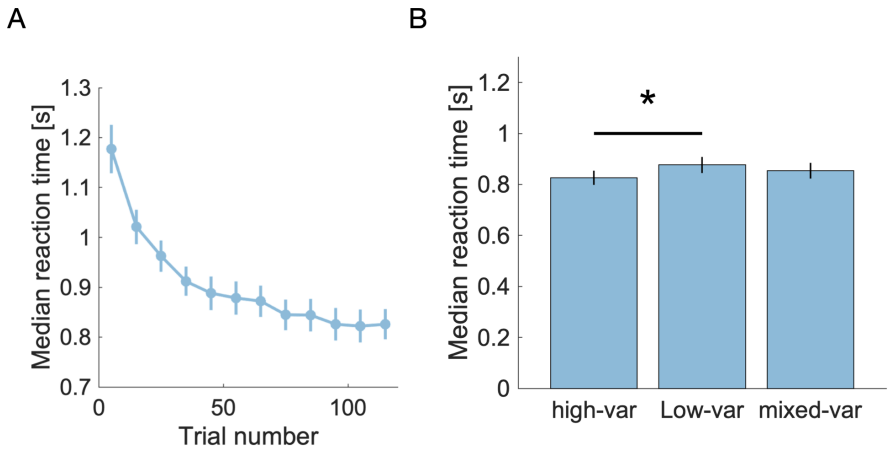

Figure 5.4: Reaction times. A Reaction times as a function of trial number. We computed the median reaction time for 10-trial buckets for each participant. We show the mean and standard error of these medians across participants. **B** Reaction times as a function of reward variance. We split the trials in the *different* condition into three categories: high-variance (high-var), mixed-variance (mixed-var) and low-variance (low-var) trials. High variance trials are those in which participants had to choose between two risky stimuli (i.e., risky-high versus risky-low); in low variance trials, they had to choose between two safe stimuli (i.e., safe-high versus safe-low). All other trials of the *different* condition are mixed variance trials (i.e., risky-high versus safe-low or safe-high versus risky-low). Again, we computed median reaction times for those categories for each participant. We show the mean and SE of these medians across participants. The star (*: $p < 0.05$) indicates that the population means differed significantly between high-variance and low-variance trials.

First, we find that reaction times decrease quickly over the first 40 trials before slowly converging towards an asymptote during the rest of the block. This is as expected: reaction times are thought to depend on value differences, with larger value differences leading to quicker choices (Pedersen et al. 2017). In our task, value differences emerge gradually as participants learn, leading to a gradual decline in reaction times. In addition, habit effects (i.e. stimulus-response links

rather than stimulus-value-response links) might speed up choices especially towards the end of a block.

Second, we find that participants are faster when choosing between risky stimuli than when choosing between safe stimuli (two-tailed paired t-test, $t(26) = -2.19$, $p = 0.038$). This is remarkable, at least according to our theory: the more reward noise a stimulus is associated with, the larger the corresponding N_i ($G_i + N_i$ encodes the spread S_i), and hence the larger the activation of the indirect pathway during the choice. We would expect that this would slow choices down—the indirect pathway has been shown to delay choices by raising the decision threshold (Herz et al. 2016).

Here, we find the opposite effect—choices between risky stimuli are slightly but significantly faster than choices between safe stimuli. Uncovering the mechanisms underlying this could be an interesting direction for future research, but goes beyond the scope of this chapter.

In summary, in data recorded from the task described above, we find the two behavioural effects we predicted, and behaviour consistent with successful learning (increasing preferences for high-value stimuli and decreasing reaction times) otherwise. Our findings thus provide initial evidence for a behavioural link between prediction errors and risk preferences. In the next section, we use computational modelling to compare our explanation of the measured effects to alternative explanations.

5.3 Modelling

Are there alternative explanations for the observed effects? Does our theory fit the data better than existing theories? In this section, we use computational modelling to answer these questions. To test our theory against alternative explanations, we use simulations as well as model comparison techniques (Palminteri et al. 2017).

5.3.1 Models

First, we introduce the models we want to consider. Here, we describe the models qualitatively. A mathematical description in given below (subsection 5.5.3). All models we use here are variants of the RW model. RW is also the first type of explanation for the effects we observed—it has been shown that even basic RW-type learning can lead to risk preferences through sampling biases (Niv, Edlund, et al. 2012).

The second type of explanation involves the utility of reward points: risk aversion, as well as risk-seeking, have been explained as consequences of nonlinear utility functions (Kahneman and Tversky 2013). We consider the s-shaped family of utility functions as particularly relevant, as it encodes important aspects of the influential prospect theory of Kahneman and Tversky (2013): outcomes are evaluated relative to a reference point, the utility function is curved in different directions beneath and above the reference point. The corresponding model is called s-shaped UTIL.

The third type of explanation is based on RW with a variable learning rate. It has been observed that humans use different learning rates for positive and negative outcomes (Gershman 2015) and that this can lead to risk preferences (Niv, Edlund, et al. 2012). We implement this in the model pos-neg RATES.

Finally, the explanation that we propose in this study is that prediction errors induce transient risk preferences, as detailed in subsection 5.1.4. The PEIRS model represents this hypothesis.

The models described so far encode what we consider to be the most relevant alternative explanations of our effect: they are all well-grounded in empirical research and have all been related to risk preferences before. We will discuss the corresponding results in some detail.

In addition to those primary alternative explanations, there is a wide array of other, less prominent models, which we will refer to as secondary alternative

explanations. We will introduce them via a short summary in the following paragraphs.

Besides s-shaped utility, several other types of utility functions exist. In particular, we consider a concave utility function (concave UTIL), a convex utility function (convex UTIL) and an inverse s-shaped utility function (inverse s-shaped UTIL). Together with s-shaped UTIL, those families of functions cover all basic curvature types, among them those that are popular in neuroeconomics (the concave type of expected utility theory and the s-shaped type of prospect theory), as well as some more exotic types (convex and inverse s-shaped).

The learning rate might depend on the valence of the prediction error, as codified in the pos-neg RATES model. However, it might also be modulated by other variables. We include one model in which the learning rate is modulated by the reward noise (the variance RATES model, which features different learning rates for the high-variance and the low-variance stimuli). We also include a model that aims to capture attention effects (the attention RATES model): here, the learning rate depends on the surprise quantified by the absolute outcome prediction error. The core idea is that very surprising outcomes might draw more attention and might hence be committed to memory more thoroughly.

Not only the learning rate can be modulated—prediction errors, too, could be subject to context-specific adaptation, as we have discussed in detail in chapter 4. We include this hypothesis as a potential explanation through the scaled PE model[3].

Finally, we include two variations of our explanation (PEIRS): first, we consider the possibility that predictions, not prediction errors, might modulate risk-seeking (the PIRS model, Predictions Induce Risk Seeking). Second, we include a model in which risk preferences are modulated by the outcome prediction error on the previous trial (the OEIRS model, Outcome Errors Induce Risk Seeking).

[3]The scaled PE model here is related with but not equivalent to the scaled prediction error model of chapter 4—the learning rules we use in this chapter are adapted slightly to fit into the PEIRS framework.

All in all, we are going to test 12 models: the base model RW, the PEIRS model that encodes our hypothesis, two other primary alternative explanations and eight secondary alternative explanations.

5.3.2 Simulations

First, we want to know which of the models can reproduce (and hence explain) the observed effects. To check this, we first fit all models to the dataset and extract maximum-likelihood parameters for each participant. We then use these parameters to simulate the task with all candidate models, generating synthetic datasets. Finally, we analyse these synthetic datasets in the same way as the experimental dataset. In that analysis, we consider the difference in risk preferences between conditions as well as the distribution of risk preferences across the population (as shown in figure 5.3B) and the evolution of risk preferences over trials (as shown in figure 5.3C).

We find that the PEIRS model reproduces the effect best: it captures the overall distribution (figure 5.5A) and the condition-specific gradual emergence of risk preferences across the population (figure 5.5B) better than RW, s-shaped UTIL and pos-neg RATES (see figure 5.12 for likelihood ratios).

For the difference in risk preferences between conditions, PEIRS further produces the most realistic effect size among those models (figure 5.6, empirical effect size: 0.122, simulated effect size with PEIRS model: 0.086, no significant difference between the experimental data and the simulated data, two-tailed paired t-test, $t(26) = -1.72$, $p = 0.0976$).

The main alternative models (RW, s-shaped UTIL, pos-neg RATES) predict very narrow distributions for the risk preferences, which is at odds with our empirical results (figure 5.5A). They also fail to reproduce the emergence of differential risk preferences over trials (figure 5.5B): preferences comparable to those observed in the experiment either emerge only in one condition (pos-neg RATES, *both-high* condition), or not at all (RW, s-shaped UTIL). With respect to the difference of

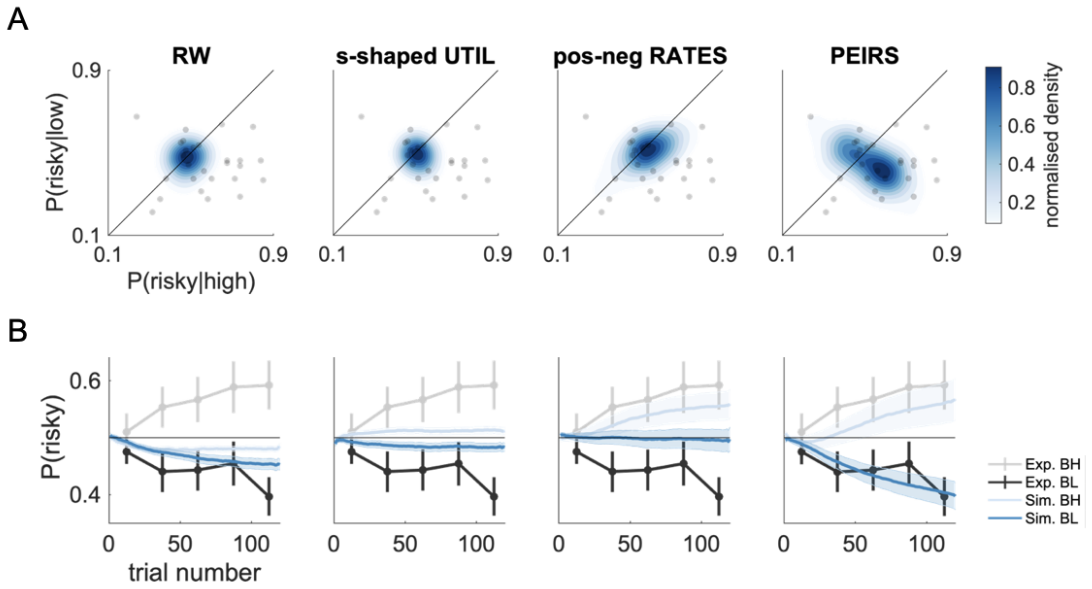

Figure 5.5: Emergence and distribution of risk preferences for the main models. A Distributions of risk preferences in simulated datasets. The risk preference distributions extracted from simulated data are plotted as blue shadings. The corresponding experimental data is superimposed as a grey scatterplot (identical in all panels). The shading corresponds to estimated probability density functions and is scaled to match each distribution's density value range for better visibility. **B** The evolution of risk preferences in simulated datasets across trials. Risk preferences were extracted from simulated (Sim.) datasets, split between conditions (BH: Both-High, BL: Both-Low) and are plotted as a function of trial number (light blue for the *both-high* condition, dark blue for the *both-low* condition). The shaded areas indicate the SE across simulated participants. The simulated data were smoothed using a moving average with a window-size of 20. The corresponding experimental (Exp.) data is superimposed in grey (light grey for the *both-high* condition, dark grey for the *both-low* condition) and represented in the same way as in figure 5.3C. This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

risk preferences between conditions, all alternative models produce effect sizes that differ significantly from the empirical findings (figure 5.6, two-tailed paired t-tests, s-shaped UTIL: $t(26) = -2.19$, $p = 0.038$, RW: $t(26) = -2.78$, $p = 0.010$, pos-neg RATES: $t(26) = -2.4$, $p = 0.025$).

We conclude that among RW, s-shaped UTIL, pos-neg RATES and PEIRS, our theory (i.e., the PEIRS model) is the best explanation for the observed risk preferences (concerning distribution, emergence and differences between conditions). All other primary alternative explanations are falsified by the simulations.

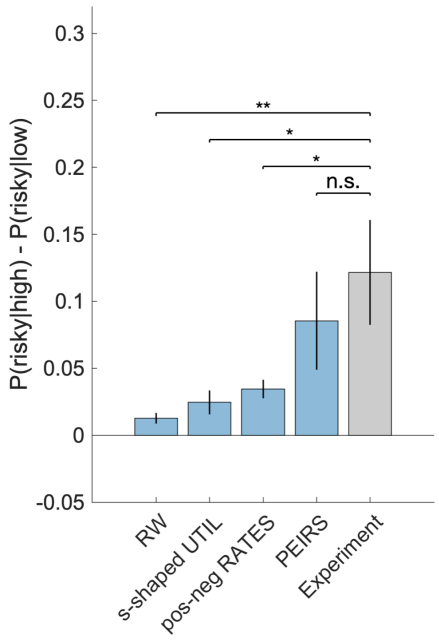

Figure 5.6: Differences in risk preferences between conditions in theory and experiment. Simulated datasets are shown as blue bars, the experimental dataset is shown as a grey bar. We show the mean (bar) and the SE (errorbar) of the difference in risk preferences between conditions across subjects, averaged over experiment repetitions for simulated datasets. The stars (*: $p < 0.05$, **: $p < 0.01$) indicate simulated differences in risk preferences that differ significantly from the empirical differences. Here, n.s. means "not significant". This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

We also test the secondary alternative explanations. We first explore how well the secondary models can reproduce the emergence and distribution of the empirical risk preferences. The results of this test are shown in figure 5.7 for the models concave UTIL, convex UTIL, inverse s-shaped UTIL and variance RATES, and in figure 5.8 for the models scaled PE, attention RATES, PIRS and OEIRS.

We also investigated whether the secondary models could reproduce the difference in risk preferences between conditions. The results of this analysis are shown in figure 5.9.

We find that most of the secondary models can be falsified as well, with the single exception of inverse s-shaped UTIL, which captures both the emergence of differential risk preferences and the difference between conditions (but is worse than PEIRS in capturing the distribution of risk preferences, see figure 5.12).

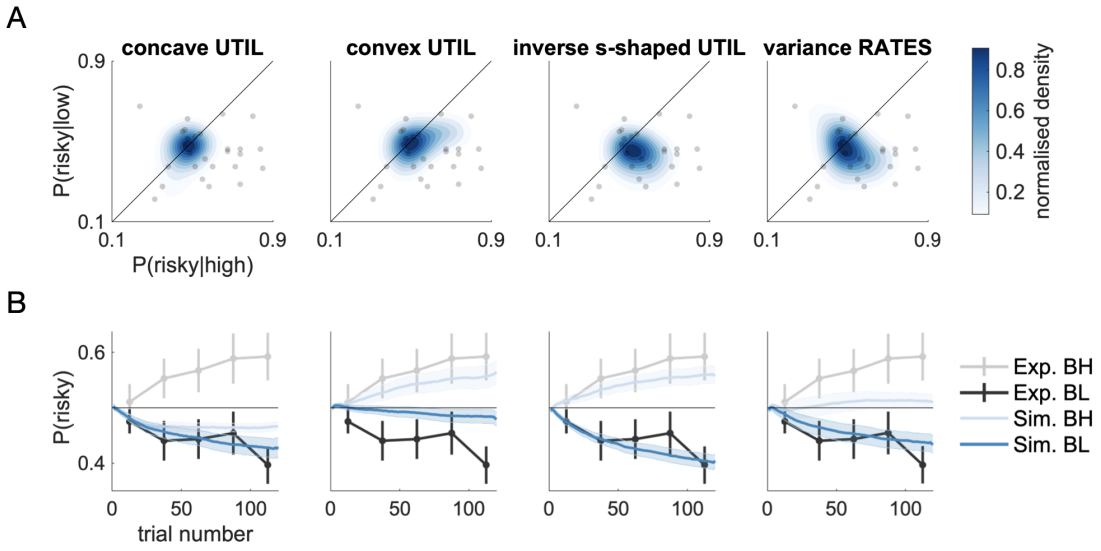

Figure 5.7: Emergence and distribution of risk preferences for secondary alternative explanations, part 1. The representation of data in panels A and B is analogous to the representation of data in panels A and B of figure 5.5. We show the same analysis as in figure 5.5, but for a different set of models: concave UTIL, convex UTIL, inverse s-shaped UTIL and variance RATES.

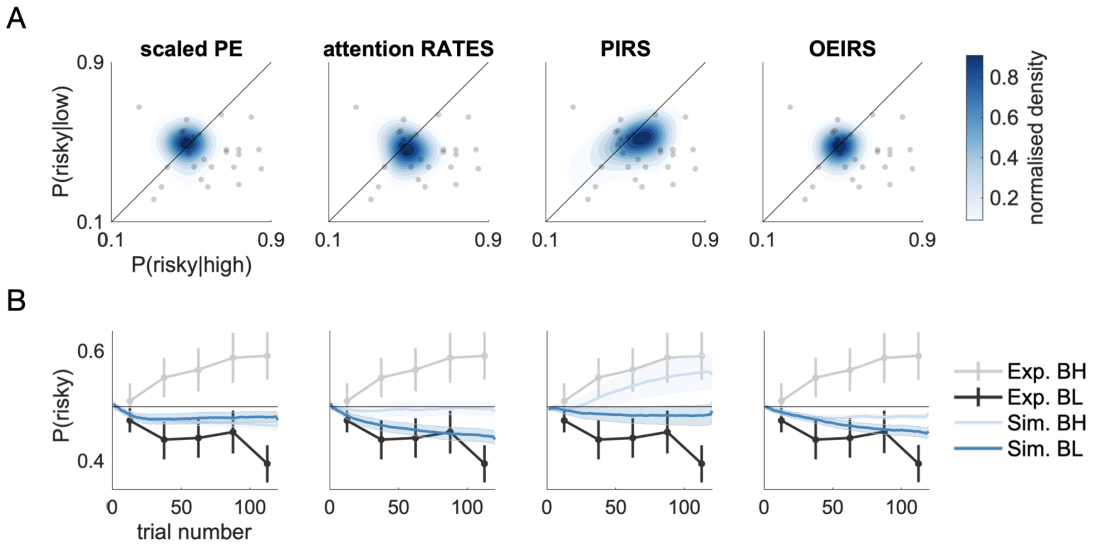

Figure 5.8: Emergence and distribution of risk preferences for secondary alternative explanations, part 2. The representation of data in panels A and B is analogous to the representation in panels A and B of figure 5.5. We show the same analysis as in figure 5.5, but for a different set of models: scaled PE, attention RATES, PIRS and OEIRS.

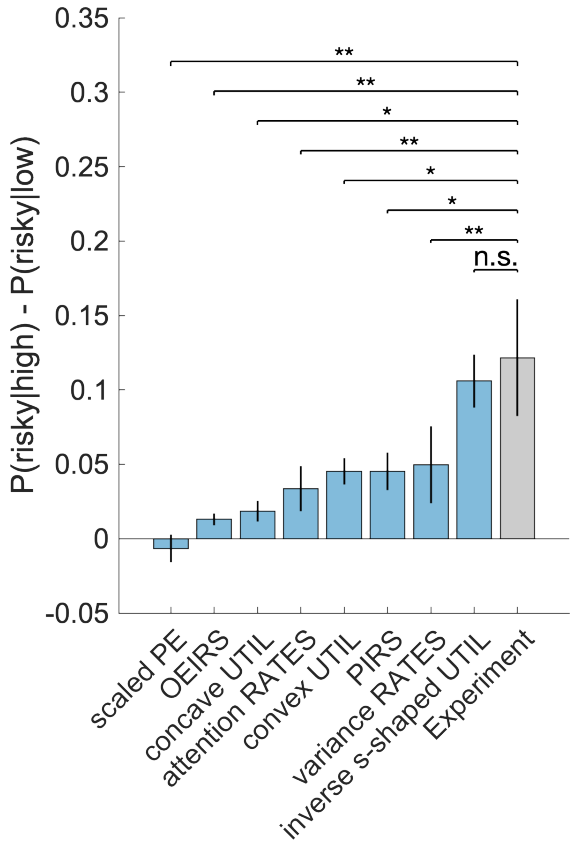

Figure 5.9: Differences in risk preferences between conditions in theory and experiment, for secondary alternative explanations. The representation of data here is completely analogous to the representation of data in figure 5.6. We show the same analysis as in figure 5.6, but for a different set of models: concave UTIL, convex UTIL, inverse s-shaped UTIL, variance RATES, scaled PE, attention RATES, PIRS and OEIRS. This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

5.3.3 Model fits

After showing that the PEIRS model can reproduce the experimental risk preferences and falsifying most alternative explanations, we now turn to using model comparison techniques to test more formally whether our dataset provides evidence for or against the PEIRS model. A model comparison can reveal which of several equally plausible explanations is the most appropriate.

We conduct a standard model comparison based on the Bayesian Information Criterion (BIC), which is a classical metric to assess how parsimoniously a model describes a dataset (Neath and Cavanaugh 2012). Our analysis is carried out on the population level: we assume that the entire population used a single one of

the candidate models, with individual sets of parameters. We then determine which candidate model is most likely to be the one that was used.

We validated this method by performing a model recovery analysis (figure 5.13). The recovery analysis suggests that for datasets like ours there is a slight risk of false negatives (i.e. attributing the dataset to RW when it was really generated from a more complicated model). The risk of false positives, on the other hand, is very low, especially for the PEIRS model (i.e., it is unlikely that we attribute a dataset to PEIRS when it was generated by another model).

The results of the model comparison suggest that the best description of our population is afforded by the pos-neg RATES model, very closely followed by the PEIRS model (see figure 5.10). Relative to the typical BIC differences in our model comparison, the difference between pos-neg RATES and PEIRS seems negligible—we may think of them as approximately equally parsimonious with respect to our dataset. The other models are worse to various degrees; the OEIRS model is the worst by a substantial margin. In particular, we can see a very clear difference in BIC between PEIRS and the inverse s-shaped UTIL model, which was the only model not falsified by the simulations.

5.3.4 Conclusions

A clear picture emerges from the results of the simulations and the model comparison: all models but PEIRS and inverse s-shaped UTIL were falsified in simulations. The model comparison suggests that out of these two, PEIRS describes our data best. This implies that the effects of interest—condition-specific risk preferences—are best explained by the PEIRS hypothesis, out of all twelve tested models. Recently, this finding has been replicated: using a very similar task, van Swieten et al. (2021) show that PEIRS describes their population better than RW.

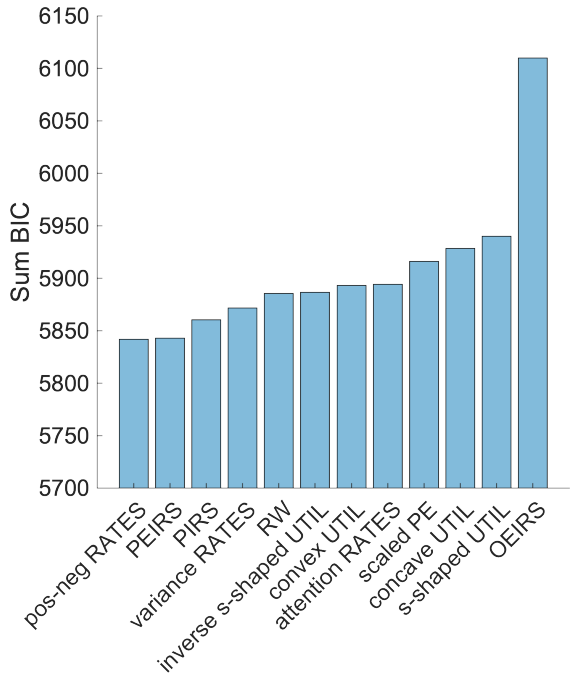

Figure 5.10: Pos-neg RATES and PEIRS dominate the overall model comparison. We show the BIC summed over all participants, for all models that we test. Lower BIC indicates a better fit. This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

5.4 Discussion

Two different behavioural phenomena—learning from errors and risk preferences—are attributed to the same neurotransmitter, dopamine. The common neural substrate led us to hypothesise that there might be an association between prediction errors and risk preferences on the level of behaviour. To investigate this, we first used the AU model to specify the mechanism that we propose, and to derive predictions from our hypothesis. We then tested our predictions using data from a task in which reward prediction errors were immediately followed by decisions that involved risk. We found that reward prediction errors and subsequent risk-taking are positively correlated: positive reward prediction errors are associated with risk-seeking, negative ones are associated with risk aversion. Finally, we used trial-by-trial modelling to compare the mechanism that we propose with various alternative explanations. A combination of simulations and model selection techniques revealed that out of all tested models, our hypothesis

is the most likely explanation for the effects that we observed.

Our results are consistent with our initial hypothesis: the two roles of dopamine (teaching signal and risk-modulator) interfere with each other. This study hence provides evidence against the conjecture that the roles of dopamine are well separated. Our conclusion fits in well with other findings: recent research has shown that phasic dopamine correlates with motivational variables (Hamid et al. 2016) and movement vigour (da Silva et al. 2018) just as well as with reward prediction errors. Together, these findings cast doubt on the separation into tonic and phasic (Niv, Daw, Joel, et al. 2007) and on separated dopamine roles in general.

In the rest of this chapter, we discuss several aspects of our findings, such as the biological plausibility of our hypothesis, its relation to other theories and further predictions.

5.4.1 Predictions and prediction errors

Our theory requires stimulus prediction errors to explain risk preferences. Generally speaking, positive stimulus prediction errors explain risk-seeking, while negative stimulus prediction errors explain risk-aversion. We indeed observe risk-seeking as well as risk-aversion (see figure 5.3C) and find that those effects are best explained by the PEIRS model, which features positive as well as negative prediction errors (see figure 5.5 and figure 5.6). While this explanation is consistent in itself and compatible with the data we collected, one might question its biological plausibility, because reward predicting stimuli are known to elicit dopamine bursts, but not dips. For example, Tobler et al. (2005) show increased dopamine activity as a response to all reward-predicting stimuli, even for stimuli that predict relatively small rewards.

There seems to be a contradiction between these results and our assumptions, but they are in fact compatible. To see this, one must consider the details of the trial structure: while our study had a fixed ITI, many classic studies (such

as the study of Tobler et al. 2005) have a variable, random ITI. This means that while the participants in our study could predict the time of stimulus onset perfectly, the onset was surprising for the subjects of Tobler et al. (2005). This is a crucial difference: if the time of stimulus presentation can be predicted, a reward expectation for that moment can be formed, and the value of the stimuli can be compared to this expectation. If the time of stimulus onset is unknown, then any stimulus that appears will first and foremost be compared to the possibility that nothing happens at that moment. The result of that comparison must always be positive, as the occurrence of reward predictors is always better than their absence.

This difference can be seen in neural recordings as well. For example, fixed ITIs were used in combination with stimuli that predicted different reward sizes in the study of Wang et al. (2021). The dopamine activity that is shown in figure 2E and figure 2F of that study clearly dips below the baseline level for stimuli that predict relatively small rewards. This suggests that the assumptions of the PEIRS model are biologically plausible for tasks with fixed ITIs, such as the one we are using in this study.

We nevertheless included a model based on reward predictions instead of reward prediction errors—the PIRS (Predictions Induce Risk Seeking) model. Reward predictions correspond to dopamine responses like those reported by Tobler et al. (2005). Simulations of that model are shown in figure 5.8 and figure 5.9. Both simulations and model comparison show clearly that PIRS does not explain the data well.

We may conclude that reward prediction errors, but not reward predictions, might induce risk-seeking in our task. We further conclude that the underlying mechanisms we assume are consistent with what is known about dopamine release in tasks with predictable timing.

Another question that might arise in this context is about the role of the outcome prediction error of the previous trial. According to our theory, the outcome prediction error should be broadcast by the dopamine system just like the stimulus prediction error, and might therefore also affect risk preferences. Of course there is a difference in timing: on average, choices follow stimulus presentations with a delay of 0.97 s, while the delay between outcome presentations and choices on the next trial is 3.47 s on average. We might thus expect that the impact of the outcome prediction error, if at all observable, might be much weaker than that of the stimulus prediction error: dopamine transients decay within hundreds of microseconds, as discussed in chapter 4, subsection 4.5.7. A computational analysis confirmed this: there is no evidence for an association between risk preference and the outcome prediction errors in the previous trial in our dataset (see figure 5.8 and figure 5.9).

5.4.2 Relation to behavioural economics

Decision making under uncertainty has been extensively studied in behavioural economics. One main finding in this field, known under the name *prospect theory*, is that humans tend to be risk-averse if decisions concern gains and risk-seeking if decisions concern losses (Kahneman and Tversky 2013). However, those classic findings rely on explicit knowledge of the probabilities involved in the decisions. Several recent studies indicate that risk preferences reverse when risks and probabilities are learned from experience: if learning is incremental and based on feedback, humans tend to make risky decisions about gains, and risk-averse decisions about losses (Wulff et al. 2018). This reversal has been termed the description-experience gap and is considered a "major challenge" for neuroeconomics (Garcia et al. 2021). In cognitive neuroscience and psychology, some studies have reproduced this phenomenon (Madan et al. 2014), while others report risk-aversion in the gain domain (Niv, Edlund, et al. 2012).

Our task differs somewhat from the tasks studied in the description-experience gap literature since we only use gains (our reward points are always positive).

However, humans often evaluate outcomes with respect to a reference point (Kahneman and Tversky 2013). In the context of our task, it seems plausible that rewards are evaluated relative to the average of previous rewards, or relative to the middle of the experienced reward range, which rapidly converges to 50 points within the first few trials. Rewards under 50 (and hence the majority of the rewards that result from choosing stimuli 3 and 4) would then be considered losses. Decisions between the two low-valued stimuli (*both-low* condition) would fall in the loss domain. From this perspective, our results are in line with the description-experience-gap and differ from those of Niv, Edlund, et al. (2012). The difference might be due to the degree of implicitness of the knowledge that is gained during the task: Niv, Edlund, et al. (2012) used bimodal reward distributions (e.g., 40 points with probability 50 %, 0 points otherwise) which participants might have been able to recognise as such after a few trials. Here, we used high-entropy reward distributions (normal distributions, see figure 5.1B), which could not be mapped onto bimodal gambles, and thus made anything but implicit learning intractable.

5.4.3 Relation to memory models

Could memory effects explain our results? Behaviour in a similar task was interpreted as the result of memory replay (Madan et al. 2014): experiences such as "obtained reward X after choosing option Y" might not only be used for immediate value updates, but might also be stored in a memory buffer. This buffer can then be used for offline learning from past experiences in times of inactivity, such as during the inter-trial interval. Madan et al. (2014) proposed that experiences are more likely to enter the buffer if they are extreme. If entering the buffer is biased in this way, then so are the values learned from replaying those experiences. In our task, "extreme" might mean that the reward was extremely high or low. The corresponding bias would drive choice towards the stimuli that produce the highest rewards, and away from those that produce the lowest, and thereby lead to a pattern similar to the one we observed.

Even though it is not feasible to represent the memory buffer model in our modelling framework, the inverse s-shaped UTIL model captures the idea of overweighting of extreme experiences. The simulations of that model (figure 5.7 and figure 5.9) show that such overweighting can indeed reproduce the risk preferences that we observed, which ties in well with the results of Madan et al. (2014). However, our model selection procedure (figure 5.10) suggests that PEIRS still explains the data better. One big difference between the two explanations is that risk preferences can flexibly appear and disappear in PEIRS. In contrast, the memory buffer theory (and equivalently the inverse s-shaped UTIL model) attribute them to distortions in the learned values and hence predict more persistent, less flexible preferences.

Another potentially relevant phenomenon based on memory effects was reported recently: Rouhani et al. (2018) show that both reward-tracking itself, as well as episodic memory, are enhanced in high-risk environments. In our context, this means that learning might be boosted for risky stimuli. We included two models that could capture such effects: the variance RATES model featured different learning rates for risky and safe stimuli, and in the attention RATES model, surprises could boost learning in all conditions. Neither of the two models could explain the effects we observed, suggesting that the mechanisms they describe are distinct from the mechanisms that drive risk preferences in our experiment.

5.4.4 Relation to utility models

Several of the models we have tested are based on the nonlinear utility of reward points. The central idea of these models is that participants might not find an outcome of 100 points twice as rewarding as an outcome of 50 points. If the perception of reward is thus distorted, risk preferences might arise as a consequence (Kahneman and Tversky 2013).

We found that none of the tested utility models could explain our effects better than PEIRS. However, one of them (inverse s-shaped UTIL) at least reproduced

the trial-by-trial emergence of risk preferences (figure 5.7B). How should this result and the other utility-related results be interpreted? We see two issues with utility models.

The first issue relates to the level of explanation: utility functions can be used to capture behavioural effects, i.e., they can provide a compact description of certain aspects of behaviour (such as risk-seeking). Such models might well be used to make predictions about behaviour. What they cannot provide us with, however, is an explanation on the level of neural processes—this distinction was phrased as *aggregate* versus *mechanistic* by Palminteri et al. (2017). In fact, it might well be that for a given neural process, one may find a utility function that can compactly summarise the effects it has on the level of behaviour. These two descriptions then concern different levels of explanation, and comparing them might not be meaningful (Palminteri et al. 2017). In this study, we derive a mechanistic model (PEIRS) and show that it is the best explanation of the effects of interest. Our level of explanation is thus the mechanistic level.

The second issue is generalisability and affects convex and inverse s-shaped utility functions in particular. Concave and s-shaped utilities are well documented and embedded in broader theories (expected utility theory and prospect theory, respectively). The effects that they describe appear in many different situations—they seem to capture a fairly general aspect of behaviour (which perhaps relates to a general mechanism in the brain). In contrast, convex and inverse s-shaped utility functions only describe behaviour in some very specific tasks—see Stauffer et al. (2015) for behaviour that is well described by convex utility—but fail to generalise to others. In these cases, one might be over-fitting the concept of utility, at the price of specificity. At the extreme, it is conceivable that one could explain most phenomena as an effect of nonlinear utility but would require a specific (and perhaps quite non-trivial) utility function for each case. From the standpoint of generalisability, it thus seems appropriate to test established theories such as

prospect theory as serious alternative explanations. Utility functions tailored to the behavioural effect in question seem problematic.

Overall, we find that neither empirical results nor general epistemological considerations indicate that much emphasis should be put on utility models in the context of our task and our goal (to understand the neural mechanisms that cause risk preferences).

5.4.5 Further experimental predictions

In this chapter we presented a theory based on neural effects and tested its behavioural predictions. The predictions were confirmed; alternative explanations were ruled out using a trial-by-trial modelling approach. What should be the next steps? Other, more elaborate tests on the level of behaviour are possible. For example, one might extend the set of stimuli. Then, one could indicate a subset of stimuli before each trial begins, and draw the options from that subset. The composition of the primed subset should affect reward expectations (now an average over the values of the stimuli in the subset, not over all stimuli) and hence stimulus prediction errors. With such manipulations, one might be able to induce opposite risk preferences in the same choice, depending which subset was primed.

Although this would be a powerful confirmation of our findings, no behavioural experiment can prove that our theory is correct on the level of neural mechanisms. More direct measurements are needed to establish this. To test our theory on the neural level, correlational studies are one possibility. The neural signal to be measured should have sufficient time resolution to differentiate the prediction errors at different times during the trial. Options are EEG, electrophysiology, voltammetry or similar (FMRI might be too slow). With those techniques, one could gain direct measurements of the prediction errors and hence resolve clearly how they are associated with subsequent risk-seeking.

Causal designs are another possibility. For example, one could use optogenetic tools to elicit or suppress dopamine transients just before choices that involve risk, for instance by stimulating or inhibiting VTA dopamine neurons at the time of stimulus onset. If our theory is correct, such manipulations should be able to increase or decrease the behavioural effects we showed.

5.4.6 Conclusions

In this chapter, we demonstrated that a variant of the AU model predicts an association between prediction errors and risk-seeking. This is based on dopamine's dual role in learning and action selection. We present behavioural data that confirms this prediction and use modelling to falsify many possible alternative explanations. Our results suggest that our explanation is the most plausible. Overall, it appears as if reward prediction errors might induce risk seeking in humans.

5.5 Methods

5.5.1 Learning performance

The dataset we used in this chapter was collected by one of the coauthors of Möller, Grohn, et al. (2021) (JG). It contains data from 30 participants. Our results are based on 27 of the 30 participants. Three participants were excluded from the analysis due to their failure to understand the task. We evaluated the participants' understanding of the task by scoring their preferences in the *different* condition. These scores are shown in figure 5.11.

To be included in our analysis, participants had to choose the high-value option over the low-value option in at least 65 % of the relevant trials across the experiment. This threshold is chosen to separate the two clusters of participants in the cohort (as can be seen in figure 5.11, 3 subjects perform substantially worse than the other 27).

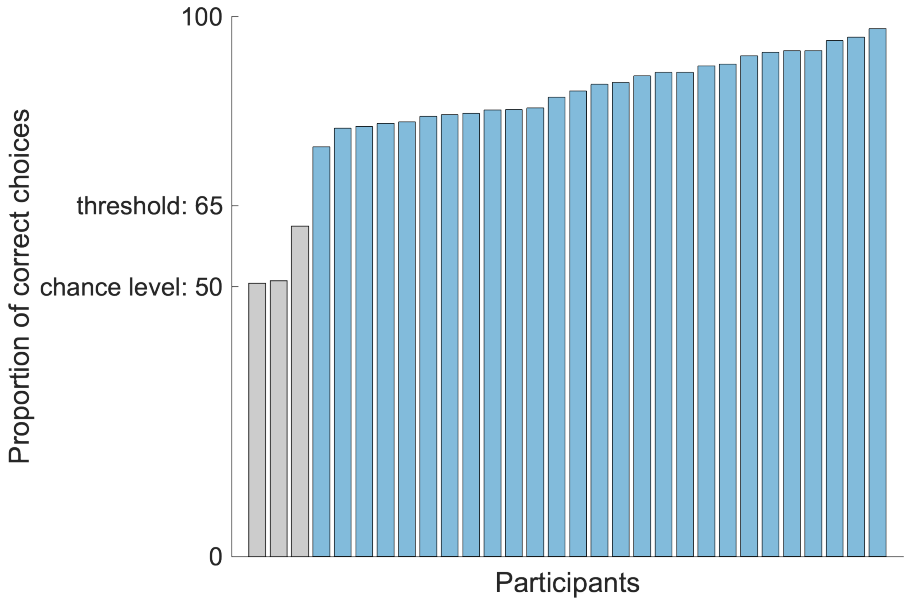

Figure 5.11: Performance. The bars indicate the participant's average performance (the proportion of choosing a high-mean stimulus over a low-mean stimulus in the *different* condition) across all four blocks. Participants with performances below 65 % were excluded from our analyses (grey bars). Participants with performances over 65 % were included in our analyses (blue bars). This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

5.5.2 Emerging preferences

To test whether value preferences emerged faster than risk preferences, we used a linear mixed-effects model for choices. As predictors, we included fixed effects of decision type, trial number, and their interaction. Here, decision type is defined as value decision (*different* condition) versus risk decision (*both-high* and *both-low* conditions). We also included random effects for all predictors and a random intercept, by participant. A likelihood ratio test for the fixed effect of the interaction between trial number and decision type reveals a significant positive effect ($p < 0.001$; based on MATLAB's *compare* function).

5.5.3 Model definitions

For all models, $i \in \{1, 2, 3, 4\}$ is the stimulus index, and Q_i is the value of stimulus i. The index j is used for the options on the screen in each trial (for example, $j \in \{1, 3\}$ if stimuli 1 and 3 are shown). The initial values at the beginning of each block are denoted by Q_0. The update rules are only applied to the values

and spreads of the chosen stimuli; the values and spreads of unchosen stimuli do not change. Rewards are denoted by r.

RW In the RW model, learning is driven by differences δ_{out} between the reward expected and received:

$$\delta_{\text{out}} = r - Q_{\text{chosen}} \tag{5.16}$$

$$\Delta Q_{\text{chosen}} = \alpha \delta_{\text{out}} \tag{5.17}$$

$$\tag{5.18}$$

Here, α denotes the learning rate. Learned values were linked to choice probabilities via a standard softmax rule (Daw et al. 2011):

$$P(i) = \frac{\exp(\beta Q_i)}{\sum_j \exp(\beta Q_j)} \tag{5.19}$$

with $P(i)$ the probability that option i is chosen. The RW model has free parameters $\alpha \in [0, 1]$, $\beta > 0$ and a fixed parameter, $Q_0 = 50$.

Concave utility To allow for concave subjective utility of reward points in our experiment, we used an exponential family of functions (Guyaguler, Horne, et al. 2001), which we adapted to our reward range through appropriate scaling (σ) and shifting (μ):

$$z = \frac{r - \mu}{\sigma} \tag{5.20}$$

$$U = \mu + \sigma \times (1 - \exp(-k_{\text{concave}} z))/k_{\text{concave}} \tag{5.21}$$

The utility of reward points was used to compute the prediction errors:

$$\delta_{\text{out}} = U - Q_{\text{chosen}} \tag{5.22}$$

Updates were computed with the RW rule (equation 5.18), choices were modelled using the softmax rule (equation 5.19). The concave UTIL model has free parameters $\alpha \in [0, 1]$, $\beta > 0$, $k_{\text{concave}} > 0$ and fixed parameters $Q_0 = 50$, $\sigma = 50$, $m = 50$.

Convex utility We also included a model that allows for convex utility functions. Although those are not mentioned as often, there is non-human primate evidence supporting them (Stauffer et al. 2015). The convex utility model is identical to the concave utility model (equation 5.21 and 5.22), with only one difference: the parameter k_{concave} is replaced by a new parameter k_{convex}, for which $k_{\text{convex}} < 0$ holds.

S-shaped utility Utilities can also be s-shaped, with different signs of curvature on both sides of a reference point (Kahneman and Tversky 2013)—convex for values below the reference point, and concave for values above. We modelled s-shaped utility functions using sign-preserving power functions (Spitzer et al. 2017), which we adapted to our reward range:

$$z = \frac{r - \mu}{\sigma} \tag{5.23}$$

$$U = \mu + \sigma \times \text{sign}(z)|z|^{k_{\text{s-shaped}}} \tag{5.24}$$

Prediction errors were calculated as in equation 5.22, updates happened according to the RW rule (equation 5.18), choices were modelled using the softmax rule (equation 5.19). The s-shaped UTIL model has free parameters $\alpha \in [0, 1]$, $\beta > 0$, $k_{\text{s-shaped}} \in [0, 1]$ and fixed parameters $Q_0 = 50$, $\sigma = 50$, $m = 50$.

Inverse s-shaped utility The last utility function needed to complete our set is inverse s-shaped. Such utility functions are not usually considered in neuroeconomics but have been used to study the perception of numerals (Spitzer et al. 2017). The inverse s-shaped utility model is identical to the s-shaped utility model, with the exception of the parameter $k_{\text{s-shaped}}$, which is substituted by $k_{\text{inverse s-shaped}} > 1$.

Different learning rates for positive and negative prediction errors It has been shown that learning from positive outcomes can differ from learning from negative outcomes (Gershman 2015). We modelled this by letting the learning rate depend on the sign of the prediction error (computed according to equation 5.17). The update rules then become:

$$\alpha(\delta) = \begin{cases} \alpha_+ & \text{if } \delta > 0 \\ \alpha_- & \text{if } \delta < 0 \end{cases} \tag{5.25}$$

$$\Delta Q_{\text{chosen}} = \alpha(\delta_{\text{out}})\delta_{\text{out}} \tag{5.26}$$

$$\tag{5.27}$$

Choices were modelled using the softmax rule (equation 5.19). The pos-neg RATES model has free parameters $\alpha_+ \in [0,1]$, $\alpha_- \in [0,1]$, $\beta > 0$, and a fixed parameter, $Q_0 = 50$.

Different learning rates for noisy and safe stimuli Similarly, it has been shown that the statistics of the reward distribution can have an effect on the learning rate (Daw et al. 2006; Behrens et al. 2007). In our task, normative theory suggests that the learning rate should depend on the variance of the signal that is being learned (Welch, Bishop, et al. 1995). We modelled this by allowing different learning rates for risky and safe stimuli. The update equations become:

$$\alpha_i = \begin{cases} \alpha_{\text{risky}} & \text{if } i \in \{1,3\} \\ \alpha_{\text{safe}} & \text{if } i \in 2,4 \end{cases} \tag{5.28}$$

$$\Delta Q_{\text{chosen}} = \alpha_{\text{chosen}} \delta_{\text{out}} \tag{5.29}$$

$$\tag{5.30}$$

Prediction errors were computed according to equation 5.17, choices were modelled using the softmax rule (equation 5.19). The variance RATES model has free parameters $\alpha_{\text{risky}} \in [0,1]$, $\alpha_{\text{safe}} \in [0,1]$, $\beta > 0$, and a fixed parameter, $Q_0 = 50$.

PEIRS This model encodes our central hypothesis: that prediction errors induce risk-seeking. It is based on the AU model; the mathematical description is given in 5.1.

The PEIRS model has free parameters $\alpha_Q \in [0,1]$, $\alpha_S \in [0,1]$, $\beta > 0$, ω, $S_0 > 0$ and a fixed parameter, $Q_0 = 50$. The initial spread estimate S_0 is allowed to vary since participants were not given any prior information about the spread magnitude.

PIRS PIRS, a variant of the PEIRS model, stands for *Predictions Induce Risk Seeking*. PIRS differs from PEIRS only in how the stimulus prediction error is computed. PIRS uses

$$\delta_{\text{stim}} = \frac{1}{2}(Q_{\text{option 1}} + Q_{\text{option 2}}) \tag{5.31}$$

The above formula is used instead of equation 5.8. Otherwise, PIRS is identical with PEIRS.

OEIRS If risk preferences are associated with stimulus prediction errors, they might also be associated with outcome prediction errors in the previous trial. This is the assumption of the OEIRS model, which differs from the PEIRS model in that the outcome prediction error δ_{out} of the previous trial (computed according

to equation 5.17) is substituted for δ_{stim}. Otherwise, OEIRS is identical to PEIRS and PIRS.

Scaled prediction errors In chapter 4 we developed a model that features scaled prediction errors. There, we discussed several studies that indicated that dopamine signalling (and hence reward prediction errors) might adapt to the magnitude of reward fluctuations (Tobler et al. 2005; Diederen and Schultz 2015; Diederen, Spencer, et al. 2016; Park et al. 2012; Rothenhoefer et al. 2021). Our own theory in chapter 4, as well as the literature, explain this adaptation by a scaling mechanism that scales reward prediction errors by the perceived reward noise, which is tracked alongside the average reward. Such scaling might affect stimulus-specific learning speeds and could hence affect risk preferences.

Here, we modelled the scaled prediction error hypothesis by implementing scaled prediction errors on top of the AU model. The model is then defined as

$$\tilde{\delta}_{\text{out}} = \frac{r - Q_{\text{chosen}}}{S_{\text{chosen}}} \tag{5.32}$$

$$\Delta Q_{\text{chosen}} = \alpha_Q \times \tilde{\delta}_{\text{out}} \tag{5.33}$$

$$\Delta S_{\text{chosen}} = \alpha_S \times \left(\left|\tilde{\delta}_{\text{out}}\right| - 1\right) \tag{5.34}$$

Choices are modelled using the softmax rule (equation 5.19). The scaled prediction error model has free parameters $\alpha_Q \in [0, 1]$, $\alpha_S \in [0, 1]$, $S_0 > 0$, $\beta > 0$, and a fixed parameter, $Q_0 = 50$. This model differs from the model in chapter 4, in that prediction errors are scaled by reward spread instead of reward standard deviation. We choose this design to make the scaled prediction error model more compatible with the PEIRS framework. However, for normal distributions the mean reward spread is proportional to the reward standard deviation (Mikhael and Bogacz 2016), so there should be no substantial difference between the model we use here and the model we developed in chapter 4.

Attention model Finally, we consider the possibility that the effects we observe might be caused by attention mechanisms—surprising outcomes (i.e., rewards that cause a high absolute prediction error) might cause subjects to be more attentive to the outcome, and memorise it more thoroughly. We modelled this as an RW learner with an additional surprise-related factor that gates learning:

$$\Delta Q_{\text{chosen}} = \alpha \times |\delta_{\text{out}}|^{k_{\text{attention}}} \times \delta_{\text{out}} \tag{5.35}$$

Choices were modelled using the softmax rule (equation 5.19). The attention model has free parameters $\alpha \in [0, 1]$, $k_{\text{attention}}$, $\beta > 0$, and a fixed parameter, $Q_0 = 50$.

5.5.4 Parameter transformations and priors

We used exponential and sigmoid transformations to constrain the parameters to their appropriate ranges. Priors were specified as multivariate normal distributions over the untransformed parameters. All but diagonal elements of the covariance matrices of those normal distributions were set to zero. Hence, the prior distributions could be factorised into univariate normal distributions (one for each parameter). Here we provide the statistics of those prior distributions. For parameters that occur in more than one model (such as learning rate α), we used the same priors across models. The notation $X \sim \mathcal{N}(\mu, \sigma^2)$ means that X has a normal distribution with mean μ and variance σ^2.

$$\text{logit}(\alpha) \sim \mathcal{N}(-1, 2) \tag{5.36}$$

$$\log(\beta) \sim \mathcal{N}(-2, 2) \tag{5.37}$$

$$\log(k_{\text{concave}}) \sim \mathcal{N}(-3, 4) \tag{5.38}$$

$$\log(-k_{\text{convex}}) \sim \mathcal{N}(-3, 4) \tag{5.39}$$

$$\text{logit}(k_{\text{s-shaped}}) \sim \mathcal{N}(3, 4) \tag{5.40}$$

$$\text{logit}(k_{\text{inverse s-shaped}} - 1) \sim \mathcal{N}(-3, 4) \tag{5.41}$$

$$\text{logit}(\alpha_+) \sim \mathcal{N}(-1, 2) \tag{5.42}$$

$$\text{logit}(\alpha_-) \sim \mathcal{N}(-1, 2) \tag{5.43}$$

$$\text{logit}(\alpha_{\text{risky}}) \sim \mathcal{N}(-1, 2) \tag{5.44}$$

$$\text{logit}(\alpha_{\text{safe}}) \sim \mathcal{N}(-1, 2) \tag{5.45}$$

$$\text{logit}(\alpha_S) \sim \mathcal{N}(-1, 2) \tag{5.46}$$

$$\omega \sim \mathcal{N}(0, 20) \tag{5.47}$$

$$\log(S_0) \sim \mathcal{N}(2, 2) \tag{5.48}$$

$$\tag{5.49}$$

5.5.5 Fitting and simulation

Fits were performed using the VBA toolbox (Daunizeau et al. 2014). This toolbox implements a variational Bayes scheme. It takes a set of measurements, a generative probabilistic model that describes how the measurements arise (which usually contains some latent, i.e., unobserved, variables) and prior distributions over the model parameters as input, and outputs among other things an approximate posterior distribution over the model parameters, an approximate posterior distribution over the latent variables, and fit statistics such as the BIC. We estimated parameters and latent variables (such as values and prediction errors) using the mean of the posterior distributions from the toolbox outputs.

Simulated risk preferences were obtained using parameters taken from fits: for each model, a fit provided us with 27 parameter sets corresponding to the 27

participants. Using those parameter sets, we simulated our task and generated datasets of the same size as the dataset obtained in our experiment. We repeated this 1000 times to obtain stable distributions.

We then extracted 27000 pairs of aggregate risk preferences and risk traces (risk preferences as a function of trial number) per model from the simulated datasets (27 simulated participants in each of 1000 simulated experiments).

For figure 5.5A, 5.7A and 5.8A, we used a kernel smoothing function to estimate the probability density corresponding to the distribution of risk preferences (MATLAB's *ksdensity* function). We then computed and visualised those estimated probability densities using MATLAB's *contour* function.

For figure 5.5B, 5.7B and 5.8B, we first split the risk preferences between conditions. We then averaged the traces over repetitions of the experiment for each simulated participant, and finally over participants. For display, the averaged simulated traces were smoothed with a 20-point moving average filter.

For figures 5.6 and 5.9, we first computed the average difference in risk preference between conditions for each simulated participant in each simulated experiment, yielding 27000 differences in risk preference. This was done for each model. We then averaged those across experiments, obtaining 27 differences in risk preferences per model. In figures 5.6 and 5.9, we compare those distributions to the empirical distribution of the mean difference in risk preference between conditions, which has 27 data points as well (one for each participant).

5.5.6 Likelihoods of risk preference

To quantify how well the simulated distributions (blue shadings in figure 5.5A, 5.7A and 5.8A) match the empirical distribution (grey dots in figure 5.5A, 5.7A and 5.8A), we used MATLAB's kernel density estimation (*ksdensity*) to interpolate the simulated distributions and hence obtain likelihoods for the empirical risk preferences. In other words, we used the simulated distributions to estimate how likely

the empirical risk preferences were under each model. We computed the log-likelihood ratios ℓ relative to PEIRS,

$$\ell(m, \text{PEIRS}) = \sum_{\text{participant } i} \log(P(r_i|m)) - \log(P(r_i|\text{PEIRS})), \tag{5.50}$$

for all models. Here, m means model and r_i denotes the risk preferences of participant i. Figure 5.12 shows that the empirical risk preferences are most likely under PEIRS, and least likely under s-shaped UTIL.

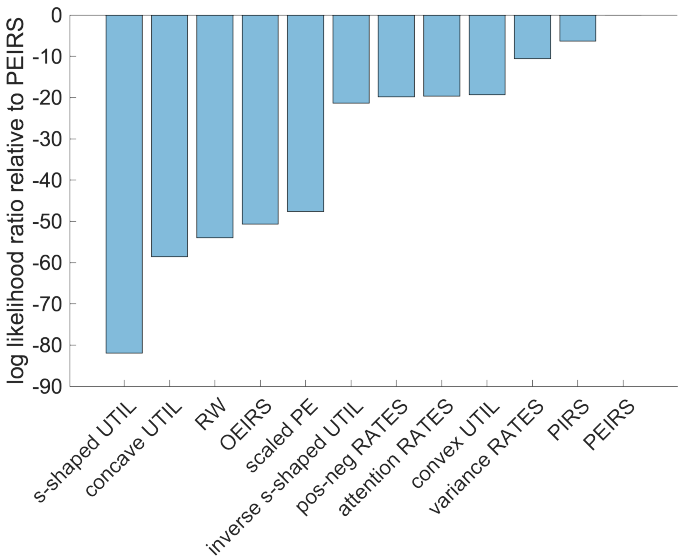

Figure 5.12: Model likelihoods of risk preferences. The blue bars indicate the log likelihood ratio between the model on the x-axis and the PEIRS model, with respect to the empirical risk preferences. The more the bar extends into the negative numbers, the more evidence there is for PEIRS relative to the model associated with the bar. This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

5.5.7 Model recovery

To validate our model selection procedures, we performed a model recovery analysis (Palminteri et al. 2017). This analysis was meant to test whether our candidate models are identifiable in the empirical parameter range. We hence performed an a-posteriori model recovery analyses: first, models were fitted to the empirical dataset, which resulted in posteriors over the parameter space.

Parameters were then sampled from those posterior distributions, and simulations were performed with the sampled parameters. Finally, all models were fitted to all simulated datasets, and model selection metrics (BIC) were extracted. This resulted in a confusion matrix. We repeated this procedure 100 times for each of our 27 participants (a total of 2700 repetitions), to account for stochastic fluctuations. The averaged confusion matrix is shown in figure 5.13.

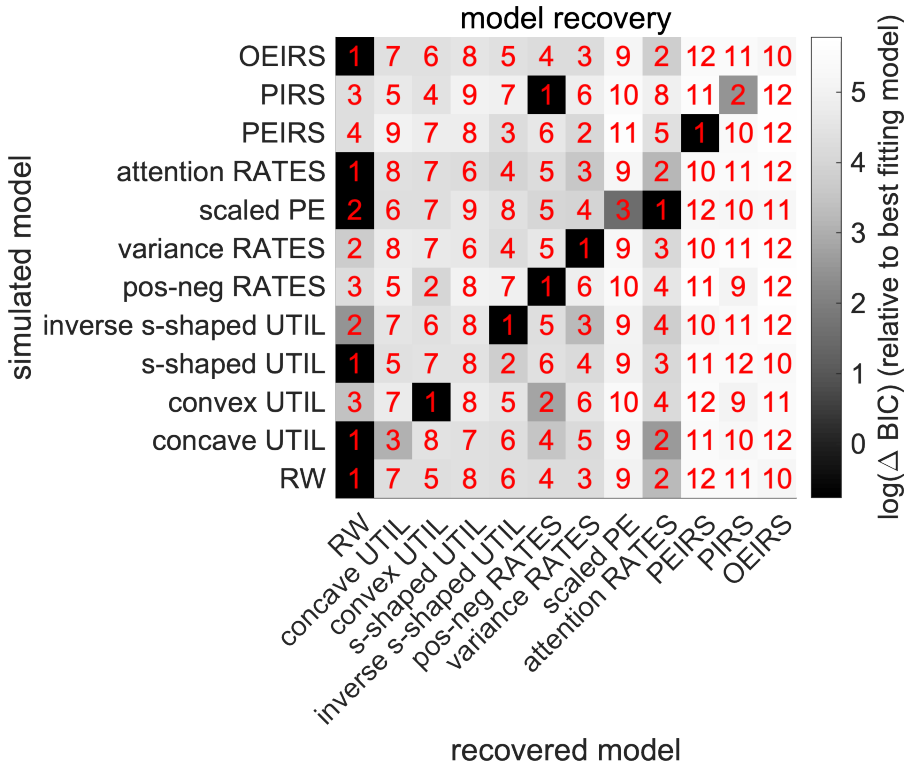

Figure 5.13: Model recovery confusion matrix. Each row corresponds to a model used to simulate datasets. Each column corresponds to a model fitted to datasets. The shading of the matrix elements represents the average BIC relative to the best fitting model (BICs were baselined per row). Higher BICs correspond to lighter colours and poorer fits. The colour scale is logarithmic to enhance contrasts. The red numbers indicate the rank of the model in the model comparison per row (1 meaning best-fitting model, 12 meaning worst fitting model). This figure was taken from Möller, Grohn, et al. (2021) and adapted for this thesis.

The confusion matrix suggests that 6 out of 12 models (RW, convex UTIL, inverse s-shaped UTIL, pos-neg RATES, variance RATES and PEIRS) can be directly identified by our procedure. For another 5 out of 12 models (Concave UTIL, s-shaped UTIL, scaled PE, attention RATES, OEIRS), RW was selected as the best model to describe the datasets they generated. Given that all those models

contain RW as a special case, this means that the additional effects they describe just do not occur in our dataset—the empirical parameter range is such that the models mimic RW and make no use of the additional features they can describe. The model comparison still penalises the additional complexity, leading to RW as the selected model. The only model that was confused with a model other than RW was PIRS, which was confused with pos-neg RATES. This means that for all models but PIRS, our model comparison will either select the correct model if an effect is there or the base model (RW) if the effect does not appear. Importantly, we find that in the empirical parameter range, PEIRS was selected unambiguously. This suggests that PEIRS being selected in the model comparison on the empirical dataset is a meaningful result.

6
Discussion

Contents

In this thesis, the usual contents of a discussion section—the relationships between new models and existing ones, the relationships between new models and experimental data, experimental predictions and directions for future work—were fairly specific to each chapter. We have therefore included them directly in the chapters themselves, instead of collecting them at the end. In this general discussion, we want to assume a broader perspective.

The models and theories we developed and tested above are all based on a certain framework used to conceptualise the basal ganglia and their function. It is the framework of reward-based decision making. The basal ganglia are thought of as a system that selects actions in response to environmental cues and adapts to

feedback, maximising an abstract quantity called reward. Striatal synapses, the dopaminergic innervation of the striatum and dopamine-dependent plasticity become synonymous to policy and value networks, reward prediction error signals and learning rules. There is sound evidence that supports this view—we have reviewed some of it in chapter 2.

But is that an exhaustive picture of dopamine and the basal ganglia? Is it the only possible perspective? This is not the case—there are alternative frameworks (albeit not as developed theoretically) through which dopamine and the basal ganglia can be understood. Here, we want to explore two of these perspectives to get a sense for directions beyond RL—this general discussion is about alternatives to the current paradigm. Our first subject will be the invigoration pathway, our second an evolutionary perspective on dopamine.

6.1 Do the basal ganglia make decisions?

Let us start by challenging an idea that we built on in all previous chapters: that the basal ganglia are the (or at least a) brain circuit that performs action selection. Simply put, we assumed that the signals that enter the basal ganglia represent cues, stimuli or the state of the environment and that the signals that come out of the basal ganglia represent one out of several possible actions—the one that the circuit computed to be the most appropriate. For example, inputs might be pictures of colourful shapes (as in chapter 5), while outputs might be high-level motor commands such as "press the left button on a controller".

Is this view realistic? Does the output of the basal ganglia encode the chosen action, and cause that action to be executed? Recent evidence challenges this view: in monkeys performing a decision-making task with speed-accuracy trade-off, it was found that the basal ganglia (in particular the output nucleus GPi) reflected information about the selected option later and less clearly than other brain regions, such as premotor and motor cortices (Thura and Cisek 2017). This contradicts the idea that the choice is made at the level of the GPi and only

then signalled to areas relevant to execution. Similar conclusions were drawn elsewhere (Turner and Anderson 1997; Turner and Desmurget 2010): based on the timing of action-specific activity in the GPi, the basal ganglia do not seem to contribute to the selection or planning of movements—the corresponding signals occur too late.

Furthermore, it was shown that a unilateral lesion of the GPi did not impair the execution of learned behaviours (Obeso et al. 2009). Together, these results indicate that at least in some cases, action or movement selection does not require or even include the basal ganglia. But what does the circuit do then? If not by mapping environmental cues to actions, how do the basal ganglia contribute to the generation of behaviour?

6.2 The Invigoration pathway

If we ignore learning for now, the literature is surprisingly consistent with respect to how the basal ganglia affect behaviour: many studies conclude that the system modulates the vigour of ongoing actions, leading to acceleration or deceleration of movements. These effects of the basal ganglia are comparable to those of the gas and brake pedals in a car—we shall call this the invigoration hypothesis. The corresponding invigoration pathway includes the dopaminergic system as well as the basal ganglia circuit. We will discuss both in turn.

6.2.1 Invigoration and dopamine

First, let us have a fresh look at dopamine. The reward prediction error hypothesis tells a very compelling story about the properties and functions of dopamine signals, but in reality, things are not as clear-cut: it was demonstrated recently that dopamine is associated with many more variables than just reward (Engelhard et al. 2019). That study shows the complexity of the information represented in the dopamine signal; the picture that emerges is a complicated one.

Intriguingly, this complexity dissolves when one focuses solely on the effects of dopamine, instead of the information that it carries. So what are the effects of dopamine? Several studies consistently found that dopamine release invigorates movements. First, it was shown that SNc activity precedes acceleration and triggers locomotion in mice on a treadmill (Howe and Dombeck 2016). The same was shown in an open field setting (da Silva et al. 2018) and replicated elsewhere (Dodson et al. 2016). The most detailed study yet used a custom device to measure head movements with unprecedented precision (Hughes et al. 2020). They found surprisingly tight correlations (R^2 values of around 0.95) between the activity of VTA populations and components of the impulse vector of head movements towards or away from rewards. Using optogenetics, it could even be shown that these movements and the activity of VTA populations were causally connected.

These novel results are consistent with older findings relating to Parkinson's disease. For example, it is known that in contrast to healthy controls, PD patients do not move faster for reward (Mazzoni et al. 2007). This could be explained by the absence of a dopaminergic gas pedal under the control of a goal-direct system.

Overall, it appears as if the effect of dopamine—the invigoration of movements—is much clearer cut than dopamine's representational content. The gas pedal metaphor helps to understand why: there may be many reasons to press a gas pedal (and hence a lot of variables represented in its activity), but the effect of the pedal is always the same.

This perspective also sheds light on results that appear somewhat enigmatic under the conventional view. For example, a go/no-go task revealed that reward predicting cues were only followed by dopamine release in the rat NAc if movement was required to obtain the reward (Syed et al. 2016). If dopamine had been signalling prediction errors, it would have responded to reward-predicting no-go cues as well, but this was not the case. The invigoration hypothesis explains

this with ease: reward predicting cues will cause a press on the dopaminergic gas pedal if and only if movement is needed to get the reward.

In summary, we find that the simple analogy between dopamine and a neural gas pedal is surprisingly effective in explaining past and recent findings. There are caveats to this (for example, learned sequences seem to require only an initial dopamine burst, but will continue without further dopamine activity, as shown by Howe and Dombeck 2016), but these go beyond the scope of this discussion.

6.2.2 Invigoration and the basal ganglia

After focusing on dopamine, we now turn to the next link in the invigoration pathway: the basal ganglia circuit. Two recent reviews by Turner and Desmurget (2010) and Dudman and Krakauer (2016) come to similar conclusions concerning the effect of basal ganglia output on behaviour: rather than selecting, planning or initiating movements, the basal ganglia seems to modulate or gate ongoing movements by controlling their vigour. Consistent with this, one finds that GPi activity represents kinematic movement variables with the appropriate timing for modulation, but not for selection. In addition, one finds that GPi lesions cause a decrease in movement speed, but no decrease in choice accuracy.

The invigoration viewpoint is further supported by striatal lesion data: stroke-caused unilateral focal lesions of the dorsal striatum in humans can cause micrographia (a strong reduction in the size of written letters, Shadmehr and Krakauer 2008). Lesions of the dorsal striatum have been shown to affect the speed of learned behaviours but not their content (Jurado-Parras et al. 2020). Overall, the recent literature seems to suggest that the basal ganglia's contribution to behaviour might be "implicit motivation expressed through movement vigour" (Dudman and Krakauer 2016).

The gas/brake-pedal view maps nicely onto the direct/indirect pathway architecture of the basal ganglia, and could also provide an explanation for the strong convergence within the circuit (for each neuron in the GPi there are about 14

neurons in the GPe and about 872 neurons in the striatum, Oorschot 1996). This reduction in signal complexity might simply reflect that a variety of context variables such as environmental cues, memories or internal states are used to compute a very low-dimensional quantity: the acceleration or deceleration of the current movement.

The invigoration hypothesis might also explain some challenges in the development of brain-machine interfaces for prosthetics: it is possible to create systems that use cortical signals to control a robotic hand, but to date this does not include graded force levels (Collinger et al. 2013; Hochberg et al. 2012). The difficulty to decode force levels from the cortical surface is consistent with them being computed and implemented in subcortical structures such as the basal ganglia. Indeed, some pioneering work in humans suggests that force levels can be decoded from the STN, which is a part of the basal ganglia network (Tan et al. 2016).

Finally, note that though the above discussion is centred around movement in physical space, the idea might be applicable more generally. The basal ganglia might also invigorate movements in more abstract, perhaps cognitive spaces—their structure determines how it processes information, but their connectivity to other brain regions determines what kind of information they processes and how the result of the computation is used. For example, one might think of decision making as a movement between the different options (think drift-diffusion model, DDM). If wired appropriately to the corresponding cortical circuits, the basal ganglia might speed up that movement or slow it down, instead of biasing it towards one option or the other, as is it is currently thought (Redgrave et al. 1999; Collins and Frank 2014). In particular, one might predict that enhanced activity in the indirect pathway should "put the brakes" on movements in decision space, perhaps causing effects akin to those of a raised DMM threshold. Phenomena of this kind have indeed been observed (Herz et al. 2016).

In summary, over the last two sections, a new picture emerged: it appears that the dopamine system and the basal ganglia circuit are part of a neural pathway dedicated to making ongoing movements quicker and more forceful, rather like a gas pedal in a car. This picture is quite different from that of the basal ganglia as a decision-making system: a gas pedal operates in continuous time and physical space, while decision making is typically modelled with discrete steps and symbolic action spaces. And while animals might use RL[1] to get better at handling their gas pedals, it seems that there might also be less complicated normative frameworks that could be applied to understand this type of control.

The reward-taxis model of Karin and Alon (2021) might be an example for such a framework. Reward-taxis is inspired by chemotaxis, which is a behavioural strategy of simple life forms such as E. Coli bacteria. Those bacteria move around in a random walk. Gradients in the concentration of desirable chemicals can bias this random motion: movements up the gradient are extended, movements down the gradient are decreased. Through this very simple control mechanism the organism is able to move to the spots with the highest concentration of the chemicals it requires.

Karin and Alon (2021) suggest that animals approach rewards in much the same way. In particular, they propose that animals perform random walks, and that dopamine biases those random walks by selectively invigorating movements towards rewards. In this theory, dopamine responds to reward gradients and hence causes reward-taxis. This is consistent with the invigoration hypothesis presented above (dopamine acts through invigoration), and it also explains why dopamine activity resembles reward prediction errors: sudden changes in reward expectation and steep reward gradients are closely related. It seems quite plausible that reward-taxis might be one function of the invigoration

[1] Yttri and Dudman (2016) might show exactly how learning to use the gas pedal works: using velocity-contingent optogentic stimulation targeted at either D1 or D2 populations in the striatum, they produced lasting increases or decreases in movement velocity.

pathway, especially in simple organisms; there might be many more waiting to be discovered.

6.3 The first role of dopamine

After challenging the idea that the basal ganglia are making decisions and describing the invigoration hypothesis, we now challenge the idea that dopamine is mainly a teaching signal. We have pointed out at various points in this thesis that dopamine has a modulator role linked to motivation, in addition to its involvement in reinforcement learning. Generally, this modulator role receives less attention than the reward prediction error role. This is especially true for modelling and theory work—the readily available computational framework of reinforcement learning dominates the literature; it is customary to try to understand new phenomena first in the context of reinforcement learning (for an example, see the RL treatment of dopamine ramps by Gershman 2014). As a consequence, the motivational aspects of dopamine are sometimes viewed just as a corollary of appropriately tailored RL models (e.g. Niv, Daw, Joel, et al. 2007).

In this last part of this thesis, we want to explore the opposite view. Could dopamine first and foremost be a modulator? The nature of the modulation could be such that it makes sense to adapt to it through plasticity, as a secondary effect. Learning could be the corollary, modulation and direct control of behaviour the priority. In the next paragraphs, we sketch out how this might look like.

6.3.1 An evolutionary perspective on dopamine

The ideas in this part have first been expressed by Cisek (2019), as part of an evolutionary perspective on brain and behaviour. In his account of the evolving brain, he suggests that dopamine's first role in the nervous system was to arbitrate between exploration and exploitation.

Cisek (2019) suggests to think of the behaviour of a simple organism (think of C. Elegans) as divided into two modes. First, there is exploration mode. This

mode entails undirected, searching locomotion, with the goal of finding a food source. When a food source has been detected, undirected locomotion stops, and approach and feeding commence. This is exploitation mode. For a simple organism survival might just require to alternate between these two modes: explore until a food source is found, exploit it as much as possible, explore further when the source is depleted, and so on. This way of conceptualising behaviour is not just a thought experiment; it is also a method to analyse behavioural data (Chen et al. 2020).

How might the nervous system control such bimodal behaviour? Cisek (2019) proposes to use two different, independent control circuits—one that produces explorative locomotion, and one that exploits a local food source through approach and feeding. In addition, one would need a modulator that can arbitrate between the two controllers. The presence of that modulator could for example inhibit the exploration controller and activate the exploitation controller. To complete this setup, one needs a food detector that controls the release of the modulator.

The resulting process might look like this: explorative locomotion takes place under the control of the explore controller. At some point, the organism finds a food source. The presence of food activates the food detector, which in turn causes the sustained release of the modulator. The presence of the modulator inhibits the explore controller and thus prevents the organism from moving on. It also activates the exploit controller; the organism hence approaches and starts eating the food. When the food is consumed, the food detector ceases to be active. As a consequence, the modulator is no longer released and both the inhibition of the explore and the activation of the exploit controller stop. The organism then stops feeding and moves on.

Cisek (2019) identifies dopamine as the modulator that responds to the presence of food and activates the exploit controller. This is not implausible: research across many species has consistently shown that dopamine responds to food stimuli, and that it changes motor patterns accordingly (Barron et al. 2010). Recent findings

in simple organisms support the idea even more directly: using behaviour tracking in combination with different neural manipulations, it could be shown that behaviour and the corresponding dopamine-related neural processes in C. Elegans closely resemble the mechanism described above (Oranth et al. 2018).

Is this also true in mammals, say rats? It is known that blocking dopamine impairs flexible approach to reward-related goals (Nicola 2010), which seems consistent with an inability to activate the exploit controller to produce approach-and-consume behaviour. A relatively simple test might involve presenting an animal with an unexpected opportunity to obtain food (for example, a light might suddenly turn on to indicate that pressing a close-by lever will now produce a food pellet). Based on the above, we would predict that the light should elicit a dopamine response which switches the animal into exploit mode. Blocking this response, for example through optogenetic inhibition of VTA populations, should prevent the subject from exploiting the opportunity, but should not disrupt exploration behaviour.

Decisions between engaging and searching (which map on our notions of exploitation and exploration) have also been studied in humans (Kolling et al. 2012). There, such decisions are referred to as foraging decisions. Relevant brain regions for foraging have been identified using FMRI techniques. For example, Kolling et al. (2012) find that the ACC encodes the value of exploration, i.e. the value of not engaging in an opportunity. They also show that stronger ACC responses to the value of exploration are correlated with a stronger negative effect of this value on the decision to engage, which supports the notion that a value estimate from the ACC might inform foraging decisions.

It is possible that the ACC is a part of the control system described above. Given that the ACC can affect the decision to engage, it might fit into our framework as part of what we called the food-detector (we would need to extend that system to include rewards other then food). If this were the case, ACC activity

should ultimately inhibit striatal dopamine release—ACC encodes the value of exploration and must therefore activate the exploration controller.

6.3.2 Form modulation to learning

We have sketched a modulator role for dopamine: to arbitrate between different control circuits that are suitable for different stages of foraging. This role is 1) easily interpretable, 2) evolutionarily valid and ecologically relevant, and 3) completely orthogonal to learning.

However, it is not difficult to see how learning could gradually evolve as an additional feature of the system. For example, let us assume that a certain chemical can often be found near food. Let us also assume that the organism can sense the chemical. The frequent coincidence of food and chemical will translate in a frequent co-activation of the chemical sensor (triggered by the chemical), the food sensor (triggered by the food) and the dopamine system (triggered by the food sensor).

Now, what if there was dopamine-dependent plasticity between food sensor and chemical sensor? The co-activation of the three systems could then forge a link between them, and the chemical could from then on activate the dopamine system (and hence the exploit controller) via the food detector. That would certainly be beneficial for the organism, which would miss less opportunities to eat—in particular those were it found the chemical but missed the food. One may easily generalise this scenario: any neural event (detected stimuli, but also sets of motor commands) that is often followed by food or other appetitive stimuli that warrant should trigger exploit behaviour itself.

We arrived at a simple form of learning and a principle reminiscent of the Bellman equation for value functions (“any state from which one reaches valuable states is itself valuable”). From here, more complex forms of learning might evolve, and RL models might start to be a good description of those processes. Still,

all learning only supports a core non-learning function—the context-dependent arbitration between different control circuits.

6.4 Concluding remarks

For the basal ganglia, and even more so for dopamine, theories are typically phrased in the language of reinforcement learning. Is this dominance of RL good for the field? In the previous sections, we saw that views exist in which the direct modulation of behaviour (through vigour and/or through the selection of the appropriate control circuit) takes centre stage, with learning only in a secondary role. Could research in these directions be accelerated by the identification of the appropriate computational (or mathematical, economical, physical, ...) framework, in the same way that RL accelerated the research on dopaminergic reward prediction errors over the last 20 years?

The application of reinforcement learning to dopamine and the basal ganglia yielded a wealth of results, including all the results in this thesis. Perhaps other, less explored avenues as the ones described in this chapter might one day do the same.

Appendices

A

A constructive explanation of the AU learning rules

In chapter 3 we develop the payoffs-costs hypothesis of Bogacz (2017b) further. There, we show that the AU model of Mikhael and Bogacz (2016) is capable of learning the payoffs and costs of actions, and of encoding them in the two pathways of the basal ganglia.

To carry out the analyses reported in chapter 3, one can simply take the AU learning rules from the literature as a starting point. However, a more intuitive understanding of the structure of these rules can be gained if one pretends to construct them from scratch for the purpose of learning payoffs and costs. We describe this construction here, as an alternative way to introduce the learning rules 2.7 and 2.8.

Note that the construction below was published by Möller and Bogacz (2019)—parts of the text as well as figure A.1 were taken from that publication and adapted for this appendix.

A.1 Constructing the AU rules step by step

We start by observing that several models of learning in the basal ganglia assume the effect of the prediction error on G to be opposite to its effect on N (Collins and Frank 2014; Schroll et al. 2014). We thus start our construction by proposing that ΔG and ΔN might simply be proportional to the prediction error and its negative, respectively. To see whether this proposal works, we test it in a simulation. We use a simple reward schedule: an alternating sequence of costs $-n$ and payoffs p.

Figure A.1A shows both the mathematical formulation and the simulation of our first proposal. We find that there is a problem: the strengthening of N due to negative prediction errors, caused by the costs, is always immediately reversed by the following positive prediction errors caused by the payoffs. The same is true for the changes in G. As illustrated by the simulation, there is no net effect of learning.

To overcome this problem, we proceed by damping the impact of negative prediction errors (which are usually caused by costs) on G, and the impact of positive prediction errors on N. This can be done by introducing a nonlinear transformation of the prediction errors. We use a piecewise-linear function f_ϵ, defined in chapter 2 and visualised in figure 2.5. For G, the transformation leaves positive prediction errors invariant ($f_\epsilon(\delta)$ is just the identity for $\delta > 0$) but reduces the impact of negative prediction errors by scaling them down (for $\delta < 0$, $f_\epsilon(\delta)$ is linear with slope $\epsilon < 1$). For N it does the opposite. Hence, f_ϵ introduces a pathway-specific imbalance between learning from positive prediction errors and learning from negative prediction errors (which, as we pointed out above in subsection 2.4.4, is consistent with the properties of dopaminergic receptors on these pathways). The nonlinearities make sure that the costs do not alter the estimate G of the payoffs too much, and vice versa.

We update our mathematical formulation accordingly, and again test our rules in a simulation—the results are shown in figure A.1B. The simulation shows that,

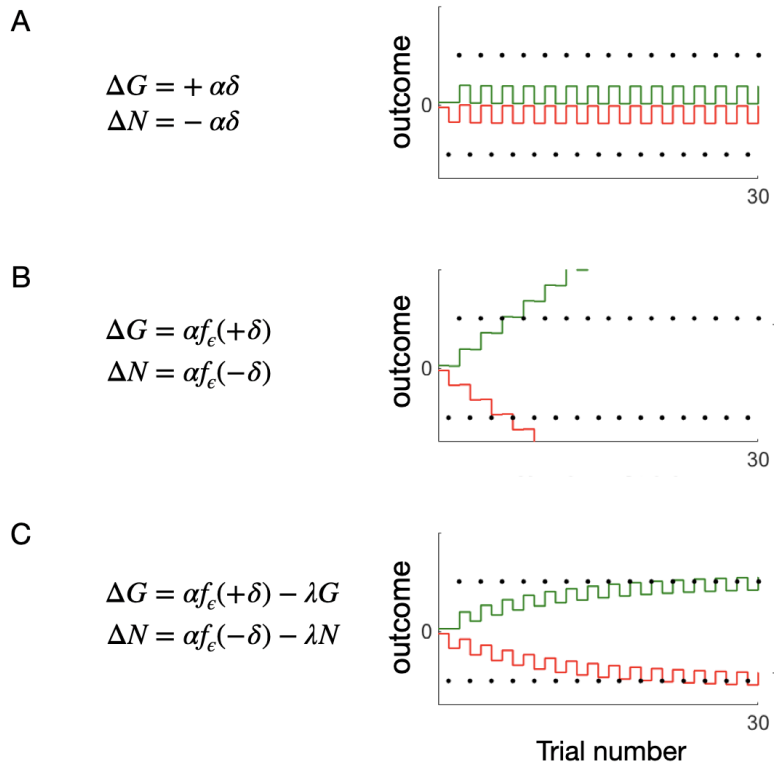

Figure A.1: The incremental construction of the AU learning rules. A - C The different stages in the construction of the learning rules. All panels feature a mathematical formulation of the rules at the given stage and a simulation of these rules. The rewards in those simulations, indicated by black dots, alternate between a fixed payoff of magnitude 20 and a fixed cost of magnitude -20. The Go weights G are shown in green, the negative No-Go weights $-N$ are shown in red. The parameters used in the simulations were $\alpha = 0.300$, $\epsilon = 0.443$ and $\lambda = 0.093$. This figure was taken from Möller and Bogacz (2019) and adapted for this thesis.

while producing the appropriate tendencies, the rules now cause unconstrained, ongoing strengthening of both weights. Such dynamics are neither biologically plausible nor useful to infer the actual payoff and cost. Another change must be made.

To stop unconstrained strengthening and stabilise the weights, we balance growth with decay. Adding unlearning terms to the mathematical formulation of the rules yields their final form (see equations 2.7 and 2.8). The simulation in figure A.1C suggests that the construction was successful: the final version of the rules allows the weights to converge to p and n respectively.

B

Derivations of actor models

Contents

In subsections 4.2.4 and 4.3.2 we use actor-critic models. We derived the actor learning rules of these models using the policy gradient method (Williams 1992; Sutton and Barto 2018). The method applies to parametrised probabilistic policies $\pi_\phi(a)$[1]. In its simplest form, it states that when a reward r is received after executing an action a, the appropriate update for the parameters of the policy is given by

$$\Delta\phi_i = \alpha(r-b)\frac{\partial \log \pi_\phi(a)}{\partial \phi_i} \tag{B.1}$$

where α is a constant learning rate and b the reward baseline (typically $b = v$, i.e. the value function is used as the baseline). Here, we show that applying the policy gradient method to the appropriate policies yields the update equations

[1] $\pi_\phi(a)$ is another notation for $P(a)$, the probability of choosing action a.

that we used for the simulations in sections 4.2.4 and 4.3.2. We will first apply it to a softmax policy for discrete action spaces, and then to a Gaussian policy for continuous action spaces.

B.1 The softmax policy for discrete action spaces

The softmax policy is used for discrete action spaces (i.e. tasks with finitely many actions a_1, a_2, ... a_n), and is given by

$$P(a) = \pi_{\phi,a} = \frac{\exp \phi_a}{\sum_b \exp \phi_b}. \tag{B.2}$$

We use it in the distracted bandit task in subsection 4.2.4, and in the for the task of Ferrucci et al. (2019) in subsection 4.3.2 (see equation 4.28), as well as in many models in chapter 5.

Computing $\frac{\partial \log \pi_{\phi,a}}{\partial \phi_i}$ yields

$$\frac{\partial \log \pi_{\phi,a}}{\partial \phi_i} = \frac{1}{\pi_{\phi,a}} \frac{\partial}{\partial \phi_i} \frac{\exp \phi_a}{\sum_b \exp \phi_b} \tag{B.3}$$

$$= \frac{1}{\pi_{\phi,a}} \left(\delta_{ia} \pi_a - \pi_i \pi_a\right) \tag{B.4}$$

$$= \delta_{ia} - \pi_i \tag{B.5}$$

with $\delta_{ia} = 1$ if $i = a$ and 0 else. Using equation B.1, we arrive at the update

$$\Delta \phi_a = \alpha(r - v)(c_a - \pi_a) = \alpha(r - v)(c_a - P(a)), \tag{B.6}$$

with $c_a = 1$ if action a was chosen and 0 otherwise. This is the update we used in equation 4.30 and equation 4.33.

B.2 The Gaussian policy for continuous action spaces

The Gaussian policy is appropriate for continuous action spaces (for example if $a \in \mathbb{R}$), and is defined by

$$\pi_\phi(a) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(a-\phi)^2}{\sigma^2}}. \tag{B.7}$$

We use it for the diminishing-returns task in subsection 4.2.4 (see equation 4.49). Computing $\frac{\partial \log \pi_\phi(a)}{\partial \phi}$ yields

$$\frac{\partial \log \pi_\phi(a)}{\partial \phi} = \frac{1}{\pi_\phi(a)} \frac{\partial}{\partial \phi} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(a-\phi)^2}{\sigma^2}} \tag{B.8}$$

$$= \frac{1}{\pi_\phi(a)} \pi_\phi(a) \frac{a-\phi}{\sigma^2} \tag{B.9}$$

$$= \frac{a-\phi}{\sigma^2} \tag{B.10}$$

Using equation B.1, we arrive at the update

$$\Delta\phi = \alpha(r - v)(a - \phi), \tag{B.11}$$

where $1/\sigma^2$ was absorbed into the learning rate α. This is the update we used in equation 4.42 and 4.46.

C

The high-noise limit of the Kalman filter

Contents

In chapter 4 we develop the scaled prediction error model. Here, we show that the scaled prediction error learning rules approximate the one-dimensional steady-state Kalman filter in the limit of high observation noise. We start by defining the Kalman filter model in section C.1. We then derive the steady-state Kalman filter in section C.2, and finally take the high-noise limit in section C.3.

C.1 The definition of the Kalman filter

The Kalman filter is a computational method for state estimation and prediction (Simon 2006). It can be derived from Bayesian principles and is optimal for tracking signals with certain characteristics. Here, we focus on a one-dimensional

Kalman filter which is used for predicting rewards, following Piray and Daw (2020a). The rules they use are

$$v_t = v_{t-1} + k_t \left(r_t - v_{t-1}\right) \tag{C.1}$$

$$k_t = (w_{t-1} + \nu^2)/(w_{t-1} + \nu^2 + \sigma^2) \tag{C.2}$$

$$w_t = (1 - k_t)(w_{t-1} + \nu^2). \tag{C.3}$$

where r_t is the reward in trial t[1]. These rules can be shown to be optimal for tracking signals such as those we used in chapter 4, subsection 4.2.3, i.e., signals that consist of samples drawn from a normal distribution with a drifting mean (Simon 2006).

C.2 The steady-state Kalman filter

The Kalman filter has several variables that must be updated on every trial. If one requires a simpler model with almost similar properties, one option is to use a Kalman filter in the limit $t \to \infty$: as $t \to \infty$, the posterior variance w_t and the Kalman gain k_t converge to limits w_∞ and k_∞.

Equation C.1 with k_∞ instead of k_t is called a *steady-state Kalman filter*. By construction, the normal Kalman filter becomes more similar to the steady-state Kalman filter the more trials pass. In practice, performance often does not differ much between the two (Simon 2006).

What are the limits w_∞ and k_∞? One may use equations C.2 and C.3 to determine them. By setting $k_t = k_{t-1}$ and $w_t = w_{t-1}$, we find

[1]Our notation differs slightly from that of Piray and Daw (2020a). We made this change to increase consistency across the thesis.

$$w_\infty = \frac{\nu^2}{2}\left(\sqrt{4\frac{\sigma^2}{\nu^2}+1}-1\right) \tag{C.4}$$

$$k_\infty = \frac{\sqrt{4\frac{\sigma^2}{\nu^2}+1}+1}{\sqrt{4\frac{\sigma^2}{\nu^2}+1}+1+2\frac{\sigma^2}{\nu^2}} \tag{C.5}$$

To use the steady-state Kalman filter, one just needs to compute k_∞ and plug it into equation C.1. One can then use this single equation to track the signal, with no other computations required. The steady-state Kalman filter is thus equivalent to the RW model in equation 2.1, parametrised with an optimal learning rate (that is to say, optimal for a signal with statistics ν^2 and σ^2).

C.3 The high-noise limit

The steady-state Kalman filter is less complex than the full Kalman filter in section C.1. However, its learning rate k_∞ is still a complex function of the signal statistics ν and σ. Can it be simplified? Let us consider of a signal with high observation noise, i.e. with σ^2 much larger than ν^2. Using equation C.5, we find that

$$k_\infty \to \frac{\nu}{\sigma} \tag{C.6}$$

for $\sigma^2/\nu^2 \to \infty$. This means that a steady-state Kalman filter with gain ν/σ is approximately optimal for signals with $\sigma^2 \gg \nu^2$. In figure C.1, we compare the optimal steady-state learning rate k_∞ with the approximately optimal learning rate ν/σ for different levels of σ, with ν fixed at $\nu = 1$.

We find that the approximation becomes very close very quickly—for $\sigma/\nu > 2$, the relative difference between the optimal learning rate and its approximation is already less than 30 %. Figure C.1 further suggests that the approximation breaks

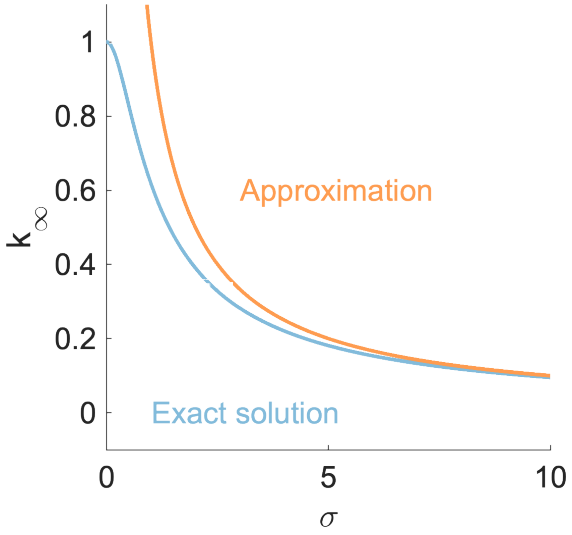

Figure C.1: The learning rate of the steady state Kalman filter. We show the learning rate k_∞ of the steady state Kalman filter as a function of the observation noise σ. We provide the exact value (blue line) and the approximation $k_\infty \approx \nu/\sigma$ (orange line).

down as ν/σ approaches unity—the optimal learning rate for signals with $\sigma = 0$ is one; any higher learning rate will be detrimental for the performance.

In summary, we find the learning rule

$$v_t = v_{t-1} + \frac{\nu}{\sigma}(r_t - v_{t-1}) \tag{C.7}$$

(C.8)

to be approximately optimal for $\sigma \ll \nu$ and large t. The rule C.7 bears striking resemblance to one of the scaled prediction error learning rules: the rule in equation 4.15. The difference between the two rules is just how the scaling is attributed: in the Kalman filter, one would perhaps speak of a scaled learning rate, while in the scaled prediction error model, one attributes the scaling to the error term. Mathematically, both formulations are equivalent.

A real difference between the Kalman filter and the scaled prediction error model is that the latter has a mechanism to track σ. No such mechanism exists in the Kalman filter. Both models require ν as an external input (for the scaled prediction error model, the corresponding parameter is α_v).

We conclude that the scaled prediction error model can be viewed as an implementation of approximately optimal one-dimensional state estimation, equipped with a mechanism to supply some of the required parameters—the observation noise σ. Other models have been proposed to track the process noise ν, for example by Piray and Daw (2020a). A combination of these approaches might be an interesting direction for future research.

D

Pupillometry

Contents

In chapter 5, we introduce a hypothesis: that prediction errors induce risk-seeking. We then test the hypothesis using data from an RL task. One central premise of the hypothesis is that two reward prediction errors occur during each trial of the task—the stimulus prediction error and the outcome prediction error. The dopamine transients related to those prediction errors play an important role in our theory.

Can we provide evidence of the occurrence of these prediction errors? In this appendix, we will use measurements of the participant's pupil diameter to show that they experienced surprise at the times when we expected reward prediction errors to occur. The surprise was proportional to the magnitude of the reward prediction errors. While this does not prove the occurrence of the

related dopamine transients, it still indicates that cognitive events related to those prediction errors took place as suggested.

To measure prediction error-related surprise, we rely on the well-documented link between pupil dilation and surprise (Preuschoff et al. 2011; Cavanagh et al. 2014; Browning et al. 2015; Lawson et al. 2017). We use a sequence of linear regressions to extract the pupil responses to prediction error magnitudes. We find pupil responses to both the stimulus prediction error and the outcome prediction error, indicating that those indeed registered as surprising events.

A paper reporting the results presented below is under review at the time of writing (a preprint was published by Möller, Grohn, et al. 2021). Text and figures were taken from that publication and adapted for this appendix.

D.1 Pupil dilation and the outcome prediction error

To test whether pupils responded to outcome prediction error magnitude, we extract trial-by-trial estimates of the outcome prediction error magnitude $|\delta_{\text{out}}|$ from the PEIRS model fits (see chapter 4 for details). We then use regression analyses to determine whether pupil dilation after reward presentation encoded $|\delta_{\text{out}}|$. We find a phasic response that peaks at about 0.9 s after reward presentation (figure D.1A). This suggests that pupil dilation indeed reflected surprise about the outcome.

D.2 Pupil dilation and the stimulus prediction error

Next, we test whether pupil dilation reflects stimulus prediction error magnitude. For this, we extract estimates of $|\delta_{\text{stim}}|$ for every trial, and use an analysis similar to that in the previous section to extract the corresponding pupil response. Here, we align the pupil traces at stimulus presentation and censor all data points collected after reward presentation to avoid confounding factors such as reward or outcome prediction errors. We find a phasic response starting about 1s after

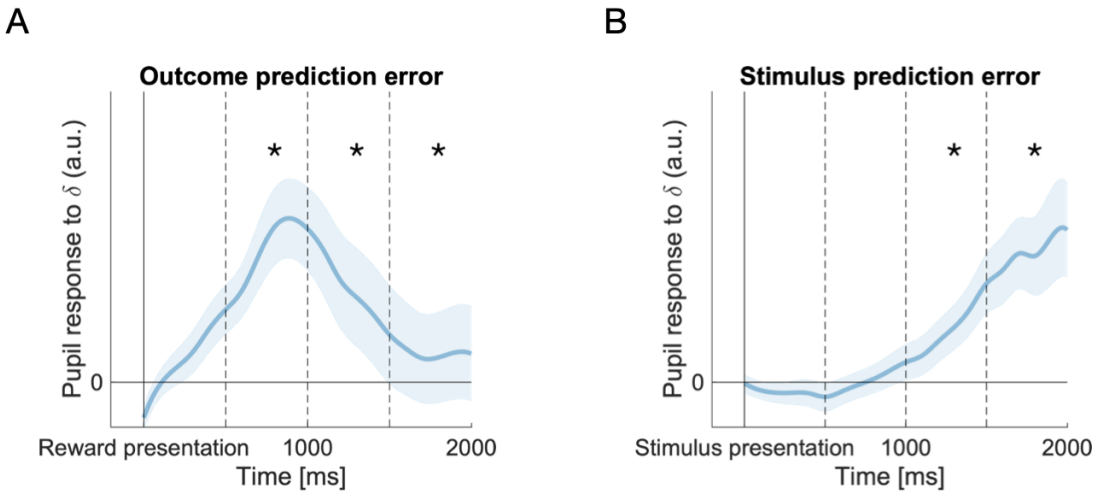

Figure D.1: Pupil responses to prediction error magnitude. A, B The blue lines and shadings represent the means and SEs of the pupil response to prediction error magnitude across participants, estimated with a mixed-effects linear model. We predicted pupil dilation traces from trial-by-trial estimates of prediction errors, which we obtained from a fitted PEIRS model. Responses are aligned at reward presentation (A) or stimulus presentation (B). For display, we smoothed the traces using spline interpolation. The dashed vertical lines indicate time bins with a width of 500 ms. Stars (*: $p < 0.5$) mark those time bins in which the pupil responses were significantly larger than zero (one-tailed t-tests, Bonferroni corrected, $\alpha = 0.05$). This figure was taken from Möller, Grohn, et al. (2021).

stimulus onset (figure D.1B). This suggests that surprise related to the stimulus prediction error is represented in the pupil during the choice period, which is consistent with the occurrence of a stimulus prediction error at that time.

The latency of the response to the stimulus prediction error is about 500 ms higher than the latency of the response to the outcome prediction error (compare figure D.1A and D.1B). There might be many reasons for this difference in latency. Among those, differences in information processing might play a role: computing a stimulus prediction error involves two stimuli, hence attention mechanisms, in addition to the retrieval of value estimates from memory. Computing the outcome prediction error, on the other hand, just requires the processing of a single number.

D.3 Discussion

We find pupil responses to the magnitude of both the stimulus prediction error and the outcome prediction error. Those responses might reflect the surprise associated with the two prediction errors, and might hence constitute an indirect physiological correlate of the mechanisms we have described in chapter 5. Note that we do not claim that pupils should be thought of as a proxy for dopamine, in terms of a direct physiological link—in fact, phasic pupil responses are perhaps more likely to reflect noradrenalin, see e.g. the study of Reimer et al. (2016).

Noradrenalin has been linked to uncertainty by Yu and Dayan (2005). In their model, they distinguish between two types of uncertainty: expected uncertainty (unexpected events follow unreliable cues) and unexpected uncertainty (unexpected events follow previously reliable cues). Yu and Dayan (2005) propose that expected uncertainty is associated with acetylcholine, which is discussed in the context of cortical and hippocampal learning, and that unexpected uncertainty is linked to noradrenaline. The uncertainty in our task is of the expected variety—our theory even suggests that participants learn how reliably the cues predict reward. Unexpected uncertainty could be introduced into our task through reversals or other sudden changes, but does not feature in the current design.

The pupil responses we found must thus be considered to encode expected uncertainty. If phasic pupil responses were linked to noradrenalin as suggested by Reimer et al. (2016), it appears that our results might be at odds with the model of Yu and Dayan (2005). Further work is required to dissociate the different kinds of uncertainty and surprise, their connection to various neurotransmitters and their relation to pupil dilation.

For the intents and purposes of this study, we propose that the pupil responses we show above might reveal cognitive states such as surprise, which in turn might be related to dopamine release according to the widely accepted reward

prediction error hypothesis. The link between pupil traces and dopamine is thus weak, which is why this analysis is a supplement to our main results, rather than a main result itself. Nevertheless, the pupil results together with our other findings yield a consistent picture.

D.4 Methods

During the task, time series of pupil diameters were recorded for every trial, using an EyeLink 1000 system. For each participant, the system was calibrated before the first block, and after subsequent blocks if required. To analyse the measurements, we first screened the raw measurements for blinks, which were identified as missing values in the recordings. Data points in the direct vicinity of blinks as well as data segments shorter than 50 ms were removed. After cleaning, missing value segments of less than 400 ms were filled using linear interpolation (Manohar 2019). Then, we aligned the traces to the relevant temporal markers (stimulus onset, or reward onset). We used the mean over the 500 ms before the alignment point to define a trial-wise baseline. All traces were divided and shifted by that baseline, resulting in traces that reflected the relative change of the pupil diameter after the alignment point. Finally, traces were downsampled to 10 Hz.

Pupil responses to prediction errors were obtained using linear mixed-effect models: we regressed pupil dilation against trial-by-trial estimates of the prediction error magnitude of interest, which we obtained from a model fit (see chapter 4). We accounted for individual differences by including random effects for the intercept and the prediction error regressor. We further included the identity of the stimulus —i.e., the identity of the fractal picture—that represented the chosen and the unchosen option as control regressors. Such regressions were run for each time bin of the pupil signal, and resulted in one pupil response time course (beta weights and standard errors as functions of time) for the whole population.

To uncover the pupil response to the stimulus prediction error, we aligned the pupil time courses at stimulus onset. After stimulus onset, participants would eventually make a choice (with variable delay; the median reaction time was 0.86 s) and receive a reward (with a 1 s delay) after their choice. Since the reward or the resulting outcome prediction error might confound our regression analysis, we censored out all data after reward presentation. This means that the number of observations on which regressions can be based rapidly declines after the median reward presentation time, which is at 1.86 s after stimulus onset. Estimates obtained later are increasingly unreliable since they are based on insufficient data. We hence conducted our analyses for the interval 0 s to 2 s after stimulus onset. This allowed us to obtain reliable estimates of the traces, while at the same time avoiding confounding effects related to reward presentation.

To test whether the pupil responses to the prediction errors are statistically significant, we followed Browning et al. (2015): we first split the pupil traces into four sequential 500 ms time bins. We then conducted one-tailed t-tests in each time bin, testing whether the mean weight in that time bin was larger than zero. Finally, we applied the Bonferroni correction for multiple comparisons to the results of the individual tests.

References

Barron, Andrew B, Eirik Søvik, and Jennifer L Cornish (2010). "The roles of dopamine and related compounds in reward-seeking behavior across animal phyla". In: *Frontiers in behavioral neuroscience* 4, p. 163.

Baunez, Christelle, Andre Nieoullon, and Marianne Amalric (1995). "In a rat model of parkinsonism, lesions of the subthalamic nucleus reverse increases of reaction time but induce a dramatic premature responding deficit". In: *Journal of Neuroscience* 15.10, pp. 6531–6541.

Bayer, Hannah M and Paul W Glimcher (2005). "Midbrain dopamine neurons encode a quantitative reward prediction error signal". In: *Neuron* 47.1, pp. 129–141.

Behrens, Timothy EJ et al. (2007). "Learning the value of information in an uncertain world". In: *Nature neuroscience* 10.9, pp. 1214–1221.

Bennett, Daniel, Yael Niv, and Angela Langdon (2021). "Value-free reinforcement learning: Policy optimization as a minimal model of operant behavior". In:

Berke, Joshua D (2018). "What does dopamine mean?" In: *Nature neuroscience* 21.6, pp. 787–793.

Berridge, Kent C and Terry E Robinson (1998). "What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience?" In: *Brain research reviews* 28.3, pp. 309–369.

Björklund, Anders and Stephen B Dunnett (2007). "Dopamine neuron systems in the brain: an update". In: *Trends in neurosciences* 30.5, pp. 194–202.

Bogacz, Rafal (2017a). "A tutorial on the free-energy framework for modelling perception and learning". In: *Journal of mathematical psychology* 76, pp. 198–211.

— (2017b). "Theory of reinforcement learning and motivation in the basal ganglia". In: *BioRxiv*, p. 174524.

— (2020). "Dopamine role in learning and action inference". In: *Elife* 9, e53262.

Browning, Michael et al. (2015). "Anxious individuals have difficulty learning the causal statistics of aversive environments". In: *Nature neuroscience* 18.4, pp. 590–596.

Buckley, Christopher L et al. (2017). "The free energy principle for action and perception: A mathematical review". In: *Journal of Mathematical Psychology* 81, pp. 55–79.

Cavanagh, James F et al. (2014). "Eye tracking and pupillometry are indicators of dissociable latent decision processes." In: *Journal of Experimental Psychology: General* 143.4, p. 1476.

Chau, Bolton KH et al. (2014). "A neural mechanism underlying failure of optimal choice with multiple alternatives". In: *Nature neuroscience* 17.3, pp. 463–470.

Chen, Cathy S et al. (2020). "Sex differences in learning from exploration". In: *bioRxiv*.

Chew, Benjamin et al. (2019). "Endogenous fluctuations in the dopaminergic midbrain drive behavioral choice variability". In: *Proceedings of the National Academy of Sciences* 116.37, pp. 18732–18737.

Cisek, Paul (2019). "Resynthesizing behavior through phylogenetic refinement". In: *Attention, Perception, & Psychophysics* 81.7, pp. 2265–2287.
Collinger, Jennifer L et al. (2013). "High-performance neuroprosthetic control by an individual with tetraplegia". In: *The Lancet* 381.9866, pp. 557–564.
Collins, Anne GE and Michael J Frank (2014). "Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive." In: *Psychological review* 121.3, p. 337.
Cui, Guohong et al. (2013). "Concurrent activation of striatal direct and indirect pathways during action initiation". In: *Nature* 494.7436, pp. 238–242.
D'Ardenne, Kimberlee et al. (2008). "BOLD responses reflecting dopaminergic signals in the human ventral tegmental area". In: *Science* 319.5867, pp. 1264–1267.
Da Silva, Joaquim Alves et al. (2018). "Dopamine neuron activity before action initiation gates and invigorates future movements". In: *Nature* 554.7691, pp. 244–248.
Daunizeau, Jean, Vincent Adam, and Lionel Rigoux (2014). "VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data". In: *PLoS Comput Biol* 10.1, e1003441.
Daw, Nathaniel D et al. (2011). "Trial-by-trial data analysis using computational models". In: *Decision making, affect, and learning: Attention and performance XXIII* 23.1.
Daw, Nathaniel D et al. (2006). "Cortical substrates for exploratory decisions in humans". In: *Nature* 441.7095, pp. 876–879.
Dayan, Peter et al. (1995). "The helmholtz machine". In: *Neural computation* 7.5, pp. 889–904.
Diederen, Kelly MJ and Wolfram Schultz (2015). "Scaling prediction errors to reward variability benefits error-driven learning in humans". In: *Journal of Neurophysiology* 114.3, pp. 1628–1640.
Diederen, Kelly MJ, Tom Spencer, et al. (2016). "Adaptive prediction error coding in the human midbrain and striatum facilitates behavioral adaptation and learning efficiency". In: *Neuron* 90.5, pp. 1127–1138.
Dodson, Paul D et al. (2016). "Representation of spontaneous movement by dopaminergic neurons is cell-type selective and disrupted in parkinsonism". In: *Proceedings of the National Academy of Sciences* 113.15, E2180–E2188.
Dreyer, Jakob K et al. (2010). "Influence of phasic and tonic dopamine release on receptor activation". In: *Journal of Neuroscience* 30.42, pp. 14273–14283.
Dudman, Joshua T and John W Krakauer (2016). "The basal ganglia: from motor commands to the control of vigor". In: *Current opinion in neurobiology* 37, pp. 158–166.
Engelhard, Ben et al. (2019). "Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons". In: *Nature* 570.7762, pp. 509–513.
Eshel, Neir, Michael Bukwich, et al. (2015). "Arithmetic and local circuitry underlying dopamine prediction errors". In: *Nature* 525.7568, pp. 243–246.
Eshel, Neir, Ju Tian, et al. (2016). "Dopamine neurons share common response function for reward prediction error". In: *Nature neuroscience* 19.3, pp. 479–486.
Ferrucci, Lorenzo et al. (2019). "Effects of reward size and context on learning in macaque monkeys". In: *Behavioural brain research* 372, p. 111983.
Ferster, Charles B and Burrhus Frederic Skinner (1957). "Schedules of reinforcement." In:
Fisher, Simon D et al. (2017). "Reinforcement determines the timing dependence of corticostriatal synaptic plasticity in vivo". In: *Nature communications* 8.1, pp. 1–13.

Frank, Michael J, Lauren C Seeberger, and Randall C O'reilly (2004). "By carrot or by stick: cognitive reinforcement learning in parkinsonism". In: *Science* 306.5703, pp. 1940–1943.

Freeze, Benjamin S et al. (2013). "Control of basal ganglia output by direct and indirect pathway projection neurons". In: *Journal of Neuroscience* 33.47, pp. 18531–18539.

Friston, Karl (2010). "The free-energy principle: a unified brain theory?" In: *Nature reviews neuroscience* 11.2, pp. 127–138.

Friston, Karl J, N Trujillo-Barreto, and Jean Daunizeau (2008). "DEM: a variational treatment of dynamic systems". In: *Neuroimage* 41.3, pp. 849–885.

Gallagher, David A et al. (2007). "Pathological gambling in Parkinson's disease: risk factors and differences from dopamine dysregulation. An analysis of published case series". In: *Movement disorders: official journal of the Movement Disorder Society* 22.12, pp. 1757–1763.

Garcia, Basile, Fabien Cerrotti, and Stefano Palminteri (2021). "The description–experience gap: a challenge for the neuroeconomics of decision-making under uncertainty". In: *Philosophical Transactions of the Royal Society B* 376.1819, p. 20190665.

Garris, Paul A et al. (1999). "Dissociation of dopamine release in the nucleus accumbens from intracranial self-stimulation". In: *Nature* 398.6722, pp. 67–69.

Gerfen, Charles R, Thomas M Engber, et al. (1990). "D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons". In: *Science* 250.4986, pp. 1429–1432.

Gerfen, Charles R and D James Surmeier (2011). "Modulation of striatal projection systems by dopamine". In: *Annual review of neuroscience* 34, pp. 441–466.

Gerfen, Charles R and Charles J Wilson (1996). "Chapter II The basal ganglia". In: *Handbook of chemical neuroanatomy*. Vol. 12. Elsevier, pp. 371–468.

Gershman, Samuel J (2014). "Dopamine ramps are a consequence of reward prediction errors". In: *Neural computation* 26.3, pp. 467–471.

— (2015). "Do learning rates adapt to the distribution of rewards?" In: *Psychonomic Bulletin & Review* 22.5, pp. 1320–1327.

— (2017). "Dopamine, inference, and uncertainty". In: *Neural Computation* 29.12, pp. 3311–3326.

— (2019). "What does the free energy principle tell us about the brain?" In: *arXiv preprint arXiv:1901.07945*.

Grewal, Mohinder S and Angus P Andrews (2010). "Applications of Kalman filtering in aerospace 1960 to the present [historical perspectives]". In: *IEEE Control Systems Magazine* 30.3, pp. 69–78.

Grillner, Sten and Brita Robertson (2015). "The basal ganglia downstream control of brainstem motor centres—an evolutionarily conserved strategy". In: *Current opinion in neurobiology* 33, pp. 47–52.

Gurney, Kevin, Tony J Prescott, and Peter Redgrave (2001a). "A computational model of action selection in the basal ganglia. I. A new functional anatomy". In: *Biological cybernetics* 84.6, pp. 401–410.

— (2001b). "A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour". In: *Biological cybernetics* 84.6, pp. 411–423.

Guyaguler, Baris, Roland N Horne, et al. (2001). "Uncertainty assessment of well placement optimization". In: *SPE annual technical conference and exhibition*. Society of Petroleum Engineers.

Hamid, Arif A et al. (2016). "Mesolimbic dopamine signals the value of work". In: *Nature neuroscience* 19.1, pp. 117–126.

Hernández-López, Salvador et al. (2000). "D2 dopamine receptors in striatal medium spiny neurons reduce L-type Ca2+ currents and excitability via a novel PLCβ1–IP3–calcineurin-signaling cascade". In: *Journal of Neuroscience* 20.24, pp. 8987–8995.

Herz, Damian M et al. (2016). "Neural correlates of decision thresholds in the human subthalamic nucleus". In: *Current Biology* 26.7, pp. 916–920.

Hessel, Matteo et al. (2019). "Multi-task deep reinforcement learning with popart". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 3796–3803.

Hochberg, Leigh R et al. (2012). "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm". In: *Nature* 485.7398, pp. 372–375.

Howe, Mark W and Daniel A Dombeck (2016). "Rapid signalling in distinct dopaminergic axons during locomotion and reward". In: *Nature* 535.7613, pp. 505–510.

Huber, Joel, John W Payne, and Christopher Puto (1982). "Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis". In: *Journal of consumer research* 9.1, pp. 90–98.

Hughes, Ryan N et al. (2020). "Ventral tegmental dopamine neurons control the impulse vector during motivated behavior". In: *Current Biology*.

Hursh, Steven R et al. (1988). "A cost-benefit analysis of demand for food". In: *Journal of the experimental analysis of behavior* 50.3, pp. 419–440.

Ikemoto, Satoshi and Jaak Panksepp (1999). "The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking". In: *Brain Research Reviews* 31.1, pp. 6–41.

Jang, Anthony I et al. (2019). "Positive reward prediction errors during decision-making strengthen memory encoding". In: *Nature human behaviour* 3.7, pp. 719–732.

Jurado-Parras, Maria-Teresa et al. (2020). "The dorsal striatum energizes motor routines". In: *Current Biology* 30.22, pp. 4362–4372.

Kahneman, Daniel and Amos Tversky (2013). "Prospect theory: An analysis of decision under risk". In: *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, pp. 99–127.

Karin, Omer and Uri Alon (2021). "The dopamine circuit as a reward-taxis navigation system". In: *bioRxiv*.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Klaus, Andreas, Joaquim Alves da Silva, and Rui M Costa (2019). "What, if, and when to move: Basal ganglia circuits and self-paced action initiation". In: *Annual review of neuroscience* 42, pp. 459–483.

Klaus, Andreas, Gabriela J Martins, et al. (2017). "The spatiotemporal organization of the striatum encodes action space". In: *Neuron* 95.5, pp. 1171–1180.

Kolling, Nils et al. (2012). "Neural mechanisms of foraging". In: *Science* 336.6077, pp. 95–98.

Kravitz, Alexxai V et al. (2010). "Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry". In: *Nature* 466.7306, pp. 622–626.

Lawson, Rebecca P, Christoph Mathys, and Geraint Rees (2017). "Adults with autism overestimate the volatility of the sensory environment". In: *Nature neuroscience* 20.9, p. 1293.

Louie, Kenway, Lauren E Grattan, and Paul W Glimcher (2011). "Reward value-based gain control: divisive normalization in parietal cortex". In: *Journal of Neuroscience* 31.29, pp. 10627–10639.

Madan, Christopher R, Elliot A Ludvig, and Marcia L Spetch (2014). "Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions". In: *Psychonomic bulletin & review* 21.3, pp. 629–636.

Manohar, SG (2019). *Matlib: MATLAB tools for plotting, data analysis, eye tracking and experiment design (Public)*.

Markowitz, Jeffrey E et al. (2018). "The striatum organizes 3D behavior via moment-to-moment action selection". In: *Cell* 174.1, pp. 44–58.

Matsumoto, Masayuki and Okihide Hikosaka (2009). "Two types of dopamine neuron distinctly convey positive and negative motivational signals". In: *Nature* 459.7248, pp. 837–841.

Mazzoni, Pietro, Anna Hristova, and John W Krakauer (2007). "Why don't we move faster? Parkinson's disease, movement vigor, and implicit motivation". In: *Journal of neuroscience* 27.27, pp. 7105–7116.

McHaffie, John G et al. (2005). "Subcortical loops through the basal ganglia". In: *Trends in neurosciences* 28.8, pp. 401–407.

Merel, Josh et al. (2019). "Deep neuroethology of a virtual rodent". In: *arXiv preprint arXiv:1911.09451*.

Middleton, Frank A and Peter L Strick (2000). "Basal ganglia and cerebellar loops: motor and cognitive circuits". In: *Brain research reviews* 31.2-3, pp. 236–250.

Mikhael, John G and Rafal Bogacz (2016). "Learning reward uncertainty in the basal ganglia". In: *PLoS computational biology* 12.9, e1005062.

Mink, Jonathan W (1996). "The basal ganglia: focused selection and inhibition of competing motor programs". In: *Progress in neurobiology* 50.4, pp. 381–425.

Mirenowicz, Jacques and Wolfram Schultz (1994). "Importance of unpredictability for reward responses in primate dopamine neurons". In: *Journal of neurophysiology* 72.2, pp. 1024–1027.

Mnih, Volodymyr, Adria Puigdomenech Badia, et al. (2016). "Asynchronous methods for deep reinforcement learning". In: *International conference on machine learning*. PMLR, pp. 1928–1937.

Mnih, Volodymyr, Koray Kavukcuoglu, et al. (2015). "Human-level control through deep reinforcement learning". In: *nature* 518.7540, pp. 529–533.

Möller, Moritz and Rafal Bogacz (2019). "Learning the payoffs and costs of actions". In: *PLoS computational biology* 15.2, e1006285.

Möller, Moritz, Jan Grohn, et al. (2021). "A Behavioral Association Between Prediction Errors and Risk-Seeking: Theory and Evidence". In: *bioRxiv*, pp. 2020–04.

Montague, P Read, Peter Dayan, and Terrence J Sejnowski (1996). "A framework for mesencephalic dopamine systems based on predictive Hebbian learning". In: *Journal of neuroscience* 16.5, pp. 1936–1947.

Montague, P Read, Samuel M McClure, et al. (2004). "Dynamic gain control of dopamine delivery in freely moving animals". In: *Journal of Neuroscience* 24.7, pp. 1754–1759.

Morris, Genela, David Arkadir, et al. (2004). "Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons". In: *Neuron* 43.1, pp. 133–143.
Morris, Genela, Alon Nevet, et al. (2006). "Midbrain dopamine neurons encode decisions for future action". In: *Nature neuroscience* 9.8, pp. 1057–1063.
Nambu, Atsushi (2008). "Seven problems on the basal ganglia". In: *Current opinion in neurobiology* 18.6, pp. 595–604.
Neath, Andrew A and Joseph E Cavanaugh (2012). "The Bayesian information criterion: background, derivation, and applications". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2, pp. 199–203.
Nicola, Saleem M (2010). "The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior". In: *Journal of Neuroscience* 30.49, pp. 16585–16600.
Niv, Yael (2007). "Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation?" In: *Annals of the New York Academy of Sciences* 1104.1, pp. 357–376.
Niv, Yael, Nathaniel D Daw, and Peter Dayan (2006). "How fast to work: Response vigor, motivation and tonic dopamine". In: *Advances in neural information processing systems*, pp. 1019–1026.
Niv, Yael, Nathaniel D Daw, Daphna Joel, et al. (2007). "Tonic dopamine: opportunity costs and the control of response vigor". In: *Psychopharmacology* 191.3, pp. 507–520.
Niv, Yael, Jeffrey A Edlund, et al. (2012). "Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain". In: *Journal of Neuroscience* 32.2, pp. 551–562.
O'Doherty, John et al. (2004). "Dissociable roles of ventral and dorsal striatum in instrumental conditioning". In: *science* 304.5669, pp. 452–454.
Obeso, JA et al. (2009). "What can man do without basal ganglia motor output? The effect of combined unilateral subthalamotomy and pallidotomy in a patient with Parkinson's disease". In: *Experimental neurology* 220.2, pp. 283–292.
Oorschot, Dorothy E (1996). "Total number of neurons in the neostriatal, pallidal, subthalamic, and substantia nigral nuclei of the rat basal ganglia: a stereological study using the cavalieri and optical disector methods". In: *Journal of Comparative Neurology* 366.4, pp. 580–599.
Oranth, Alexandra et al. (2018). "Food sensation modulates locomotion by dopamine and neuropeptide signaling in a distributed neuronal network". In: *Neuron* 100.6, pp. 1414–1428.
Padoa-Schioppa, Camillo (2009). "Range-adapting representation of economic value in the orbitofrontal cortex". In: *Journal of Neuroscience* 29.44, pp. 14004–14014.
Palminteri, Stefano, Valentin Wyart, and Etienne Koechlin (2017). "The importance of falsification in computational cognitive modeling". In: *Trends in cognitive sciences* 21.6, pp. 425–433.
Park, Soyoung Q et al. (2012). "Adaptive coding of reward prediction errors is gated by striatal coupling". In: *Proceedings of the National Academy of Sciences* 109.11, pp. 4285–4289.
Paz, Jeanne Tamar et al. (2007). "Activity of ventral medial thalamic neurons during absence seizures and modulation of cortical paroxysms by the nigrothalamic pathway". In: *Journal of Neuroscience* 27.4, pp. 929–941.

Pedersen, Mads Lund, Michael J Frank, and Guido Biele (2017). "The drift diffusion model as the choice rule in reinforcement learning". In: *Psychonomic bulletin & review* 24.4, pp. 1234–1251.

Pessiglione, Mathias et al. (2006). "Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans". In: *Nature* 442.7106, pp. 1042–1045.

Piray, Payam and Nathaniel D Daw (2020a). "A simple model for learning in volatile environments". In: *PLoS computational biology* 16.7, e1007963.

— (2020b). "Unpredictability vs. volatility and the control of learning". In: *bioRxiv*.

Preuschoff, Kerstin, Bernard Marius t Hart, and Wolfgang Einhauser (2011). "Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making". In: *Frontiers in neuroscience* 5, p. 115.

Rangel, Antonio and John A Clithero (2012). "Value normalization in decision making: theory and evidence". In: *Current opinion in neurobiology* 22.6, pp. 970–981.

Redgrave, Peter, Tony J Prescott, and Kevin Gurney (1999). "The basal ganglia: a vertebrate solution to the selection problem?" In: *Neuroscience* 89.4, pp. 1009–1023.

Reimer, Jacob et al. (2016). "Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex". In: *Nature communications* 7.1, pp. 1–7.

Rescorla, Robert A (1972). "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement". In: *Current research and theory*, pp. 64–99.

— (1988). "Pavlovian conditioning: It's not what you think it is." In: *American psychologist* 43.3, p. 151.

Reynolds, John NJ, Brian I Hyland, and Jeffery R Wickens (2001). "A cellular mechanism of reward-related learning". In: *Nature* 413.6851, pp. 67–70.

Rothenhoefer, Kathryn M et al. (2021). "Rare rewards amplify dopamine responses". In: *Nature Neuroscience*, pp. 1–5.

Rouhani, Nina, Kenneth A Norman, and Yael Niv (2018). "Dissociable effects of surprising rewards on learning and memory." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44.9, p. 1430.

Salamone, John D and Mercè Correa (2012). "The mysterious motivational functions of mesolimbic dopamine". In: *Neuron* 76.3, pp. 470–485.

Salamone, John D, Mercè Correa, et al. (2009). "Dopamine, behavioral economics, and effort". In: *Frontiers in behavioral neuroscience* 3, p. 13.

Salamone, John D, RE Steinpreis, et al. (1991). "Haloperidol and nucleus accumbens dopamine depletion suppress lever pressing for food but increase free food consumption in a novel food choice procedure". In: *Psychopharmacology* 104.4, pp. 515–521.

Samejima, Kazuyuki et al. (2005). "Representation of action-specific reward values in the striatum". In: *Science* 310.5752, pp. 1337–1340.

Saunders, Benjamin T et al. (2018). "Dopamine neurons create Pavlovian conditioned stimuli with circuit-defined motivational properties". In: *Nature neuroscience* 21.8, pp. 1072–1083.

Schroll, Henning, Julien Vitay, and Fred H Hamker (2014). "Dysfunctional and compensatory synaptic plasticity in P arkinson's disease". In: *European Journal of Neuroscience* 39.4, pp. 688–702.

Schultz, Wolfram, Peter Dayan, and P Read Montague (1997). "A neural substrate of prediction and reward". In: *Science* 275.5306, pp. 1593–1599.

Shadmehr, Reza and John W Krakauer (2008). "A computational neuroanatomy for motor control". In: *Experimental brain research* 185.3, pp. 359–381.
Shen, Weixing et al. (2008). "Dichotomous dopaminergic control of striatal synaptic plasticity". In: *Science* 321.5890, pp. 848–851.
Silver, David et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587, pp. 484–489.
Simon, Dan (2006). *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons.
Smith, Y et al. (1998). "Microcircuitry of the direct and indirect pathways of the basal ganglia." In: *Neuroscience* 86.2, pp. 353–387.
Sommer, Marc A (2003). "The role of the thalamus in motor control". In: *Current opinion in neurobiology* 13.6, pp. 663–670.
Spitzer, Bernhard, Leonhard Waschke, and Christopher Summerfield (2017). "Selective overweighting of larger magnitudes during noisy numerical comparison". In: *Nature Human Behaviour* 1.8, pp. 1–8.
St Onge, Jennifer R and Stan B Floresco (2009). "Dopaminergic modulation of risk-based decision making". In: *Neuropsychopharmacology* 34.3, pp. 681–697.
Stauffer, William R et al. (2015). "Economic choices reveal probability distortion in macaque monkeys". In: *Journal of Neuroscience* 35.7, pp. 3146–3154.
Steinberg, Elizabeth E et al. (2013). "A causal link between prediction errors, dopamine neurons and learning". In: *Nature neuroscience* 16.7, pp. 966–973.
Stewart, Neil, Nick Chater, and Gordon DA Brown (2006). "Decision by sampling". In: *Cognitive psychology* 53.1, pp. 1–26.
Surmeier, D James et al. (2007). "D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons". In: *Trends in neurosciences* 30.5, pp. 228–235.
Sutton, Richard S, Andrew G Barto, et al. (1998). *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge.
Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.
Suzuki, Shosuke et al. (2020). "Distinct regions of the striatum underlying effort, movement initiation and effort discounting". In: *Nature Human Behaviour*, pp. 1–11.
Syed, Emilie CJ et al. (2016). "Action initiation shapes mesolimbic dopamine encoding of future rewards". In: *Nature neuroscience* 19.1, pp. 34–36.
Tan, Huiling et al. (2016). "Decoding gripping force based on local field potentials recorded from subthalamic nucleus in humans". In: *Elife* 5, e19089.
Tecuapetla, Fatuel et al. (2014). "Balanced activity in basal ganglia projection pathways is critical for contraversive movements". In: *Nature communications* 5.1, pp. 1–10.
Thura, David and Paul Cisek (2017). "The basal ganglia do not select reach targets but control the urgency of commitment". In: *Neuron* 95.5, pp. 1160–1170.
Thurley, Kay, Walter Senn, and Hans-Rudolf Luscher (2008). "Dopamine increases the gain of the input-output response of rat prefrontal pyramidal neurons". In: *Journal of neurophysiology* 99.6, pp. 2985–2997.
Tobler, Philippe N, Christopher D Fiorillo, and Wolfram Schultz (2005). "Adaptive coding of reward value by dopamine neurons". In: *Science* 307.5715, pp. 1642–1645.
Tsai, Hsing-Chen et al. (2009). "Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning". In: *Science* 324.5930, pp. 1080–1084.

Turner, Robert S and Marjorie E Anderson (1997). "Pallidal discharge related to the kinematics of reaching movements in two dimensions". In: *Journal of neurophysiology* 77.3, pp. 1051–1074.
Turner, Robert S and Michel Desmurget (2010). "Basal ganglia contributions to motor control: a vigorous tutor". In: *Current opinion in neurobiology* 20.6, pp. 704–716.
Ungless, Mark A, Peter J Magill, and J Paul Bolam (2004). "Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli". In: *Science* 303.5666, pp. 2040–2042.
Van Swieten, Maaike MH, Rafal Bogacz, and Sanjay G Manohar (2021). "Gambling on an empty stomach: Hunger modulates preferences for learned but not described risks". In: *bioRxiv*.
Voon, V et al. (2006). "Prospective prevalence of pathologic gambling and medication association in Parkinson disease". In: *Neurology* 66.11, pp. 1750–1752.
Wang, Yawei et al. (2021). "Tonic firing mode of midbrain dopamine neurons continuously tracks reward values changing moment-by-moment". In: *Elife* 10, e63166.
Watabe-Uchida, Mitsuko, Neir Eshel, and Naoshige Uchida (2017). "Neural circuitry of reward prediction error". In: *Annual review of neuroscience* 40, pp. 373–394.
Weintraub, Daniel et al. (2010). "Impulse control disorders in Parkinson disease: a cross-sectional study of 3090 patients". In: *Archives of neurology* 67.5, pp. 589–595.
Welch, Greg, Gary Bishop, et al. (1995). *An introduction to the Kalman filter*.
Williams, Ronald J (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8.3-4, pp. 229–256.
Wilson, Hugh R and Jack D Cowan (1972). "Excitatory and inhibitory interactions in localized populations of model neurons". In: *Biophysical journal* 12.1, pp. 1–24.
Wise, Roy A (2004). "Dopamine, learning and motivation". In: *Nature reviews neuroscience* 5.6, pp. 483–494.
Wise, Roy A and P-P Rompre (1989). "Brain dopamine and reward". In: *Annual review of psychology* 40.1, pp. 191–225.
Wolff, Steffen BE, Raymond Ko, and Bence P Ölveczky (2019). "Distinct roles for motor cortical and thalamic inputs to striatum during motor learning and execution". In: *bioRxiv*, p. 825810.
Wulff, Dirk U, Max Mergenthaler-Canseco, and Ralph Hertwig (2018). "A meta-analytic review of two modes of learning and the description-experience gap." In: *Psychological bulletin* 144.2, p. 140.
Yttri, Eric A and Joshua T Dudman (2016). "Opponent and bidirectional control of movement velocity in the basal ganglia". In: *Nature* 533.7603, pp. 402–406.
Yu, J Angela and Peter Dayan (2005). "Uncertainty, neuromodulation, and attention". In: *Neuron* 46.4, pp. 681–692.
Zaghloul, Kareem A et al. (2009). "Human substantia nigra neurons encode unexpected financial rewards". In: *Science* 323.5920, pp. 1496–1499.
Zalocusky, Kelly A et al. (2016). "Nucleus accumbens D2R cells signal prior outcomes and control risky decision-making". In: *Nature* 531.7596, pp. 642–646.
Zénon, Alexandre, Sophie Devesse, and Etienne Olivier (2016). "Dopamine manipulation affects response vigor independently of opportunity cost". In: *Journal of Neuroscience* 36.37, pp. 9516–9525.