On the computational and neural characterisation of reward learning behaviour

Kitti Ban

MA Economics, MSc Psychological Science

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

Institute of Neuroscience and Psychology

College of Medical, Veterinary and Life Sciences

University of Glasgow

May 2024

Abstract

Do we learn differently from better- or worse-than-expected decision outcomes? Over the past decades, converging evidence emerged about the crucial role of the dopaminergic system in guiding learning through signalling reward prediction errors. However, a complete characterisation of how this learning process is influenced by feedback valence, surprise, and uncertainty is still lacking. The current thesis focuses on exploring the differential behavioural and neural mechanisms related to learning from positive versus negative decision outcomes whilst examining the influence of uncertainty on these processes.

In our first experiment, we collected simultaneous EEG and eye-tracking data during a probabilistic reversal learning task. Using multivariate EEG analysis, we replicated the two distinct spatiotemporal reward learning systems reported by Fouragnan and colleagues (2015). Given that locus-coeruleus-noradrenaline (LC-NA) activity is difficult to directly measure non-invasively in humans, we used the pupil response as a proxy for LC-NA activity. We showed that the increased feedback-related pupil response to negative compared to positive outcomes is exclusively driven by increased negative feedback processing in the early and the late system. Additionally, a stronger coupling between early, but not late, system activity and the feedback-evoked pupil response was linked to reduced performance, increased uncertainty as well as exploration propensity. In line with existing research indicating the LC-NA network in uncertainty signalling and network resets, we propose that when internal estimates of environmental uncertainty surge in response to negative feedback, the early system, regulated by noradrenergic activity, interrupts processing in structures of the late system. Such network resets may aid flexible adaptation to changing environments by simultaneously reducing the influence of learned value representations and increasing the neural gain of new information.

Our second experimental chapter extended the above study by examining post-feedback response adaptation as a function of early and late system activity. Specifically, we utilised hierarchical drift diffusion modelling, in which the drift rate and boundary separation were constrained by trial-wise and valence-specific early and late system activity. We hypothesised that an LC-NA-induced interruption in reward learning structures would reduce subsequent evidence accumulation as learned value representations become less influential and participants consider a reversal in reward contingencies more likely. Consistent with this hypothesis, we found that increased negative feedback processing by the early and late system reduced evidence accumulation in the next trial. Furthermore, a stronger association between the feedback-locked pupil response and early system activity following negative outcomes was significantly associated with the degree of drift rate reduction prompted by the early system. This result implies that LC-NA mediated network resets may be primarily associated with the early system, which in turn may down-regulate late system activity.

Our final study explored differential value learning in the Balloon Analogue Risk Task (BART) under varying levels of uncertainty. By deriving differential learning rates from the newly developed Scaled Target Learning model, we showed that participants preferentially learn from positive compared to negative feedback under increased levels of uncertainty. Furthermore, the degree of this learning bias was negatively related to performance under the highest level of uncertainty. These results provide further evidence for differential mechanisms implicated in positive and negative feedback processing and indicate the important modulatory role of uncertainty in reward learning.

Together, this thesis provides novel insights on the valence-specific neural and behavioural characteristics associated with feedback processing. Our results also highlight the important modulatory role uncertainty and noradrenaline play in reward learning and thus provide a more complete depiction of reward learning behaviour.

Table of contents

List of tables

List of figures

Acknowledgement

First and foremost, I am profoundly grateful to my supervisors, Dr Martin Lages and Professor Marios Philiastides – your expertise, integrity, and guidance have motivated me to improve day by day. Thank you both for your patience, even during the most challenging times. Martin – I am forever thankful for making this thesis possible, your integrity and compassion has encouraged me to do my best and gain the confidence needed to carry on.

I would like to further thank Dr Andrea Kóbor, Dr Eszter-Tóth-Fáber, Professor Roshan Cools and Dr Hanneke den Ouden for their willingness and trust to collaborate. Thank you for making me feel welcome in your teams and showing an excellent example of how research ought to be pursued. Andi, Eszti – I have genuinely enjoyed working with you, thank you for this opportunity and for bringing joy into research. Roshan, Hanneke – I am truly grateful for accepting me in your research groups, your enthusiasm and commitment towards research has never ceased to inspire me.

Thank you to the wonderful colleagues I had the fortune to meet in Glasgow, Nijmegen, and Budapest. I feel lucky to share this experience with you. Desi – thank you for the inspiring academic (and non-academic) discussions. Professor Frank Pollick and Professor Peter Ulhaas - I am thankful for your professional feedback, advice, and encouragement throughout the years. Dr Andrew Tyler Morgan - I am grateful to have learned from such a smart and compassionate researcher. A special thank you to Dr Emanuele de Luca - you have made my MSc project an exciting and fulfilling journey and an important bridge towards my doctoral studies.

I would not have been able to carry on without the enduring support from my friends and family. Thank you for encouraging and believing in me, even when I did not believe in myself. Marjan - thank you for broadening my perspectives and encouraging me to be brave. Avka - thank you for your care, patience, and sacrifice during the toughest of times, I truly appreciate you.

Author’s declaration

I declare that, except where explicit reference is made to the contribution of others, this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

List of abbreviations

ACC	Anterior cingulate cortex
AIC	Akaike information criterion
BART	Balloon analogue risk task
BOLD	Blood oxygenation level dependent
BSR	Bayesian sequential risk-taking
DA	Dopamine
DDM	Drift diffusion model
DIC	Deviance information criterion
EEG	Electroencephalogram
fMRI	Functional magnetic resonance imaging
GLM	General linear model
HDDM	Hierarchical drift diffusion model
HDI	Highest density interval
LC	Locus-coeruleus
LC-NA	Locus-coeruleus-noradrenaline
LDA	Linear discriminant analysis
LOOIC	Leave-one-out information criterion
MCMC	Markov Chain Monte Carlo
NA	Noradrenaline
OFC	Orbitofrontal cortex
PuRF	Pupil response function
RL	Reward learning
RPE	Reward prediction error
RT	Reaction time
SSM	Sequential sampling model
STL	Scaled target learning model
STL-D	Scaled target learning model with decay
TD	Temporal difference
vmPFC	Ventromedial prefrontal cortex
VTA	Ventral tegmental area

Chapter 1. General introduction

1.1 Background

In our modern era, we face increasingly more decisions as we contemplate which of the myriad of items to purchase or how to best spend our precious free time. For a long time, mainstream economics preoccupied itself with the aim to predict how humans make value-based choices given a set of inputs. This prediction was based on the premise that each human agent is a 'homo economicus' – a completely rational agent primarily concerned with utility maximisation. However, research over the last decades revealed that humans are frequently unable to escape emotional affect and are prone to cognitive and personal biases such heuristics and stereotypes (Kahnemann & Tversky, 1979). Kahnemann and Tversky's research revealed nuanced aspects of human decision making that contradicted the concept of the homo economicus and paved the way for the inclusion of psychological factors into the study of decision making. These efforts produced the field of behavioural economics, in which the choice paradigms and formal models of decision computations from economics were augmented with cognitive theories and behavioural data from psychology.

A new wave of progress in decision science was afforded by the significant advances in neuroscience and computer science over the last few decades, including breakthroughs in functional neuroimaging, animal research, machine learning algorithms, and computational modelling (Glimcher & Fehr, 2014). The incorporation of these new tools and techniques into the study of decision making brought about the multidisciplinary field of neuroeconomics, which explores the neural substrates linked to value-based decisions. Although these efforts substantially contributed to an improved understanding of the behavioural and biological components of decision making, a complete characterisation of the decision making process is still lacking. In the current thesis, I combine state-of-the-art computational modelling, electrophysiological and eye-tracking data analyses, and machine learning

classification to contribute to an improved understanding of the neural and behavioural characteristics of decision making.

1.2 The decision making process

Rangel and colleagues (2009) first proposed a unified framework for value-based decision making that takes into account various fields of research within neuroeconomics. They described five distinct steps that make up the value-based decision process (Fig.1-1). First, the brain must compute a representation of the decision problem, including relevant internal (e.g., level of hunger) and external (e.g., threat level) states as well as potential courses of action. Next, the value linked to each of these courses of action and environmental states needs to be calculated. During action selection, these different values are compared in order to make a decision, after which the action with the highest subjective value is executed.

Once a decision outcome becomes available, the brain needs to evaluate its desirability, which once again requires the computation of value associated with the outcome. Finally, learning serves to make better decisions in the future and affects various steps in the decision process as relevant value representations of states (internal, external) and actions need to be updated. The above presented five steps involved in decision making are not considered rigid and their relative temporal order is not yet fully accounted for. The major advantage of this framework lies in breaking down the decision process into separate sub-processes so that research in neuroeconomics can easier create and test specific hypotheses.

1.2.1 Different valuation systems

Three different valuation systems have been proposed to facilitate decision making; the Pavlovian system, the habitual system, and the goal-directed system (Rangel et al., 2008). The Pavlovian system assigns value to evolutionarily relevant behaviours elicited by specific environmental cues and concerns a small set of stimulus-driven, reflexive behaviours (Clark et al., 2012). Unlike the other two systems, the Pavlovian system is not instrumental;

it is concerned with environmental stimuli innately predictive of rewards rather than learning related to potential courses of actions. In comparison, the habitual system learns to slowly assign value through generalisation by computing stimulus-outcome associations through repeated exposure. Finally, the goal-directed system computes the value of action-outcome associations, based on which it evaluates the reward linked to different decision outcomes. Thus, behaviours driven by this system are conducted in favour of a specific reward or 'goal', the approach of which is flexibly influenced by internal or external state changes.

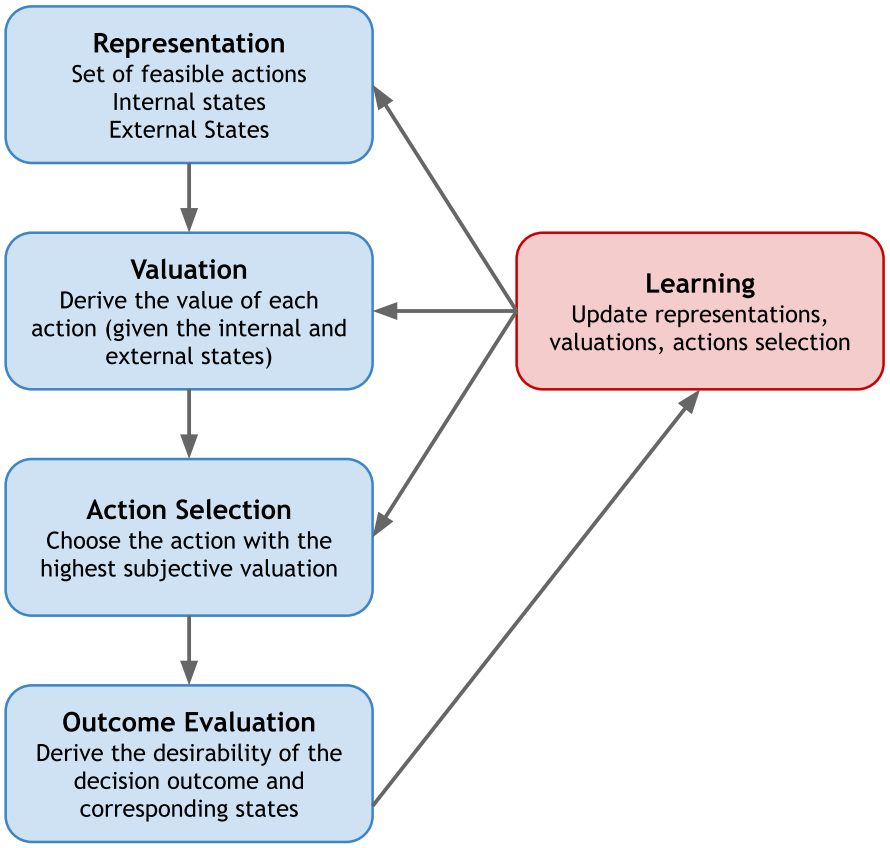

Figure 1-1. The decision making process. Rangel and colleagues (2008) broke down the value-based decision making process into five basic steps. First, the representations associated with the decision problem, including potential courses of action and related internal and external states, are generated. Next, value is assigned to these representations and a choice is made by selecting the action with the highest subjective value. Following the implementation of the decision, the desirability of the outcomes is evaluated, which is used to update the first three processes through learning to improve future decision making.

Whilst these systems differ in their respective speed-accuracy trade-offs, they can be recruited simultaneously to optimise behaviour and maintain flexibility. Consequently, these systems can assign distinct values to the same decision alternatives, which can occasionally lead to a conflict during action selection. For example, when considering a slice of cake, a person on a diet might be conflicted as the appetitive Pavlovian system assigns a high value to food, whilst the goal-directed system focused on weight loss pushes to avoid the cake. Thus, the co-existence of these systems can explain several of the inconsistencies we often see in choice.

Daw and colleagues (2005) made an analogous differentiation in value systems to explain decision discrepancies. They proposed that a model-free system is responsible for habitual or reflexive behaviours akin to the Pavlovian and habitual systems, and takes only previous choices into account when deriving value. A model-based system, which is comparable to the goal-directed system, assigns value in line with a learned model of the world, which includes knowledge of both internal and external states. The two systems compete to assign value for choice options and represent opposite ends of a cost-benefit trade-off. The model-free system sacrifices potential accuracy for computational simplicity and the model-based system demands more computational power to achieve higher levels of accuracy. In line with this dual system, Daw and colleagues showed that participants completed tasks using a mixture of these two systems.

Although different systems appear to assign value to choice options, I presume that primarily the habitual or model-free system is associated with behaviour in our experimental tasks given that participants had to make fast decisions driven by stimulus-outcome associations to maximise rewards. Consequently, in the remainder of this thesis, I will examine decision making and reward learning primarily in the habitual, model-free system.

1.2.2 What is subjective value?

To better understand the decision making process, it is important to interpret subjective value and its intricate relationship with learning. Studying

subjective value has been a focus of several disciplines, including philosophy, economics, and psychology. Philosophers have studied subjective value through the concept of pleasure, which can vary in the amount or intensity, without a focus on quantification. Economists, on the other hand, have used the term utility to describe the worth of goods and services. Bernoulli (1954) was the first to introduce the subjectiveness of utility through the notion of diminishing marginal utility – the phenomenon that additional units of the same good result in progressively smaller increases in subjective value. Although utility cannot be precisely quantified, economists have utilised the ranking of goods and services (ordinal utility) and the amount of effort or willingness to pay (cardinal utility) to indirectly measure it.

Psychologists have taken a behavioural approach to study subjective value and focused on the concept of reward – an experience with positive affect that influences behaviour (White, 2011). Thorndike's Law of Effect (1905) proposed that behaviours producing positive affect are more likely to be repeated and responses that produce discomfort are more likely to be avoided. Pavlov's (1927) formalisation of classical conditioning described automatic associations between specific behaviours and stimuli and laid the foundation for differentiating anticipated and experienced rewards. Rescorla and Wagner's (1972) reinforcement learning theory built on existent work on classical conditioning and described learning as the difference in expected and experienced rewards. Consequently, increased deviations between expected and experienced were considered to more significantly influence subsequent behaviour. Overall, subjective value over time has been increasingly understood as a reinforcer that promotes behavioural change and has become closely coupled with the notion of learning.

1.3 Reinforcement learning

Given that learning is closely intertwined with subjective valuation and decision making (Fig.1-1), understanding the latter processes requires the adequate characterisation of learning. As we have seen, reinforcement learning (RL; Rescorla and Wagner, 1972) posits that discrepancies between

expected and experienced rewards, termed reward prediction errors δ (RPEs), drive learning according to

$$\delta_k = r_k - V_k(s_k), \tag{1.1}$$

where r_k is the actual reward at trial *k* and V_k is the expected reward at trial *k* associated with stimulus *s* at trial *k*. Learning is then formalised by the Rescorla-Wagner-rule, given by

$$V_{k+1}(s_k) = V_k(s_k) + \beta \cdot \delta_k, \tag{1.2}$$

where learning is further influenced by the rate of learning β. Here, β controls the steepness of decay during learning; higher learning rates produce values more strongly weighted towards recent rewards. Rescorla and Wagner were not particularly concerned with the effect of the rate of learnig, which they considered to be constant, and focused primarily on how the difference between anticipated and actual rewards controls learning (Yau & MacNally, 2023).

However, Rescorla-Wagner-rule fails to explain some empirical findings, including choices with a sequential structure or generated by second-order conditioning, the phenomenon where neutral stimuli consistently paired with primary reinforcers predict rewards (Daw & Tobler, 2014). To account for such intra-trial influences on choices and learning, Sutton (1988) and Sutton and Barto (1990) introduced temporal difference (TD) learning as an extension of the Rescorla-Wagner model. In this framework, agents choose actions by considering the consequences of each action and make decisions that maximise aggregate reward in the long run. Thus, learning is guided by temporally successive predictions according to

$$V(s_t) = r(s_t) + E[V(s_{t+1})|s_t]. \tag{1.3}$$

Here, agents predict the value function *V* from a starting state s_t, where value is the sum of the rewards in each state s_{t+1} given the previous state s_t. TD

learning modifies the Rescorla-Wagner model based on how the prediction error is derived;

$$\delta_t = r_k + V(s_{t+1}) - V(s_t). \quad (1.4)$$

This temporal difference prediction error thus reflects agents' own predictions $V(s_t)$ and $V(s_{t+1})$ and is used to update the prediction $V(s_t)$ in Equation 1.3. If the temporal difference prediction error is positive, present reward expectations exceed those predicted in the previous state, and predicted value increases. Similarly, negative temporal difference signals that previous reward prediction was too optimistic and expected value drops. Consequently, unlike in the Rescorla-Wagner model, agents actively engage with their environment in TD learning and learn the values of different states over time.

Q-learning models (Watkins & Dayan, 1992) build on the idea of TD learning and directly estimate action values in the current state, rather than estimating the value linked to individual states. Consequently, by estimating the value of particular actions ('Q-values'), learning and choice become more compatible as Q-values directly inform decisions (Gureckis & Love, 2015).

With the implementation of the above ideas, computational models of RL have become increasingly complex and reliant on input from computer science to model adaptive action control. Importantly, reinforcement learning and neuroscience have mutually informed each other over the last three decades. This collaboration resulted in an improved understanding of how decision making is implemented in the brain as well as the development of new RL algorithms (Fan et al., 2023).

1.4 Neural underpinnings of reinforcement learning

The basal ganglia include several subcortical nuclei, most notably the striatum, that communicate with cortical and subcortical structures via excitatory ("Go") and inhibitory ("No-Go") pathways. This part of the brain is not only implicated in motor control but is also involved in many complex

cognitive processes, including decision making and reward learning. Three anatomical and functional basal ganglia-thalamo-cortical circuits have been proposed based on the connections with cortical regions and their activation pattern; oculo-motor, associative, and limbic (Lanciego et al., 2012). Within each circuit, structures from the cortex and the basal ganglia send and receive projections from each other, forming a cortico-basal ganglia-cortical loop. The oculo-motor network is associated with motor preparation and execution, the associative circuit is related to executive functions such working memory and cognitive flexibility, and the limbic loop is linked to emotional and reward responses. The latter network is associated with the ventral striatum and medio-dorsal thalamus, which receive projections from the orbitofrontal cortex (OFC) and the anterior cingulate cortex (ACC).

1.4.1 Dopamine

A large proportion of the medium spiny neurons of the striatum receive dopaminergic input from the midbrain, which has been associated with movement control, motivation, and reward (Gepshtein et al., 2013; Mazzoni et al., 2007). Most dopamine (DA) neurons in the brain can be found in the substantia nigra pars compacta and the ventral tegmental area (VTA), which send diffuse projections to the striatum, subcortical limbic structures, and (mainly frontal) cortical regions. DA can affect other neurons through two different modes of activity; phasic or tonic. Excitatory or inhibitory phasic activity marks transient DA release in response to specific events, whilst the tonic mode corresponds to overall, background DA activity.

DA neurons have a homogenous firing profile as they are electronically coupled to one another, which allows for strong inter-neuronal synchrony (Vandecasteele et al., 2005). The wide-spread dopaminergic projections across the brain and the homogenous firing profile of DA neurons makes them an ideal candidate to signal a single numerical value such as a reward prediction error (Daw & Tobler, 2014). Although target structures receive the same response from DA neurons, this signal can have distinct effects on different regions in the brain depending on the different DA release and

reuptake characteristics, varying effects of DA on receptors, neuronal types and circuits, or additional input to the region (Schultz, 2007).

Schultz and colleagues (1997) were the first to show that dopaminergic neurons in the macaque VTA increased their phasic firing rates in response to unpredicted reward delivery and cues predicting rewards but not to predicted reward delivery. Furthermore, when a predictive cue appeared without a subsequent reward delivery, DA neurons decreased their phasic firing rates. These observations provided evidence that phasic activity in dopaminergic neurons in the VTA scale with the Rescorla-Wagner RPEs. This finding has since been consistently replicated in a variety of electrophysiological studies in different species (Bayer & Glimcher, 2005 Bayer et al., 2007; Cohen et al., 2012; Matsumoto & Hikosaka, 2009; Roesch et al., 2007). Human functional magnetic resonance imaging (fMRI) experiments confirmed these conclusions as blood oxygenation level dependent (BOLD) response in the striatum was found to reflect Rescorla-Wagner RPEs (O'Doherty et al., 2003; Tobler et al., 2006; 2007).

Furthermore, the firing rate of DA neurons is consistent with the temporal difference RPE (Eq.1.4) as higher reward probabilities elicited a stronger dopaminergic response (Fiorillo et al., 2003). Similarly, temporal difference learning consistently explains dopaminergic responses to both unexpected rewards and secondary reinforcers that provide novel information about rewards and therefore produce a reward prediction error (Daw et al., 2011; Seymour et al., 2004; Tobler et al., 2006; Waelti et al., 2001).

In line with Q-learning models that explain learning through action-outcome associations, dopamine has been associated not only with RPEs, but also human choice behaviour (Cools et al., 2006; Frank et al., 2004; Seymour et al., 2012) and action-related decision variables (Samajima et al., 2005). Animal studies provided causal evidence for the role of DA in coding action values and promoting choices; animals have been found to prefer locations where their DA neurons were activated (Tsai et al., 2009) and avoid those where their DA neurons were inhibited (Tan et al., 2012). Additionally,

dopamine-enhancing medication treating the motor deficits in Parkinson's disease have been associated with both improved motor control and effects on complex cognitive functions (Cools et al., 2001; 2003; Frank et al., 2004). Although DA does not determine movements, tonic DA levels in the dorsal striatum appear to directly modulate response vigour, which determines the speed and effort related to actions (Niv et al., 2006; 2007). Dopamine may provide such implicit motivation for movement by coding action energy costs (Gepshtein et al., 2013; Mazzoni et al., 2007), with higher costs associated with more rapid actions. Overall, reinforcement learning and response vigour appear to be naturally interconnected through motivation and dopamine; dopamine provides both explicit motivation for behaviour aimed at obtaining rewards via phasic firing and implicit motivation for movement by encoding response vigour via tonic activity.

1.4.2 Ventromedial prefrontal cortex

Apart from the dopaminergic system and the striatum, the ventromedial prefrontal cortex (vmPFC) is another crucial component of the brain's valuation system. Activity within both the striatum and the vmPFC have been consistently found to predict virtually all types of human preferences and choices (for a meta-analysis, see Levy & Glimcher, 2012), including consumable rewards, monetary and social rewards, gains, losses, and even abstract rewards. These results imply that the valuation system encodes all reward types in a shared neural currency, termed subjective value signals. Both single cell recording in animals and human fMRI studies suggest that several cortical areas contribute to these value signals (Glimcher, 2014). For example, the dorsolateral prefrontal cortex provides valuation for social cooperation and rewards requiring self-control (Kim et al., 2008; Knoch et al., 2006), the orbitofrontal cortex appears to be critical for valuation stemming from consumable rewards (O'Doherty et al., 2001; Padoa-Schioppa & Assad, 2006), whilst the amygdala provides input for affective valuation (De Martino et al., 2010; Sokol-Hessner et al., 2012).

Overall, dopaminergic teaching signals have been consistently found to modulate existing value representations in both the striatum and the vmPFC by altering synaptic strength (Wickens et al., 2007). However, it is still unclear how these value representations are exactly mapped onto the vmPFC or broadcasted to parietal regions for comparison. Furthermore, apart from dopamine, other neurotransmitter systems, such as noradrenaline and serotonin, are known to project onto and receive input from similar prefrontal regions (Briand et al., 2007), and are often released simultaneously (Briand et al., 2007; Devoto & Flore, 2006). Although some progress has been made to investigate how these different systems impact adaptive learning (Bouret & Sara, 2005; Cool et al., 2008; Homberg, 2012; Yu & Dayan, 2005), our current understanding of these mechanisms is still greatly limited.

1.5 Components of reinforcement learning

As we have seen, there is strong interdisciplinary evidence that dopamine encodes RPEs and that a reinforcement learning rule nudges the prediction of stimulus values in a direction that decreases RPEs. Importantly, the rate of value updating in reinforcement learning depends on the product of the RPE and the learning rate. The learning rate determines the extent to which stimulus value is modified following decision feedback and falls between 0 and 1. Higher learning rates indicate faster learning, whereby recent outcomes are weighed more strongly, whilst lower learning rates take into account decision outcomes for a longer period of time. Unlike RPEs, learning rates are not independent of context, and can be influenced by a variety of factors, including the level of uncertainty and surprise or outcome valence (positive or negative). The exact mechanisms in which these factors guide reward learning in the brain are unclear and can be difficult to decouple due to the strong interdependence of these processes and technical difficulties in measuring (especially subcortical) neural signals with adequate temporal and spatial resolution in vivo. Nevertheless, recent studies using a variety of research methods were able to shed some light on the complex processes influencing learning.

1.5.1 Uncertainty

In real life, we are often not aware of the true reward contingency associated with a specific stimulus as decision outcomes are frequently uncertain and can fluctuate over repeated decisions. For adaptive learning and decision making, it is crucial not to significantly modify behaviour in response to outcomes that are representative of the typical outcome range. At the same time, behaviour should be substantially adjusted in response to outcomes that signal a change in reward contingencies, either with regards to the magnitude, probability, or delay associated with rewards. Thus, effective learning necessitates differentiation between inconsequential variability in outcomes due to the probabilistic nature of rewards, termed 'expected uncertainty' or 'risk', and environmental changes, labelled 'unexpected uncertainty'.

Expected uncertainty has been related to the variance of RPEs in different computational frameworks. Preuschoff and Bossaerts (2007) incorporated risk prediction into the Rescorla-Wagner learning rule by scaling RPEs based on the covariance between optimal predictions and past RPE changes. Soltani and Izquierdo (2019) built on this idea to account for both types of uncertainty as well as their interaction in reward learning. Unlike others (Behrens et al., 2007; Diederen & Schultz, 2015; Preuschoff & Bossaerts, 2007), they assumed that unexpected uncertainty is inherently subjective as uncertainty can only be truly unexpected if it is unpredictable, whilst environmental volatility indexes the true rate of environmental change. In this framework, expected uncertainty can be estimated by averaging the unsigned RPEs (absolute value of the RPEs) over trials to approximate variability in outcomes. Expected uncertainty thus provides a baseline level of variability against which the degree of surprise can be evaluated. Surprising, unexpected events therefore arise when RPEs exceed this baseline level and directly increase the learning rate to weigh recent RPEs more strongly.

In the Bayesian framework, overall uncertainty associated with a stimulus can be quantified by assigning a probability to each possible value being the

correct one conditional on the previous stimuli and rewards (Daw, 2014). Following Bayes' rule, the mean and variance of stimulus values are updated given each new outcome. This way, uncertainty is a distribution, the variance of which characterises the spread of belief, i.e., uncertainty. In an environment with stable reward contingencies, uncertainty is initially high but declines as agents collect more information via feedback, which leads to a decaying learning rate over time. An environment with changing reward probabilities can be modelled via the Kalman filter (Daw, 2014), which includes an additional uncertainty term accounting for the possibility that reward contingencies have changed between trials. This essentially means that agents are never fully certain about stimulus values. Moreover, the faster the environment is changing (higher volatility), the less relevant past experiences become, resulting in higher learning rates to assign more weight to recent observations.

These different mathematical formulations accurately capture the phenomenon that learning rates decay over time in stable environments (Daw, 2014; Diederen & Schultz, 2015) but increase as a function of environmental change (Behrens et al., 2007, Browning et al., 2015; Palminteri et al., 2017). However, it is unclear which formulation best approximates the neural computations that implement uncertainty processing in the brain. Evidence indicates that dopaminergic activity in the striatum is scaled by the degree of risk, i.e., the range of rewards expected (Fiorillo et al., 2003; Preuschoff & Bossaerts, 2007; Tobler et al., 2005). This suggests that DA encodes RPEs scaled by the learning rate rather than the raw RPE. Nevertheless, it remains unclear how the brain exactly tracks risk prediction. Evidence from both animal and human studies indicate that the anterior insula (Ishii et al., 2012; Jo & Jung, 2015; Kuhnen & Knutson, 2005; Preuschoff et al., 2008) and the OFC (Jo & Jung, 2015; Mobini et al., 2002; Tobler et al., 2007) play a crucial role in the computation and representation of risk. At the same time, the ACC has been implicated in the processing of both expected and unexpected uncertainty (Behrens et al., 2007; Hayden et al., 2011; Hyman et al., 2017; Monosov, 2017), whilst the mediodorsal thalamus (Chakraborty et al., 2016;

Parnaudeau et al., 2015) has been linked to learning under unexpected uncertainty.

To reconcile the various neural cortico-limbic correlates of uncertainty processing, Soltani and Izquierdo (2019) proposed a distributed network processing uncertainty, including the OFC, ACC, striatum, amygdala, and hippocampus. They contemplated that the OFC and the ACC may provide complementary inputs for the computation of unexpected uncertainty. Considering that the OFC is involved in estimating value representation and expected uncertainty, it could determine a baseline for computing unexpected uncertainty. At the same time, the ACC could provide a spectrum of learning rates that may be utilised to adjust synaptic plasticity, while the hippocampus may supply cognitive maps with predictions about outcomes. These signals are in turn proposed to be integrated in the amygdala to compute unexpected uncertainty (Soltani &Izquierdo, 2019). Indeed, the amygdala has reciprocal connections with the OFC, ACC, and the dopaminergic system and has been associated with detecting surprise in the internal and external environment (Wassum & Izquierdo, 2015). The mediodorsal thalamus is thought to support rapid value updating due to its association with learning in changing environments (Chakraborty et al., 2016; Parnaudeau et al., 2015). Finally, in line with the role of the striatum in value updating and vigour (see section 1.4.1), converging input about value and uncertainty estimates from the above structures are suggested to be integrated here to implement flexible learning. Although this proposed network specification is promising, it is solely based on research in the appetitive domain, necessitating further research to determine whether uncertainty is processed similarly in the brain independently of valence. Moreover, research examining the timing of neural activity in each of the above structures is crucial to confirm their proposed role in uncertainty processing.

1.5.2 Surprise

Most studies on value-based decision making have associated outcome surprise (or salience) with the absolute difference between experienced and expected reward. Consequently, surprise has been quantified as the unsigned RPE, without a simultaneous consideration of environmental uncertainty. Intuitively, unexpected outcomes can result from the inherent probabilistic nature of rewards or from environmental change, in turn signalling increased expected or unexpected uncertainty. At the same time, RPEs of the same magnitude could be perceived as more surprising when uncertainty about stimulus-outcome associations is low. This line of thought, where surprise is subjective and dependent on outcome variability, has been recently formalised by Soltani and Izquierdo (2019). In this framework, surprising events emerge when RPEs exceed a baseline level, quantified by the mean of past RPEs. Surprise in turn increases the learning rate to place more weight on recent outcomes when updating existing value representations. Indeed, both higher environmental uncertainty (Behrens et al., 2007, Browning et al., 2015; Palminteri et al., 2017) and surprise (den Ouden et al., 2012; Niv et al., 2015; Roesch et al., 2012) have been linked to elevated learning rates, suggesting that these processes may be interrelated.

Recent neural evidence indicates that dopaminergic neurons show a two-component phasic activation response, where a short-latency initial response codes physical and motivational surprise generating stimulus-driven attention, followed by a longer-lasting response associated with the reward prediction error (Schultz et al., 2017). Similarly, other studies found that outcome surprise is encoded in a distributed network of structures independent of outcome valence processing (Fouragnan et al., 2017; 2018; Metereau & Dreher, 2012). The brain structures consistently indicated to encode surprise across these studies include the ACC, striatum, anterior insula, and the midbrain. Other studies on surprise processing also frequently indicated the ACC (Lin et al., 2012; Vassena et al., 2020), the anterior insula (Loued-Khenissi et al., 2020; Preuschoff et al., 2008), the amygdala (Preuschoff & Bossaerts, 2007; Roesch et al., 2012), and prefrontal regions

(Fouragnan et al., 2017; 2018; Vassena et al., 2020) in processing surprise. The close overlap in the neural structures encoding surprise and uncertainty provides further support to the notion that surprise and uncertainty are inherently intertwined, and in turn necessitates the consideration of uncertainty when investigating the behavioural and neural underpinnings of surprise.

1.6.3 Valence

The categorical valence of outcomes signals whether an outcome is better or worse than expected and can be quantified as the signed RPE. Valence thus determines the direction in which learning will take place; negative RPEs induce a downward adjustment in stimulus values and positive RPEs lead to an increase in stimulus-value representations. Interestingly, a wide range of behavioural and neural evidence indicates that positive and negative feedback is processed differently in the brain.

Studies using instrumental learning tasks (den Ouden et al., 2013; Frank et al., 2007; Lefebvre et al., 2017; Niv et al., 2012; Palminteri et al., 2017) indicated differential learning associated with positive and negative outcomes, with increased learning linked to positive compared to negative feedback, evidenced by higher positive than negative learning rates. However, it is unclear how this learning bias is linked to performance, with contradictory results across studies (Harada, 2020; Lefebvre et al., 2017; Palminteri et al., 2017). Negative feedback has also been found to decrease response times over the next trial, termed post-error slowing, associated with increased caution following errors (Dutilh et al., 2012; Fievez et al., 2022; Goldfarb et al., 2012; Hofling et al., 2018).

The neural underpinnings mediating these differential valence-specific effects are not yet fully understood. The dopaminergic system is known to increase and decrease its phasic firing activity following better-than-expected and worse-than-expected (resulting from either the omission of reward or an aversive outcome) outcomes, respectively (Schultz et al., 1997). A recent fMRI meta-analysis (Fouragnan et al., 2018) showed that the brain regions

associated with increased positive compared to negative feedback processing included prominent structures linked to value representation and reward learning, including the ventral striatum, vmPFC, OFC, and the VTA. At the same time, the ACC, anterior insula, thalamus, amygdala, and motor structures were associated with a greater BOLD response for negative compared to positive feedback.

Indeed, the ACC has been frequently implicated in negative feedback processing (Bush et al., 2002; Daniel & Pollmann, 2010; Ullsperger & von Cramon, 2003;) and is thought to primarily contribute to the feedback-related negativity event-related potential (ERP) reported in response to negative feedback (Hajcak et al., 2007; Holroyd et al., 2004). However, after balancing outcome probability and valence type, the ACC was found to signal unexpectedness independently of outcome valence (Ferdinand & Opitz, 2014; Ferdinand et al., 2012). In most learning paradigms, negative outcomes tend to occur more unexpectedly as positive outcomes are increasingly expected with learning. As such, negative feedback is confounded with uncertainty in many reward learning tasks, which explains the tight overlap between brain structures indicated in uncertainty (Soltani & Izquierdo, 2019) and negative feedback processing (Fouragnan et al., 2018).

Whilst dopamine has been consistently indicated in positive and negative feedback processing, the role of other neurotransmitter systems is less clear. Seminal work by Schultz et al. (1997) demonstrated that DA neurons increase and decrease their phasic firing activity following better-than-expected and worse-than-expected outcomes, respectively. However, DA neurons have a limited firing range to signal unexpected negative events (Daw, 2014), implying that other systems may be involved in processing (un)expectedness and negative outcomes.

Pharmacological studies on serotonin found that serotonin improved reversal learning by diminishing perseverance behaviour (Bari et al., 2010; Cools et al., 2008; Evers et al., 2005; Homberg, 2012). Consequently, reviews on serotonin's role in decision making proposed that it could modulate vigilance

(Homberg, 2012) and promote adaptive learning by increasing sensitivity to punishments and inhibiting punished behaviours that were previously rewarded (Cools et al., 2008; Homberg, 2012). As different neurotransmitter systems could be activated by the same event and project to similar cortical structures, it is difficult to disentangle their unique contributions to learning and decision making (Briand et al., 2007; Devoto & Flore, 2006). Additionally, these systems are also likely to be activated in response to activity in other systems, e.g., serotonergic activity has been found to modulate the dopaminergic system (Homberg, 2012). In the next section, I will review the growing body of research implicating noradrenaline in decision making and reward learning.

1.6 Noradrenaline in reward learning

Noradrenaline (NA) is primarily synthesised in the locus coeruleus (LC) of the brainstem, with widely distributed, ascending projections to the cortex. NA has been primarily implicated in regulating arousal (Berridge & Waterhouse, 2003), although recent evidence indicates this system to be involved in more intricate cognitive processes. Specifically, increased locus-coeruleus-noradrenaline (LC-NA) system activity has been linked to higher levels of uncertainty and surprise (Colizoli et al., 2018; de Gee et al., 2014; Filipowicz et al., 2020; Lavín et al., 2014; Nassar et al., 2012; Preuschoff et al., 2011; Urai et al., 2017; van Slooten et al., 2017; 2018), exploration (Aston-Jones & Cohen, 2005; Jepma & Nieuwenhuis, 2011; Pajkossy et al., 2017), a reduction in choice bias (Krishnamurty et al., 2017; Urai et al., 2017), attentional set shifting (Jepma & Nieuwenhuis, 2011; Pajkossy et al., 2017), enhanced information processing (de Gee et al., 2017; Joshi & Gold 2020; Zenon, 2019), boosted behavioural flexibility (Eldar et al., 2013), and increased reversal probability in reward learning tasks (Jepma & Nieuwenhuis, 2011; Nassar et al., 2012).

These results are congruent with the proposed role of NA in signalling unexpected uncertainty (Dayan & Yu, 2006; Yu & Dayan, 2005). As discussed above, uncertainty is closely connected to surprise (Soltani & Izquierdo, 2019)

and the rate of learning (Behrens et al., 2007, Browning et al., 2015; Palminteri et al., 2017), is often confounded with negative outcome processing (Ferdinand & Opitz, 2014; Ferdinand et al., 2012), and uncertainty reduction has been shown to underlie motivation for explorative behaviour (Cavanagh et al., 2012; Cockburn et al., 2022; Gershman, 2019; Walker et al., 2022; Wilson et al., 2014).

Non-invasive, direct neural recording of LC-NA activity in humans is difficult to obtain. Consequently, the majority of the above research investigating the link between the LC-NA system and cognition did not directly measure noradrenergic activity but utilised pupillometry as a proxy for LC-NA activity. Indeed, studies with both monkey (Joshi et al., 2016; Joshi & Gold) and human (Murphy et al., 2014; Reimer et al., 2016) participants indicated that rapid pupil dilations are associated with phasic LC-NA activity. Given the considerable evidence that links the LC-NA system to the pupil dilation and the relative ease of obtaining pupillometry data, an increasing number of research has utilised the pupil response as a proxy for LC-NA activity. However, it is important to keep in mind that other brainstem nuclei, including dopaminergic, cholinergic, and serotonergic structures may partly control this pupil effect (de Gee et al., 2017; Reimer et al., 2016; Urai et al., 2017; van Slooten et al., 2018), and more research is needed to disentangle the unique contribution of each system.

1.6.1 Network reset hypothesis

Successful adaptation to environmental changes depends on the down-regulation of learned influences and increasing the relevance of novel information. The LC-NA system has been proposed to contribute to task-relevant behavioural adaptation in response to unexpected events and increased uncertainty through the implementation of cortical network resets (Fig. 1-2; Bouret & Sara, 2005, Dayan & Yu, 2006; de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009; Urai et al., 2017). Such noradrenaline-induced network resets are suggested to prompt task-relevant cortical network reorganisation in order to reduce top-down influences and

increase the relevance of newly acquired information. High phasic LC firing is thought to promote, whilst low phasic activity is considered to inhibit neural shifts in response to external events, consequently facilitating behavioural adaptation (Bouret & Sara, 2005). In order for the LC to carry out such network resets, it requires access to both bottom-up and top-down information (Filipowicz et al., 2020). The ACC is a prime candidate for supplying these inputs to the LC (Aston-Jones & Cohen, 2005; de Gee et al., 2017) given its strong reciprocal connection with the LC (Briand et al., 2007; Joshi & Gold, 2020) and its access to both top-down and bottom-up information due to its robust ties with prefrontal and sensorimotor areas.

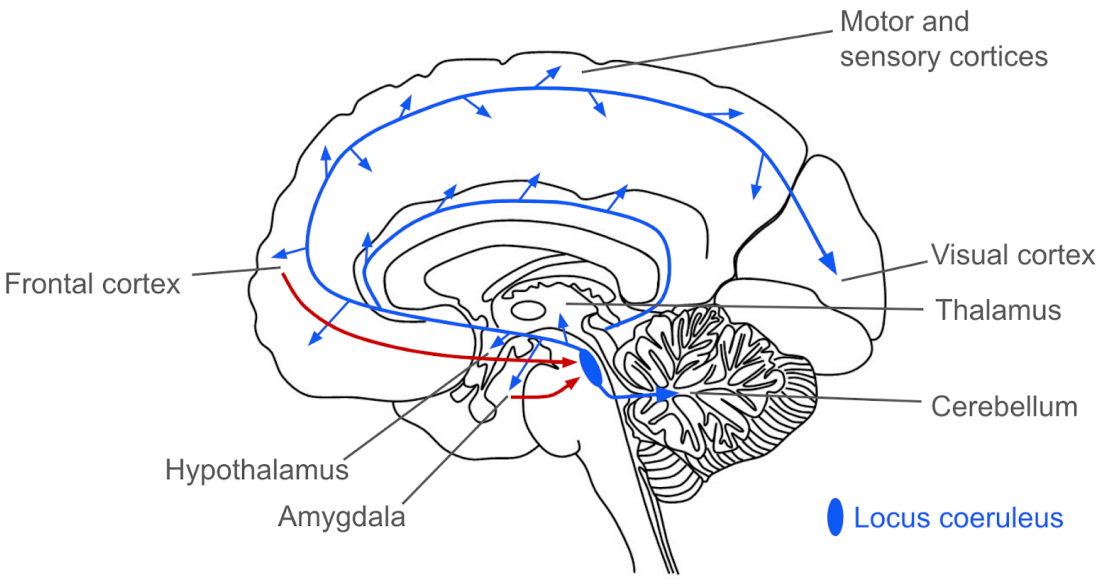

Figure 1-2. LC-NA projections. The LC sends diverse connections (blue arrows) to virtually all parts of the brain, which makes it an ideal candidate for implementing cortical network resets in different neural networks. At the same time, it receives input (red arrows) from a limited number of structures, including the amygdala and the prefrontal cortex, which can supply the LC with task-relevant information.

Consistent with this interpretation, Fouragnan and colleagues (2015) found two spatiotemporally distinct networks implicated in probabilistic reversal learning. An 'early' network was linked to arousal and motor preparation, including the ACC, centromedial thalamus, and motor regions, whilst a 'late' network was functionally and structurally consistent with the dopaminergic valuation system. Consistent with the role of the dopaminergic system in RPE signalling, late system activity scaled with trial-wise value updating independently of feedback type, whilst activity in the early system was mainly triggered by negative outcomes. Furthermore, in response to negative feedback, the early system down-regulated the late system through a thalamo-striatal coupling, the strength of which predicted choices and avoidance learning. Early system structures, including the thalamus, ACC, and motor regions, overlap with ascending projections sites of the LC-NA network (Benarroch, 2017; Sara & Bouret, 2012), suggesting that the LC may regulate early system activity. Such an interaction between the early system and the LC-NA network is compatible with the LC-NA network reset hypothesis (Bouret & Sara, 2005; Dayan & Yu, 2006; de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009). Specifically, negative feedback likely increases environmental uncertainty estimates, necessitating the down-regulation of learned reward contingency representations in the reward network. Such a reset in cortical value representations would presumably decelerate subsequent decision making as evidence in favour of a change in reward contingencies is accumulated, in turn increasing exploration propensity.

It is important to keep in mind that the strong interdependence across neurotransmitter systems makes it difficult to disentangle the unique contribution of each system. For example, the LC receives input from the serotonergic raphe nuclei and projects to the dopaminergic VTA (Benarroch, 2017). Furthermore, there is no in vivo, non-invasive neural recording method that has both adequate temporal and spatial resolution to unequivocally characterise the neural activity linked to the complex cognitive processes that occur at a fast timescale. Consequently, it is precarious to measure neural signals from subcortical nuclei that contain the majority of dopaminergic,

noradrenergic, or serotonergic cell bodies in humans. Additionally, increasing evidence indicates the involvement of both the dopaminergic and noradrenergic systems in mediating exploration (Aston-Jones & Cohen, 2005; Chakroun et al., 2020; Cinotti et al., 2019; Frank et al., 2009; Jepma & Nieuwenhuis, 2011; Pajkossy et al., 2017; van Slooten et al., 2019), pointing to the importance of across-system regulation of behaviour. Thus, a focus on between-system interaction is likely necessary to reveal the more intricate mechanisms involved in the complex cognitive processes implicated in value learning and decision making.

1.7 Aims of the thesis

The main goal of the current thesis is to contribute to disentangling the behavioural and neural underpinnings of valence-based learning mechanisms in humans. Considering the significance of uncertainty in the reward learning process, we considered the link between uncertainty and valence-specific outcome processing in our analyses.

Our first study, presented in Chapter 2, was designed to explore whether the early reward learning system (Fouragnan et al., 2015) may be under noradrenergic control. Specifically, we utilised multivariate electroencephalogram (EEG) analysis to replicate the early and late reward learning systems during probabilistic reversal learning. By using the pupil response as a proxy for LC-NA activity (Joshi et al., 2016; Joshi & Gold, 2020; Murphy et al., 2014; Reimer et al., 2016), we showed that the increased feedback-evoked pupil response for negative compared to positive outcomes is exclusively driven by negative feedback encoding. In line with the proposed role of the LC-NA system in mediating uncertainty and implementing cortical network resets (Bouret & Sara, 2005; Dayan Yu, 2006; de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009), we found that the increased coupling between the early system and the phasic pupil response was linked to increased uncertainty and exploration tendency.

Our second study (Chapter 3) extends the work introduced in Chapter 2. We hypothesised that LC-NA-induced cortical resets in reward learning structures

would reduce evidence accumulation in subsequent decisions as evidence is accumulated towards a reversal in reward contingencies. To test this notion, we used drift diffusion modelling (Ratcliff, 1978; Ratcliff & McKoon, 2008) to explore how single-trial EEG-derived regressors indicating the degree of positive versus negative feedback encoding affected post-feedback response adaptation. In line with the LC-NA network reset hypothesis, we found that increased negative feedback encoding in both the early and late systems reduced evidence accumulation on the next trial. To our knowledge, this is the first study utilising neurally informed sequential sampling modelling to characterise post-feedback response adaptation in reward learning.

Finally, Chapter 4 presents our last study, which was born out of a collaboration with Dr Andrea Kóbor and Dr Eszter Tóth-Fáber at the HUN-REN Research Centre for Natural Sciences. In this study, we utilised the Balloon Analogue Risk Task (BART; Lejuez et al., 2002) to investigate whether humans preferentially learn from positive compared to negative outcomes in a more complex and ecologically valid experimental paradigm with different levels of uncertainty. We applied the newly developed Scaled Target Learning model (STL; Zhou et al., 2021), and its extension with decay (STL-D), to our data to derive differential learning rates as a means to quantify learning bias. For the first time, we report a learning bias in the BART in conditions with increased levels of uncertainty. Furthermore, this bias only showed a negative association with task performance only under the highest level of uncertainty. Moreover, learning bias was positively related to risk-seeking propensity under all levels of uncertainty, implying that it may play a crucial role in shaping risk preferences.

Overall, the three experimental chapters of this thesis provide an improved characterisation of the neural and behavioural underpinnings of valence-specific decision outcome processing, whilst considering the important modulatory role of uncertainty on these mechanisms.

Chapter 2. Pupil modulation of early reward learning system

2.1 Summary

The ability to learn to reinforce rewarding actions and avoid repeated mistakes is crucial for survival in dynamic environments. Yet, the exact neural mechanisms and neurotransmitter systems implementing reward-based decisions remain undetermined. In the current study, we obtained simultaneous EEG and pupillometry data from 48 participants engaging in a probabilistic reward learning task with reversals. By exploiting the single-trial variability in the EEG signal, we replicated the previously established two spatiotemporally distinct systems implicated in reward learning (Fouragnan et al., 2015). Moreover, by using the pupil response as a proxy for locus coeruleus-noradrenaline activity, we show that the feedback-evoked pupil response difference between positive and negative feedback trials is exclusively driven by negative feedback encoding, most likely originating in the early system. Increased coupling between the feedback-evoked pupil response and the early, not the late, component following negative outcomes was linked to increased uncertainty and exploration tendency as well as reduced accuracy. Consistently with previous research indicating the noradrenergic system in uncertainty signalling and network resets, we propose that when internal estimates of contextual uncertainty are high following negative feedback, the early system, regulated by locus coeruleus activity, implements a network reset in reward learning structures of the late system. This interruption may simultaneously increase the neural gain related to the processing of novel information and decrease the influence of existing representations in reward learning structures, in turn improving performance by creating new, more accurate internal representations of the external world.

2.2 Introduction

Choosing actions that maximise value amongst alternatives is crucial for the survival of an organism. During the value-based decision making process, cost and benefit value representations of competing actions are integrated into action values, serving the basis of action selection. Upon feedback, the organism must consider the extent to which the outcome resulting from the chosen action was desirable, which outcome evaluation process drives reward learning. Yet, the exact neural mechanisms, including the role of different neurotransmitter systems, involved in reward learning remain elusive.

2.2.1 Spatiotemporally distinct systems in reward learning

Considerable empirical evidence suggests the involvement of the dopaminergic system in reinforcement learning, which activates or suppresses the brain's reward network in response to positive or negative decision outcomes (Fouragnan et al., 2015; Schultz et al., 1997). This network includes structures such as the substantia nigra, ventral tegmental area, striatum, nucleus accumbens, amygdala, dorsal posterior cingulate cortex, and ventromedial prefrontal cortex (D'Ardenne et al., 2008; Fouragnan et al., 2015; Gläscher, et al.,, 2008; Matsumoto & Hikosaka, 2009; O'Doherty et al., 2007; Schultz, et al., 1997; Zaghloul et al., 2012).

In addition to this well-established dopaminergic network, Fouragnan and colleagues (2015) suggested that two spatiotemporally distinct but interacting value systems encode decision outcomes. Specifically, linear discriminant analysis on the single-trial feedback-locked EEG data revealed two temporally distinct components discriminating between positive and negative feedback. An 'early' component peaked on average at 219 ms following outcomes presentation, while a 'late' component peaked at 319 ms after feedback. Furthermore, the mapping of single-trial electroencephalogram (EEG) responses onto the functional magnetic resonance imaging (fMRI) activations associated with positive versus negative outcome value allowed for the

successful assignment of temporal order to relevant blood oxygenation level dependent (BOLD) responses. This analysis revealed the early system to be mainly activated by negative outcomes and engage brain structures related to arousal and motor preparation, such as clusters in the anterior medial cingulate cortex, centromedial thalamus, and neighbouring premotor regions. The late system was found to be functionally and structurally consistent with the dopaminergic system; it was engaged by both positive and negative outcomes, in response to which activity in dopaminergic structures (striatum and vmPFC) as well as in the amygdala and posterior cingulate cortex increased or decreased, respectively. Late system activity also reliably predicted the degree of value updating, consistent with the role of the dopaminergic system in reward learning.

The early system was further found to down-regulate the late system; as thalamic activity in the early system increased in response to negative outcomes, striatal activity in the late system reduced. Additionally, the strength of this coupling predicted participants' switching behaviour and avoidance learning. It is worth noting that the encoding of outcome valence and salience, defined in terms of the magnitude of the absolute reward prediction errors, was spatially, but not temporally, distinct (Fouragnan et al, 2017; 2018). Whilst the valence processing system correlated primarily with the reward network promoting approach or avoidance learning, the salience processing network overlapped with the human attentional network determining the speed of learning.

Taken together, these results indicate that beyond the well-established dopaminergic system, an additional early system is implicated in outcome value processing. Consistently with previous reports that indicated the thalamo-striatal pathway in avoidance control (Kerns et al., 2004; Minamimoto et al., 2005; Seifert et al., 2011), Fouragnan and colleagues speculated that the early network provides a fast alertness signal in response to undesirable outcomes to facilitate behavioural adjustment towards alternative options. As structures of the early system, including the thalamus, ACC, and neighbouring motor regions, overlap with ascending projections sites

of the locus-coeruleus-noradrenaline (LC-NA) system (Sara & Bouret, 2012), we hypothesised that the LC-NA network might play a role in regulating the activity of the early system.

2.2.2 The role of the LC-NA system in reward learning

The LC-NA system has been implicated in shaping global cognition through wide-spread cortical and subcortical norepinephrine release (Aston-Jones & Cohen, 2005; Bouret & Sara, 2005; Dayan & Yu, 2006; de Gee et al., 2017; Maness et al., 2022; Yu & Dayan, 2005). As pupil diameter has been found to be a reliable peripheral marker of LC-NA-driven arousal state (Joshi et al., 2016; Joshi & Gold, 2020; Murphy et al., 2014; Reimer et al., 2016), it has been increasingly used to investigate noradrenergic involvement in the decision making process. Specifically, increased pupil dilation and LC-NA activity have been associated with increased levels of uncertainty and surprise (Colizoli et al., 2018; de Gee et al., 2014; Filipowicz et al., 2020; Lavín et al., 2014; Nassar et al., 2012; Preuschoff et al., 2011; Urai et al., 2017; van Slooten et al., 2017; 2018), exploration (Aston-Jones & Cohen, 2005; Jepma & Nieuwenhuis, 2011; Pajkossy et al., 2017), a reduction in choice bias (Krishnamurty et al., 2017; Urai et al., 2017), attentional set shifting (Jepma & Nieuwenhuis, 2011; Pajkossy et al., 2017), enhanced information processing (de Gee et al., 2017; Joshi & Gold 2020; Zenon, 2019), boosted behavioural flexibility (Eldar et al., 2013), and signalling change point probability in reward learning tasks (Jepma & Nieuwenhuis, 2011; Nassar et al., 2012).

These results are consistent with the proposed role of noradrenaline (NA) in signalling unexpected uncertainty, i.e., gross changes in the environment that violate prior reward-contingency beliefs (Dayan & Yu, 2006; Yu & Dayan, 2005). Efficient adaptation to contextual changes requires the down-regulation of habitual, top-down influences and enhancing the relevance of newly acquired information. Phasic NA release has been proposed to facilitate such widespread network resets by triggering task-relevant cortical network reorganisation as a form of an internal

interrupt signal (Bouret & Sara, 2005; Dayan & Yu, 2006; de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009).

By using pupil dilation as a proxy for LC-NA activity (Joshi et al., 2016; Joshi & Gold, 2020; Murphy et al., 2014; Reimer et al., 2016), the current study utilised pupillometry and EEG recordings to investigate whether activity in the LC-NA and early systems are systematically correlated. Specifically, we expected the feedback-evoked pupil response to be exclusively driven by the trial-by-trial variations in early system activity induced by negative outcomes. In line with previous research linking the LC-NA system to cortical resets, we hypothesised that increased phasic noradrenergic activity following negative outcomes, mirrored by increased pupil dilation, signals the early system to down-regulate late system activity by introducing a cortical reset. As these LC-mediated cortical resets presumably attempt to increase the neural gain related to the processing of new observations by reducing the influence of top-down, learnt value representations, we hypothesised that they signal an increased probability of reward contingency reversals. Consequently, we expected increased early system activity, driven by noradrenergic activity attempting to adapt to contextual changes, to be associated with an increased likelihood of choosing the option that was previously considered to be inferior (i.e., exploring) as well as to decelerate subsequent decision making as evidence is accumulated in favour of new reward contingency representations (i.e., reversals). Finally, consistent with the role of the dopaminergic system in reward learning and the association of late system activity with value updating (Fouragnan et al., 2015), we expected activity predominantly in the late, rather than the early, system to drive the value updating process.

2.2.3 Current study

To test the hypothesis that the early system covaries systematically with the noradrenergic system, we used the pupil diameter as a proxy for LC-NE activity. Specifically, we implemented a probabilistic reversal learning task, during which we collected simultaneous EEG and pupillometry data. Using

linear discriminant analysis (LDA) on the feedback-locked EEG data, we replicated the two spatiotemporally distinct valence systems implicated in reward learning (Fouragnan et al., 2015).

As expected, the evoked pupil response was significantly higher for negative compared to positive outcomes. To confirm the driving force behind this feedback-induced pupil response, we built participant-specific regressions that utilised the single-trial discriminant amplitudes of the feedback-locked EEG data to predict the full course pupil data. This analysis demonstrated that the more negatively an outcome is encoded by the early and the late systems, the larger feedback-related pupil response is produced. Furthermore, a stronger coupling between the feedback-induced pupil response and the early, but not the late, system following negative outcomes was associated with an increased level of decision uncertainty, higher exploration propensity, and reduced task performance. These results support our hypothesis that negative feedback elevates internal estimates of contextual uncertainty (reversal probability), signalled to the early system by increased LC-NA activity, which consequently disrupts processing in the late system in order to decrease the influence of learnt reward contingency representations in favour of increasing the neural gain related to the processing of new observations.

2.3 Method

2.3.1 Participants

A total of 73 student participants took part in the experiment. Data of 10 participants were excluded from successive analyses due to poor behavioural performance. This was evaluated by a binomial test statistic with a significance threshold of $p = 0.01$ in order to determine whether participants were choosing between the two symbols randomly. Our criteria for pupil data exclusion included excessive blinking compromising the pre-processing pipeline (causing unreliable recovery of data points during the interpolation of blinks) and cases where more than 10% of all epoched pupil trials were determined to be outliers. We also decided to exclude EEG data with excessive noise that led to poor outcome valence (positive versus negative) discrimination. Consequently, we excluded 9 participants based on meeting pupil exclusion criteria, a further 3 participants as a result of inadequate EEG data, and 3 participants that had both poor quality pupil and EEG data. Following the exclusion of 25 participants based on all the above criteria, we retained data from 48 participants in our analyses. All participants were right-handed, had corrected-to-normal vision, had no existing psychiatric conditions, and were not taking psychoactive medication at the time of the study. Each participant gave written, informed consent in accordance with the School of Psychology and Neuroscience Ethics Committee at the University of Glasgow.

2.3.2 Stimuli display

The experiment was carried out through Presentation software (Neurobehavioral Systems, Inc., Berkeley, CA). Crucially, we designed all stimuli to be equiluminant across all trials and trial events of the experiment to ensure that pupil responses were not confounded by changes in luminance (Mathot & Vilotijevic, 2022). Stimulus symbols entailed a line crossing a circle, where the angle of the cuts determined stimulus type (delay symbol, and two stimulus symbols; Fig.2-1a). The symbols were placed on the left and right

side of a small vertical line enclosed from the top and the bottom by short horizontal lines. During the feedback screen of each trial, this central figure, surrounded by the delay symbols on each side, was transformed into an arrow pointing upwards or downwards, informing participants of the outcome of each trial. A representation of a typical trial in the experiment is displayed on Figure 2-1a. To reduce pupil data distortions, participants were seated opposite the eye-tracker camera at an angle of 180 degrees. Participants sat in a sound-proof booth with intermediate levels of ambient lighting in order to minimise distractions and discomfort during the study.

2.3.3 Reversal learning task

The experiment involved a total of 300 trials, broken down into four blocks and separated by breaks. In each trial, participants had to choose from two stimulus symbols, with the aim to discover the one with the higher reward probability. Participants were informed that changes in reward contingencies may occur during the blocks and that they would need to modify their choice behaviour accordingly. Participants received a fixed compensation of £6 per experimental session as well as a further performance-based payment of up to a maximum of £6.

Each trial commenced with the same delay symbol appearing on both the left and right side of the screen for a random interval between 2 to 3 seconds. Next, two stimulus symbols replaced the delay symbols for 1.25 seconds, during which participants picked one of the symbols by pressing designated buttons on a response box. After signalling their choice, participants saw a second delay screen for 1.5-2 seconds. The outcome of their choice was issued by an arrow in the middle of the screen signalling either a positive outcome or a negative outcome (i.e., the arrow pointing upwards or downwards, respectively). Figure 2-1a outlines this sequence of trial events. In case participants failed to choose a symbol on time during stimulus representation, the feedback screen displayed a message reminding participants to make a faster choice next time. Such lost trials were excluded from all analyses.

At any point during the experiment, one of the stimulus symbols was corresponding to a high reward probability of 0.7, whilst the other symbol had a reward probability of 0.3. Participants were naïve about the reward probabilities allocated to stimulus symbols and were instructed to discover the symbol with the higher reward probability through trial and error by considering the outcome in each trial. We determined reversals points in a way that ensured that participants have a sufficient amount of time to learn and exploit the ongoing reward contingencies based on data from previous reversals learning experiments from the lab. In order to become familiar with the task, participants practised a block of 75 trials prior to the experiment.

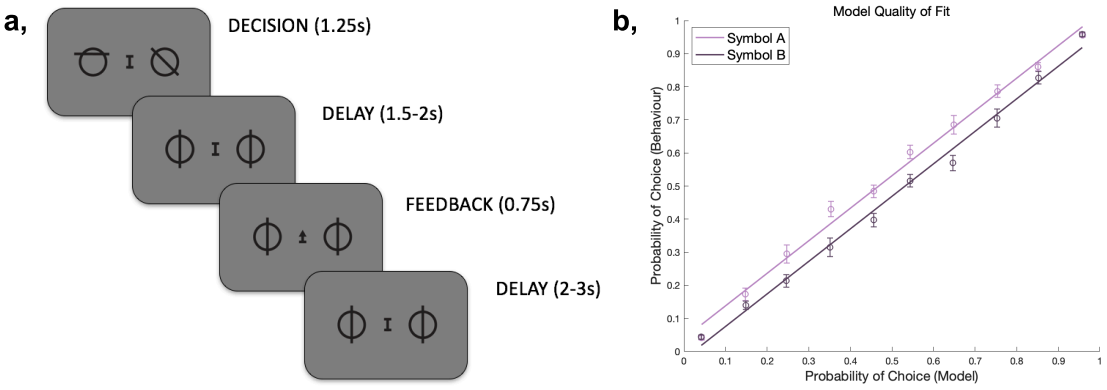

Figure 2-1. Experimental design and behavioural model fit. a, A representation of a typical trial from the experiment. On each trial, participants had 1.25 seconds to choose from two abstract symbols, with the goal of selecting the one with the higher reward probability (0.7 versus 0.3). Following a random delay of 1.5-2 seconds, an arrow in the middle of the screen indicated either a positive or a negative (i.e., an arrow pointing upwards or downwards, respectively) decision outcome for 2-3 seconds. A further random delay of 2-3 seconds preceded the next trial. Participants performed 4 blocks of 75 trials each. **b,** Illustration of the behavioural model fit. Predicted choice-probabilities from the reward learning algorithm (*x*-axis), which utilised a softmax method (binned into ten bins, with a bin size of 0.1, and averaged across all participants for both symbols A and B), strongly corresponded to observed choices (*y* axis), derived as the proportion of trials in which participants chose symbol A or B. For each bin, error bars represent the standard error of the mean around the observed choice probabilities. The solid lines show the degree of correlation between the observed and predicted choice probabilities for each of the choice options.

2.3.4 Reinforcement learning model

We used a model-free reinforcement learning model optimised for reversal learning paradigms (Krugel et al., 2009) to derive trial-by-trial reward prediction errors. In reinforcement learning models, each choice option is assigned an expected reward value $q_A(t)$, which in turn is used to obtain choice probabilities $p_A(t)$. This is derived from choosing option *A* in trial *t*, following the softmax choice rule:

$$p_A(t) = \frac{1}{1 + exp[-\beta \cdot q_A(t) - q_B(t) - a]}, \quad (2.1)$$

where β is the sensitivity parameter controlling the weight of reward expectations on choice probabilities, $q_A(t)$ and $q_B(t)$ are the expected values of each of the two choice options, and *a* is the indecision point, representing equiprobable choice between the stimuli (Hampton et al., 2006). After the participant selected one option, the observed outcome $r_A(t)$ is compared with the expected reward $q_A(t)$, where the discrepancy produces the reward prediction error δ (RPE):

$$\delta_A(t) = r_A(t-1) - q_A(t-1). \quad (2.2)$$

Reinforcement learning models postulate that the deviations expressed by RPEs drive learning as expected choice values are updated as follows:

$$q_A(t) = q_A(t-1) + \alpha \cdot \delta_A(t-1), \quad (2.3)$$

where α is the learning rate that controls the influence of the RPE on the updating of the stimulus expected value. The expected value of the unselected stimulus on trial *t* is not updated. A dynamic learning rate $\alpha(t)$, instead of a constant learning rate α, was used to capture the participant- and trial-wise fluctuations in the task introduced by reversals. This dynamic learning allows for both rapid adaptations after reversal and the stabilisation of behaviour once the better option has been discovered. It is modulated by

the slope *m* of the smoothed and unsigned RPE according to the following update rule:

$$\alpha(t) = \alpha(t-1) + f(m(t)) \cdot (1 - \alpha(t-1)), \quad if\ m > 0 \qquad (2.4)$$

$$\alpha(t) = \alpha(t-1) + f(m(t)) \cdot \alpha(t-1), \quad if\ m < 0. \qquad (2.5)$$

Accordingly, the learning rate increases when the slope of the unsigned RPE is positive and decreases when the slope is negative. The slope was estimated over the smoothed and unsigned RPEs, where smoothing is determined as follows:

$$\delta\, abs(t) = \delta\, abs(t-1) \cdot (1 - \alpha(1)) + \delta(t) \cdot \alpha(t). \qquad (2.6)$$

Here, $\alpha(1)$ serves as the learning rate for the first trial but also represents the learning rate used to update the unsigned RPE. Consequently, high values for $\alpha(1)$ suggest that only the most recent RPEs determine present estimate of the RPE weight. The PE slope was normalised over the last two RPEs in order to produce a slope that is independent of the scale of payoffs:

$$m(t) = \frac{\delta abs(t) - \delta abs(t-1)}{(\delta abs(t) + \delta abs(t-1))\ /\ 2}. \qquad (2.7)$$

Finally, the weight of the slope of the RPE on the learning rate was transformed by a double sigmoid function that allows for the slope to take on values between 0 and 1, according to

$$f(m) = sign(m) \cdot (1 - exp(-(m/\gamma)^2)). \qquad (2.8)$$

Here, the free parameter γ controls the extent to which the RPE slope influences the learning rate. Therefore, if γ takes on a high value ($\gamma > 3$), the updating of the dynamic learning rate is negligible, so that the learning rate becomes a constant that is guided by $\alpha(1)$.

For each participant, we estimated four parameters: the sensitivity parameter of choice β, the dynamic learning rate $\alpha(t)$, the sensitivity of the learning

rate with regards to the RPE slope γ, and the indecision point a. We used the following predetermined starting points for the above parameters:

$$a = 0$$

$$\alpha(t) = 0.5$$

$$\beta = 2$$

$$\gamma = 1,$$

These values constituted the starting points of the maximum likelihood estimation fitting procedure for sets of parameters θ_j, such that

$$\theta_j^{ML} = argmax_\theta \, logL(\theta_j), \tag{2.9}$$

where the likelihood *logL* was calculated according to

$$logL(\theta_j) = \frac{C_A logP_A(\theta_j)}{N_A} + \frac{C_B logP_B(\theta_j)}{N_B}. \tag{2.10}$$

Here, $P(\theta_j)$ stands for the choice log probability given the model parameters θ_j, C is a binary vector for observed choice, N is the number of observed choices, and the subscripts mark each of the available two choices. During the optimisation procedure, we constrained the free parameters according to:

$$0.9 > a > 0.1.$$

$$10 > \alpha(t) > -10$$

$$20 > \beta > 0$$

$$10 > \gamma > 0.01.$$

Lastly, to obtain estimates of trial-by-trial estimates of RPEs and dynamic learning rates, the final set of parameters were reintroduced into the reinforcement learning algorithm.

The value of the free parameter γ was estimated as < 3 for the majority of participants (35 out of 48), indicating that learning was regulated by a dynamic learning rate. Accordingly, in subsequent analyses, we used RPE estimates derived by the algorithm utilising dynamic learning rates. To illustrate the model's goodness of fit, we binned the participant-specific choice probabilities into ten groups based on distribution quintiles and calculated the participant-specific average choice probability for each bin. This allowed us to recover the participant- and group-wise mean observed choice probabilities as well as the corresponding Pearson's correlation coefficient (r = 1, p = 0), both of which are shown on Figure 2-1b.

2.3.5 EEG data acquisition and pre-processing

We simultaneously acquired continuous EEG and pupillometry data from participants performing the experimental task in an electrostatically shielded and sound-attenuated booth. We used a 64-channel EEG amplifier system (BrainAmps MR-Plus, Brain Products GmbH, Germany) with Ag/AgCl scalp electrodes situated following the international 10-20 system on an EasyCap (Brain Products GmbH, Germany). A chin electrode served as a ground and all EEG channels were referenced to the left mastoid. Input impedance was adjusted to under 20 kΩ. Data were recorded in Brain Vision Recorder (BVR; Version 1.10, Brain Products, Germany) at a sampling rate of 1000 Hz and subjected to online (hardware) filtering by an analog band-pass filter of 0.016 - 250 Hz. Experimental event trigger codes of stimulus presentation and participant responses were synchronised with the EEG data and collected via Brain Vision Recorder. These data were stored for offline analysis in MATLAB (version 2018b, The Mathworks Inc., 2018). We implemented a band-pass filter with cutoff frequencies between 0.5 and 40 Hz to the data to eliminate slow direct current drifts and high frequency noise. Beyond this, a notch filter at 50 Hz was used for noise reduction. Finally, the data were re-referenced to the average of all electrodes.

To remove eye-movement artefacts, participants were asked to complete an eye-movement calibration task before the main experiment. During this

exercise, participants were required to blink repeatedly while viewing a fixation cross at the centre of the screen, after which they were asked to make a number of horizontal and vertical saccades in accordance with the location of the fixation cross. The timing of these visual cues was recorded, allowing us to identify linear EEG sensor weights linked to blinks and saccades using principal component analysis (Parra et al., 2005), which in turn were projected onto and subtracted out from the broadband data from the main task. Due to the lack of eye calibration data for two participants, we employed independent component analysis implemented in EEGLab (Delorme et al., 2007), which utilised the data from the main task to remove eye-movement artefacts.

2.3.6 Multivariate EEG data analysis

We aimed to exploit the single-trial variability in the EEG-derived, temporally-specific representations of outcome valence to create parametrically-modulated regressors for our EEG-informed behavioural and pupil analyses, designed to disentangle brain networks associated with reward learning (see below for details). Similarly to previous work (Fouragnan et al., 2015; 2017; Franzen et al., 2020; Parra et al., 2005, Sajda et al., 2009), we utilised a linear multivariate classifier to EEG data locked to the time of decision feedback, using the sliding window method. This allowed us to identify a projection (y_i) of the multidimensional EEG signal, $x_i(t)$, where i = (1…N trials), within a short time window that achieved the greatest level of discrimination between positive and negative feedback trials. All time windows had a width of 60 ms and the window centre was moved from -100 ms to 600 ms relative to feedback onset, in increments of 10 ms. Thus, our discriminator was trained to map positive component amplitudes to positive feedback and negative component amplitudes to negative feedback.

Specifically, the classifier produced a 64-channel spatial weighting $w(\tau)$ via logistic regression (Parra et al., 2005) that maximally discriminated between positive and negative feedback trials within each time window, arriving at the one-dimensional projection $y_i(\tau)$, for each trial i and time window τ:

$$y_i(\tau) = w(\tau)^T \cdot x(\tau) = \sum_{i=1}^{D} w_i(\tau) \cdot x_i(\tau). \quad (2.11)$$

Here, *T* is the transpose operator, *D* is the number of EEG channels, $y_i(\tau)$ is a vector of single trial discriminator amplitudes (1 x trials), $w(\tau)$ is a vector with a number of weights corresponding to the number of EEG channels (1 X 64), and $x_i(\tau)$ is a matrix (Channels X Trials/Samples).

We assessed the performance of the discriminator for each time window using the area under a receiver operating characteristic curve, specified as the A_z value, coupled with a leave-one-trial-out cross-validation method to control for overfitting (Philiastides & Sajda, 2006; Philiastides, Ratcliff, & Sajda, 2006). Accordingly, we used N-1 trials for each iteration to determine a spatial filter *w*, which was in turn applied to the left out trial to derive out-of-sample discriminant component amplitudes, y_i, and calculate the A_z curve. To evaluate the significance of the discriminator, we applied a bootstrapping technique that implemented leave-one-out tests following a randomisation of positive and negative feedback trial labels. Repeating this randomisation procedure a thousand times (at 200 ms following feedback presentation to ensure reliable discrimination performance) allowed us to derive a random probability distribution for A_z values, which we in turn utilised as the reference to estimate the original A_z value leading to a significance level of $p < 0.001$ on the random distribution (population $A_z sig = 0.598$). All the above discriminant analysis steps were performed on the participant level, whereby each participant served as their own replication unit.

The linearity of this model allowed us to derive the scalp topographies of the discriminating components resulting from equation(2.11) by estimating a forward model according to:

$$a(\tau) = \frac{x(\tau) \cdot y(\tau)}{y(\tau)^T \cdot y(\tau)}. \quad (2.12)$$

Here, for each time window τ, the discriminating component y(τ) is vectorised (1 X Trials) and the EEG data $x(\tau)$ is shown in a matrix form (Channels X

Trials). The forward model $a(\tau)$ can be considered a scalp plot and interpreted as the coupling between the discriminating component amplitudes and the observed EEG. Thus, $a(\tau)$ represents the electrical coupling of the discriminating component $y(\tau)$ that describes the majority of activity in $x(\tau)$. Consequently, a stronger coupling is associated with a low attenuation of the component and reflects the intensity of $a(\tau)$. Using this approach, we merged information spatially across the multidimensional electrode space, allowing us to retain the trial-by-trial variability in the discriminating component and increase the signal-to-noise ratio of our data.

In order to establish the early and late valence components, we defined time windows for each component based on the results reported by Fouragnan and colleagues (2015) and ensured that the participant-specific topographies for each component consistently match these previous results. To evaluate the spatial characteristics of the discriminating activity, we examined the resulting scalp plots, using the corresponding participant-specific $a(\tau)$ vectors. We selected peak times of the participants-specific accumulating activity (A_z) within the predefined time windows of 150-290 ms and 300-550 ms for the early and late component, respectively, that matched the early and late component topographies reported by Fouragnan et al. Thus, by utilising the participant-specific scalp topographies and A_z values, we selected the participants-specific early and late feedback components that varied within the predefined time windows for each component. We found both components in all 48 participants, suggesting the robustness of these components. In subsequent analyses, we utilised the participant-specific discrimination amplitudes (y_i) linked to the individually-selected early and late components within each trial as an index of how decision outcomes are perceived in individual trials (Fouragnan et al., 2025; 2017; Franzen et al., 2020).

2.3.7 Analyses of EEG and behavioural data

Next, we aimed to investigate the hypothesis that increased early system activity is associated with increased exploration tendency and response slowing, whilst the late system is predominantly engaged in the

value-updating of reward representations. To test this conjecture, we carried out regression analyses to predict four different behavioural metrics using the participant-specific trial-wise component discrimination amplitudes. We created single-trial predictors reflecting the participant-specific discriminating amplitudes of the early and the late components for each trial, both of which were further broken down into positive and negative outcomes. We used the resulting four regressors, i.e., early positive, early negative, late positive, late negative, to predict switching behaviour, exploration, value updating, and response slowing separately for each participant. Therefore, our regression analyses effectively removed overall unspecific valence effects across positive and negative outcomes, and only the trial-by-trial variability within each positive and negative component was used to predict behaviour. To test the significance of the effects of the four components in predicting behaviour, we used t-tests to assess whether the population of regression coefficients resulting from the participant-specific regressions for each regressor type came from a distribution with mean zero. We repeated this step for each coefficient derived from each of the regressions.

The first model incorporated logistic binomial regression using a probit link function to test the link between the discriminator amplitudes of the four EEG components and participants' switching behaviour (i.e., a binomial variable indicating whether the participant chose the other symbol on the following trial). Next, a similar binomial regression was used to disassociate the influence of four EEG-derived component amplitudes on exploration. Exploration was signified by choices where participants chose the symbol with the lower reward value as determined by our reward learning model (Daw et al., 2006; Harada, 2020; Warren et al., 2017). To test the late component's hypothesised role in value-updating, our third analysis utilised linear regression to establish the degree to which our component amplitudes were predictive of participants' value-updating of the chosen symbol on the next trial. To quantify trial-wise value-updating, we utilised the value difference of the chosen symbol between the current and next trials as derived from our reward learning model. Finally, a linear regression assessed whether

discriminating amplitudes linked to the components predicted response slowing, calculated as the difference in the z-scored reaction times between consecutive trials.

To assess the across-participants correlation between the value-updating regression coefficients and overall accuracy, we used robust bend correlations. For this, we utilised the *'bendcorr'* function in MATLAB from the robust correlation toolbox devised by Pernet and colleagues (2013), which down-weighs bivariate outliers by 20% of all data points in each dimension (i.e., bending constant = 0.2). This function outputs a correlation coefficient (*r*), as well as *t*- and *p*-values, the latter of which was evaluated against an alpha level of 0.05 to determine correlation significance.

In the above regression analyses, we utilised parametric EEG regressors, which were broken down into early and late components for each feedback type, which removed overall valence effects. In these analyses, our primary focus was to explore the extent to which trial-wise variability in each of our EEG-based components provides explanatory power in predicting behaviour, independently of an added valence-specific effect. Nevertheless, we carried out control analyses for each of our regression models, whereby the four EEG-based predictors were replaced by a binary feedback-type predictor, with values +1 for positive feedback and -1 for negative feedback, to predict each of our behavioural measures (switching, value updating, exploration, and RT change). To compare the explanatory power of our EEG-based model with that utilising feedback-valence predictor, we calculated and contrasted the adjusted R^2 value associated with each model (Table 2-1).

2.3.8 Pupil data acquisition and pre-processing

We used the Tobii Pro x3-120 system (Tobii AB, Stockholm, Sweden) to record participants' pupil responses from both eyes at 40 Hz during the experiment. Pupil data were pre-processed and analysed in MATLAB (version 2018b, The Mathworks Inc., 2018). We used the averaged pupil diameter across both eyes during all our analyses. We removed invalid data points, resulting from cases when the eye-tracker was unsuccessful in detecting the eyes. Lost data

resulting from eye blinks was replaced by linearly interpolated values in the range of -100 to +100 milliseconds around missed events. Finally, we applied a bandpass filter of 0.01 - 4 Hz to remove measurement noise resulting from non-physiological sources (Hoeks & Levelt, 1993; van Slooten et al., 2017) and z-scored each participant's data.

2.3.9 Analysis of the feedback-evoked pupil response

To derive the participant-wise feedback-evoked pupil response, we epoched the pupil data with a baseline correction of 500 milliseconds relative to outcome presentation and considered data up until 3000 milliseconds following outcome display. Next, we computed the average pupil response for 25 ms intervals within each epoch, yielding 141 data points per trial. To account for outliers, we removed trials in which the pupil response was more than 3 standard deviations away from the mean pupil response or the trial standard deviation surpassed the 0.11 cutoff value (the latter was performed to exclude trials with missing data). Using this outlier-removal method, on average, 7.65 trials *(SD* = 6.32 trials) were removed from each participant's dataset (2.55% of all trials).

To assess the difference in the evoked pupil response due to feedback valence, we calculated the population mean evoked response separately for positive and negative feedback trials for each 25 ms interval within our epoch time frame. All trials were individually baseline-corrected by calculating the mean pupil response in the 500 ms time window preceding outcome presentation and then subtracting this baseline value from at each point in the pupil time series. Subsequently, we compared the mean response for positive and negative trials in each of the intervals using pairwise t-tests. To reduce the likelihood of Type I errors, we used Bonferroni-correction, whereby we evaluated each resulting *p*-value against $\alpha = 0.05/141$, as we evaluated 141 separate hypotheses against the 0.05 alpha level.

2.3.10 EEG-informed pupil analysis

Next, we probed the hypothesised link between pupil dilation, a proxy for LC-NA activity (Joshi et al., 2016; Joshi & Gold, 2020; Murphy et al., 2014; Reimer et al., 2016), and the early system. As the pupil response at any given point may reflect the influence of various ongoing internal signals linked to perceptual and cognitive processing, merely contrasting the feedback-evoked pupil response between positive and negative feedback cannot disentangle the differences associated with specific cognitive processes. In order to link the pupil response to distinct internal signals evoked by specific events, we used the general linear model (GLM) approach (de Gee et al., 2014; Dension et al., 2020) to model the linear combination of the early and late component-associated pupil responses. In this model, the pupil responses linked to components are the internal signal time series linked to individual trial events. Thus, this regression analysis exploited the trial-by-trial variability in the LDA-derived discrimination amplitudes linked to the early and late components broken down by positive and negative outcomes in order to disentangle the unique contribution of these two systems in explaining the feedback-related pupil response. This allowed us to test the hypothesis that early system activity is, at least partially, driven by the LC-NA system, and exclusively drives the increased feedback-related pupillary effect following negative outcomes.

More specifically, we utilised the LDA-derived discrimination amplitudes ($y_{i,}$ Eq. 2.11) linked to the participant-specific early and late components, broken down by feedback type (positive and negative), to build four parametric pupil regressors (EarlyPos, EarlyNeg, LatePos, LateNeg). To remove nonspecific arousal effects across positive and negative outcomes, we created a parametric salience predictor, the amplitudes of which were modulated by the unsigned RPE obtained from our reward learning model. To absorb any nonspecific effects related to the presentation of the feedback, stimulus onset, and the time of the decision, we created boxcar predictors, with

amplitudes set to 1, at the time of feedback, stimulus display, and choice, respectively.

All above regressors were convolved with participant-specific pupil response functions (PuRF), which were modelled based on the canonical Erlang gamma function (Dension et al., 2020; Hoeks & Levelt, 1993). The PuRF spanned from -500 to 3000 ms relative to trial events and the width, peak time, and undershoot of each PuRF was adjusted to match the mean participant-wise feedback-evoked pupil response (Figure 2-2b). Furthermore, the PuRF for each participant was normalised to a maximum value of 1, so that if a regressor component amplitude value of 1 in each predictor amounted to a 1% increase in pupil size compared to baseline (Dension et al., 2020). Using these 8 predictors and the pupil time series as the dependent variable, we fit a linear model to each participant's data using MATLAB's *robustfit* function. We employed t-tests to evaluate the significance of the effect of the 4 EEG-derived components on the pupil response. As such, for each regressor type (earlyPos, earlyNeg, latePos, lateNeg), we assessed whether the group of participant-wise coefficients came from a distribution with mean zero.

In order to confirm the above GLM results, we also implemented a linear mixed effects model (utilising MATLAB's *fitlme* function). In this, we treated participants as random effects by specifying participant-specific model intercepts as well as slopes for the four EEG-derived components (Baayen et al., 2008). Similarly to the participant-specific GLMs, we used the above 8 regressors to predict the full-course pupil data. To evaluate the degree to which each regressor contributed to the pupil response, we utilised the built-in coefficient evaluation function in *fitlme*, which determined the *p*-value associated with the t-statistic for two-sided hypothesis tests.

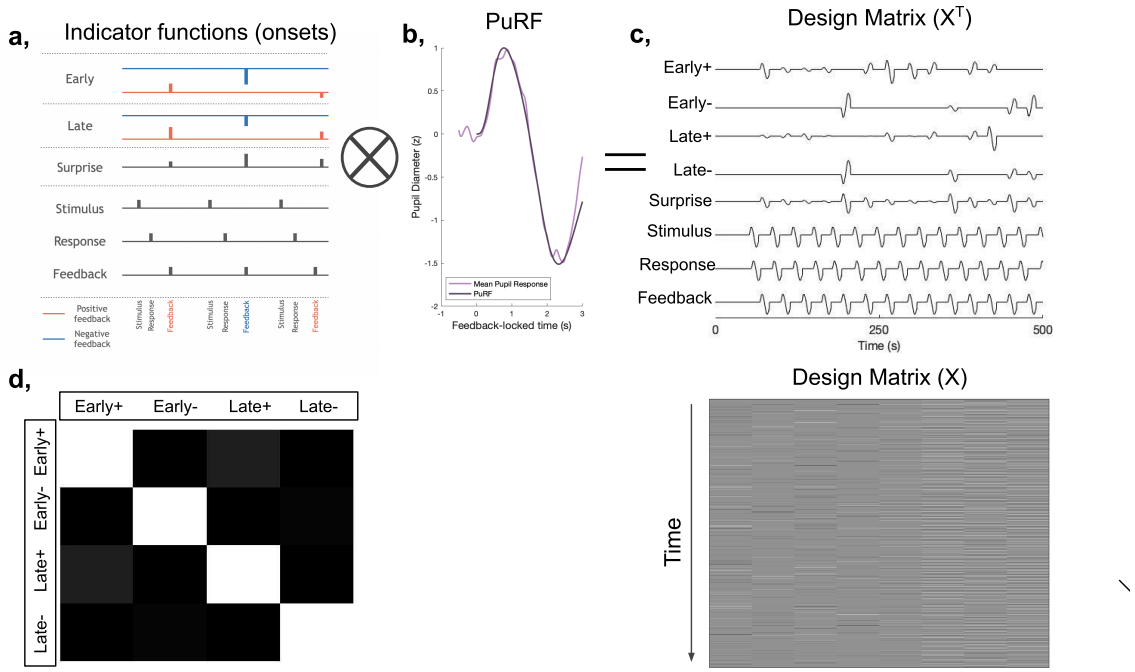

Figure 2-2. EEG-informed pupil analysis. a, Our participant-wise GLM predicted the full course pupil data and included 8 regressors. Four parametric regressors were modulated by the amplitudes of the *y* values derived from our feedback-locked linear discriminant analysis (yielding early negative, early positive, late negative and late positive predictors). To represent the effect of surprise, we created a parametric regressor modulated by the absolute value of the prediction error estimated by our reward learning model. To account for variance linked to other task-related processes, we included non-parametric, boxcar regressors at the times of stimulus display, response, and feedback presentation, the value of which was set to 1 at the appropriate time points. Finally, we convolved each regressor with the **b,** participant-specific pupil response function (PuRF) that was adjusted to each participant's mean feedback-locked pupil response. **c,** The transposed design matrix (X^T) is shown for the first 500 ms of the experiment, after the predictor indicator functions have been convolved with the participant-specific PuRF (top). The design matrix (X) is shown for the full experiment. Columns denote predictor vectors,whereby the indicator functions are convolved with the participant-specific PuRF. Darker shades represent lower values. **d,** Covariance matrix depicting the degree of correlation among the four EEG-based single trial predictors (i.e., the first 4 columns in X). Darker shades represent lower correlation. For all graphs, data from a randomly selected participant is shown.

Next, we aimed to determine how the the component regression coefficients with significant predictive power related to behavioural markers, including choice ambiguity, exploration tendency, and mean accuracy. The former was derived from our reward learning model by calculating the absolute difference in probabilities linked to choosing either symbols A (p_A) or B (p_B). As lower values indicate less substantial differences between the likelihood of choosing one option over the other, they index decisions that are more ambiguous (Bland & Schaefer, 2012). To test the hypothesis that early-system-mediated network resets in the late system increase exploration tendency, we tested whether increased coupling between the early system and the pupil response is associated with a boosted exploration tendency. To quantify exploration tendency, we used two different measures. First, we calculated the proportion of decisions in which participants chose the symbol with the lower reward probability (Daw et al., 2006; Harada, 2020; Warren et al., 2017). As an additional measure, we utilised the inverse temperature parameter obtained from our reward learning model (van Slooten et al., 2019). This free parameter marks reward sensitivity — the degree to which participants' decisions are guided by the difference in reward values. Higher values indicate a more substantial influence of symbol value deviations on choice, and therefore mark reduced exploratory propensity. We used robust bend correlations (see above) to measure the association between the three behavioural markers and the regressions coefficients obtained from the EEG-informed pupil GLMs.

In the above EEG-informed pupil regression analyses, we tested the ability of each of our single-trial EEG predictors to explain variability in the feedback-related pupil response, independently of an added outcome valence effect. Additionally, we employed a baseline version of this model to test for outcome valence effects. This baseline model replaced the four parametric EEG predictors by a binary outcome valence predictor (containing +1 for positive feedback and -1 for negative feedback at feedback onset times), and was otherwise identical to the EEG-based model. We utilised the adjusted R^2

statistic to compare the explanatory power related to the EEG-informed and baseline models (Table 2-1).

2.3.11 Pupil deconvolution

In order to visualise the unique contribution of early and late negative component activity on the pupil response, we implemented deconvolution analyses based on linear regression (Dimigen & Ehinger, 2021; Wierda et al., 2012). These models utilised the temporal variability in the overlap across the EEG component amplitudes, which allowed us to disentangle the unique contribution of each component to the pupil response. Thus, despite the close temporal proximity of the EEG components, this method allowed us to reveal the temporal dynamics of the early and late negative components on the feedback-related pupil response by linearly combining the different component pupil responses.

We implemented two deconvolution models for each participant; either the early or the late component offset times were used to divide negative feedback trials into high and low discrimination groups based on a participant-specific median split in trial-wise discrimination values *(y* values, Eq.2.11). Specifically, we built a design matrix *X*, which spanned all samples of the continuous pupil data. Each row of *X* coded the condition of the event type, whereby individual regressors modelled each of the 25 ms sections in the 3 second post-feedback interval, yielding 121 regressors per event type (low *y*, high *y*). The regressors for all trials belonging to the relevant condition were set to 1 at feedback onset time and to 0 otherwise. We made no assumptions regarding the shape of the resulting pupil response functions. Solving the regression formula

$$pupil = \beta \cdot X + e \tag{2.13}$$

for β, where *X* is the design matrix and *e* represents the error term, we derived a unique β for each of the 121 time points within each event type. The resulting time series of betas represent the non-overlapping average pupil

response for low and high discrimination event types and therefore can be interpreted as the average pupil response evoked by the specific event in the time window of interest. To examine how the degree of valence discrimination following negative outcomes affected the pupil response, we averaged the resulting time series of betas for both the low and high discrimination conditions across participants and formally contrasted the response at each of the 121 time points with two-tailed t-tests (both for early and late offset).

Finally, we cross-checked the event-related pupil responses derived from deconvolution against the observed pupil response. Specifically, we extracted the pupil response for each 25 millisecond interval in the 3 second period following feedback separately for low and high discrimination trials following negative feedback for both early and late offset. We used the same procedure to derive low and high discrimination trials as in the deconvolution analysis. Once again, we used two-tailed *t*-test to contrast the evoked response at each of the 121 time time points between low and high discrimination trials, with separate analyses carried out for early and late offset.

2.4 Results

2.4.1 Behaviour

During the 300 trials of the experiment, participants reached an accuracy level of 66.89% (SD = 7.07%), suggesting a high level of engagement. As Fig.2-3a shows, mean accuracy was significantly higher ($t(47) = 38.13$, $p < 0.001$) in positive ($M = 81.84\%$, $SD = 5.51\%$) compared to negative feedback trials ($M = 47.51\%$, $SD = 8.59\%$). Consistently, participants also experienced significantly more ($t(47) = 13.25$, $p < 0.001$) positive (mean = 168.3, SD = 10.58) compared to negative outcomes ($M = 128.27$, $SD = 10.52$). On average, 11.46 ($SD = 0.58$) reversals took place during the experiment. As Fig.2-3b depicts, mean accuracy on the trial before reversals reached 68.93% ($SD = 16.48\%$), dropped to 32.82% ($SD = 15.83$) on the trial where reversals were introduced, and plateaued on the 9th trial following reversals at 73.11 % ($SD = 13.34\%$).

There was no significant difference ($t(47) = -1.02$, $p = 0.31$) between the mean reaction times (RTs) in positive ($M = 614.42$ ms, $SD = 81.55$ ms) and negative ($M = 617.41$ ms, $SD = 87.13$ ms) feedback trials (Fig2-3.c). Similarly, the change in reaction times in consecutive trials (RT_{t+1} - RT_t) did not show compelling difference ($t(47) = 0.35$ $p = 0.73$) between positive ($M = 23.42$ ms, $SD = 0.09$ ms) and negative ($M = -1.82$ ms, $SD = 19.89$ ms) outcomes (Fig.2-3d). As expected, value updating, obtained from our reward learning model (Eq. 2.3), was significantly higher ($t(47) = 19.60$, $p < 0.001$) following positive ($M = 23.42$, $SD = 0.09$) than negative ($M = -0.31$, $SD = 0.11$) outcomes (Fig.2-3e). Thus, positive outcomes increased the chosen symbol's internal value representation by simultaneously de-valuing the unchosen symbol. Conversely, negative feedback produced the opposite pattern; the value associated with the unchosen symbol was raised at the expense of the chosen symbol. Finally, participants were significantly more likely ($t(47) = 12.81$, $p < 0.001$) to switch to the other symbol following negative (M: 44.07%, $SD = 20.82\%$) compared to positive (M: 8.61%, $SD = 9.30\%$) feedback (Fig.2-3f).

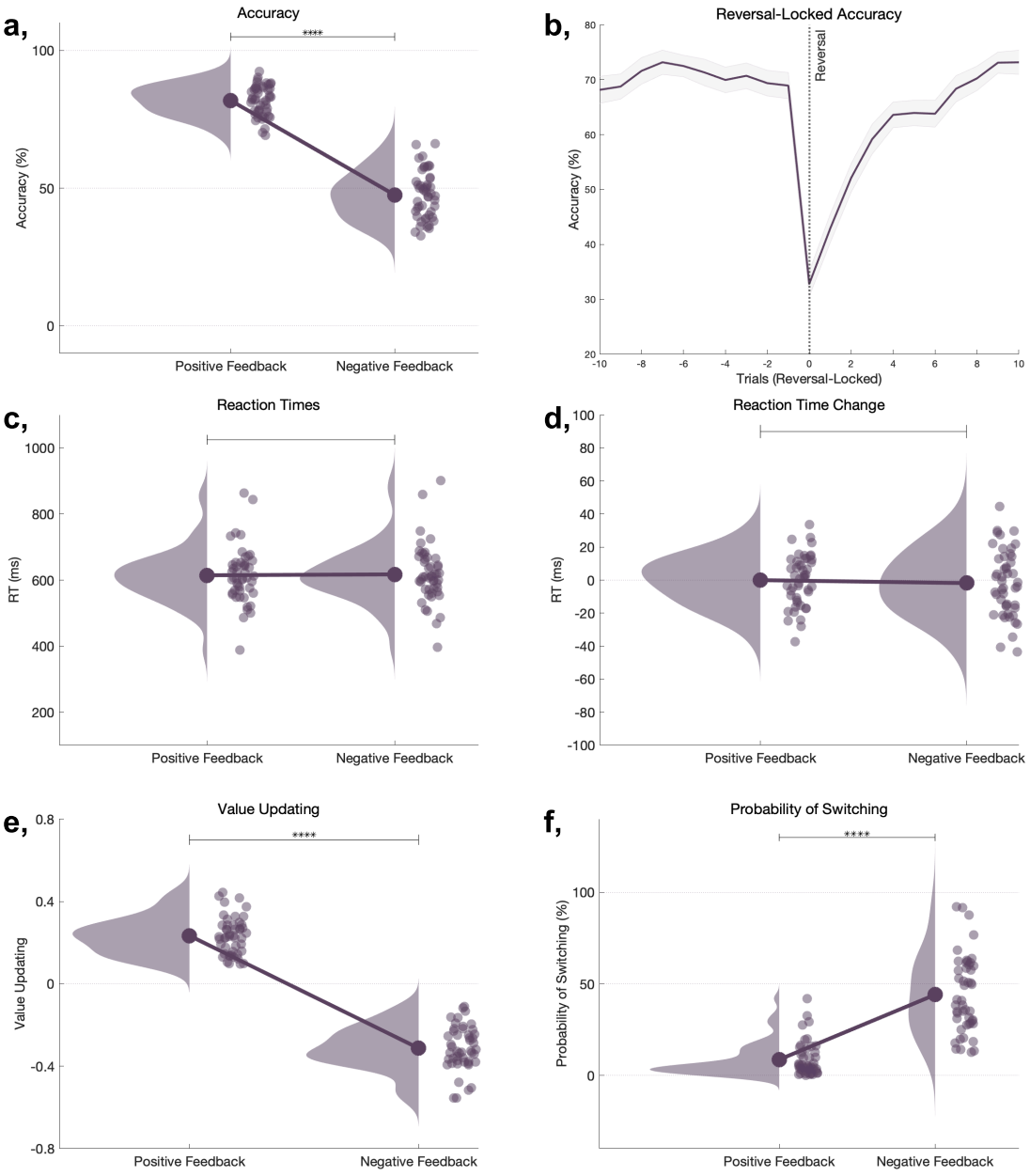

Figure 2.3. Behavioural results. a, Population accuracy, defined as selecting the symbol with the higher reward probability, broken down into trial type (positive or negative feedback). **b,** Population accuracy as a function of trial position locked to reversal points. The shaded error bars show across-participant standard errors. **c,** Population reaction times broken down into trial type (positive or negative feedback). **d,** Population reaction time changes from trial(t) to trial(t+1), shown for positive and negative outcome trials. **e,** Value updating from trial(t) to trial(t+1) derived from our reward learning model, broken down by trial type (positive and negative feedback trials). **f,** Population probability of switching symbols on the next trial following positive and negative feedback.

2.4.2 Multivariate EEG analysis

To establish temporally distinct neuronal components linked to outcome valence, we utilised single-trial multivariate discriminant analysis on the feedback-locked EEG signal (Parra et al., 2005). Specifically, using the sliding window method, we estimated the linear weight of the multidimensional EEG electrode signals that achieved maximal discrimination between positive and negative trials within each time window for each participant. Projecting the data through linear weights produced an aggregate representation of the component amplitude (y_i, Eq.2.11.) for each trial, which we in turn used as a proxy of how the participants experienced the feedback. These component discriminant amplitudes illustrate the distance of each trial from the discriminating hyperplane and can be interpreted as a proxy for the neural signal variability elicited by positive and negative decision outcomes, with the shared signal variance removed (Fouragnan et al., 2015; 2017; Franzen et al., 2020; Parra et al., 2005, Sajda et al., 2009). To assess the performance of the discriminator, we utilised the area under a receiver operating characteristic curve combined with a leave-one-trial-out cross-validation method for each participant (Philiastides & Sajda, 2006; Philiastides, Ratcliff, & Sajda, 2006). Due to the linearity of our model, we were able to derive scalp topographies (Eq. 2.12) representing the electrical coupling between the discriminating component amplitudes and the observed EEG signal.

Population discriminator performance over time and the scalp topographies linked to its significant peak points are shown in Figure 2-4a. Consistent with previous work (Fouragnan et al., 2015; 2017), the discriminator was significant between 150 and 590 ms after feedback presentation. We found a clear transition of topographies around 300-320 ms post-feedback, which we in turn used as a cutoff for defining the time windows for the early and late components. To replicate the two components, we selected the peak times of the participants-specific accumulating discrimination activity A_z within these predefined time windows (150-290 ms and 300-550 ms post-feedback for the early and late components, respectively) and based on the spatial resemblance of the corresponding scalp maps to the average topographies of

the early and late components described by Fouragnan and colleagues (2015; 2017). Using this approach, we replicated the early and the late valence components (Fig.2-4b), which revealed the same broad and distinct spatial profiles as previously established. The early component peaked, on average, at 227 ms (*SD* = 36 ms) following feedback presentation with an A_z value of 0.68 (*SD* = 0.07), whilst the late component peaked, on average, at 414 ms (*SD* = 75 ms) following feedback presentation with an A_z value of 0.70 (*SD* = 0.07). The robustness of these components, and consequently the corresponding neural systems producing them, is highlighted by the presence of the components in all 48 participants in our experiment. The temporal and spatial resemblance of our components to those isolated by Fouragnan and colleagues further substantiates this robustness.

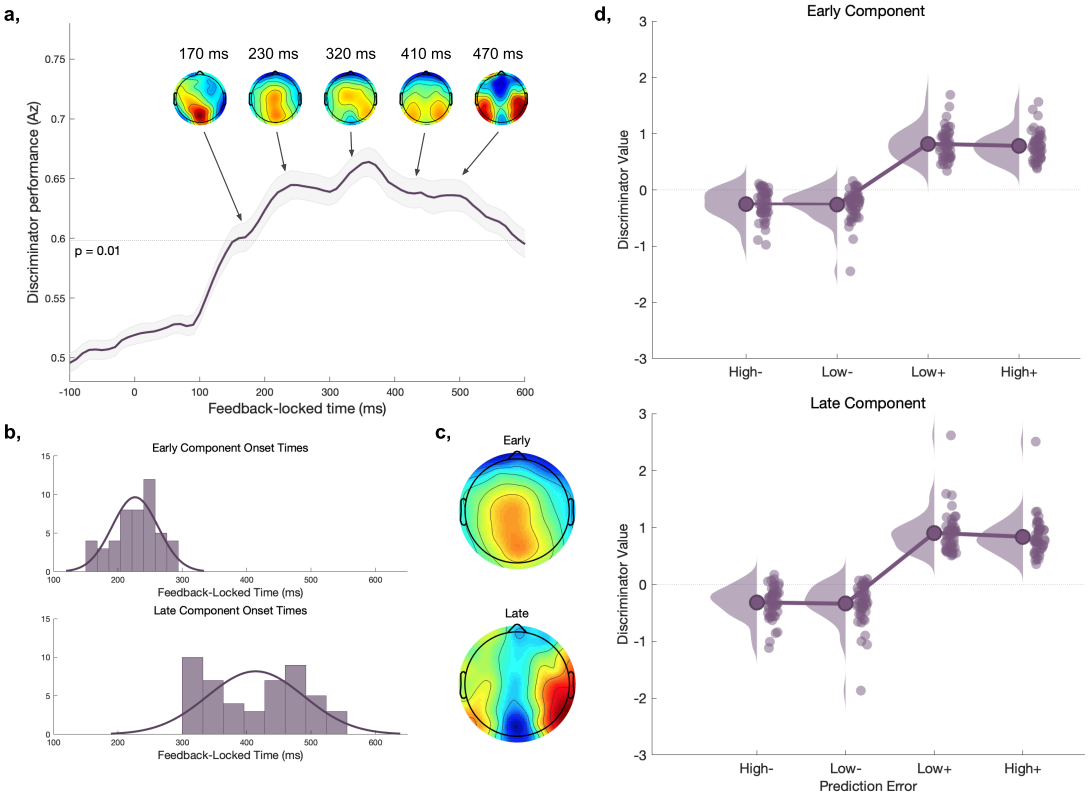

Figure 2-4. Single-trial EEG analyses. a, Single-trial discriminator performance (cross-validated A_z) during valence discrimination (positive vs negative feedback) of the feedback-locked EEG signal, averaged across participants (*n*=48). The horizontal line illustrates the A_z associated with a significance level of p = 0.001, approximated by using a bootstrap test. We

focused on recreating the early and late valence components that were established by Fouragnan and colleagues (2015). The shaded error band shows across-participant standard errors. The scalp maps on top represent the evolution of the population feedback-related EEG response at significant peak points of the A_z curve. **b,** Illustration of participant-specific early (top) and late (bottom) component onset times. **c,** Scalp maps illustrating the across-participant spatial topographies of the early (top) and the late components (bottom). **d,** Mean participant-specific discriminator output values (y_i, Eq.2.11) linked to the early (top) and late (bottom) components are binned by RPE magnitude, which was derived from our reward learning model. Both components reveal predominantly categorical response characteristics, without modulation by RPE magnitude.

To verify that the early and late components are not linked to salience (surprise) effects, we split the component-wise mean discriminator output (y_i) values into four bins based on the magnitude of the RPEs estimated by our reward learning model. Consistent with the results by Fouragnan et al. (2015; 2017), both components revealed a categorical RPE response profile (with higher discrimination values for positive compared to negative RPEs), with no additional modulation by RPE magnitude. This is confirmed by the low within-participant Pearson's correlation coefficients for both the early (Positive RPE: r = -0.03, Negative RPE: r = -0.01) and the late (Positive RPE: r = -0.05, Negative RPE: r = -0.02) components and RPEs, suggesting that neither components correlate with salience. In subsequent analyses, we utilised the participant-specific discrimination amplitudes (y_i, Eq.2.11) linked to the early and late components as an index of how decision outcomes are perceived in individual trials (Fouragnan et al., 2025; 2017; Franzen et al., 2020). Trial-wise amplitude variations in the early and late components were generally uncorrelated as revealed by Pearson's correlation coefficients (r = 0.09 and p = 0.33 for positive outcomes, r = 0.15, p = 0.24 for negative outcomes). This allowed us to utilise the participant-specific discrimination amplitudes (y_i) linked to the early and late components as an index of how decision outcomes are perceived in individual trials (Fouragnan et al., 2025; 2017; Franzen et al., 2020) and build parametric EEG-informed regressors to predict behavioural and pupil data.

2.4.3 Single-trial EEG components and behaviour

To investigate the link between our EEG components and behavioural measures, we carried out four separate regression analyses using the single-trial early and late components discriminant amplitudes. We broke down each component by outcome type (positive and negative) in order to test the hypothesis that the early component following negative feedback is associated with behavioural counterparts of an LC-driven cortical reset, whilst the late component was expected to play a crucial role in reward processing and value updating following all outcomes (Fouragnan et al., 2015). Consequently, each participant-specific regression analysis included four predictors; early positive, early negative, late positive, and late negative, the values of which were determined by the trial-wise linear discriminant amplitudes (y_i) differentiating between positive and negative feedback. Higher discriminant amplitudes (more positive values) of the early and late positive components correspond to increased neural reactivity to positive outcomes, whilst lower values indicate reduced encoding of positive outcomes. Conversely, lower discriminant amplitudes (more negative values) of the early and late negative components correspond to boosted neural reactivity to negative outcomes, whilst higher values indicate decreased encoding of negative outcomes.

First, we performed a binomial logistic regression to determine whether the four single-trial EEG components on each trial are predictive of participants' switching behaviour on the following trial. Our results (Fig.2-5a) suggest that all four components predict switching behaviour ($t(47) = -7.80$, $p < 0.001$ for early positive, $t(47) = -5.20$, $p < 0.001$ for early negative, $t(47) = -9.21$, $p < 0.001$ for late positive, and $t(47) = -3.92$, $p < 0.001$ for late negative). In line with the findings of Fouragnan and colleagues (2015), the more positive decision outcomes were neurally encoded, suggested by higher positive discrimination amplitudes, the probability of selecting the same symbol over the next trial increased. Similarly, the more negative the decision outcomes were neurally encoded, as reflected by more negative discrimination amplitudes, the more probability of selecting the same symbol over the next

trial decreased. These findings imply a valence-specific component influence, whereby positive feedback processing depresses and negative feedback processing promotes avoidance learning.

Next, a binomial logistic regression tested whether any of the four EEG components predicted participants' explorative choices, that is, whether they chose the symbol with the lower associated value. Our results (Fig.2-5b) showed that higher early positive ($t(47) = -7.80$, $p < 0.001$) and the late positive ($t(47) = -5.20$, $p < 0.001$) component amplitudes significantly reduced the likelihood of choosing the symbol with the lower value. At the same time, increased negative feedback encoding by the early ($t(47) = -5.20$, $p < 0.001$) and late ($t(47) = -5.20$, $p < 0.001$) systems increased participants' inclination to explore over the next trial. This is consistent with the hypothesis that increased negative feedback encoding, reflected by more negative discriminant amplitudes, could promote an LC-induced cortical reset in reward learning structures, potentially driven by the early system, which signals reversals (i.e., unexpected uncertainty arising from negative feedback), and in turn promotes exploratory behaviours appropriate for establishing a new model of the prevailing reward contingencies.

Additionally, we investigated how our EEG component amplitudes affected the value updating process. Value updating was defined as the absolute value difference of the chosen symbol between the current and the next trial. Figure 2-5c depicts the participant-specific linear regression coefficients by predictor type. All four EEG components were highly predictive of value updating ($t(47) = 12.50$, $p < 0.001$ for early positive, $t(47) = 4.31$, $p < 0.001$ for early negative, $t(47) = 12.67$, $p < 0.001$ for late positive, and $t(47) = 5.59$, $p < 0.001$ for late negative). Similarly to Fouragnan et al. (2015), we found a positive relationship between the extent of up-and downregulation of value and the amplitudes of the EEG components. Accordingly, more positive component EEG discrimination amplitudes were associated with an increasing likelihood that the value of the chosen symbol would increase compared to the previous trial. Similarly, more negative EEG component discrimination

amplitudes were linked to a higher likelihood that the value of the chosen symbol would decrease compared to the previous trial.

Interestingly, stronger across-participant associations between value updating and the early (r = -0.50, p < 0.001) or the late (r = -0.61, p < 0.001) negative predictor amplitudes were linked to lower accuracy levels, as assessed by robust bend correlations (see Methods). The degree to which early (r = -0.10, p = 0.52) or late (r = -0.03, p = 0.84) positive component amplitudes predicted value updating had no significant correlation with accuracy. Specifically, the more participants' negative feedback encoding predicted value updating, the worse they performed in the task. Due to the probabilistic nature of the task, participants occasionally received negative feedback despite having selected the symbol with the higher reward probability (expected uncertainty), rendering significant value adjustments unnecessary. Nonetheless, negative feedback could also signal a reversal of reward contingencies (unexpected uncertainty), in which case larger alterations in value representations are adaptive. Thus, the more strongly participants associate negative outcomes with a reversal, the larger adjustments in value representations are expected. Consequently, our correlational results are consistent with the interpretation that a more robust link between negative feedback processing and value updating reflects the overinterpretation of negative feedback (suboptimally large internal estimates of unexpected uncertainty), which in turn results in lower performance. These results are consistent with the hypothesis that an LC-NA-driven early system, which down-regulates the late system following negative outcomes, signals unexpected uncertainty.

Our last, linear regression assessed whether our EEG component amplitudes predicted response caution, indexed by the difference in the z-scored RTs between consecutive trials. As Fig.2-5d shows, only the early component following negative feedback was predictive of response caution (t(47) = -2.05, p = 0.046). The early positive (t(47) = 1.76, p = 0.085), late positive (t(47) = 0.73, p = 0.47), and late negative (t(47) = 0.66, p = 0.51) component amplitudes showed no significant association with response caution. This suggests that the more strongly negative outcomes are encoded by the early

system, the more response caution increases. This result matches the proposed function of the early system in network resets, which in turn could slow down evidence accumulation over the next trial as evidence in favour of a new hypothesis (i.e., reversal) is accumulated.

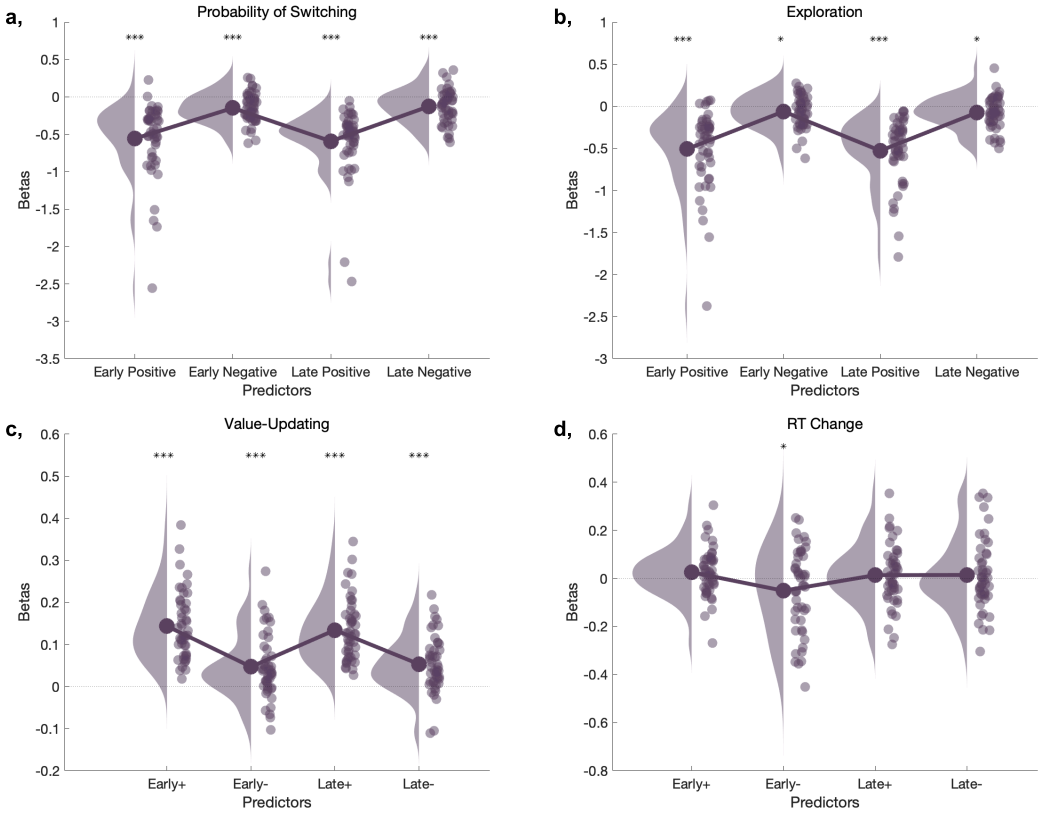

Figure 2-5. Single-trial EEG component amplitudes predict behaviour. a, Results of the binomial logistic regression, using four EEG-derived regressors (early positive, early negative, late positive, late negative) at trial *t*, predicting the probability of switching at trial *t*+1. *T*-tests indicate all four components to be significant predictors of switching ($p < 0.001$ in all cases). **B,** Results of the binomial regression, using four EEG-derived regressors (early positive, early negative, late positive, late negative) at trial *t*, predicting exploration (choosing the lower-valued option) over the next trial. Increasing early and late positive component amplitudes were associated with decreased explorative choices over the next trial ($p < 0.001$), whilst increased negative feedback encoding by the early ($p = 0.03$) and late ($p = 0.01$) systems boosted exploration on the next trial. **c,** Results of the linear regression, using the four EEG-derived regressors (early positive, early negative, late positive, late negative) at trial *t*, predicting value updating at trial *t*+1. *T*-tests indicate all four components to be significant

predictors of value updating ($p < 0.001$ in all cases). **d,** Results of the linear regression, using four EEG-derived regressors (early positive, early negative, late positive, late negative) at trial *t*, predicting response caution between consecutive trials. *T*-tests indicate that only the early component linked to negative outcomes is a significant predictor of response caution ($p = 0.046$), unlike the early component associated with positive outcomes ($p = 0.09$), or the late component linked to positive ($p = 0.47$) or negative ($p = 0.51$) outcomes.

By breaking down the feedback-related EEG response into an early and a late component for each feedback type, our four parametric EEG predictors utilised in the above regression analyses effectively removed overall valence effects. Nevertheless, we were interested in contrasting the explanatory power associated with each of the above regression analyses with baseline models including a single feedback-type predictor (specifying positive feedback as +1 and negative feedback as -1). Table 2-1 shows the adjusted R^2 values associated with each model. Our EEG-based predictors had an improved ability in explaining switching behaviour and between-trial RT change compared to the baseline model. At the same time, the binary feedback-type predictor explained exploration and value-updating better than the EEG-based predictors. Nevertheless, as our EEG-based regression models revealed, valence-independent variability in each of our EEG component predictors still significantly explained both exploration and value-updating, which implies their behavioural relevance.

Table 2-1. Model comparison. Adjusted R^2 values associated with our EEG-based and baseline regression models predicting switching, exploration, value-updating, RT change, and the pupil data. The higher adjusted R^2 value is bolded, illustrating higher explanatory power.

Predicted measure	Baseline model	EEG-informed model
Switching	-93.09	**-22.50**
Exploration	**-9.02**	-18.94
Value-updating	**.56**	.27
RT change	.001	**.009**
Pupil data	.039	**.040**

2.4.4 Pupil data analysis

Following pre-processing, we baseline-corrected and averaged the feedback-locked pupil responses by outcome type. The population pupil response peaked at around 1 second post-feedback, after which it plummeted at around 2.2 milliseconds, followed by a modest rise. This is compatible with previous findings indicating peak pupil dilation around 1 second following stimulus onset (van Rij et al., 2019). As expected, the obtained pupil response was larger following negative compared to positive feedback (Fig.2-6a).

We utilised two-tailed t-tests to quantify the difference in participants' pupil dilation per outcome type for each 25 ms time window from the 500 ms before the onset of feedback presentation until 3000 ms post-feedback. To account for multiple comparisons on the same dependent variable, we applied Bonferroni correction ($p_{Bonferroni}$ = 0.05/141 = 0.00036). Pupil response following negative feedback significantly exceeded pupil dilation after positive feedback in two temporal segments; 725 to 2000 ms and 2850 to 3000 ms post-feedback. This outcome valence-induced difference in pupil dilation can be reconciled with our hypothesis that negative feedback prompts elevated LC-NA system activity reflecting increased levels of contextual

uncertainty and a neural interrupt signal. However, our comparison of population-level pupil responses might only be caused by a general valence effect, whereby negative feedback is broadly perceived as more salient. To unmask latent trends in the data potentially concealed by averaging across trials, we utilised four single-trial EEG component discriminant amplitudes to predict the pupil response.

2.4.5 EEG-informed pupil analysis

To further test the hypothesised link between the early feedback processing and the LC-NA systems, we capitalised on the single-trial EEG discriminator amplitudes associated with our EEG components. Focusing on the across-trial component amplitude fluctuations within each component type allowed us to establish component-pupil coupling effects beyond any added unspecific effects of feedback valence. To predict the full course pupil data using a general linear model approach, we built neural regressors corresponding to the trial-specific linear discriminant amplitudes for the early and late systems organised by outcome type. This yielded the following four predictors; early positive, early negative, late position, and late negative. To account for the effect of salience, we constructed a surprise regressor based on the unsigned RPEs estimated by our reward learning model. Additionally, our model included three unmodulated regressors to control for nonspecific effects elicited by stimuli, choice, and feedback (Fig.2-2).

Figure 2-6b depicts the resulting regression coefficients for the four neural regressors. Secondary analyses revealed that unlike the early ($t(47) = 0.38$, $p = 0.70$) and late ($t(47) = 0.35$, $p = 0.72$) positive, the early ($t(47) = -2.37$, $p = 0.02$) and late negative ($t(47) = -2.36$, $p = 0.02$) components significantly predicted the pupil response. Increasingly negative discriminant amplitudes of the early and late negative components correspond to boosted neural reactivity to negative feedback. Therefore, the negative association between the early and late negative component amplitudes and the pupil response suggests that more negative discriminant amplitudes (increased negative feedback encoding) results in a more pronounced pupil response.

These effects subsist beyond the impact of salience ($t(47) = 4.07$, $p < 0.001$) as well as the additional non-specific effects related to stimuli presentation ($t(47) = 8.24$, $p < 0.001$), choice ($t(47) = 9.24$, $p < 0.001$), and feedback display ($t(47) = 2.38$, $p = 0.02$). Our mixed effects linear model confirmed these results; the early ($t(47) = -3.39$, $p < 0.001$) and late ($t(47) = -2.44$, $p = 0.01$) negative component amplitudes significantly predicted the feedback-evoked pupil response, unlike the early ($t(47) = 1.28$, $p = 0.20$) and late ($t(47) = 0.24$, $p = 081$) positive component amplitudes. These effects were sustained beyond the impact of surprise as well as the three event-related dummy variables ($t(47) = 0.38$, $p = 0.70$).

The significant correlation between the early negative component and the pupil response is consistent with the hypothesis that early system activity is guided by the LC-NA arousal network. We speculate that the late negative component explains the pupil response as a result of a down-regulation from the early system. Consistent with this account, we did not find a significant difference between the early and late negative component betas ($t(47) = -0.49$, $p = 0.63$), suggesting that the two components may represent similar neural processes.

In line with the proposed role of the LC-NA system in signalling unexpected uncertainty and implementing network resets (Bouret & Sara, 2005; Yu & Dayan, 2005; 2009), we expected that diminished performance and increased uncertainty regarding which option to choose would increase the likelihood of initiating such interrupt signals. To test that a stronger coupling between the early system and the feedback-related pupil response following negative outcomes reflects network resets, we correlated the early and late negative component regressions coefficients derived from the above EEG-informed pupil GLM with participant-wise task accuracy, uncertainty defined as the overall choice probability difference (i.e., the absolute deviation in choice probabilities, see Methods), and exploration tendency.

Conforming to our expectations, we found a significant correlation between task accuracy and the early ($r = 0.32$, $p = 0.02$), but not the late ($r = 0.10$, $p =$

0.48) negative component regression coefficients (Fig.2-6c). In addition, the early (r = 0.29, p = 0.046) but not the late (r = 0.10, p = 0.49) negative regression coefficients significantly correlated with difference in overall choice probabilities between the two symbols (Fig.2-6d). The strong correlation between participant-specific mean accuracy and choice probability difference (r = 0.80, p < 0.001) further indicates that participants who experienced less uncertainty regarding which symbol to choose performed better. Overall, these results suggest that lower accuracy levels and increased choice uncertainty are linked to a stronger coupling between the feedback-related pupil response and negative feedback processing by the early, but not the late, system. This can be readily reconciled with the proposition that lapses in accuracy and increased uncertainty result in NA-induced network resets mediated by the early system that aid the late system in adjusting to the prevailing reward contingencies.

Reversal-induced network resets may also facilitate exploratory behaviour in order to support the establishment of a new model of the external world. Accordingly, we found that the increased coupling between the early (r = -0.29, p = 0.049), but not the late (r = -0.06, p = 0.67), negative component regression coefficients and the pupil response were linked to an increased propensity to choose the symbol with the lower reward value (Fig.2-6e). Correlation coefficients of similar magnitude (r = 0.26, p = 0.08 and r = 0.03, p = 0.86 for the early and late negative components, respectively) were obtained when using the participant-wise inverse temperature parameter from our reward learning model as a measure of exploration tendency. Lower values in the inverse temperature parameter index reduced sensitivity to reward deviations in the available options, and consequently, they mark an increased exploration propensity. We found a strong correlation (r = 0.90, p < 0.001) between the two measures of exploration tendency, suggesting they represent the same underlying construct. This increased exploration propensity linked to increased pupil and early system coupling further implies that LC-driven network resets driven by the early system contribute to increased exploratory behaviour.

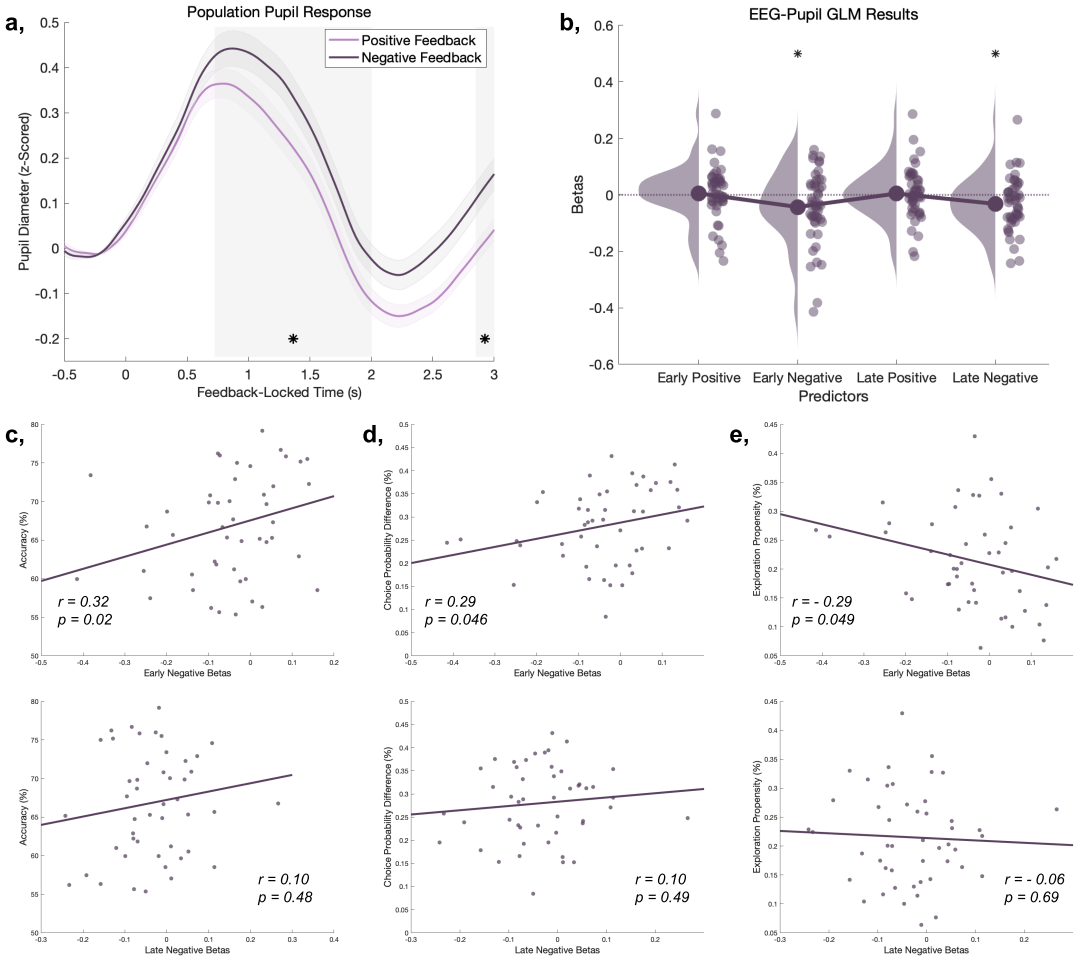

Figure 2-6. Pupil analyses. a, Population pupil response linked to positive and negative feedback trials. The grey shaded area corresponds to periods of time where the positive and negative pupil response significantly differed, derived from two-tailed t-test at each time point, with a Bonferroni-corrected *p*-value of 0.05. **b,** Results of our GLM predicting the full course pupil data using the four EEG-derived predictors (early positive, early negative, late positive, late negative). T-tests ($p = 0.05$) revealed that unlike the early and late positive, the early and late negative component amplitudes significantly predicted the pupil response. **c,** Across-participant correlation between the early (top) and late (bottom) negative EEG-pupil GLM coefficients and mean accuracy. **d,** Across-participant correlation between the early (top) and late (bottom) negative EEG-pupil GLM coefficients and absolute choice probability difference. **e,** Across-participant correlation between the early (top) and late (bottom) negative EEG-pupil GLM coefficients and exploration tendency (i.e., the proportion of choices in which the symbol with the lower reward probability was selected).

Finally, we employed a baseline version of the EEG-informed pupil GLM, in which the trial-wise EEG predictors were replaced by a binary feedback-type predictor, to evaluate the effect of outcome valence on the feedback-related pupil response. To compare the explanatory power related to the EEG-informed and baseline models, we calculated the adjusted R^2 statistic, which revealed that the former better captured variability in the pupil response (Table 2-1). Additionally, secondary significance testing on the binary feedback coefficients of the baseline model showed that feedback valence was not a significant predictor of the pupil response ($t(47) = -1.3$, $p = 0.20$). This result further confirms that the early and late negative components individually contribute to variance within the feedback-evoked pupil response, independently of a valence effect.

2.4.6 Deconvolution of the pupil response

To disentangle the influence of the overlapping early and late EEG component amplitudes on the pupil data, we employed deconvolution analyses based on linear regression (Dimigen & Ehinger, 2021). This method allowed us to exploit the temporal variability in the overlap across component-linked EEG events to derive the unique contribution of each event type. Using this method, we aimed to visualise the significant effect of the early and late negative component amplitudes on the feedback-related pupil response revealed by our EEG-informed pupil GLM. Accordingly, we hypothesised that more negative outcome encoding by the early and the late systems would be associated with a more pronounced pupil response compared to lower component discriminant amplitudes.

To disentangle the unique contribution of high and low EEG discrimination following negative outcomes, we divided all trials into a low or high discrimination condition based on a participant-specific median split of the single-trial component amplitudes, separately for early and late offset. Thus, our design matrix consisted of the continuous pupil data as well as two distinct event types, each of which were broken down into 121 standalone regressors representing each 25 millisecond time section in the three-second

post-feedback interval. This regression-based deconvolution analysis yielded 121 unique regression coefficients for both event types, which can be interpreted as the time series depicting the non-overlapping mean pupil response for the relevant event type (de Gee et al., 2014; Dimigen & Ehinger, 2021; Wierda et al., 2012).

The resulting population time series for the low and high conditions for both early and late offset are shown on Figure 2-7. In congruence with our hypothesis, the high discrimination condition is associated with a more pronounced pupil response from around 300 to 800 ms post-feedback for early offset and from around 300 to 1300 ms post-feedback for late offset. However, on the declining phases of the pupil response the opposite is true for both offset types. Formally contrasting the regression coefficients for the low and high negative conditions during each time point in the post-feedback interval did not reveal any significant difference (all two-tailed *t*-tests contrasting the high and low *y* conditions at each time point showed (*DF* = 47), $p > 0.05$, for both the early and the late offset conditions).

Finally, to cross-check our results from the deconvolution analyses, we extracted the feedback-evoked pupil response for each 25 millisecond time window in the post-feedback interval separately for high and low discrimination trials following negative feedback for both early and late offset. This analysis yielded similar results as the deconvolution analyses; high discrimination amplitudes following negative feedback lead to a stronger pupil response until around 1500 mspost-feedback for both the early and late offset categories. In line with the deconvolution results, a comparison of the evoked pupil response between low and high discrimination associated with each time point revealed no significant distinction (for all two-tailed t-tests contrasting the high and low *y* conditions at each time point (*DF* = 47), $p > 0.05$, for both early and late offset conditions).

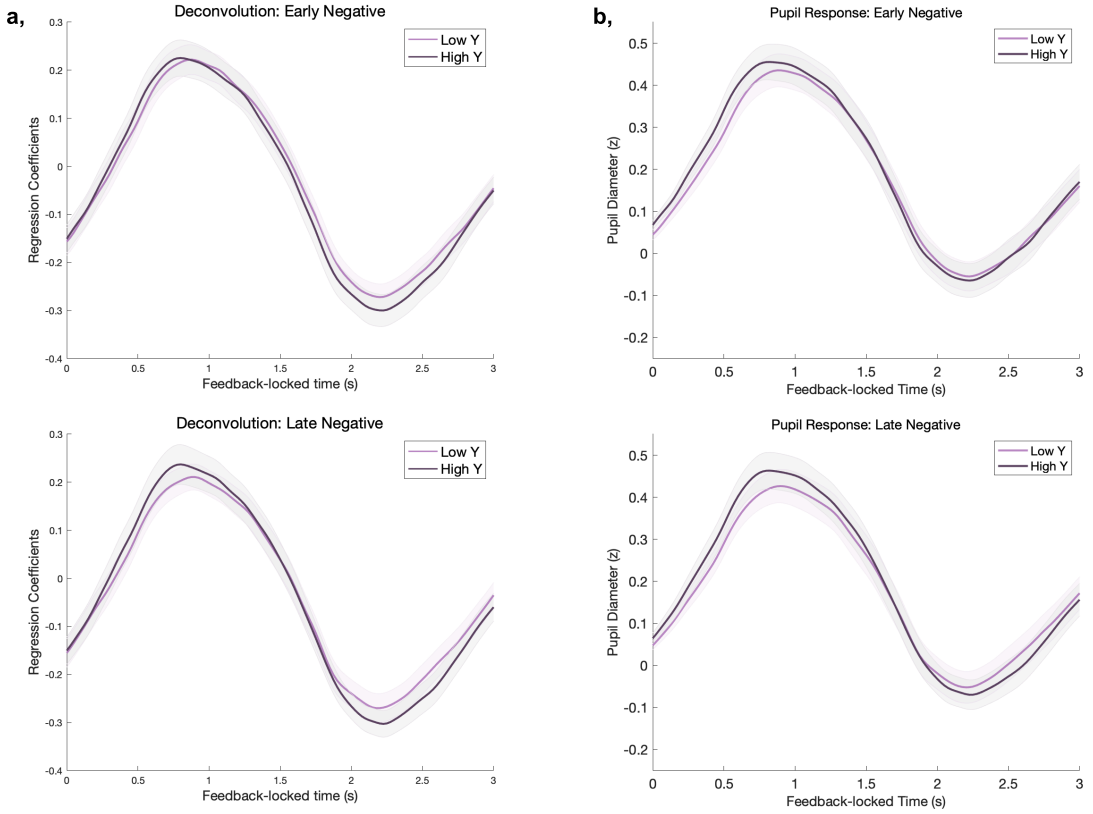

Figure 2-7. Pupil Deconvolution. Feedback-locked pupil responses were split by the magnitude of the EEG component amplitudes following negative feedback, separately for early and late offset. **a,** Population regression coefficients from the deconvolution analysis decoupling the impact of low versus high EEG discrimination amplitudes following negative outcomes for the early (top panel) and late (bottom panel) valence components. **b,** The feedback-evoked pupil response following negative feedback is categorised by the magnitude of the EEG discriminant amplitudes (low or high) associated with the early (top) and late (bottom) components.

2.5 Discussion

By exploiting the trial-by-trial variability in the feedback-locked EEG signal produced during probabilistic reversal learning, we replicated the two spatiotemporally distinct neural systems associated with feedback evaluation during reward learning (Fouragnan et al., 2015; 2017). Importantly, correlating trial-wise electrophysiological and pupillometry data allowed us to isolate the neural networks responsible for generating the differential feedback-valence-related pupil response. Moreover, associations across measures in the electrophysiological, pupillometry, and behavioural domains provided additional explanation about the function of the two systems involved in reward learning.

Whilst the LDA-derived scalp maps of the early and late components and the timing of the early component closely matched the results reported by Fouragnan and colleagues (2015), our late component showed increased variability and a later mean latency (300 vs 410 ms post-feedback). This discrepancy is likely due to the difference in the type of feedback stimuli used during the experiments; there was only a subtle visual divergence between positive and negative feedback stimuli in our task in order to minimise the impact of visual processing on the pupil response. This reduced visual salience of feedback valence could have contributed to a more sluggish feedback evaluation process. Nevertheless, the latency of the early component is consistent with the amount of time it takes for the LC to phasically respond to events and circulate signals to frontal structures (150-200 ms; Laeng et al., 2012), consistent with the hypothesis that LC-NA activity modulates the early system. On the other hand, the late component, due to its spatial characteristics, latency, and significant role in reward value updating, is likely under dopaminergic control (Fouragnan et al., 2015). Consistently, prominent structures of the late system, such as the OFC and the vmPFC have been repeatedly associated with the reward valuation network and exploitative behaviour (de A Marcelino et al., 2023). Overall, the

successful replication of these two components is further evidence for the robustness of two spatiotemporally distinct reward learning systems.

In line with previous results (de Gee et al., 2021; Schneider et al., 2018; Urai et al., 2017), the feedback-evoked pupil response was greater for negative than positive outcomes. Results from our general linear and linear mixed effects modelling approaches likewise implicated both the early negative and the late negative components to drive the feedback-related pupil response. Notably, despite the link between feedback salience and the stimulus-evoked pupil response (de Gee et al., 2014; 2017; 2021; Filipowicz et al., 2020; Preuschoff et al., 2011; Urai et al., 2017, Van Slooten et al., 2017; Varazzani et al., 2015), our analysis accounted for the impact of salience, indicating that the impact of the early and late negative components on the pupil response occurs independently of this effect. This result is further supported by recent evidence (van Slooten et al., 2018) that the feedback-related pupil response signals the level of uncertainty rather than salience.

The strength of the coupling between the early, but not the late, system and the feedback-evoked pupil response were strongly associated with behavioural markers associated with network resets. Consistent with the finding that the early component down-regulates the late component following negative outcomes (Fouragnan et al., 2015), it is likely that the coupling between the late system and the pupil response following negative outcomes merely reflects the regulatory control originating in the early system. This interpretation is supported by a recent finding that influence in the striatum over exploratory behaviour is likely exerted via thalamocortical disinhibition originating in frontal structures linked to exploration such as the ACC, anterior insula, and the dorsolateral prefrontal cortex, all of which are prominent structures of the early system. Other studies have similarly suggested that exploration is accomplished by frontal structures overriding the dopamine-mediated exploitation tendency, which process is crucial for effective behavioural adaptation (Cavanagh et al., 2011; Daw et al., 2006; Hassal et al., 2013; de Marcelino et al., 2023).

Correspondingly, increased activity in the early system following negative outcomes was linked to response slowing, whilst larger associations between early negative component amplitudes and the pupil response were negatively correlated with overall accuracy and positively correlated with choice ambiguity as well as exploration tendency. Finally, participants whose negative feedback encoding by the early system more strongly predicted value updating, reflecting increased internal estimates of unexpected uncertainty, were characterised by poorer performance. These results support the proposition that increased estimates of contextual uncertainty, produced by negative feedback, are signalled to early system structures by increased LC-NA system activity, in turn interrupting processing in the late system in an attempt to reduce the influence of learnt reward contingency representations. Such a network reset would consequently decelerate subsequent decision making as evidence is accumulated in favour of a new, competing hypothesis (i.e., reversal), in turn increasing the likelihood of exploring (i.e., choosing the option that was previously considered inferior).

Consistent with this interpretation, phasic LC-NA activity and the feedback-evoked pupil response have been linked to within-task uncertainty signalling (Colizoli et al., 2018; Dayan & Yu, 2006; de Gee et al., 2014; 2017; Lavín et al., 2014; Nassar et al., 2010; Pajkossy et al., 2017; Preuschoff et al., 2011; Urai et al., 2017; Yu & Dayan, 2005; van Slooten et al., 2017; 2018). Moreover, increased pupil evoked responses have likewise been linked to diminished accuracy and post-error slowing (Aston-Jones & Cohen, 2005; Murphy et al., 2016; Urai et al., 2017). This is compatible with the positive association found between phasic pupil responses and higher hazard rates (i.e., change-point probability; Nassar et al., 2012) and increased neural gain (Aston-Jones & Cohen, 2005; Bouret & Sara, 2005; Dayan & Yu, 2006; Eldar et al., 2013; Filipowicz et al., 2020; Gilzenrat et al., 2010; Krishnamurty et al., 2017; Nassar et al., 2012; Zenon, 2019), indicating the LC-NA system in adaptively regulating the cortical influence of new versus learnt information.

Similarly, separate accounts have proposed that the LC-NA network achieves such behavioural flexibility in response to unexpected events through the implementation of a cortical network reset (Bouret & Sara, 2005, Dayan & Yu, 2006; de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009; Urai et al., 2017). Such internal resets could dynamically interrupt the current neuronal organisation in the relevant network, thereby promoting rapid behavioural adaptation, which is conducive to optimal performance in a changing environment. Whilst high phasic LC firing is considered to promote, low phasic activity is thought to prevent neural and behavioural shifts as a response to external events, thereby facilitating behavioural adaptation (Bouret & Sara, 2005). As the LC-driven interrupt signal is hypothesised to reflect within-task state change, the detection of which has been associated with increased exploratory behaviour (Bland & Schaefer, 2012; Daw et al., 2006; Jepma et al., 2020), it could also promote explorative behaviour as reversals would disrupt recent value representations in reward processing structures. Indeed, despite the noradrenergic network and the pupil response have been repeatedly linked to explorative behaviour (Aston-Jones & Cohen; 2005; Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011; Yu & Dayan, 2005; 2009; van Slooten et al., 2018), it is yet unclear how phasic, tonic, or perhaps intermediate (Urai et al., 2017) LC activity regulates the above effects.

In order for the LC to implement such network resets, it needs to have access to a dynamic combination of bottom-up and top-down information (Filipowicz et al., 2020). The ACC is a prime candidate for providing these inputs to the LC (Aston-Jones & Cohen, 2005; de Gee et al., 2017) as it has strong reciprocal connection with the LC (Briand et al., 2007; Joshi & Gold, 2020), it has access to both top-down and bottom-up information via its robust connections with prefrontal structures and sensorimotor areas, and it is a prominent structure of the early system, which takes part in the down-regulation of the late system (Fouragnan et al., 2015). ACC activity has also been found to mediate pupil responses (Joshi et al., 2016; Joshi & Gold, 2020; Murphy et al., 2014), which effect appeared to be exerted via the LC (Joshi & Gold, 2020; Schneider et al., 2018). Moreover, the ACC has also been

functionally associated with cognitive control (Shackman et al., 2011) and the promotion of adaptive behaviour, such as task difficulty monitoring (Ullsperger & von Cramon, 2001), conflict processing (Botvinick et al., 2001; 2004; Etkin et al., 2011), error detection (Rushworth et al., 2007; Yeung et al., 2004), task volatility tracking (Behrens et al., 2007), exploration (Chakroun et al., 2020; Daw et al., 2006; Forstmann et al., 2006; Marcelino et al., 2023), and influencing the weight of decision feedback on upcoming decisions (Rushworth & Behrens, 2008), all of which processes are crucial for the implementation of network resets. Accordingly, adaptive gain theory (Aston-Jones & Cohen, 2005) proposes that ACC signalling to the LC produces phasic noradrenergic firing, which in turn increases cortical NA levels that enhance cognitive processing, perhaps by resetting reward learning networks. Further research is nevertheless needed to determine the precise role the ACC plays in facilitating such network resets.

Whilst our results indicate that the increased link between the early component and the feedback-evoked pupil response is associated with increased exploration tendency, our experimental design did not allow for a definite conclusion regarding exploration subtype. Recent evidence suggests that only random, and not directed, exploration tendency is affected by noradrenaline-reuptake inhibiting medication (Warren et al., 2017; Zajkowski et al., 2017). Whilst direct exploration promotes information-seeking and in turn facilitates optimal decision making, random exploration (decision noise) requires less neural computational power and is therefore more tenable in complex tasks and uncertain environments. Hence, effective problem solving requires dynamic transformations across these two forms of explorations (Wilson et al., 2014). Dopamine has also been implicated in mediating the exploration-exploitation trade-off (Cinotti et al., 2019; van Slooten et al., 2019), with research implicating prefrontal dopamine in guiding directed exploration (Frank et al., 2009). Whilst it is enticing to attribute direct exploration to dopamine and random exploration to noradrenaline, more research is vital for an improved characterisation of the roles these neurotransmitters play in different exploration types.

In the current study, we examined reward learning in the appetitive domain. However, animal research indicates that avoidance learning is implemented differently by midbrain neurons depending on whether feedback represents a true loss or an omission of reward (Fiorillo 2013;, Fiorillo et al., 2013). Consistently, recent evidence suggests increased pupil responses following aversive compared to appetitive conditions (Finke et al., 2021), suggesting that the LC-NA network might differentially respond under conditions of positive versus negative reinforcement. Nonetheless, work in progress in our lab replicated the early and late feedback valence systems during both appetitive and aversive learning conditions. In line with our results, the feedback-evoked pupil response was significantly more pronounced for negative compared to positive outcomes, independently of learning type. Thus, it appears that the mechanisms reported here may be fundamental constituents of the reward learning process, during both reward and punishment learning.

Despite considerable evidence indicates the LC in generating the changes in pupil size related to cognitive processes (Joshi et al., 2016; Joshi & Gold, 2020; Murphy et al., 2014; Reimer et al., 2016), other brainstem nuclei, including dopaminergic, cholinergic, and serotonergic structures, as well as their connections with the LC, could partly moderate this pupil effect (de Gee et al., 2017; Reimer et al., 2016; Urai et al., 2017; van Slooten et al., 2018). It has been shown that the amount of NA circulated in the brain depends on the level of acetylcholine (Preschoff et al., 2011) and cholinergic activity in the basal forebrain (Murphy et al., 2014; Nelson & Mooney, 2016; Reimer et al., 2016). Furthermore, serotonergic (Schmid et al., 2015; Vitiello et al., 1997) and dopaminergic structures (de Gee et al., 2017; Joshi et al., 2016; Reimer et al., 2016; Wang & Munoz, 2015) could also play a role in the pupil dilation effect.

As with pupil dilation, different neurotransmitters have been associated with adaptive learning processes. Apart from its well-established role in value updating and approach learning (Seymour et al., 2007; Schultz et al., 1999;

Wise, 2004; Varazzani et al., 2015), dopamine has been found to mediate the explore-exploit trade-off (Chakroun et al., 2020; Cinotti et al., 2019; Frank et al., 2009; van Slooten et al., 2019). Additionally, while noradrenaline has been suggested to signal unexpected uncertainty, acetylcholine has been linked to indicating the degree of expected certainty in the task at hand (Yu & Dayan, 2005). Furthermore, each neurotransmitter system is known to send projections to and receive modulatory input from similar prefrontal target structures, in turn modulating each other via prefrontal feedback loops (Briand et al., 2007). Noradrenaline and dopamine appear particularly interconnected as dopamine has been detected to be released from the LC, and noradrenergic and dopaminergic neurons responds to similar environmental stimuli, causing these neurotransmitters to be simultaneously released in frontal cortex (Briand et al., 2007; Devoto & Flore, 2006). These results are consistent with the proposition that uncertainty-dependent exploration is mediated through an interaction across the dopaminergic, cholinergic, and noradrenergic systems (Bland & Schaefer, 2012; Cohen et al., 2007). Thus, both pupil dilation and the rapid network reconfigurations needed for adaptive learning are most likely achieved by these different neurotransmitter systems operating in concert. Further research is needed to disentangle the complementary contribution as well as mode of action (i.e., phasic vs tonic) of each neurotransmitter system to these processes. Whilst pupillometry is an inexpensive and non-invasive tool for such investigation, high-resolution imaging of brainstem nuclei as well as single-cell recordings in animals could provide particularly influential insights.

Our study incorporating single-trial electrophysiological, pupillometry, behavioural, and modelling measures replicated the previously reported two spatiotemporally distinct reward learning components (Fouragnna et al., 2015). We showed that the increased feedback-related pupil response for negative compared to positive outcomes is exclusively driven by negative feedback encoding, likely originating in the early system. In line with previous accounts implicating the LC-NA system in uncertainty signalling and interrupting cortical organisation, we propose that when internal estimates of

contextual uncertainty are high following negative feedback, the early system, regulated by LC-NA activity, implements a network reset in reward-processing structures of the late system. This interruption serves to improve performance by creating new, more accurate internal representations of external reward values. Additionally, our EEG- and pupillometry-based research methods into the neural correlates underlying reward-based decision making could be of further value in providing insights about decision making mechanisms altered in non-verbal populations (infants, patients with aphasia or locked-in syndrome, animals) as well as an array of neuropathological conditions characterised by altered noradrenergic activity, such as dementia (Hermann et al., 2004), mood and anxiety disorders (Brunello et al., 2003; Ehlers & Todd, 2017; Leonard, 1997), addiction (Torregrossa, 2019), or schizophrenia (Yamamoto & Hornykiewicz, 2004).

Chapter 3. Increased negative feedback processing reduces evidence accumulation during value-based decisions

3.1 Summary

For the first time, we characterised post-feedback response adaptation during value-based learning via neurally informed sequential sampling modelling. Using linear discriminant analysis on the feedback-locked EEG data, we replicated the early and late reward learning components reported by Fouragnan and colleagues (2015). Next, we utilised drift diffusion modelling to link these components, broken down by feedback valence type (positive and negative), to inter-trial behavioural adaptation. We found that increased negative feedback processing by the early and late systems reduced evidence accumulation in the next trial, whilst stronger positive feedback processing by these systems increased both evidence accumulation and boundary separation. Moreover, increased coupling between the feedback-locked pupil response and early system activity following negative outcomes was positively linked to the degree of drift rate reduction induced by the early system. These results support the hypothesis that increased noradrenergic activity in response to increased uncertainty is signalled to the early system, which in turn down-regulates late system activity. This down-regulation presumably promotes behavioural adaptation through decelerating decision making as a change in reward contingencies is considered more plausible and learned value representations become more obsolete.

3.2 Introduction

In the previous chapter, we showed that the increased feedback-related pupil response for negative compared to positive outcomes is exclusively driven by enhanced negative feedback encoding, likely originating in an early feedback system regulated by locus-coeruleus noradrenergic (LC-NA) activity. Consistent with the proposed role of the LC-NA system in cortical network resets (Bouret & Sara, 2005, de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009; Yu & Dayan, 2005), we found that increased coupling between early system activity and the pupil response following negative outcomes was associated with reduced accuracy, response slowing, and increased exploratory behaviour.

To disassociate the contribution of the early and late feedback systems (Fouragnan et al., 2015) to successive decision making and probe the hypothesis that LC-NA-induced network resets in reward learning structures reduce evidence accumulation in subsequent decisions, we utilised drift diffusion modelling (DDM; Ratcliff, 1978; Ratcliff & McKoon, 2008; Wiecki et al., 2013) on our reversal learning data set introduced in the previous chapter. The DDM belongs to the family of sequential sampling models (SSMs; Ratcliff & Smith, 2004; Ratcliff et al., 2016), which jointly account for accuracy and reaction time (RT) data in decision tasks. SSMs conceptualise decision making as an information accumulation process, whereby momentary evidence is accumulated in favour of one of the choice alternatives over time until evidence reaches a threshold and a response is initiated. A major advantage of this framework lies in its capacity to decompose the relationship between RT and accuracy data into meaningful psychological constructs.

The DDM is perhaps the most prominent SSM for two-alternative forced choice decisions; it has been successfully applied to a large number of tasks, including perceptual, social, value-based, memory, lexical, or categorisation paradigms (for a review, see Ratcliff et al., 2016). The model assumes that a single accumulator integrates evidence in favour of each choice alternative

according to a stochastic drift diffusion process. The process terminates when the accumulated evidence crosses one of the decision thresholds corresponding to each choice option. The DDM decomposes choice and RT data into four latent parameters mapped onto internal psychological processes. The drift rate *v* reflects the rate of evidence accumulation, with higher values implying higher efficiency, and in turn faster and more accurate responses. Boundary separation *a* denotes the amount of evidence required to make a decision, with higher values indicating increased response caution, leading to slower but more accurate decisions as more evidence is accumulated until the threshold is reached. The starting point *z* expresses *a priori* bias, whereby evidence accumulation does not start halfway between the boundaries, but is biassed towards one of the choice options. Lastly, non-decision time *t* denotes the time taken for stimulus encoding and motor response initiation. As previous results indicated that the inclusion of inter-trial variability parameters for the drift rate, starting point, and non-decision time improved model fits (Ratcliff & McKoon, 2008), we also incorporated these into our model.

At the neural level, distinct brain structures have been shown to reflect the different psychological constructs of the DDM. Drift rate has been linked to activity in the striato-thalamo-cortical network (Frank et al., 2015; Turner et al., 2015; Mulder et al., 2014), whilst the fronto-basal ganglia network has been suggested to control the degree of boundary separation (Cavanagh et al., 2011; Forstmann et al., 2010; Mulder et al., 2014; Turner et al., 2015). Given that largely distinct neural activity appears to reflect elements of the drift diffusion process, trial-wise measurements of the related brain activity can be utilised to explain variability in parameters and constrain the model based on these neural processes (Frank et al., 2015; Franzen et al., 2020; Mattes et al., 2022; Verdonckhove et al., 2011).

Although several studies have explored the computational and neurophysiological mechanisms that transform sensory evidence into a choice (Basten et al., 2010; Cavanagh et al., 2011; 2014; Chakroun et al., 2023;

Forstmann et al., 2010; Frank et al., 2015; Franzen et al., 2020; Krajbich et al., 2010; Nunez et al., 2017; Ratcliff et al., 2009; Stock et al., 2016; Turner et al., 2015; White et al., 2012; 2014; Wiehler et al., 2021), to our knowledge, only a single study (Mattes et al., 2022) utilised neurophysiological measures to characterise post-feedback response adaptation within the DDM framework. In this perceptual decision study, trial-wise and feedback-locked event-related potentials (correct response negativity and error positivity) were found to balance the opposing demands of responding fast and accurately, with opposite effects on the decision threshold. Other studies investigating post-response adaptation within the DDM framework utilised behavioural markers of feedback type (Dutilh et al. 2012; Fontanesi et al., 2019; Cohen Hoffing et al., 2018), leaving it unexplored how neural mechanisms implicated in response monitoring affect subsequent choices.

3.2.1 Current study

To address this gap, we implemented drift diffusion modelling, constrained by the trial-wise, feedback-locked electroencephalogram (EEG) components (early positive, early negative, late positive, late negative) discriminant amplitudes derived in the previous chapter (section 2.4.2). As in the preceding chapter, we hypothesised that if the early system implements LC-induced network resets in reward learning structures of the late system following negative outcomes, subsequent decision making would be decelerated as evidence is accumulated towards a new, competing hypothesis (i.e., reversal in reward contingencies). Consequently, we hypothesised that the drift rate parameter in the DDM will be reduced on trials following negative outcomes (Fig.3-1). In line with the hypothesised role of the early system in cortical resets and subsequent role of the late system in updating value representations, we expected the reduction in the drift rate to scale with both early and late system activity. To our knowledge, this is the first value-based decision making study that utilises a neurally informed SSM to characterise post-feedback response adaptation.

To disentangle the effects of our feedback-locked EEG components on the drift rate and boundary separation processes in subsequent decisions, we fit three neurally informed DDMs to our reversal-learning data. In these models, either the drift rate, boundary separation, or both varied as a function of the single-trial EEG discriminant component amplitudes from the previous trial. We implemented these three models to ensure that the accuracy drop following negative versus positive feedback (Fig.2-3b) is not solely explained by increased boundary separation, but is, at least partially, generated by a diminished evidence accumulation process (Fig.3-1). The best fitting model supported our hypothesis; increased negative feedback processing by the early and late systems reduced evidence accumulation in the following trial.

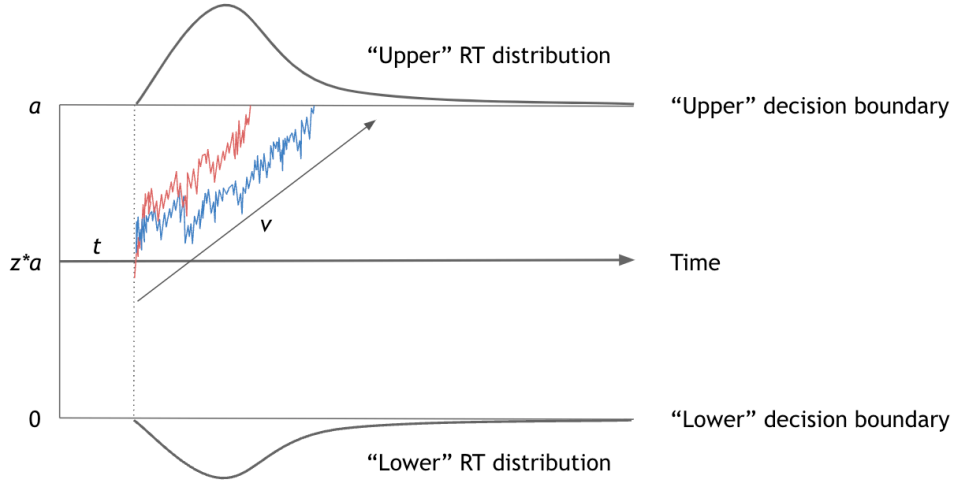

Figure 3-1. Hypothesised drift rate effects. Drift trajectories are shown following positive and negative feedback within the drift diffusion modelling (DDM) framework. Evidence is stochastically accumulated over time (*x*-axis) with average drift rate *v* until one of the two boundaries (at 0 and *a*, representing the "upper" and "lower" boundaries, respectively) is crossed, and a response is initiated. The diffusion process begins at the starting point between the two boundaries (marked by the proportion of *a* by *z*). We hypothesised that evidence accumulation (i.e., drift rate) following negative (blue line) compared to positive (red line) feedback will be reduced. Total response time is the sum of the non-decision time *t*, denoting the time taken for stimulus encoding and motor processing, and the duration of the diffusion process. The upper and lower plots represent RT distributions for drift diffusion processes hitting the upper and lower boundaries, respectively.

Across-participant correlations further revealed that the increased coupling between the feedback-locked pupil response and early system activity following negative outcomes (see our EEG-informed pupil analysis in section 2.4.5) was positively associated with the early system induced drift rate reduction. This result further supports the conjecture that increased uncertainty produced by negative feedback is signalled to the early system by increased LC-NA system activity, which in turn interrupts processing in late reward learning structures as a means to reduce the influence of learnt reward contingency representations. This network reset would consequently decelerates decision making as evidence for a new, competing hypothesis is accumulated. Finally, results from a behavioural-only model further confirmed that negative feedback processing results in diminished evidence accumulation in the next trial, further substantiating our findings.

3.3 Method

3.3.1 Participants, stimuli, and experimental task

The participants, stimuli, experimental task, and procedure referred to in this chapter are described in sections 2.3.1-3.

3.3.2 EEG data acquisition, pre-processing, and analysis

EEG data acquisition, pre-processing, and multivariate analysis methods are identical to those specified in sections 2.3.5-7.

3.3.3 Hierarchical drift diffusion model

We fit hierarchical drift diffusion models to participants' choice and reaction time data using the HDDM toolbox (version 0.9.8; Wiecki et al., 2013) in Python (version 3.7.12, van Rossum & Drake, 2009) via Jupyter Notebook (version 6.5.5; Kluyver et al., 2016). HDDM implements a hierarchical Bayesian estimation method, which allows for the simultaneous estimation of individual and group-level parameters at different hierarchical stages by assuming that model parameters for individual participants are randomly drawn samples from the group-level distribution. Consequently, the more similar participants behave to each other, the smaller the variance in group-level parameter distribution becomes, in turn more strongly constraining parameter estimates of individual participants. This approach has been shown to lead to more reliable parameter estimation, especially when less data is available or single-trial neural data is used for parameter estimation (Matzke et al., 2013; Wiecki et al., 2013). Additionally, the Bayesian method generates joint posterior distributions for all parameters, which not only provides an estimate for the most likely value of the given parameter, but also quantifies uncertainty in parameter estimation.

The DDM (Ratcliff 1978) assumes that during two-choice decisions, responses are determined by a noisy evidence accumulation process. Evidence in favour

of a choice alternative can fluctuate from time point to time point based on the level of stimulus noise, its neural representations, or the attention paid to task and stimulus features. Observed choice and RT data $x_{i,j}$ for participant *i* on trial *j* is represented by the DDM likelihood function (Navarro & Fuss, 2009);

$$x_{i,j} \sim F(a_i, z_i, v_i, t_i, sv, st, sz). \quad (3.1)$$

The rate of evidence accumulation is determined by the drift rate parameter *v* and specifies the amount of evidence accumulated per unit of time. The drift rate can reflect participants' task efficiency as well as task difficulty, such that higher values of *v* produce faster and more accurate responses. A response is executed when enough relative evidence is accumulated in favour of a choice option, i.e., when the amount of relative evidence in favour of a stimulus crosses the decision threshold *a*. This parameter controls the speed-accuracy trade-off (Ratcliff & Rouder, 1998; Zhang & Rowe, 2014) and reflects response caution; higher values of this free parameter lead to slower but more accurate responses. The non-decision time parameter *t* captures RT components linked to the perceptual encoding of stimuli and motor processing after a choice is selected. Consequently, this parameter is independent of evidence accumulation, and affects RTs without influencing accuracy. The starting point bias parameter *z* captures bias towards one of the choice alternatives and determines the start of the drift diffusion process relative to the two boundaries. Thus, values of $z > 0.5$ suggest an *a priori* bias towards the upper boundary and values of $z < 0.5$ imply a bias towards the lower boundary. In case of no bias (i.e., when there is an equal amount of evidence in favour of both choice options), evidence accumulation begins halfway between the two boundaries, with $z = 0.5$. In all our models, we implemented accuracy coding; the upper boundary reflected choices where the symbol with the higher reward probability was selected, whilst the lower boundary reflected responses where the symbol with the lower reward probability was chosen by participants.

We implemented the 'full DDM' (Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998), which includes three inter-trial variability parameters *st*, *sv*, and *sz* to capture the variability in non-decision time, drift rate, and starting point, respectively. The full DDM has been shown to outperform the simpler version of the model as it is able to account for the behavioural phenomena that errors are either faster or slower then correct responses (Ratcliff & Rouder, 1998). We estimated only the group-level values for these three parameters as the influence of the individual inter-trial variability parameters is often modest or even absent, with a large dataset required for reliable estimation (Wiecki et al., 2013).

3.3.4 Model fitting

To estimate the influence of the participant-specific EEG feedback components on subsequent decision making, we utilised the regression functionality within the HDDM toolbox. This functionality allowed us to model the linear link between our neural predictors and HDDM parameters on a trial-by-trial basis by incorporating single-trial EEG component discriminant amplitudes into the parameter estimation process. We estimated posterior distributions for HDDM parameters as well as for regression coefficients specifying the degree to which model parameters are influenced by the trial-to-trial variations in the electrophysiological measures.

As Figure 3-2a shows, EEG component discriminant amplitudes were derived from our linear discriminant analysis (for more details, see section 2.3.6), whereby we extracted single-trial discriminator amplitudes from the participant-specific temporal windows corresponding to the early and late feedback-locked components (Fouragnan et al., 2015). The early and late components were further broken down by feedback valence type (positive and negative), yielding four components in total; early positive, early negative, late positive, late negative (Fig.3-2b). These components reflect the participant-specific feedback valence discriminating amplitudes (i.e., y_i values from Eq.2.2.11) linked to each component.

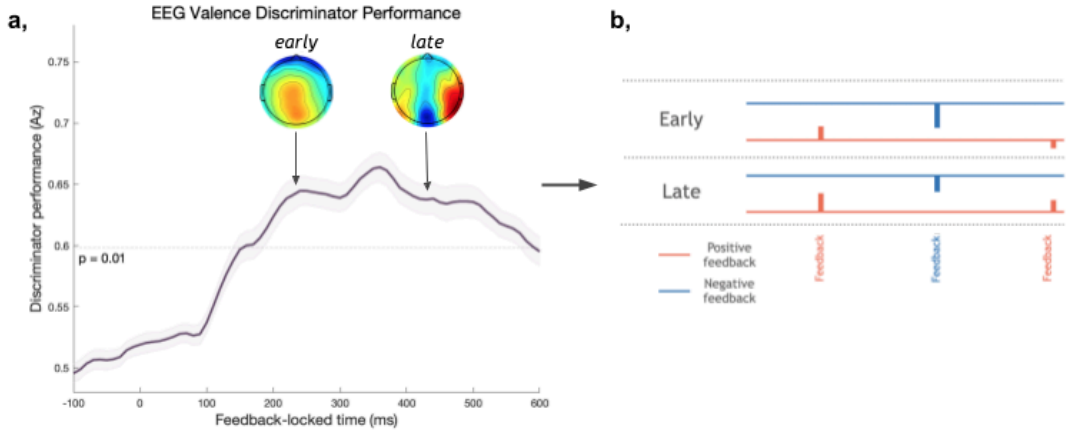

Figure 3-2. EEG-informed regressors. Regressors for the DDM were derived from our single-trial linear discriminant analysis on the feedback-locked EEG signal (see section 2.3.6). **a,** Single-trial discriminator performance (cross-validated *Az*) during valence discrimination (positive vs negative feedback) of the feedback-locked EEG signal, averaged across participants (n = 48). The horizontal line illustrates the *Az* leading to a significance level of p = 0.01, approximated using a bootstrap test. The shaded error band shows across-participant standard errors. We replicated the early and late valence components that were established by Fouragnan and colleagues (2015). The two scalp maps illustrate the population average spatial topographies of the early (left) and the late (right) components. **b,** The neural regressors for the HDDM included the component discriminant amplitudes (y_i, Eq.2.2.11) derived from our feedback-locked multivariate EEG analysis. The early and late components were broken down by feedback valence type (positive, negative), yielding four predictors in total; early negative, early positive, late negative, and late positive.

To test our hypothesis that increased feedback processing by the early and late systems following negative outcomes reduced evidence accumulation in subsequent trials, we utilised the above four single-trial component amplitudes as predictors to explain parameter variability. Specifically, component discriminating amplitudes from trial j-1 were used to predict the drift rate and/or boundary separation on trial j according to

$$v_i = v_{i0} + v_1 * Y_{EarlyPos} + v_2 * Y_{EarlyNeg} + v_3 * Y_{LatePos} + v_4 * Y_{LateNeg}$$

$$a_i = a_{i0} + a_1 * Y_{EarlyPos} + a_2 * Y_{EarlyNeg} + a_3 * Y_{LatePos} + a_4 * Y_{LateNeg} \quad (3.3)$$

Here, v_{i0} and a_{i0} represent the participant-specific intercept terms for the drift rate and boundary separation parameters, respectively. Additionally, $v_{1\text{-}4}$ and $a_{1\text{-}4}$ weigh the slope of the drift rate and boundary separation parameters, respectively, by the values of the four EEG component discriminant amplitudes on the previous trial (Fig.3-3). We utilised three separate neural models; either the drift rate, boundary separation, or both varied according to the four EEG component amplitudes. To more reliably compare the three different model variants, we estimated neural regression coefficients on the group-level only (Fig.3-3) as individual-level estimation of HDDM regression coefficients has been suggested to increase parameter collinearity (Frank et al., 2015), and could therefore bias results.

To further probe the hypothesised relationship between LC-NA system mediated network resets in reward learning structures and reduced evidence accumulation in subsequent decisions, we adjusted the best-fitting neurally informed HDDM to estimate both individual- and group-level parameters for the neural regression coefficients. This in turn allowed us to explore the relationship between how the early or late negative components predict the pupil response during feedback (Section 2.4.5) and affect evidence accumulation on the following trial. We hypothesised that participants with a stronger coupling between the early/late negative component and the pupil response would also exhibit a stronger reduction in evidence accumulation in the next trial as a function of early/late system activity. To evaluate this hypothesis, we utilised robust bend correlations (for more details, see section 2.3.10) to quantify the relationship between the regression coefficients linking early/late negative discriminant amplitudes to the feedback-related pupil response (Section 2.4.5) and the regression coefficients determining the influence of the early/late negative component amplitudes on evidence accumulation in the next trial.

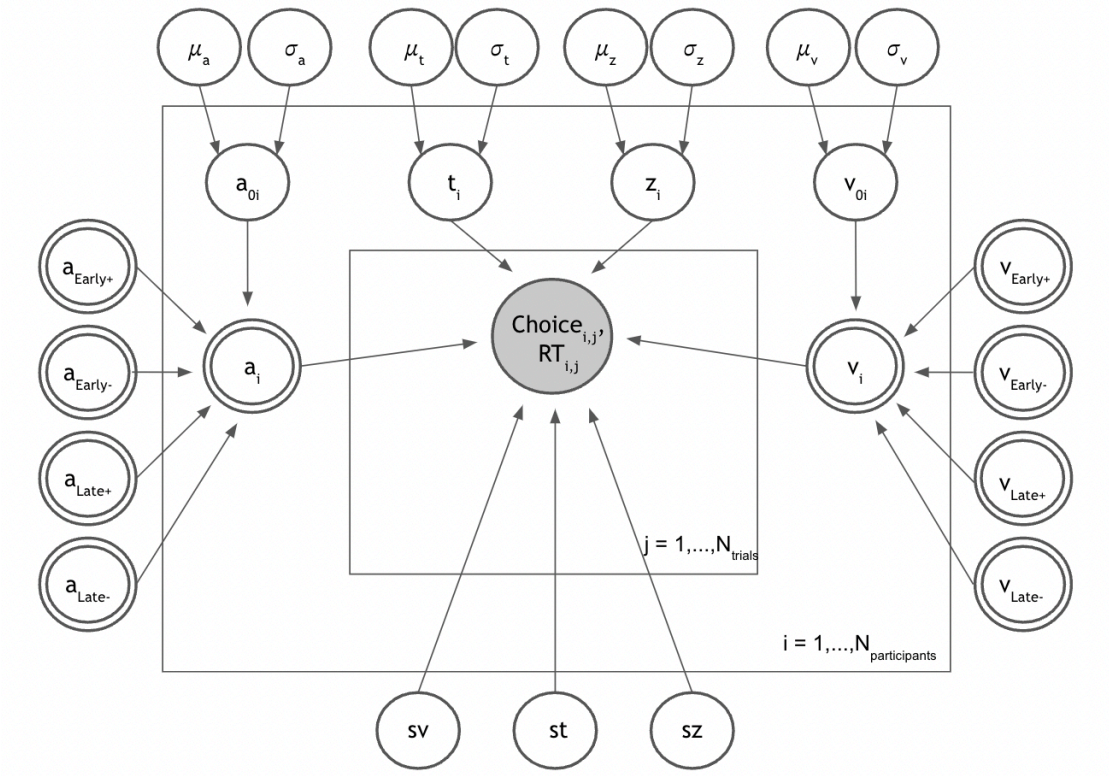

Figure 3-3. Graphical model illustration. Our hierarchical drift diffusion model (HDDM) with trial-wise neural regressors is shown**.** Round nodes illustrate continuous random variables and double-bordered nodes illustrate deterministic variables, defined in terms of other variables. Shaded nodes represent observed data (RT, choice), whilst unshaded notes represent latent variables. Participant-specific parameters t_i, v_i, a_i, and z_i are estimated from individuals drawn from a group distribution with inferred mean μ and variance σ. Trial-to-trial variations in the decision threshold a and drift rate v are determined by the amplitudes of four EEG discriminating components (y_i, Eq.2.11). Multiple random variables share the same parents and children (e.g., each participant specific threshold parameter a_i shares the same parents that determine the group distribution). The inner plate is over trials j and the outer plate is over participants i. Trial-by-trial neural regression coefficients were estimated on the group-level only to increase estimation reliability and avoid parameter explosion.

To verify our results from these neurally informed models, we fit an HDDM to our data with behavioural-only feedback type regressors. Specifically, we modelled the drift rate and boundary separation as a linear function of a binary predictor specifying feedback valence type (positive or negative) according to

$$v_i = v_{i0} + v_{Neg}$$
$$a_i = a_{i0} + a_{Neg}\,. \qquad (3.4)$$

Here, v_{i0} and a_{i0} represent the participant-wise drift rate and boundary separation, respectively, following positive outcomes trials, whilst v_{Neg} and a_{Neg} represent the effect of negative feedback on the drift rate and boundary separation, respectively. Similarly to the neurally informed models, we implemented the full HDDM, with inter-trial variability parameters and the effect of negative feedback on the drift rate and boundary separation estimated on the group-level only. All other parameters were derived on both the group- and individual-levels. We expected our results from the behavioural-only and neurally informed models to align, with a similar effect of feedback (encoding) on the drift rate and boundary separation parameters.

For all our models, we utilised Markov Chain Monte Carlo (MCMC) sampling (Gamerman & Lopes, 2006) within the HDDM package to derive the joint posterior distribution for all parameters. For each model, we generated 10,000 samples, discarded the first 5,000 as burn-in, and thinned the remaining 5,000 samples by keeping every fifth draw, resulting in a total of 1,000 samples of the joint posterior distribution of the parameters. Following the recommendation by Wiecki and colleagues (2013), we used the default, informative priors implemented in the HDDM. These priors are based on results from a set of 23 studies, and were found to improve parameter recovery and reduce issues related to collinearity (Matzke & Wagenmakers, 2009). We used the default outlier setting in the HDDM, whereby 5% of RT outliers are removed to account for responses produced by processes other than drift diffusion, such as lapses in attention. Thus, we estimated a mixture model, whereby 5% of the trials are considered to be distributed according to

a uniform distribution and the remaining 95% of trials are distributed according to the DDM likelihood function. This method of categorising each trial as a guess or a delayed startup has been found to improve parameter recovery in the presence of outliers (Frank et al., 2015; Verdonckhove & Tuerlinckx, 2007; Wiecki et al., 2013).

3.3.5 Model convergence, selection, and predictive accuracy

To formally test model convergence, we ran four chains of each model and utilised the Gelman-Rubin $\widehat{R}$ statistic (Gelman & Rubin, 1992) to compare within- and between-chain variances. Model convergence is commonly considered satisfactory when $\widehat{R}$ is close to 1 (when samples of the different chains are indistinguishable) but below 1.1 (Franzen et al., 2020; Pedersen & Frank, 2020; Pedersen et al., 2017). Additionally, we visually inspected all group-level sample traces, autocorrelation levels, and the marginal posteriors to further ensure appropriate model convergence. As such, we watched out for signs of poor convergence, including the presence of drifts and major jumps in the posterior sample traces, high autocorrelation levels that surpass the recommended level of 50 for any of the parameters (Wiecki et al., 2013), and non-normal marginal posterior distributions.

We used the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) to compare our models, as other measures, such as the Akaike Information Criterion (Akaike, 1978) or Bayesian Information Criterion (Schwarz, 1978) are not appropriate for hierarchical model comparison. The DIC is widely used for assessing the fit of hierarchical models by selecting the model that achieves the most optimal trade-off between model complexity and goodness-of-fit (Spiegelhalter et al., 2002). Lower DIC values are preferred as they are linked to models with the highest likelihood and the least degrees of freedom.

To evaluate the ability of our models to reproduce key elements of the data, we carried out posterior predictive checks. Specifically, we simulated 500 data sets from the posterior distribution of the parameters of the best fitting neurally informed model, and compared the distance between the summary

statistics of the predicted and observed data. By simulating multiple data sets from the posterior, we derived both summary statistics and a distribution of the simulated data, allowing us to capture uncertainty in our model estimates. The application of the HDDM to reward learning tasks entails that the model should capture trends in both choice probabilities and RT distributions.

To examine RT trends across simulated and observed data, we plotted observed and predicted RT probability distributions alongside each other (Fig.3-5a). To further probe that simulated RTs and choices captured key trends in our data, we generated quintile probability plots (Ratcliff & McKoon, 2008; Frank et al., 2015). Figure 3-5b shows the mean observed and simulated RTs for each quintile (10th, 30th, 50th, 70th, and 90th) separately for the upper and lower boundaries. The shaded areas reveal estimation uncertainty as it represents predicted RTs within one standard deviation from the mean. Before averaging across the population, RT quintiles were calculated for each participant separately.

3.3.6 Hypothesis testing

To evaluate our hypotheses regarding the impact of feedback (processing) on subsequent evidence accumulation and boundary separation, we utilised Bayesian hypothesis testing on the group-level HDDM regression coefficients. Carrying out significance tests directly on the model posteriors is a major advantage of the Bayesian model estimation method (Kruschke, 2010). We computed the proportion of the posterior distribution of the regression coefficients that were above or below 0. In case at least 95% of the proportion of the group-level coefficient posteriors were above (below) 0, we concluded that they had a significant positive (negative) association with the parameter of interest (Frank et al., 2015; Mattes et al., 2021; Weicki et al 2013).

3.4 Results

To decompose the simultaneous effects of our feedback-locked EGG component amplitudes on subsequent decision processes, we utilised hierarchical Bayesian estimation of DDM parameters (HDDM; Wiecki et al., 2013). Specifically, we used neurally informed models, in which we constrained the estimation of model parameters by incorporating single-trial EEG component discriminant amplitudes (y_i values, Eq.2.11) into the parameter estimation process. These regression models allowed us to estimate regression coefficients, which quantified the link between HDDM model parameters and trial-to-trial variations in the electrophysiological measures derived from our linear discriminant analysis (see Section 2.3.6). Importantly, we utilised the feedback-locked EEG component amplitudes in each trial to characterise subsequent decision making on the following trial. Given the presumed role of the early system in implementing network resets in reward learning structures of the late system following negative outcomes, we hypothesised that increased negative feedback encoding by the early and late systems would reduce evidence accumulation in the next trial as evidence towards the alternative choice strategy (i.e., reversal) is accumulated.

3.4.1 Behaviour

During the experiment, participants reached an accuracy level of 66.89% (SD = 7.07%), suggesting a high level of engagement. As Figure 3-4a depicts, mean accuracy was significantly higher ($t(47) = 37.81$, $p < 0.001$) in trials following positive (mean accuracy = 81.87%, SD = 5.51%) compared to negative feedback (mean accuracy = 47.65%, SD = 8.62%). There was no significant difference ($t(47) = -0.94$, $p = 0.27$) between the participant-specific mean reaction times following positive (mean RT = 614 ms , SD = 84 ms) and negative (mean RT = 617 ms, SD = 88 ms) feedback (Fig.3-4b). For more behavioural results, please see section 2.4.1.

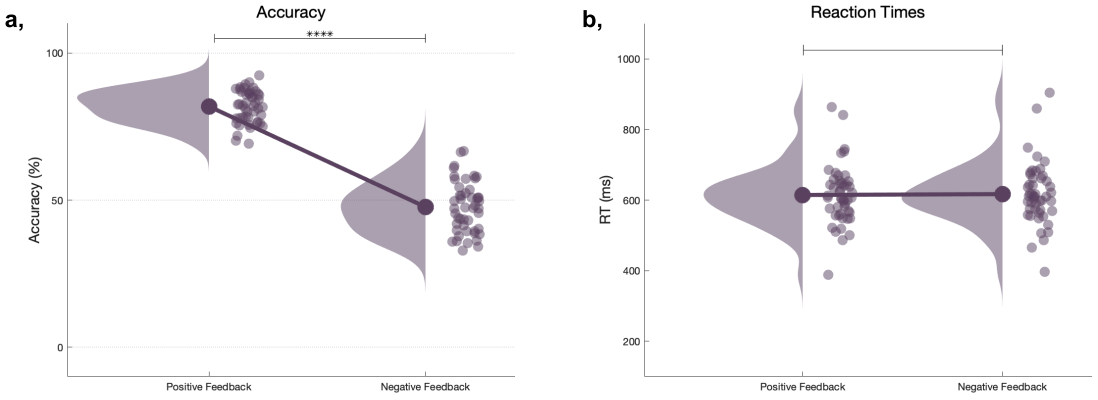

Figure 3-4. Behavioural results. a, Population accuracy, reflecting the participant-specific mean accuracy level by the trial type of the previous trial (positive or negative feedback). Accuracy was defined as selecting the symbol with the higher reward probability. **b,** Participant-specific mean reaction times broken down by the trial type of the previous trial (positive or negative feedback).

3.4.2 Model selection

We estimated three different neurally informed regression models within the HDDM framework; we either varied the drift rate, boundary separation, or both according to our four single-trial EEG component (early positive, early negative, late positive, late negative) discriminant amplitudes. The model with both drift rate and threshold varying dynamically as a function of our EEG component amplitudes achieved a better fit (DIC = 1908) compared to the model with only the drift rate (DIC = 1948) or boundary separation (DIC = 2202) varying as a function of the component amplitudes. To ensure that our models converged, we ran four chains of each model and utilised the Gelman-Rubin $\widehat{R}$ statistic (Gelman & Rubin, 1992) to compare within- and between-chain variances. All three models appeared to converge appropriately, with all parameter $\widehat{R}$ values lying under the recommended value of 1.1 (Franzen et al., 2020; Pedersen & Frank, 2020; Pedersen et al., 2017). Additionally, all parameter $\widehat{R}$ values of the best-fitting model were below 1.01, suggesting excellent model convergence. Group-level parameter estimates for this model are shown in Table 3-1.

Table 3-1. Descriptive statistics for group-level HDDM parameters. Mean, stand deviation, and the 2.5th and 97.5th quantiles for all group-level parameters in the best-fitting neurally informed HDDM. In this model, both the drift rate and boundary separation varied as a function of four EEG component (early positive, early negative, late positive, late negative) discriminant amplitudes from the previous trial.

Parameter	Mean	SD	2.5th & 97.5th Quantiles
a_0	.88	.03	[.83, .93]
$a_{EarlyPos}$	.03	.009	[.008, .04]
$a_{EarlyPos}$	-.007	.007	[-.02, .008]
$a_{LatePos}$	.01	.007	[-.002, .03]
$a_{LateNeg}$	-.005	.008	[-.02, .01]
v_0	.78	.08	[.64, .94]
$v_{EarlyPos}$	.29	.05	[.20, .39]
$v_{EarlyPos}$	.13	.05	[.04, 24]
$v_{LatePos}$	.33	.05	[.24, .43]
$v_{LateNeg}$	.11	.05	[.03, .21]
t	.43	.009	[.41, .45]
z	.49	.006	[.48, .50]
sv	.84	.15	[.51, 1.11]
st	.19	.004	[.18, .20]
sz	.12	.074	[.004, .27]

To evaluate that the best-fitting model correctly captured RT and choice trends in our data, we simulated 500 sets of RT and choice data by drawing samples from the joint posterior distribution of model parameters for each participant. Predicted RT data for the higher- and lower-valued options were then pooled together from individual participants and plotted alongside the experimental data as posterior model predictions (Fig.3-5a; Frank et al., 2015; Franzen et al., 2020). Predicted and observed RTs followed a similar

trend for both response types, implying that the model provided a good fit to the experimental data. To further examine model predictions for RT and choice data, we generated quintile-probability plots (Frank et al., 2015; Ratcliff & McKoon, 2008) depicting RT quintiles separately for each response type (Fig.3-5b). In all cases, observed RT quintiles were within one standard deviation of the predicted RT quintiles, confirming that our best fitting model provided a good fit to observed choice proportions and RT distributions.

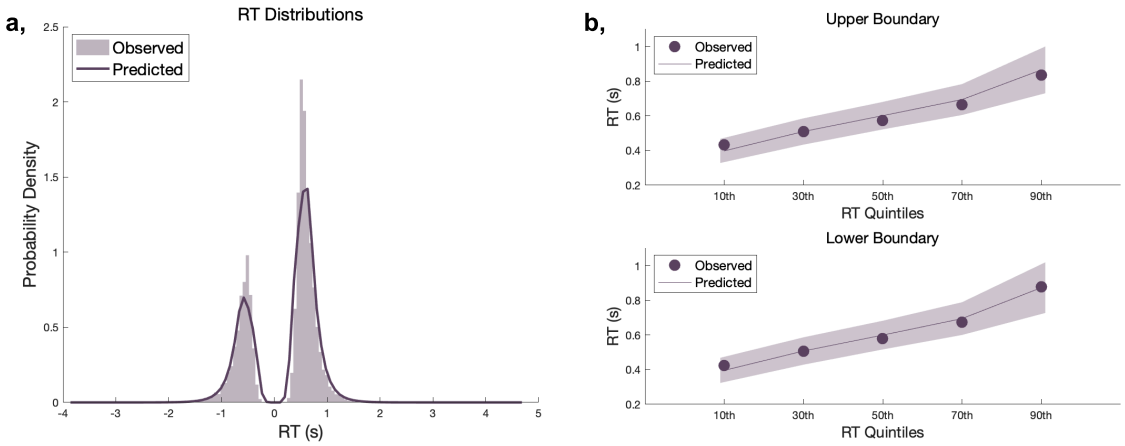

Figure 3-5. Observed and predicted RT distributions. a, Probability density distributions are shown for the observed RTs (histogram) and the RTs predicted by the posterior predictive simulations from the best-fitting HDDM (density curve). RTs on the negative scale portray RTs for the lower-valued options (RTs were sign-flipped), whilst the RTs on the positive scale portray RTs for the higher-valued option. The relative area under the histogram and probability density curve represent choice proportions, demonstrating a higher proportion of choices for the higher- compared to lower-valued option. **b,** Quintile probability plots towards the upper boundary reflecting the higher-valued option (top) and towards the lower boundary reflecting the lower-valued option (bottom). RT quintiles (10th, 30th, 50th, 70th, 90th) are shown on the *x*-axes, with their corresponding observed (dots) and model-predicted (line) values on the *y*-axes. The shaded areas represent one standard deviation above and below the posterior predictive RT distributions from the best-fitting HDDM and capture estimation uncertainty. Predicted data were simulated from the posteriors of the best-fitting neurally informed model.

3.4.3 EEG feedback components shape subsequent decision making

We utilised Bayesian hypothesis testing (Frank et al., 2015; Mattes et al., 2021; Weicki et al 2013) on the regression coefficients linked to the four EEG components of the best-fitting neurally informed model to explore whether they significantly explained variation in the drift rate and boundary separation over the next trial. If at least 95% of the posterior distribution of a neural regression coefficient were above (below) 0, we concluded that there was a significant positive (negative) association between the component and the parameter of interest (Frank et al., 2015; Mattes et al., 2021; Weicki et al 2013).

As Figure 3-6a shows, all regression coefficients estimating the impact of the EEG component amplitudes on the drift rate were positive. The entire posterior distribution is shifted away from zero for the early positive and late positive EEG components, whilst most (99.9%) of the posterior distribution was above zero for the early negative and late negative EEG components. Examining the mean posterior estimates for the group-level regression coefficients suggested that positive feedback processing by the early (mean = 0.27, SD = 0.04) and late (mean = 0.30, SD = 0.03) systems had a larger impact on evidence accumulation compared to the effect of the early negative (mean = 0.14, SD = 0.04) and late negative (mean = 0.10, SD = 0.04) components.

Increasingly positive amplitude values of the early and late positive components indicate that feedback was more positively encoded by the early and late systems, respectively. Conversely, the more negatively feedback is encoded by the early and late systems, the more negative the early and late discriminant amplitudes become (for more details, see Section 2.3.6). Consequently, the positive regression coefficients linking the early and late positive components to the drift rate suggest that the more positively feedback is encoded by the early and the late systems, the more efficient evidence accumulation becomes in the next trial. Conversely, the positive regression coefficients determining the relationship between the early and late negative components and the drift rate imply that the more negatively

feedback is encoded by the early and late systems, the more evidence accumulation is reduced on the next trial. These results support the hypothesis that the early and late systems may reflect network resets in the reward learning structures following negative outcomes, which in turn reduce evidence accumulation on the next trial.

The four EEG components explained a markedly smaller extent of the variation in boundary separation compared to the drift rate (Fig.3-6). There was strong evidence that decision threshold was associated with modulations in the early positive (*M* = .03, *SD* = .009, 99.9% of the posterior > 0) and late positive (*M* = .01, *SD* = .07, 98.5% of the posterior > 0) component amplitudes. In contrast, we did not find convincing evidence that the early (*M* = -.007, *SD* = .07) or late (*M* = 3.6b-.005, *SD* = .08) negative components explained a significant proportion of the variation in this parameter (Fig.3-6b).

The positive regression coefficients linking the early and late positive components to decision threshold imply that the more positively feedback is encoded by the early and the late systems, the more boundary separation widens on the following trial. Despite their lack of a significant effect on decision threshold, the early and late negative component regression coefficients suggest a similar influence; the more negatively feedback was encoded by either system, the more boundary separation increased on the next trial. The EEG component amplitudes explained a larger proportion of the variation in the drift compared to boundary separation, suggesting that the difference in accuracy following positive versus negative feedback trials is mainly driven by fluctuations in the drift rate. Overall, boosted drift rates and boundary separation following increasing positive feedback encoding together with the reduced drift rates following negative trials explain the phenomenon that accuracy is higher following positive compared to negative feedback trials, whilst RTs remain similar.

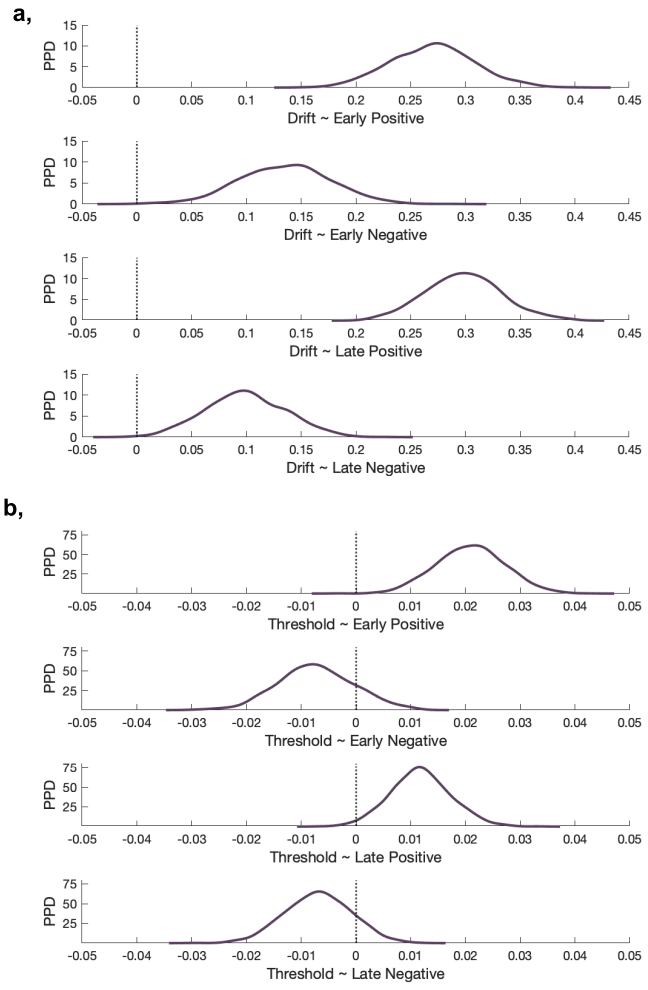

Figure 3-6. Regression coefficient posterior probability distributions. Peak values of each distribution portray the best estimate for the relevant regression coefficient. The width of each distribution represents estimation uncertainty. **a,** Posterior distributions show the estimated effect of the four EEG component discriminant amplitudes on the drift rate (early positive, early negative, late positive, late negative, from top to bottom, respectively). Drift rate was significantly related to all four EEG components demonstrated by the posterior distributions being shifted away from zero. **b,** Posterior distributions show the estimated effect of the four EEG component discriminant amplitudes on boundary separation (early positive, early negative, late positive, late negative, from top to bottom, respectively). Boundary separation was significantly modulated by the early and late positive component amplitudes, but not by the early and late negative component amplitudes. The EEG component amplitudes explained a larger proportion of the variation in the drift rate compared to boundary separation, suggesting that the feedback-related accuracy difference following positive and negative feedback trials is primarily driven by drift rate fluctuations.

Next, we extended the above neurally informed model to estimate not only group- but also individual-level parameters for the influence of our four EEG component discriminant amplitudes on subsequent evidence accumulation and boundary separation. Similar to previous models, convergence was evaluated by calculating $\widehat{R}$ values based on four chains. All parameter $\widehat{R}$ values (maximum: 1.08) were below the recommended value of 1.1 (Franzen et al., 2020; Pedersen & Frank, 2020; Pedersen et al., 2017), suggesting appropriate model convergence. The extended model achieved a better fit (*DIC*: 1851) compared to the model with the group-level-only estimates (*DIC*: 1908), and showed good predictive accuracy, with all quantiles of predicted data lying within the credible interval of the respective quantiles of the observed data. Reassuringly, Bayesian hypothesis tests on the group-level neural regression coefficients gave similar results to those derived from the previous model with group-level only estimates, suggesting reliable effects.

We hypothesised that the more early and late system activity is coupled with the feedback-related pupil response following negative outcomes (see section 2.4.5), the more evidence accumulation would be reduced in response to negative outcomes, as both processes are hypothesised to reflect LC-NA mediated network resets in reward learning structures. To evaluate these hypotheses, we utilised robust across-participant bend correlations (Pernet et al., 2013) to quantify the link between the regression coefficients determining the coupling between the early (late) negative discriminant amplitudes and the feedback-evoked pupil response on one hand, and the HDDM parameter estimates linking the early or late negative component amplitudes to the evidence accumulation on the following trial on the other hand. Although we found this relationship to be in the predicted direction for both the regression coefficients linked to the early *(r* = -.34, *p* = .01) and late *(r* = -.13, *p* = .36) components (Fig. 3-7), only the former correlation was statistically significant.

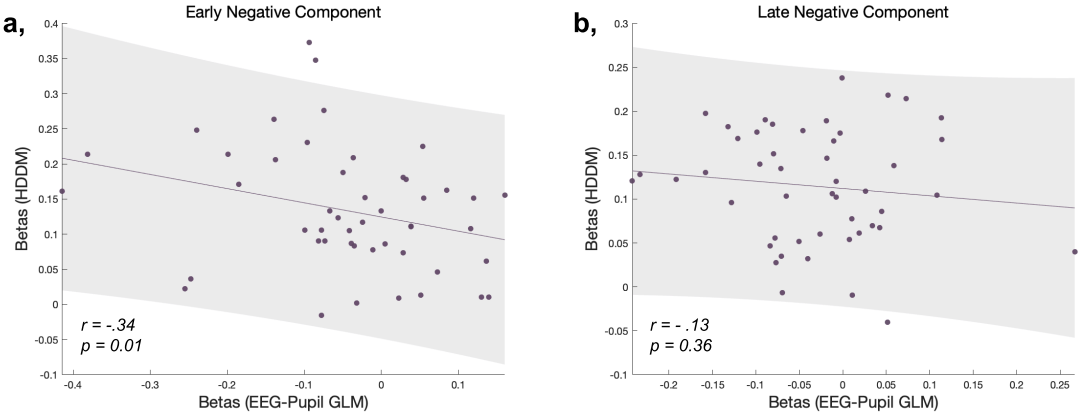

Figure 3-7. EEG-pupil and EEG-drift rate correlations.
Across-participant bend correlations between individual-level regression coefficients linking the early/late negative components to the feedback-locked pupil response and individual-level HDDM parameters linking the respective component to evidence accumulation on the next trial. **a,** We found a significant, negative bend correlation (r = -.34, p = .01) between regression coefficients linking the early negative component to the feedback-locked pupil response on one hand, and evidence accumulation on the following trial on the other hand. **b,** We found a non-significant bend correlation (r = -.13, p = .36) between regression coefficients linking the late negative component to the feedback-locked pupil response on one hand, and evidence accumulation on the following trial on the other hand. The shaded area on each graph shows the least-squares fit line and its 95% confidence band.

3.4.4 Converging behavioural and neural modelling results

Finally, we compared whether our results from the neurally informed models align with those from a model based on behavioural data only with simple binary predictors. To confirm this, we fit an additional model to our data, in which the drift rate and boundary separation varied as a function of a binary predictor specifying feedback valence type (positive or negative). The model successfully converged, with all parameter $\widehat{R}$ values (maximum: 1.03) below 1.1 and with a DIC value of 1665. Furthermore, posterior predictive checks revealed that the model provided a satisfactory fit to our data, with model-predicted and observed RT and choice trends following a similar pattern.

Consistent with the results from the neurally informed models, Bayesian hypothesis testing on the group-level regression coefficients showed that positive feedback increased (*M* = 1.49, *SD* = .09, 100% of the posterior > 0), whilst negative feedback decreased (*M* = -1.03, *SD* = 05, 100% of the posterior < 0) the drift rate in the next trial. Further in alignment with our neurally informed models, boundary separation increased following positive feedback (*M* = .93, *SD* = .02, 100% of the posterior > 0). However, whilst the neurally informed models showed a weak association between negative feedback processing and subsequent boundary separation, the behaviour-only model revealed negative feedback significantly reduced boundary separation during subsequent decisions (*M* = -.05, *SD* = .007, 100% of the posterior < 0). Crucially, results from our neurally informed and behavioural-only models converged in supporting our hypothesis that negative feedback and increased negative feedback processing by the early and late systems reduce evidence accumulation on the next trial, which suggests the reliability of this effect.

3.5 Discussion

In this chapter, we utilised hierarchical drift diffusion modelling to investigate how feedback-related EEG components affect the decision process in the next trial during probabilistic reward learning. Choice and RT data were successfully accounted for by a neurally informed model, in which both the drift rate and boundary separation were parametrically modulated by single-trial, feedback-related EEG component discriminant amplitudes. Specifically, increased negative feedback processing reduced the drift rate on the next trial, without a significant change in boundary separation, in turn resulting in reduced accuracy. At the same time, increased positive feedback processing produced higher drift rates and boundary separation on the next trial, which contributed to improved accuracy. Furthermore, we found a significant across-participant correlation between the degree of coupling between the early negative component and feedback-related pupil response at feedback on one hand, and the early negative component and reduced evidence accumulation on the subsequent trial on the other hand. These results support the hypothesis that LC-induced network resets are signalled to the early system, which reduces the influence of existing value representations in the late system, as reflected by the reduced evidence accumulation following increased negative feedback processing.

In this study, we explored how negative feedback processing, as reflected by our early and late discriminant component amplitudes, affected evidence accumulation on the next trial. In line with the proposed role of LC-NA-mediated cortical resets (Bouret & Sara, 2005, de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009; Urai et al., 2017; Yu & Dayan, 2005), our results in Chapter 2 indicated that increased negative feedback processing by the early and the late systems is associated with an increased feedback-induced pupil response. Furthermore, the strength of the coupling between the early EEG component and the feedback-evoked pupil response was associated with reduced accuracy, as well as increased uncertainty and exploration (increased tendency to choose the lower-valued option). These

results are consistent with the account that increased LC-NA activity signals elevated uncertainty to the early system, which in turn implements a reset in the late system in an attempt to aid adaptation to environmental changes.

Such network resets are presumed to decelerate evidence accumulation in subsequent trials as evidence in favour of alternative options is increasingly accumulated. Consequently, we hypothesised that stronger negative feedback processing by the early and late systems, is likely to reflect network resets, which are expected to progressively reduce evidence accumulation on the next trial. Importantly, results from our best-fitting neurally informed HDDM provided support for this hypothesis; the more strongly negative feedback was encoded by both the early and late systems, as evidenced by more negative discriminant amplitudes, the more evidence accumulation decreased on the following trial.

Additionally, we found that participants who showed a stronger coupling between early system activity and the pupil response following negative outcomes also reduced evidence accumulation more in response to increased negative feedback processing by the early system. Whilst we found a similar association related to the late system, the correlation was significantly weaker. We speculate that as LC-NA mediated network resets, generated by increased uncertainty following negative feedback, are directly signalled to the early system, which in turn modulates the late system to reduce existing value representation. These results support the interpretation that early system activity, rather than a direct LC-NA signal, communicates network resets to the valuation system, which in turn progressively updates existing value representations. This mediatory role of the early system is consistent with results by Fouragnan and colleagues (2015), who reported that following negative outcomes, the early system down-regulated late system blood oxygenation level dependent (BOLD) activity.

Our neurally informed model also revealed that the more strongly positive feedback was encoded by the early and late systems, the more boundary separation and the drift rate increased on the next trial. This result is

consistent with the increasing drift rates reported over the course of reward learning tasks (Fontanesi et al., 2019a; 2019b). Positive feedback in our task may have increased the drift rates on subsequent trials as it signalled participants that they had successfully learnt the prevailing reward contingencies. Indeed, the increased drift rates following stronger positive feedback processing could have resulted from an increasing difference in internal stimulus value representations across choice options, which has been found to increase the rate of evidence accumulation (Fontanesi et al., 2019a; 2019b).

This interpretation also accounts for the opposite effect of negative feedback processing on the drift rate. As negative feedback is likely to cause increased uncertainty regarding the optimal choice strategy (Ferdinand & Opitz, 2014; Ferdinand et al., 2012), it presumably undermines participants' belief that they successfully identified the symbol with the higher reward probability (i.e., negative feedback increases the similarity of internal value representations across choice options). This interpretation is further supported by results from a neurally informed DDM by Stock and colleagues (2016). They found that increased N2 event-related potential amplitudes, associated with boosted levels of within-trial decision conflict (defined in terms of the value difference between choice options), were associated with a reduced drift rate. This is consistent with our results insofar as a negative outcome is likely to increase uncertainty and decision conflict, which may trigger a network reset in order to increase adaptation to new reward contingencies.

Our result that increased positive feedback processing boosts boundary separation on the following trial contradicts results from previous reward learning studies. Within-trial decision conflict, defined in terms of the similarity between choice option values, has been shown to vary linearly with boundary separation; as decision conflict increased, so did boundary separation, presumably in an effort to buy more time for evidence accumulation during heightened decision difficulty (Cavanagh et al., 2011; Fontanesi et al., 2019b; Frank et al., 2015). As positive feedback likely

decreases decision conflict, it is expected to decrease boundary separation in the next trial. However, unlike the current study, previous research examined within-trial, not inter-trial, effects and did not use a reversal learning paradigm. Such differences between the above studies and our research likely contributed to the conflicting outcomes. It is possible that participants in our study increasingly expected reversals following positive feedback, which could have produced widening boundary separation in the next trial. Future research based on reversal learning paradigms should confirm whether positive feedback indeed boosts boundary separation in subsequent trials.

To confirm our result from the neurally informed HDDM, we implemented a behavioural-only model with binary feedback type (positive, negative) predictors for the drift rate and boundary separation. Consistent with the results from the best-fitting neurally informed model, positive feedback increased both boundary separation and the drift rate in the next trial, whilst negative feedback reduced subsequent evidence accumulation. However, the two models produced contradictory results regarding the modulatory effect of negative feedback (processing) on boundary separation in the next trial. Whilst the neurally informed model showed that the strength of negative feedback processing insignificantly increased decision threshold in the following trial, the behavioural-only model indicated negative feedback to reduce boundary separation in the successive trial. Previous results are more consistent with the neurally informed model as they found that the post-error response slowing generated by negative feedback resulted from an increased response threshold (Dutilh et al., 2012; Goldfarb et al., 2012; Hofling et al., 2018).

We speculate that the conflicting results from the neurally informed and behavioural-only models are introduced by the difference in the amount of information the neural and behavioural predictors carry. Neural constraints can account for a significant proportion of noise in cognitive models and ignoring these trial-to-trial variations in neural activity can lead to the overestimation of noise in the decision process (Franzen et al., 2020; Nunez et al., 2017; Turner et al., 2015). The inclusion of neural predictors in our

model presumably resulted in a more accurate characterisation of decision dynamics by reducing noise inherent in a simplistic, binary categorisation of feedback type. If further confirmed, this would further emphasise the need to incorporate neural measures into models of cognitive processing and support the growing field of model-based cognitive neuroscience (Forstmann & Wagenmakers, 2015; Turner et al., 2015).

Both structures of the early and late system (Fouragnan et al., 2015) have been linked to adjustments in the drift rate and decision threshold. Consistent with the key role that dopamine plays in reward learning, the striatum of the late system, together with fronto-parietal parietal structures, were linked to adjustments in both boundary separation (Cavanagh et al., 2011; Forstmann et al., 2008; Mulder et al., 2014) and the drift rate (Forstmann et al., 2010; Frank et al., 2015; Mulder et al., 2014; Verdonck et al., 2021). Among early system structures, the ACC (Chakroun et al., 2023; Rowe et al., 2010; Mulder et al., 2014; Stock et al., 2016; Turner et al., 2015), the anterior insula (Chakroun et al., 2023; Liu & Pleskac, 2011; Philiastides & Sajda, 2007), the dorsolateral prefrontal cortex (Philiastides & Sajda, 2007; Rolls et al., 2010) as well as the (pre-)supplementary motor area (Chakroun et al., 2023; Mulder et al., 2014; Verdonck et al., 2021) have all been related to similar processes. Notably, both Chakroun et al. (2023) and Mulder et al. (2014) found that the ACC adjusts response threshold via the striatum. This one-way relationship between the ACC and the striatum during decision making coincides with how the ACC of the early system takes part in the down-regulation of the striatum of the late system during reversal learning (Fouragnan et al., 2015). These converging results further stress the crucial role of early-late system interactions, presumably mediated by the ACC, during decision making. Nevertheless, further research is necessary to specify how these interactions shape subsequent decision dynamics.

Overall, our results from the current and previous chapters support the proposition that increased internal estimates of environmental uncertainty following negative outcomes are signalled to the early reward learning system by increased LC-NA system activity. Our findings are also consistent with the

hypothesis that the early system interrupts late system activity in order to reduce the impact of learned reward value representations and increase the neural gain related to the processing of new information. As expected, increased negative feedback processing by both the early and late systems in reward learning structures reduced evidence accumulation over the next trial as evidence presumably accumulated towards a new hypothesis (i.e., a reversal in reward contingencies). This reduced evidence accumulation can improve adaptation to a changing environment by promoting exploration over exploitation. Overall, our results support the growing body of research indicating that the LC-NA system in signalling uncertainty and cortical network resets (Bouret & Sara, 2005, de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009; Yu & Dayan, 2005).

Nevertheless, given the limited number of research investigating response adaptation within neurally constrained sequential sampling models (SSMs), it is crucial to establish the extent to which our results can be generalised. Replications within other experimental paradigms, including reward learning tasks with and without reversals, incorporating experimental manipulations (of reward magnitude, level of risk, approach and avoidance learning), and a wide range of neural measures (functional imaging, electrophysiological recordings), could help to further elucidate the impact of the early and late feedback components on subsequent decisions. Establishing the generalisability of these effects is essential given the popularity of drift diffusion modelling in a wide range of research domains, including decision making, perception, memory, attention, ageing, and neuropathology (for an overview, see Ratcliff & McKoon, 2008 and Ratcliff et al., 2016). Interestingly, it has been recently suggested that evidence accumulation could be a transdiagnostic vulnerability factor in psychopathology (Sripada & Weigard, 2021), which will undoubtedly pave the way for further studies investigating neuropathology within the SSM framework.

For the first time, we utilised drift diffusion modelling to link reward learning components, as reflected by single-trial, feedback-locked EEG component discriminant amplitudes, to inter-trial behavioural adaptation. We found that

increased negative feedback processing by both the early and the late systems progressively decreased drift rates. At the same time, stronger positive feedback processing by these systems boosted drift rates and boundary separation in the next trial. Furthermore, increased coupling between the feedback-locked pupil response and early system activity following negative outcomes was positively associated with the early-system-induced drift rate reduction. These results support our hypothesis that the early system, driven by LC-NA activity in response to increasing uncertainty, implements network resets in the late reward learning network following negative outcomes. Such cortical resets would adaptively decelerate evidence accumulation as evidence for a new, competing hypothesis is accumulated and learned reward contingency representations become less plausible.

Chapter 4. Uncertainty-dependent learning bias in the Balloon Analogue Risk Task

4.1 Summary

Do we preferentially learn from positive rather than negative decision outcomes? A number of studies indicated that such bias characterises learning during simple reward learning tasks. However, no research has yet confirmed whether learning bias is also present in the Balloon Analogue Risk Task (BART), which measures risk-taking propensity under uncertainty and closely resembles everyday decision making. Comparing learning from positive and negative outcomes in the BART has recently been made possible through the development of the Scaled Target Learning (STL; Zhou et al., 2021) model, which characterises both risk-taking propensity and sensitivity to wins and losses. We applied this model, and its extension with decay, to a dynamic BART paradigm manipulating the level of uncertainty throughout the experiment. We found that both models successfully fit our data, with their learning parameters adequately capturing the learning process. Crucially, our analyses revealed learning bias in the BART during high levels of uncertainty, under which condition it appeared to be maladaptive. Furthermore, increased sensitivity to wins compared to losses was significantly linked to more risk-seeking behaviour across all conditions, suggesting that learning bias could mediate risky behaviour. Overall, our results contribute to a more accurate characterisation of learning bias in reinforcement learning and suggest that both the presence and maladaptiveness of bias depends on the level of uncertainty.

4.2 Introduction

As our earlier results indicated differential neural and behavioural correlates linked to learning from wins and losses in value-based decisions, we aimed to further explore whether similar discrepancies persist in an experimental paradigm more closely approximating real-life decision making. More complex tasks can reveal crucial aspects of decision making, otherwise missed by the more simple and abstract paradigms commonly utilised in cognitive neuroscience, most of which measure specific component processes such as reversals (Yechiam et al., 2005). Whilst disentangling the various component processes, such as learning or risk-taking propensity, involved in complex tasks can be challenging, the development of novel methodologies can promote breakthroughs.

Optimism bias, whereby people overestimate the probability of positive future events and underestimate the probability of negative future events, has been evidenced in several different areas of life (Kuzmanovic & Rigoux, 2017; Sharot et al., 2012; Shepherd et al., 2013; Weinstein, 1980) and is presumed to originate from an asymmetry in belief updating. Evidence from simple instrumental learning tasks with (den Ouden et al., 2013; Palminteri et al., 2017) and without reversals (Frank et al., 2007; Lefebvre et al., 2017; Niv et al., 2012) demonstrated that this asymmetry also characterises basic reinforcement learning, as indicated by higher positive compared to negative learning rates. Additionally, a recent study by Palminteri (2023) revealed this learning rate bias to be present in 9 different two-armed bandit tasks with binary probabilistic outcomes and feedback, further suggesting that learning bias is a universal reinforcement learning phenomenon. Finally, Harada (2020) also revealed such a bias in learning in the Iowa Gambling Task, a more complex and ecologically valid paradigm compared to the more abstract instrumental learning tasks used in the above studies. However, this bias disappeared with the introduction of dynamic, trial-wise learning rates.

To our knowledge, no studies yet reported whether this positivity bias in learning also exists in the Balloon Analogue Risk Task (BART; Lejuez et al.,

2002), a popular and intuitive paradigm that simulates sequential decision making under uncertainty. The BART measures real-life risk-taking propensity by emulating an uncertain decision context with probabilistic rewards. In each experimental trial, participants are presented with a sequence of virtual balloons and have to repeatedly decide whether to take a risk in order to earn a higher reward ("pump" the balloon) and potentially burst the balloon or collect the already accumulated sum. Each successful pump increases both the amount of reward in the temporary bank and the probability of a balloon burst. A trial ends either by a balloon burst, in which case the temporary bank gets emptied and participants lose their earnings from the trial, or if participants choose to collect and transfer their earnings from the temporary to the permanent bank. The goal of the task is for participants to maximise the reward earned by the end of the experiment. A major advantage of the BART lies in its external validity; the adjusted (BART) score, which indexes risk-taking propensity by marking the average number of pumps on unexploded balloons, has been consistently linked to naturalistic risk-taking behaviours such as smoking or substance use (Aklin et al., 2005; Lejuez et al., 2003; Wallsten et al., 2005).

Until recently, there was no available model to reliably estimate differential learning rates in the BART. The Bayesian Sequential Risk-Taking (BSR) model (Pleskac, 2008), originally referred to as "model 3" by Wallsten and colleagues (2005), has been the most prominent model in the field. Despite the original, four-parameter BSR model incorporates learning, the parameters representing initial belief and the updating of subjective burst probability were found to be unreliable to estimate (Pleskac, 2008; van Ravenzwaaij et al., 2011). This led to a simplification of the model, dubbed BSR-2, with only two parameters; the optimising number of pumps and risk-taking propensity. However, although these two parameters can be recovered reliably, the BSR-2 (and the original BSR model) makes two simplifying assumptions that limit its applicability to a variety of BART paradigms. First, it presumes that burst probabilities are constant across all balloon inflation steps, suggesting that the model may not generalise well to paradigms with gradually increasing

burst probabilities across pumps. Second, participants are assumed to be informed about burst probability, which is incompatible with both real life decision making and most studies where participants are expected to learn through trial and error. As a final drawback, whilst the risk-taking propensity parameter in the BSR model showed good external validity against real-life risk-taking measures, it provided little information about risk-taking propensity beyond that of the adjusted score, which is significantly more straightforward to derive (Wallsten et al., 2005).

To capture differential learning in the BART, Zhou and colleagues (2021) developed the Scaled Target Learning (STL) model, which characterises both participants' risk-taking propensity and the extent to which they learn from past experiences. The model encompasses four parameters, which estimate participants' target number of pumps, behavioural consistency, and the degree to which they adjust their target number of pumps in response to positive and negative feedback. The extension of STL, the Scaled Target Learning model with Decay (STL-D), includes an additional parameter estimating decay, i.e., how fast participants' adjustments of their target number of pumps decay across trials.

Zhou and colleagues found both models to have satisfactory parameter recovery and predictive accuracy, with STL-D outperforming STL in most data sets. Furthermore, STL and STL-D's parameter estimate for the target number of pumps and behavioural consistency showed improved external validity compared to the adjusted BART score and the corresponding parameters in the BSR and BSR-2 models, suggesting an improved ability to capture individual differences in risk-taking propensity. Crucially, both learning parameters in the STL(-D) showed good external validity, implying that they adequately characterise learning from one trial to the next. When comparing the STL(-D) against the BSR and BSR-2 models, the former outperformed both models in terms of parameter recovery, predictive accuracy, and external validity. With improved model performance compared to the prominent BSR models, STL(-D) seems a promising tool for improving our understanding of the

complex psychological processes underlying the BART, including both risk-taking propensity and differential learning.

4.2.1 Current study

In this study, we fit hierarchical versions of the STL and STL-D models to a modified version of the BART paradigm to investigate the degree to which learning bias characterises sequential decision making under different levels of uncertainty. Each participant completed three phases of the BART, characterised by varying levels of burst probabilities (uncertainty). To increase ecological validity, burst probabilities gradually increased across pumps, mirroring real balloons that are more likely to explode with more and more inflation. First, all participants completed a baseline phase, characterised by an intermediate level of balloon burst probability function, followed by a lucky or an unlucky phase. The lucky phase had a more moderate and the unlucky phase had a steeper increase in their respective balloon burst probability functions compared to the baseline phase. To measure potential order effects confounding behaviour, the order of the lucky and unlucky phases was counterbalanced across participants. Our collaborators at the HUN-REN Research Centre for Natural Sciences, Andrea Kóbor and Eszter Tóth-Fáber, determined the experimental design and collected the data (the project was supported by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the FK scheme, grant no. 124412, PI: A.K.). Besides investigating differences in risk-taking propensity arising from phase and order manipulations, the original study design was influenced by hypotheses not pursued in this chapter.

As humans have been shown to flexibly adapt their decision making under different levels of risk and uncertainty in the BART (Kóbor et al., 2023), we did not expect significant differences in participants' risk-taking propensity or learning depending on the order in which the phases were completed. Given the well-established phenomenon whereby learning rates surge with increasing levels of environmental uncertainty (Behrens et al., 2007; Browning

et al., 2015; Palminteri et al., 2017), we expected the learning rates in the unlucky phase to exceed those in the lucky condition. Crucially, evidence for this effect would provide further support that the learning parameters in STL(-D) accurately capture learning in the BART, which was the primary reason behind developing these models. As to the best of our knowledge, this is the first application of these models since their original publication, additional evidence that these models can account for differential learning in a dynamic BART paradigm would provide further support for the credibility of these models.

In line with results indicating a learning bias in instrumental learning tasks (den Ouden et al., 2013; Frank et al., 2007; Harada, 2020; Lefebvre et al., 2017; Niv et al., 2012; Palminteri, 2023; Palminteri et al., 2017), we expected to find higher positive than negative learning rates in the BART. Additionally, we were also interested in how learning rate bias is related to risk-taking propensity and performance. Consistent with the result that higher positive versus negative learning rates resulted in the overestimation of the value of riskier decision alternatives (Niv et al., 2012), we hypothesised that the magnitude of learning rate bias would be positively associated with risk-taking propensity. However, evidence on the relationship between learning rate bias and performance has been mixed. Whilst Harada (2020) found a negative association between these measures, Lefebvre and colleagues (2017) reported no correspondence. At the same time, Palminteri et al. (2017) reported a negative association between learning bias and performance only following the introduction of reversals in reward contingencies but not in the stable period of their two-armed bandit task. By investigating how learning bias is linked to performance under different levels of uncertainty, we aim to further characterise the circumstances in which learning rate bias is (mal)adaptive during reinforcement learning.

In the first application of the STL and STL-D models since their development, we successfully fit these models to a dynamic BART paradigm with varying levels of uncertainty. We showed that the learning rate parameters in these models are sensitive to experimental manipulations of the learning process,

suggesting that they accurately characterise learning in sequential decision-making. Importantly, we found evidence for a learning bias under high uncertainty, suggesting that it is contingent on underlying reward contingencies. Furthermore, learning bias was positively associated with risk-taking propensity, and only showed a significant, negative association with performance under high uncertainty. Overall, our study suggests that learning bias is contingent on elevated levels of uncertainty, under which condition it appears to be maladaptive. Additionally, our results indicate that the balance between learning from positive and negative outcomes may be a chief determinant of risky behaviour.

4.3 Method

4.3.1 Participants

A total of 50 student participants (age: *M* = 21.3 years, *SD* = 2.5 years) took part in the experiment, who were recruited from university courses. Participants were randomly assigned to either of the order conditions (Order 1: 6 males, 19 females, Order 2: 4 males, 21 females, Order 1: *M* = 21.6 years, *SD* = 2.8 years, Order 2: *M* = 21 years, *SD* = 2.2 years). All participants had normal or corrected-to-normal vision, reported no existing psychiatric or neurological conditions, and were not taking psychoactive medication at the time of the experiment. Before enrollment in the study, all participants provided written informed consent. The experiment was approved by the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary, and was conducted in accordance with the Declaration of Helsinki. Participants received course credit in exchange for participation as well as a supermarket voucher. Whilst participants were told that the value of this voucher would vary between 1000-2000 HUF depending on their performance in the task, all participants received a 2000 HUF voucher at the end of the study. All participants were retained in all of our analyses.

4.3.2 Stimuli and task

We utilised a modified version of the Balloon Analogue Risk Task (BART; Fein & Chang, 2008; Kóbor et al., 2015; 2023) to explore whether a learning bias characterises behaviour in a complex decision making task. The BART was first proposed by Lejuez and colleges (2002) and has since been established as a widely used and well-validated measure of risk-taking propensity (Aklin et al., 2005; Lejuez et al., 2003). The modified version of the task allowed for shorter trial lengths, which kept the duration of the experiment within a reasonable time range. Furthermore, the implementation of incremental potential reward values allowed for a more accurate assessment of individual differences in risk-taking propensity (Élteto et al., 2019). The task was

implemented in Presentation (Version 21.1, Neurobehavioral Systems, Inc.,Berkeley, CA) and responses were recorded via a Cedrus RB-540 response device (Cedrus Corporation, San Pedro, CA).

During the task, participants had to repeatedly decide whether to continue ("pump") or stop inflating a virtual balloon that could either increase in size or explode following each inflation step. Successful balloon pumps increased the size of the balloon as well as the reward, but also the likelihood of a balloon burst. Participants used two response keys of the response device to indicate their decision to further pump a balloon or finish the trial and collect their accumulated score in the trial (cash-out). A balloon inflation could result in two outcomes; the balloon would increase in size together with the accumulated score (positive feedback) or the balloon would burst (negative feedback). In case participants decided to stop inflating the balloon, their score earnt in the trial would be transferred to a virtual permanent bank. If a balloon burst ended the trial, the accumulated score in that trial was lost without a decrease in the participant's score in the permanent bank. Participants were instructed to maximise their total score in the task, reflected by the accumulated score in their virtual permanent bank.

During the task, participants could continuously see their accumulated score in the current trial, which was displayed in the middle of the balloon. The accumulated score in the permanent bank, the score collected in the previous trial, and the response key options for inflating the balloon and collecting the accumulated score were also displayed throughout the experiment. The feedback of a balloon burst was represented by a fragmented balloon and the cash-out screen informed participants about the score they earnt in the trial (Fig.4-1). Each feedback screen was presented for 3000 ms and participants' responses were not limited in time.

Participants completed a total of 270 trials in the experiment, divided into three 90-trial phases, each characterised by different balloon burst probabilities. The baseline phase was characterised by an intermediate level

of balloon burst probability, which probability increased in the unlucky phase and decreased in the lucky phase. Each participant started the task with a baseline phase, after which half of the participants first completed the lucky phase followed by the unlucky phase (Order 1) or continued with the unlucky phase and finished the task with the lucky phase (Order 2). In the baseline and lucky phases, the maximum number of balloon inflation steps was 20, whilst this was limited to 10 in the unlucky phase.

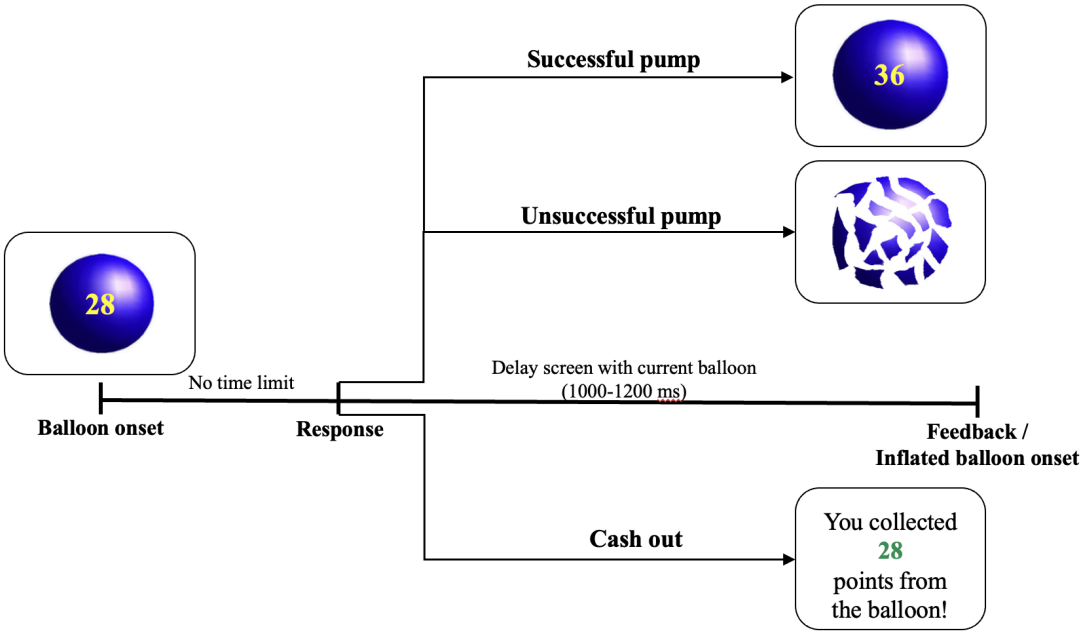

Figure 4-1. An example trial of the modified BART. Participants had to repeatedly decide whether to continue inflating a virtual balloon, which could either increase in size or explode, or stop inflating and cash out their accumulated score from the trial. Each inflation step increased the balloon's size and the reward (Eq.4.2) as well as the likelihood of a balloon burst (Eq.4.1). Participants could see their score accumulated in the current trial in the balloon and had no time limit to respond.

The balloon would not explode following the first two inflations in any of the phases, after which point the probability of balloon burst increased. The burst probabilities for each inflation steps were determined according to

$$P(e_i) = 0, \qquad \text{if } p_i \leq 2, \text{ in all phases}$$

$$P(e_i) = \frac{1}{21 - p_i}, \quad \textit{if } 3 \leq p_i \leq 19 \textit{ in the baseline phase}$$

$$P(e_i) = \frac{1}{31 - p_i}, \quad \textit{if } 3 \leq p_i \leq 19, \textit{ in the lucky phase}$$

$$P(e_i) = \frac{1}{11 - p_i}, \quad \textit{if } 3 \leq p_i \leq 9 \textit{ in the unlucky phase}$$

$$P(e_i) = 1, \quad \textit{if } p_i = 20 \textit{ in the baseline and lucky phases}$$

$$P(e_i) = 1, \quad \textit{if } p_i = 10 \textit{ in the unlucky phase}, \qquad (4.1)$$

where $P(e_i)$ is the probability that the balloon explodes on the *i*th inflation step, and p_i represents the number of inflation steps in the *i*th trial. Thus, the balloon would surely explode on the 20th inflation step in the baseline and lucky phases and on the 10th pump in the unlucky phase (Fig.4-2). As a result of the distinct burst probabilities in the different phases of the experiment, the optimal number of pumps differed in each phase. For baseline balloons, it was most advantageous to inflate the balloon 13 times, whilst the highest expected return was associated with 19 and 6 pumps in the lucky and unlucky phases, respectively (for more details, see the Supplementary Methods in Kóbor et al., 2023). Participants were naïve regarding the burst probabilities in the experiment, including the zero probability of balloon burst in the first two inflation opportunities. Participants were also unaware that burst probabilities would change during the experiment, and the beginning of a new phase was not signalled to participants.

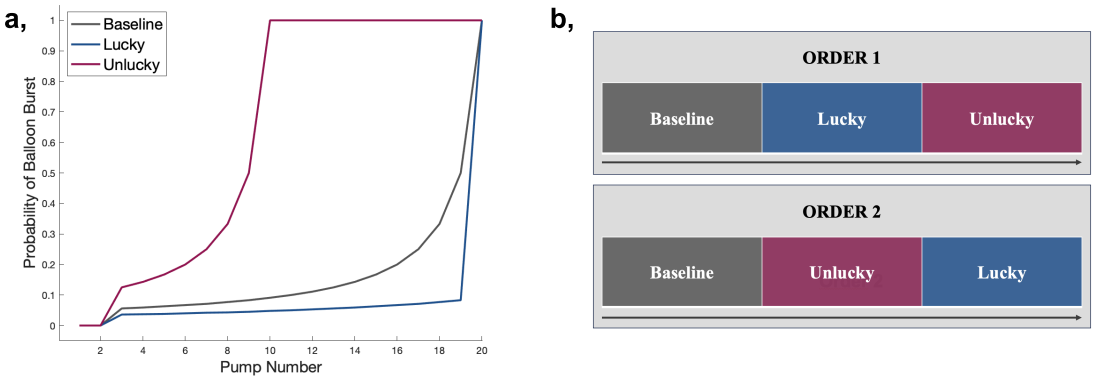

Figure 4-2. Experimental design a, Illustration of the balloon burst probabilities in each phase of the experiment. Balloon bursts were disabled for the first two balloon pumps in each phase, after which burst probability was controlled by a separate truncated power function for each phase (Eq.4.1). The balloon was certain to explode on the 20th pump in the baseline and lucky phases and on the 10th pump in the unlucky phase. **b,** Representation of the two experimental groups. In both groups, participants first completed 90 trials of baseline balloons. After these, participants in Order 1 faced 90 lucky balloons, followed by 90 unlucky balloons. Order 2 counterbalanced the order of the lucky and unlucky balloons; baseline balloons were succeeded by a set of 90 unlucky balloons, then a set of 90 lucky balloons.

For each successful balloon inflation, the score participants earnt in each trial was set to

$$score = \sum_{n}^{p_i} n, \tag{4.2}$$

where p_i is the number of inflation steps in the *i*th trial and n is the within-trial inflation step index (i.e., p_i-1). Thus, participants could gain one point for the first successful balloon inflation, two for the second successful inflation (with an accumulated score of 3 in the trial), three for the third successful inflation (with an accumulated score of 6 in the trial), and so on.

Participants could freely decide whether to inflate the balloon or collect their accumulated score in 80 trials in each phase. In the remaining 10 trials of each phase, participants were instructed to either inflate the balloon to a

predetermined point or until it exploded. These forced-choice trials were included in the experiment to guide participants towards the optimal number of pumps in each phase, and were presented in the same predetermined trial positions to all participants in order to control for across-participant variance.

4.3.3 Procedure

The experiment consisted of two separate sessions. On the first day, participants were asked about any existing medical conditions and medication regimens, their consumption of cognitive performance enhancing drugs, and their motor skills and alertness levels were evaluated. Next, participants took part in an approximately one-hour neuropsychological assessment in order to evaluate potential factors that could alter their performance in the main task. During this process, participants' impulsivity, emotional affect, and cognitive performance (such as working memory and executive functions) were evaluated. These measurements were collected to pursue hypotheses not explored in this work.

On the second day, participants were questioned about factors that could affect their cognitive performance such as sleep quality, their current mood, and how rested they felt. Before beginning the BART, participants had the chance to practise the task with the experimenter. During the main task, participants could take predetermined short breaks every 20-25 trials, and there was an additional larger break halfway through the experiment. The BART was followed by a short verbal interview to assess participants' strategies throughout the task and the degree to which they had awareness of the presence of the different phases in the experiment. These results are not described in this chapter. During the task, continuous electroencephalogram (EEG) data were also recorded; however, as the analysis of electrophysiological data is outside the scope of this study, details on the recording and analysis of the EEG data are omitted. Altogether, the second experimental session took approximately 2-2.5 hours.

4.3.4 Scaled Target Learning model

The Scaled Target Learning (STL) model (Zhou et al., 2021) characterises learning in the BART through adjustments in participants' optimising number of pumps; positive feedback increases, whilst negative feedback decreases the optimising number of pumps. This kind of learning originates in the principle of The Law of Effect (Thorndike, 1898), according to which people are prone to repeating choices that have resulted in desirable outcomes and tend to scale down choices that have led to undesirable outcomes. Therefore, STL predicts that participants would increase their target number of pumps following the collection of a reward, and reduce their target number of pumps following an unwanted balloon burst. Crucially, STL implements separate learning rates for wins (*vwin*) and losses (*vloss*) to account for the distinct degrees of sensitivity to rewards and punishments (Cazé & van der Meer, 2013; Corr, 2004; Frank et al., 2007; Gray, 1975; Lefebvre et al., 2017; Niv et al., 2011; Sharot et al., 2011) and the differential neural mechanisms that implement approach and avoidance learning (Daw et al., 2002; Fouragnan et al., 2015; O'Doherty et al., 2001; Schultz, 2016; Palminteri & Pessiglione, 2017; Seymour et al., 2007).

STL does not make the assumption that an intrinsic risk-taking propensity guides behaviour in the BART. Instead, it assumes that participants begin the task with a target number of pumps (ω_k) in mind and adapt this value after each trial according to

$$\omega_k = \omega_{k-1} \times (vwin \cdot \frac{npump_{k-1}}{nmax}), \quad \textit{if participant collects in trial k-1}$$

$$\omega_k = \omega_{k-1} \times (1 - vloss \cdot (1 - \frac{npump_{k-1}}{nmax})), \quad \textit{if balloon explodes in trial k-1} \tag{4.3}$$

with *vwin*, *vloss* > 0. In STL, ω_k is scaled by the design parameter *nmax*, representing the maximum pump number possible in each trial, so that the value of ω_k falls between 0 and 1. This was implemented in order to account

for two phenomena observed in the BART. First, participants tend to increase their pumps more following a win with a larger compared to a smaller reward value. Second, participants are apt to pump more following a loss with a higher compared to a lower reward that could have been obtained (Schmitz et al., 2016; Zhou et al., 2021). Thus, adjustments after a win ($vwin \cdot \frac{npump_k}{nmax}$) imply a larger increase in ω_k after a larger collection in the previous trial ($\frac{npump_{k-1}}{nmax}$), whilst adjustments after a loss ($vloss \cdot (1 - \frac{npump_k}{nmax})$) imply a smaller reduction in ω_k following a loss with a larger potential reward ($\frac{npump_{k-1}}{nmax}$). Additionally, as the amounts of reward collected or lost due to a balloon burst are scaled by *nmax*, model estimates across various experimental designs of the BART (i.e., different burst probabilities) can be directly compared (Zhou et al., 2021) as long as differences in *nmax* are not excessively broad.

STL further assumes that human behaviour entails a degree of randomness; participants' decisions are probabilistic and are not solely based on their target number of pumps ω_k, but are also determined by participants' behavioural consistency β which influences the degree to which participants behave rationally. Thus, the probability that participants will pump on trial *k* for a given pump opportunity *l (= 1, 2, …)* is given by

$$p_{kl}^{pump} = \frac{1}{1+e^{\beta \cdot (l-\omega_k)}} , \tag{4.4}$$

with $\beta \geq 0$. Thus, *l* increases with more pumps on trial *k*, and the probability that participants will further pump declines until *l* reaches the target number of pumps ω_k, when the probability of pumping equals chance. As participants with higher β rely more on their target number of pumps ω_k, behavioural consistency β can be also understood to reflect participants' prior evaluation of options (Wallsten et al., 2005).

The Scaled Target Learning with Decay (STL-D) model builds on STL by including an additional decay parameter α, which reflects how fast

adjustments in ω_k decay across trials. STL-D characterises decay as a linear function according to

$$\omega_k = \omega_{k-1} \times \left(1 + \frac{vwin \cdot \frac{npump_{k-1}}{nmax}}{1+\alpha \times (k-1)}\right), \textit{ if participant collects in trial k-1}$$

$$\omega_k = \omega_{k-1} \times \left(1 - \frac{vloss \times \left(1 - \frac{npump_{k-1}}{nmax}\right)}{1+\alpha \times (k-1)}\right), \textit{ if balloon explodes in trial k-1}$$

(4.5)

with *vwin*, *vloss*, $\alpha > 0$. Similarly to STL, STL-D assumes that participants adjust their target number of pumps as a function of past outcomes. However, the degree of adjustment decreases across trials *k* and is reflected by the decay parameter α. Finally, both STL and STL-D assume the same choice process, given by Equation (4.4), whereby ω_k controls the probability of pumping p_{kl}^{pump} given each pumping opportunity *l*. Thus, whilst STL has four free parameters reflecting participants' target number of pumps ω_k, behavioural consistency β, and learning rates following wins and losses (*vwin* and *vloss*, respectively), STL-D implements a fifth free parameter α reflecting decay.

4.3.6 Model fitting, comparison, and parameter validity checks

We implemented Hierarchical Bayesian Analysis (Gelman et al., 2013; Zhou et al., 2021) to estimate individual and group-level parameters for the STL and STL-D models, whereby individual-level parameters were drawn from normally-distributed group-level distributions with weakly informative priors. Analysis was performed in the *Rstan* package (version 2.17.2; Stan Development Team, 2019) in R (version 3.3.3; R Core Team, 2019) and utilised Markov Chain Monte Carlo (MCMC; Gamerman & Lopes, 2006) sampling to derive the joint posterior distribution of parameters. For each model, we generated 5000 samples after discarding the first 1000 observations as burn-in.

We fit both the STL and STL-D models separately to each of the three phases within each order, resulting in 6 runs in total (Order 1 baseline, Order 2 baseline, Order 1 lucky, Order 2 lucky, Order 1 unlucky, Order 2 unlucky) for each model. We used the 80 free-choice trials within each phase as we considered the inclusion of the forced-choice trials in the model conceptually problematic as participants had to carry out external instructions in these trials. However, when the models were implemented utilising all 90 trials of each phase, change in the resulting parameter estimates and model fits were negligible. We monitored model convergence by calculating the $\widehat{R}$ statistic (Gelman & Rubin, 1992) to compare within- and between-chain variance across four chains of each model. All model parameters successfully converged with $\widehat{R}$ <1.01 at the group level. Estimated levels of the decay parameter α fell in the recommended range between 0 and 0.1 (Zhou et al., 2021) for the STL-D model in all cases. Values of α above this range have been associated with reduced recovery of the learning parameters *vwin* and *vloss*.

Similarly to Zhou and colleagues (2021), we evaluated the predictive accuracy of our models by comparing the leave-one-out information criterion (*LOOIC*; Vehtari et al., 2017) for both STL and STL-D in each phase of the experiment. This commonly used measure of leave-one-out cross-validation estimates the out-of-sample predictive accuracy of Bayesian models, with lower *LOOIC* values representing improved predictive accuracy. It is considered more accurate compared to other information criteria such as the *Akaike Information Criterion* (*AIC*; Akaike, 1978) or the *Deviance Information Criterion* (*DIC*; Spiegehalter et al., 2002). We computed *LOOIC* via the *loo* R package (Vehtari et al., 2017). This comparison revealed that STL-D fit our data slightly better (Table 4-1), with lower *LOOIC* values for 4 out of the 6 phases. Consequently, we used the parameter estimates from STL-D for further analyses.

Table 4-1. Model comparison. Leave-one-out information criterion (LOOIC) for the STL and STL-D models in the different experimental conditions. The lower LOOIC value in each condition is bolded, illustrating a better model fit.

Order	Phase	STL	STL-D
Order 1	Baseline	**5600**	5621
	Lucky	6776	**6753**
	Unlucky	4561	**4530**
Order 2	Baseline	5447	**5423**
	Lucky	**6936**	6944
	Unlucky	4303	**4240**

To evaluate whether learning bias is present in the BART, we compared group-level estimates of *vwin* and *vloss* from STL-D in the Bayesian framework. Specifically, we calculated the 95% highest density intervals (HDIs) to assess the group-level difference in learning rates, quantified by subtracting the group-level estimates of *vloss* from the group-level estimates of *vwin*. We considered the difference in the group-level learning rates to be credible if the 95% HDIs did not contain zero (Kruschke, 2014). We carried out this analysis separately for each experimental phase (Fig.4-6b-d).

4.3.7 Secondary analyses

To analyse the effect of phase and order manipulations, we utilised *lm()* in R (version 3.3.3; R Core Team, 2019) to perform a two-way analysis of variance (ANOVA) on each of the individual-level STL-D parameter estimates. Although directly comparing group-level parameter estimates in the Bayesian framework (see above) would be preferred, this would be overly complicated considering the complexity of our experimental design. Consequently, we opted to carry out secondary analyses on the individual-level parameters to make the comparisons and their interpretation more straightforward. We baseline-corrected each parameter by subtracting the baseline value from the

corresponding parameter estimates from the lucky and unlucky conditions. We utilised Bonferroni-correction to reduce the likelihood of Type I errors in our analyses. Consequently, we divided the original *p* value of .05 by the number of tests carried out (5, one for each STL-D parameter) and evaluated the analysis of variance effects against this adjusted *p*-value of .01.

Given the flexible adaptation to reward contingencies observed in the BART (Kóbor et al., 2023), we expected no significant difference in model parameters across the two orders. In line with previous results indicating a positive association between learning rates and uncertainty (Behrens et al., 2007; Browning et al., 2015; Palminteri et al., 2017), we predicted that learning rates would be higher in the unlucky compared to the lucky condition, as the former introduced considerably more change in reward contingencies. Evidence in favour of this effect would provide further support for the credibility of the learning parameters in these models.

As we found no learning rate difference across corresponding phases across the two order conditions, we combined parameter estimates phase-wise in all subsequent analyses. To examine whether a learning rate bias is present in our data, we first quantified learning rate bias by normalising the learning rate difference (Niv et al., 2012; Palminteri et al., 2017) for each participant and phase, following

$$bias = \frac{vwin - vloss}{vwin + vloss}. \quad (4.6)$$

To decrease the likelihood of Type I errors when, we utilised Bonferroni-correction, whereby we evaluated each *t*-test against the adjusted *p*-value, obtained by dividing the *p* value by the number of tests carried out, i.e., $p = .05/3 = .017$ (Fig.4-6a).

Finally, we examined how individual-level learning rates *vwin* and *vloss* and learning bias are linked to risk-taking propensity and performance. We utilised the adjusted BART score (mean number of pumps across unexploded balloons) and the total number of points earned by each participant as a proxy for risk-taking propensity and performance, respectively. We used

across-participant Pearson's correlations to evaluate the degree of association across these variables separately for each experimental phase. In line with the findings by Niv and colleagues (2012), we hypothesised that there would be a positive link between learning bias and risk-taking propensity and a negative relationship between each learning rate and risk-taking propensity (Fig.4-7). Due to the mixed results regarding the association between learning rate bias and performance (Lefebvre et al., 2017; Palminteri et al., 2017; Harada, 2020), related analyses were exploratory (Fig.4-8).

4.4 Results

To our knowledge, this is the first application of the STL and STL-D models (Zhou et al., 2021) to a dynamic BART paradigm with varying levels of uncertainty. Our results indicate that these models accurately capture differences in learning resulting from experimental manipulations. Consistent with earlier findings in instrumental learning tasks, we found evidence for a learning bias under high levels of uncertainty (Fig.4-6). In line with previous results (Niv et al., 2012), learning bias was positively linked to risk-taking propensity independently of the level of uncertainty (Fig.4-7c). Moreover, learning bias only demonstrated a significant, negative association with performance under the highest level of uncertainty (Fig.4-8c).

4.4.1 Behaviour

To evaluate behaviour across the different experimental phases, we calculated each participant's adjusted score (i.e., the mean number of balloon inflations on unexploded balloons) and the total points earned. There were no significant differences in either of these measures between corresponding phases across the two orders ($p > .05$ for all two-tailed t-tests). Consequently, we aggregated data across the orders to examine behavioural differences across conditions.

As Figure 4-3a shows, the adjusted score was significantly higher in the lucky compared to both the baseline ($t(98) = -2.79$, $p = .006$) and unlucky ($t(98) = 14.43$, $p < .001$) conditions. As expected, the adjusted score was significantly higher in the baseline compared to the unlucky condition ($t(98) = 15.14$, $p < .001$). Figure 4-3b illustrates that participants achieved a significantly higher number of points in the lucky than in the baseline ($t(98) = -7.25$, $p < .001$) or unlucky ($t(98) = 23.22$, $p < .001$) conditions, with significantly more points earned in the baseline compared to the unlucky phase ($t(98) = 17.51$ $p < .001$).

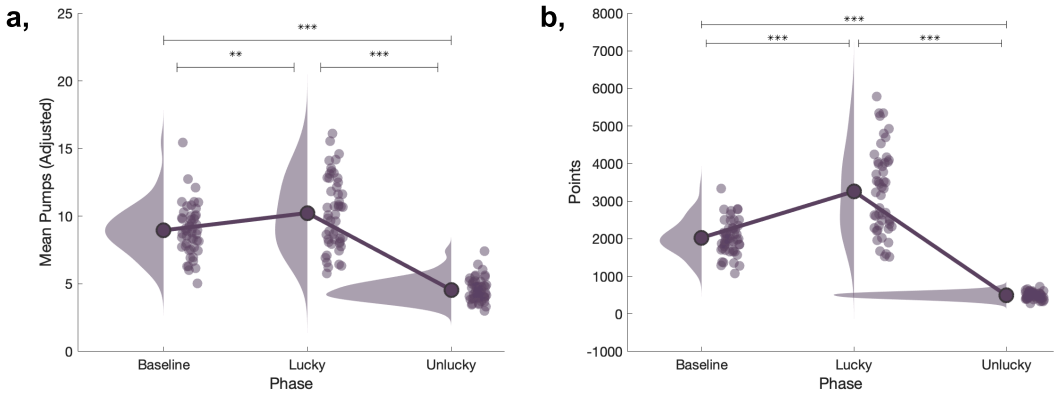

Figure 4-3. Behavioural results. a, Adjusted score in each phase of the experiment. The adjusted score was calculated for each participant separately for trials in which the balloon did not burst. We aggregated data for each phase across the two experimental orders. **b,** Points earnt by each participant across the different experimental conditions. We aggregated data phase-wise from the two experimental orders.

4.4.2 Modelling results

We implemented Hierarchical Bayesian Analysis (Gelman et al., 2013; Zhou et al., 2021) to estimate individual and group-level parameters of the STL and STL-D model. We fit the models separately to each experimental phase of each order group of our BART data, resulting in six independent runs per model. To compare the two models, we utilised the leave-one-out information criterion (LOOIC; Vehtari et al., 2017; Zhou et al., 2021), which estimates the out-of-sample predictive accuracy of Bayesian models, with lower values indicating improved predictive accuracy. The two models produced similar parameter estimates across phases. However, as Table 4.1 shows, STL-D slightly outperformed STL, with lower LOOIC values for 4 out of the 6 phases. Consequently, we utilised the parameter estimates from STL-D for further analyses. Posterior distributions of all group-level STL-D parameters, broken down by phase and order type, are displayed on Figure 4-4.

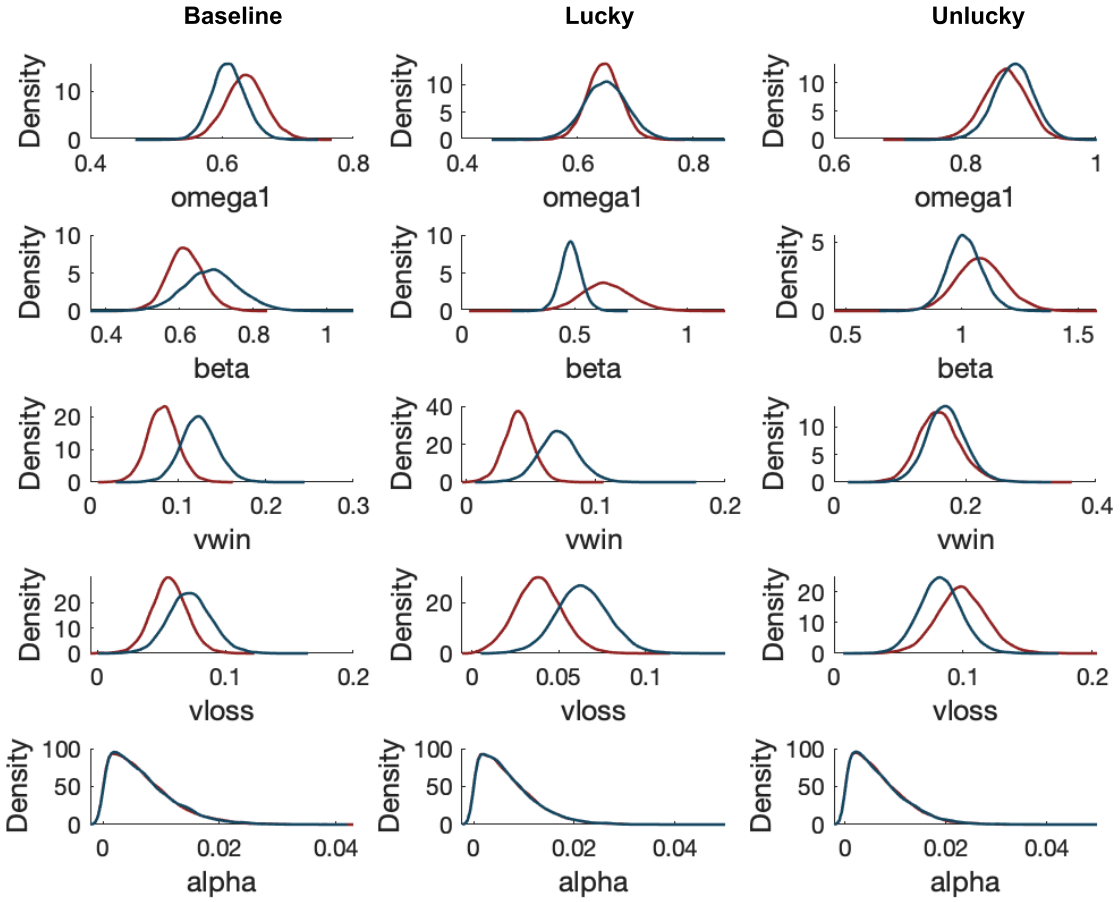

Figure 4-4. STL-D parameter posterior distributions. Group-level parameter posterior distributions are shown separately for each phase column-wise. The red (blue) line indicates parameter distributions for Order 1 (Order 2).

4.4.3 Phase and order effects

To evaluate potential differences across the experimental phases and orders, we performed a 2 (Phase) x 2 (Order) mixed ANOVA on each of the individual-level parameters from the STL-D model. We baseline-corrected parameter estimates from the lucky and unlucky phases by subtracting their corresponding baseline value. The baseline-corrected, individual-level parameter estimates, broken down by phase and order, are depicted on Figure 4-5. For the parameter estimating participants' target level of pumps ω_k, we found a main effect of phase ($F(1,48) = 75.75$, $p < .001$). We did not find a main effect of order ($F(1,48) = 1.81$, $p = .18$) or an interaction effect ($F(1,48) = .05$, $p = .82$). For the parameter β, reflecting participants' behavioural

consistency, we identified a main effect of both phase ($F(1,48) = 55.62$, $p < .001$) and order ($F(1,48) = 8.67$, $p = .004$), without a significant interaction ($F(1,48) = .48$, $p = .49$). Please note that both ω_k and β are scaled by the maximum possible number of pumps *nmax* in each phase, which differed in the lucky and unlucky phases. Whilst comparison within the STL and STL-D models is possible across conditions with different maximum burst points (Zhou et al., 2021), large differences in *nmax* may distort results.

We found a similar pattern for how learning from wins and losses, captured by the parameters *vwin* and *vloss*, respectively, changed throughout the task. Specifically, there was a main effect of phase for both *vwin* ($F(1,48) = 44.21$, $p < .001$) and *vloss* ($F(1,48) = 21.32$, $p < .001$). However, we found no main effect of order for either *vwin* ($F(1,48) = 1.76$, $p = .19$) or *vloss* ($F(1,48) = 1.85$, $p = .18$). The interaction effect was not significant for either *vloss* ($F(1,48) = 5.83$, $p = .018$) or *vwin* ($F(1,48) = .45$, $p = .51$). Similarly to the learning parameters, the ANOVA on the decay parameter α revealed a significant main effect of phase ($F(1,48) = 14.76$, $p < .001$), without a significant effect of order ($F(1,48) = .84$, $p = .84$) or a significant interaction effect ($F(1,48) = .56$, $p = .46$). Larger values for the learning parameters in the unlucky compared to lucky phase provide evidence for the external validity of these parameters. Higher levels of environmental volatility, as introduced by the unlucky phase, have been consistently associated with boosted learning reflected by increased learning rate estimates (Behrens et al., 2007; Browning et al., 2015; Palminteri et al., 2017).

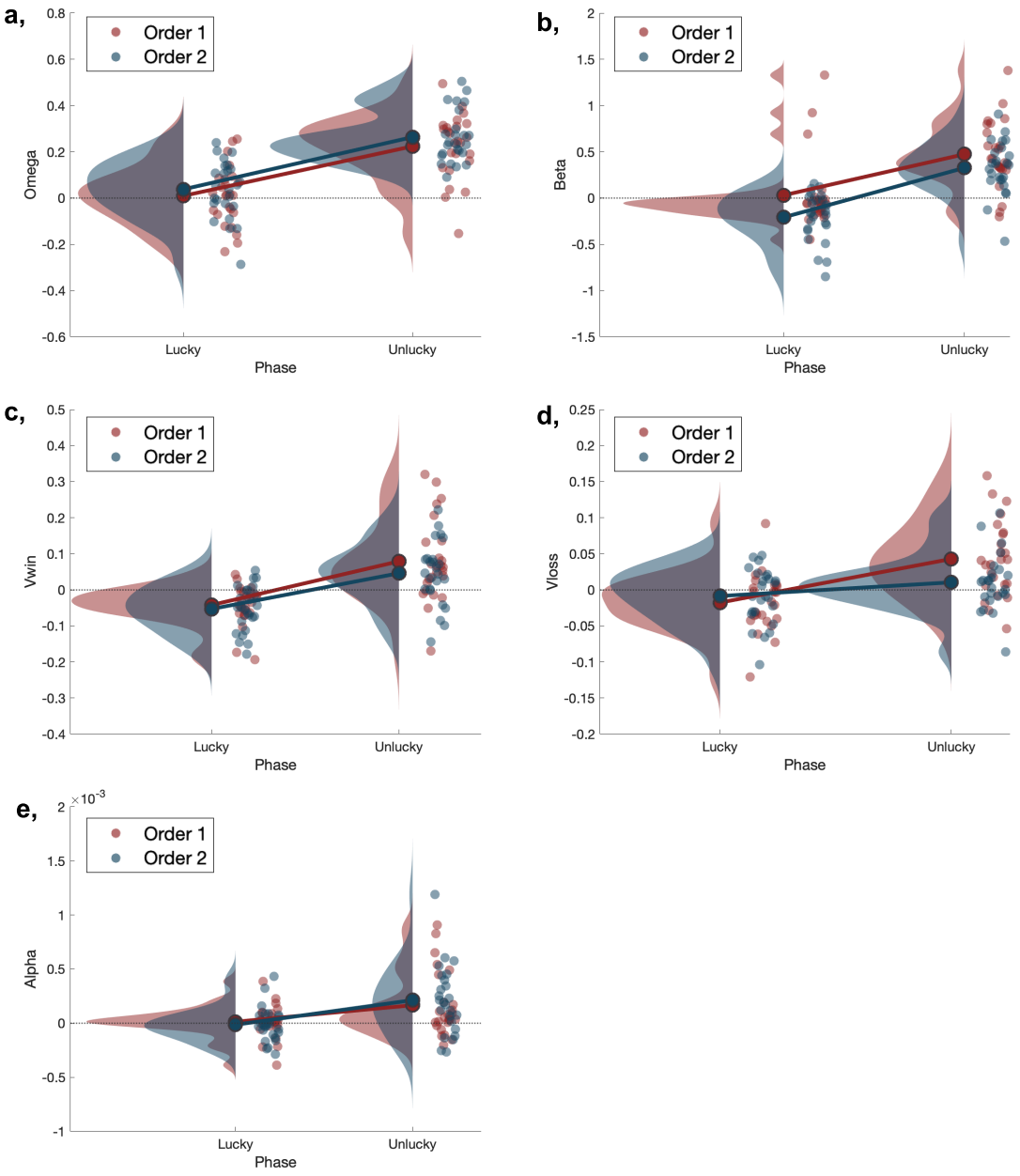

Figure 4-5. Phase and order effects. Individual-level parameter estimates for the target number of pumps ω_k (**a**), behavioural consistency β (**b**), learning from wins *vwin* (**c**) and losses *vloss* (**d**), and decay (**e**) from the Scaled Target Learning Model with Decay (STL-D) are shown separately for each phase (lucky and unlucky) and order (Order 1, Order 2). All parameters were baseline-corrected by subtracting the parameter estimates linked to the baseline phase from the corresponding parameter estimates from the lucky and unlucky conditions.

4.4.4 Learning rate analyses

We utilised both individual- and group-level analyses to determine whether learning bias is present in the BART under varying levels of uncertainty. First, we carried out one-tailed, undirected t-tests assessing whether the participant-wise normalised learning rate difference (Eq.4.6) is significantly different from zero in each phase of the experiment (Fig.4-6a). When evaluated across the adjusted p value of 0.17, we found that the normalised learning bias was significantly higher than zero in the baseline (M = .25, SD = .21, $t(49)$ = 8.30, p < .001) and unlucky (M = .28, SD = .31, $t(49)$ = 6.40 , p < .001) phases, but not in the lucky phase (M = .07, SD = .24, $t(49)$ = 2.07, p = .44).

Next, we employed Bayesian hypothesis testing to evaluate the difference in the group-level estimates of *vwin* and *vloss*. We considered evidence for the difference in learning rates (*vwin - vloss)* to be credible if the 95% HDI for this difference did not include zero (Kruschke, 2014). We found the learning rate difference to be credible in the unlucky (95% HDI: .01 to .15 with a mean of .07), but not in the baseline (95% HDI: -.01 to .09 with a mean of .04) or lucky (95% HDI: -.03 to .04 with a mean of .01) phases (Fig.4-6b-d). These results indicate that the presence of learning bias is contingent on the level of uncertainty in the BART, with bias emerging under increased uncertainty.

To examine the link between learning and risk-taking propensity, we carried out across-participant Pearson's correlations. To measure risk aversion, we utilised the adjusted score, i.e., the mean number of pumps on unexploded balloons. We correlated this participant-specific adjusted score with individual-level estimates of *vwin*, *vloss*, and the normalised learning rate difference (learning rate bias, Eq.4.6) in each phase. As Figure 4-7a shows, the adjusted score significantly and negatively correlated with *vloss* in all phases (baseline: r = -.64 , p < .001, lucky: r = -.47 , p < .001, unlucky: r = -.66 , p < .001). Similarly, we found a significant negative correlation between the adjusted score and *vwin* in the baseline (r = -.36, p = .01) and lucky (r =

-.28, p = .045) phases, whilst the negative correlation in the unlucky phase was not significant (r = -.22, p = .11; Fig.4-7a).

These results are consistent with previous findings that indicated a negative association between risk-taking propensity and learning rates (Niv et al., 2012; Palminteri et al., 2017), suggesting that elevated learning rates are linked to more risk-averse behaviour. Conversely, we found that the magnitude of the learning bias was positively associated with the mean number balloon pumps in all phases (Fig.4-7c; baseline: r = .53 , p < .001, lucky: r = .35 , p = .01, unlucky: r = .47, p < .001). In line with previous results (Niv et al., 2012), this suggests that participants with larger learning bias are characterised by increased risk-taking propensity.

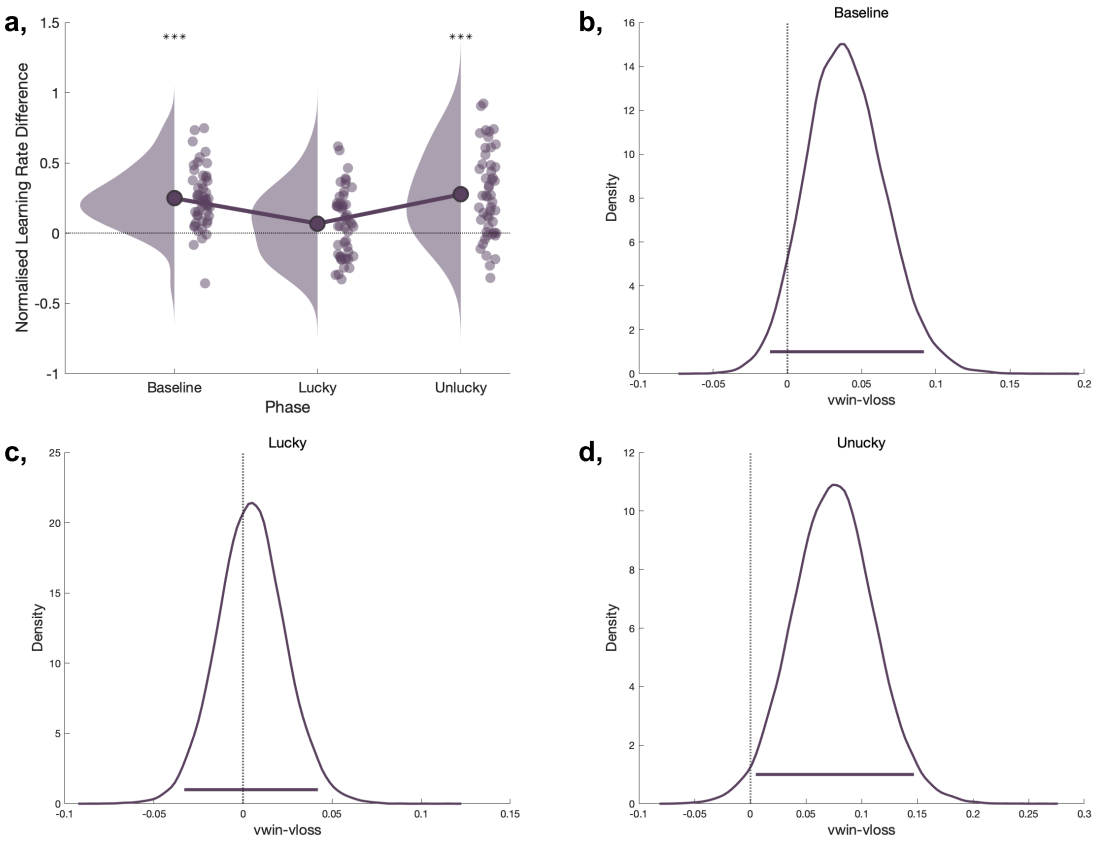

Figure 4-6. Learning rate bias. a, Participant-wise normalised learning rate difference is shown for each phase. One-way t-tests revealed a significant bias in the baseline and unlucky phases, but not in the lucky phase. Significance was evaluated against the Bonferroni-corrected p-value, adjusted by the number of tests carried out. Distribution for the differences in the

group-level estimates of *vwin* and *vloss* are shown for the baseline (**b**), lucky (**c**), and unlucky phases (**d**). The horizontal line within each distribution represents the 95% HDI. The difference in learning rates was only credible in the unlucky (the vertical dotted line at 0 does not cross the bar reflecting the 95% HDI), but not in the baseline or lucky phases.

Our final analyses explored the link between learning and performance. We quantified participants' performance as the number of points earned in each phase. We correlated this measure with the individual-level learning rates *vwin* and *vloss*, as well as the normalised learning rate difference (learning rate bias, Eq.4.6) in each phase (Fig.4-8). We found a negative link between performance and *vwin* in all conditions (baseline: r = -.49 , p < .001, lucky: r = -.34 , p = .002, unlucky: r = -.41 , p = .003). Similarly, performance and *vloss* were negatively correlated in the baseline (r = -.55, p < .001) and lucky phases (r = -.37, p = .01), whilst this association was not significant in the unlucky phase (r = .13, p = .38). This implies that the more weight participants attributed to recent feedback, the more they overestimated environmental fluctuations, which in turn resulted in worse performance. On the other hand, in the unlucky phase, it is adaptive to learn predominantly from negative feedback, which explains the lack of a negative correlation between performance and *vloss*. Performance significantly and negatively correlated with learning bias in the unlucky phase (r = -.51, p < .001), whilst this association was not significant in the baseline (r = .21, p = .14) and lucky (r = .11, p = .43) phases.

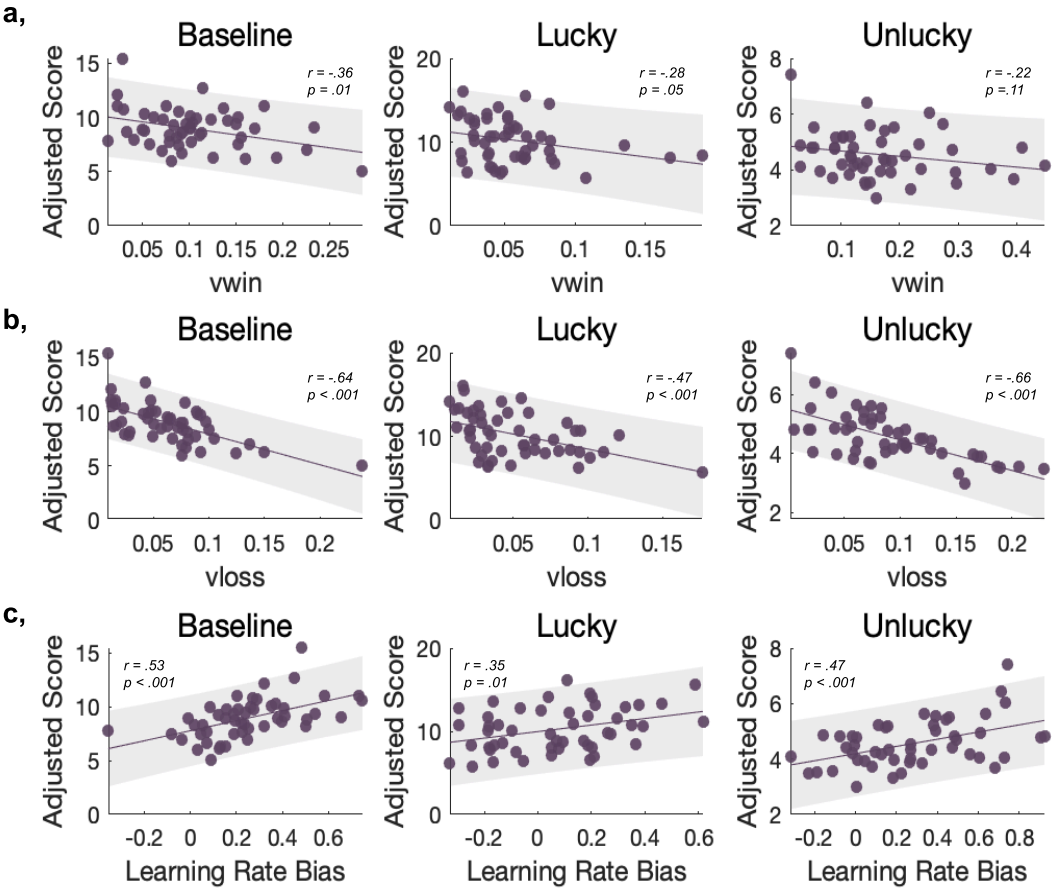

Figure 4-7. Learning rates, learning rate bias, and pumping behaviour. **a,** Across-participant Pearson's correlation between individual-level parameter estimates for learning from wins *vwin* from the STL-D model and the participant-wise adjusted score, shown separately for each phase of the experiment. **b,** Correlation between individual-level parameter estimates for learning from losses *vloss* from the STL-D model and the participant-wise adjusted score, shown separately for each phase of the experiment. **c,** Correlation between participant-wise learning bias and the participant-wise adjusted score, shown separately for each phase of the experiment. Each graph shows the least-squares fit line and its 95% confidence band as the shaded area. The Pearson's correlation coefficient and its corresponding *p*-value are shown on the top left/right of each graph. Data were aggregated across corresponding phases of Orders 1 and 2.

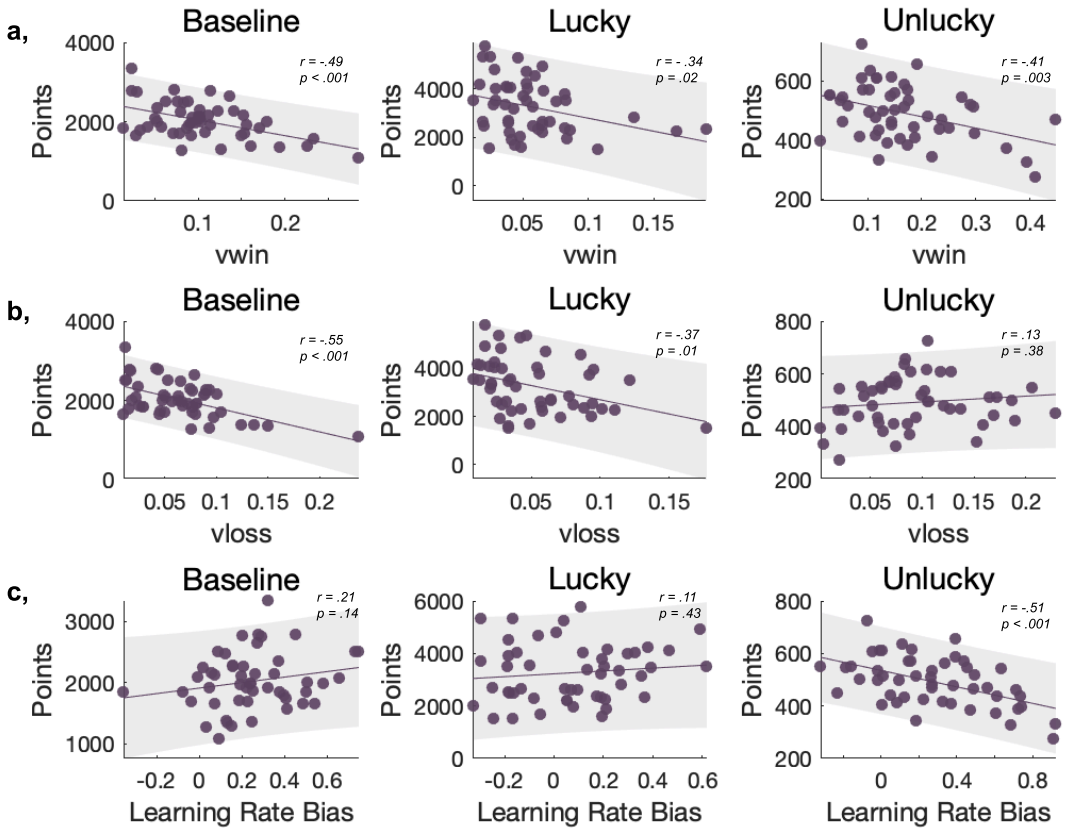

Figure 4-8. Learning rates, learning rate bias, and performance. a, Across-participant Pearson's correlation between individual-level parameter estimates for learning from wins *vwin* from the STL-D model and points earned, shown separately for each phase of the experiment. **b,** Across-participant Pearson's correlation between individual-level parameter estimates for learning from losses *vloss* from the STL-D model and points earned, shown separately for each phase of the experiment. **c,** Across-participant Pearson's correlation between the participant-wise learning bias and the points earned, shown separately for each phase of the experiment. Each graph shows the least-squares fit line and its 95% confidence band as the shaded area. The Pearson's correlation coefficient and its corresponding *p*-value are shown on the top left or right/each graph. Data were aggregated across corresponding phases of Orders 1 and 2.

4.5 Discussion

To our knowledge, this study constitutes the first application of the STL and STL-D models since their original development (Zhou et al., 2021). We successfully fit both models to a dynamic BART paradigm, with experimental conditions manipulating the level of uncertainty. STL-D marginally outperformed STL, and its learning rates appeared sensitive to experimental manipulations, suggesting that they accurately capture the learning process. Crucially, our results suggest that a learning bias was present in the BART under increased levels of environmental uncertainty. Similarly, learning bias was only linked to reduced performance under high uncertainty, suggesting that both the presence of learning bias and its correspondence with performance is contingent on the level of uncertainty in the environment. At the same time, learning bias and risk-taking propensity were positively associated in all conditions, implying that the degree of learning bias may universally shape risk-taking preferences.

We successfully applied the STL and STL-D model to a dynamic BART paradigm with different burst probabilities (i.e., uncertainty) across conditions. These models were originally developed to reliably and meaningfully characterise learning during sequential decisions. Crucially, the model distinguishes learning from positive and negative feedback by estimating differential learning rates to account for the distinct sensitivity to rewards and punishments (Cazé & van der Meer, 2013; Corr, 2004; Frank et al., 2007; Gray, 1975; Lefebvre et al., 2017; Niv et al., 2011; Sharot et al., 2011) and the separate neural processes facilitating approach and avoidance learning (Daw et al., 2002; Fouragnan et al., 2015; O'Doherty et al., 2001; Schultz, 2016; Palminteri & Pessiglione, 2017; Seymour et al., 2007).

Indeed, the differential learning rates estimated via STL-D reflected the change in response to manipulations of environmental uncertainty in accordance with the well-established phenomenon that learning rates increase under heightened levels of uncertainty (Behrens et al., 2007; Browning et al., 2015; Palminteri et al., 2017). Both Zhou and colleagues'

(2021) and our findings suggest that the STL(-D) reliably and meaningfully characterises learning in the BART. This is a major improvement compared to the prominent BSR (Wallsten et al., 2005) and BSR-2 (Pleskac, 2008; van Ravenzwaaij et al., 2011) models as the former cannot reliably recover its learning parameter (Pleskac, 2008; van Ravenzwaaij et al., 2011) and the latter does not take learning into account. Unlike the BSR models, STL(-D) can be applied to paradigms with gradually increasing burst probabilities and does not require participants to be aware of the underlying burst probabilities. As such, they can be applied to a greater variety of experimental paradigms, which flexibility provides a further advantage compared to the BSR models.

Consistent with Zhou and colleagues' (2021) findings, the STL model appears to be improved on by its extension, STL-D, implying that in contrast to assuming constant learning, a non-stationary and linearly decaying learning process better characterises behaviour in the BART. These results parallel other findings in the reinforcement learning literature that indicate improved model fit for Q-learning models including decay (Geana et al., 2022; Radulescu et al., 2016; Yechiam & Busemeyer, 2005). The decay parameter in STL-D indicates how fast adjustments to pumping behaviour decline with experience (Eq.4.5). That is, higher decay reflects a larger weight of past experiences as participants change their pumping behaviour the most in the beginning of the experiment and adjust their behaviour progressively less across trials. Indeed, given that reward contingencies changed with the introduction of a new experimental phase and each phase is modelled separately, it is adaptive to integrate feedback information over a longer period of time to avoid behaviour being overly influenced by the most recent outcomes throughout the experiment (Frank et al., 2007). Accordingly, the higher decay in the unlucky compared to the lucky phase suggests increased emphasis on learning in the beginning of the phase. Considering that higher decay is adaptive with the introduction of a larger change in reward contingencies (as in the unlucky phase), the decay parameter in STL-D constitutes a meaningful addition to modelling the learning process in the BART.

In line with our initial expectation that humans flexibly adapt their decision making in response to the level of environmental uncertainty (Kóbor et al., 2023), we did not find significant differences in the STL-D parameters reflecting participants' target level of pumps, learning rates, or decay across the two orders. Whilst we observed a significant order effect in the behavioural consistency parameter, this was largely driven by three outlier individual estimates in the lucky phase of Order 1 (Fig.4-5b), which questions the generality of this effect. Furthermore, our results suggest increased behavioural consistency and target level of pumps in the unlucky compared to the lucky phase. Although this may seem counterintuitive, both parameters are proportional to the maximum number of pumps, which differed across the two conditions. Despite participants pumped less in the unlucky compared to the lucky phase, the lower number of possible pumps in the unlucky phase generated higher parameter estimates for the target number of pumps and behavioural consistency.

This counter-intuitive conclusion likely stems from modifications to the original BART paradigm (Lejuez et al., 2002; Wallsten et al., 2005), which had a substantially higher maximum burst point as well as constant burst probabilities across balloons. In fact, advantages of STL and STL-D include the applicability to paradigms with gradually increasing burst probabilities and meaningful comparison across conditions with different maximum burst probabilities (Zhou et al., 2021). However, it appears that simultaneous adjustments in these aspects of the task, including large differences in maximum burst points across conditions, may give rise to paradoxical conclusions across conditions or experiments. To reliably compare participants' behavioural consistency and target number of pumps, future studies should implement similar maximum burst points across conditions or utilise cognitive models without a scaling property.

It is also worth bearing in mind that the current version of the BART included forced choice trials to guide participants towards the optimising number of pumps in each phase. This manipulation was systematic; all participants followed the same instructions in the same trials throughout the experiment.

Reassuringly, modelling only free choice or both free and forced-choice trials resulted in similar STL(-D) parameters estimates, suggesting that the inclusion of forced-choice trials did not obscure our results. Nevertheless, it would be reassuring to see converging results from other BART studies.

For the first time, we report evidence for learning bias under high levels of uncertainty in the BART. Both our Bayesian analyses at the group-level and our frequentist analyses at the individual-level implicated learning bias in the unlucky but not in the lucky condition. The two lines of analyses diverged when it comes to the baseline phase; the individual-level analysis suggested that learning bias was present in this phase, whilst the group-level analysis could not credibly confirm this. It is worth noting that group-level comparison is inherently more conservative as it reflects the behaviour of an entire group, including participants with both low and high bias. At the same time, although drawn from a population distribution, participant-specific parameters can accommodate individual differences in behaviour. As such, individual-level analyses are more accurate in capturing participants' underlying behaviour, implying that learning bias was indeed present in the baseline phase.

The presence of learning rate bias under increased uncertainty in the BART is consistent with previous studies reporting learning bias during instrumental learning (den Ouden et al., 2013; Frank et al., 2007; Lefebvre et al., 2017; Niv et al., 2012; Palminteri, 2023; Palminteri et al., 2017). Whilst Harada (2020) found learning bias in the Iowa Gambling Task, a similar paradigm to the BART compared to two-armed bandit tasks, when estimating static learning rates, this bias disappeared with the introduction of time-varying learning rates in their Q-learning model. This suggests that the time-dependent nature of learning rates may give rise to a pseudo learning bias. Although STL-D does not estimate distinct learning rates for each trial, it depicts learning as a non-stationary, decaying process with linearly decreasing learning rates across trials. Consequently, our results from the BART provide some evidence that learning bias is not merely a by-product of static learning rates. Although further research would be helpful to confirm our results in the BART, evidence to date (den Ouden et al., 2013; Frank et al., 2007; Harada,

2020; Lefebvre et al., 2017; Niv et al., 2012; Palminteri, 2023; Palminteri et al., 2017) indicates a universal learning bias in reward learning.

Despite learning bias appearing to be a robust phenomenon across different instrumental learning tasks, its purpose and behavioural implications remain more elusive. Palminteri et al. (2017) and Palminteri's (2023) meta-analysis showed that learning bias in two-armed bandit tasks arises from a confirmation bias, rather than a positivity bias. In other words, participants seemed to preferentially learn from positive outcomes because the outcome confirms their choice strategy, not on the grounds that they are positively valenced. However, these studies utilised cognitive models with static learning rates, which may not be perfectly suitable to account for the stationary reward contingencies of the task. As current BART models do not allow for the differentiation of confirmation or valence-induced bias, meaningful assessment suggests the development of further cognitive models.

Learning from decision outcomes is inherently linked to balancing exploration and exploitation as they both function to enhance adaptation and maximise subjective value (Harada, 2020). As we have shown in Chapter 2, enhanced negative feedback encoding can trigger exploration in an attempt to adjust to perceived changes in the environment. Indeed, learning rates have been consistently associated with the exploration-exploitation trade-off (Cazé & van der Meer, 2013; Harada, 2020; Smith et al., 2021; Sutton & Barto, 2018), with higher learning rates contributing to a faster switch from explore to exploit mode (Smith et al., 2021). Research on the relationship between learning bias and exploration-exploitation is limited and inconsistent. Harada (2020) found bias to be positively associated with exploration in the Iowa Gambling Task, whereas Lefebvre et al. (2017) reported a positive link between bias and exploitation in a two-armed bandit task. As paradigmal and contextual differences may play a crucial role in mediating the relationship between learning bias and exploration-exploitation, more research across a variety of tasks and contexts is essential.

In line with the account that reinforcement learning is inherently related to risk-sensitivity (Niv et al., 2012), our results indicate a consistent positive association between the adjusted score, used to index risk-taking propensity (Aklin et al., 2005; Lejuez et al., 2003; Wallsten et al., 2005), and learning bias across all experimental conditions. At the same time, both learning rates showed the opposite relationship with the adjusted score. These results are consistent with previous findings (Niv et al., 2002; 2012) and can be explained within reinforcement learning theory. Specifically, higher learning rates cause more fluctuation in estimated value and therefore lead to more risk-aversion. If positive feedback increases stimulus value more than negative feedback decreases it, then stimulus value will be higher than the mean nominal outcome, leading to increased risk-seeking. Consequently, biassed learning towards positive outcomes leads to the overestimation of value, resulting in a higher target number of pumps, as borne out by our results. Our results also indicated that compared to *vwin*, *vloss* is more consistently and strongly linked to risk-taking propensity. If learning bias indeed arises from reduced negative feedback processing (Lefebvre et al., 2017), it is plausible that the inverse relationship between *vloss* and risk-aversion on the one hand, and learning bias and risk-aversion on the other hand, represent the same cognitive process. Thus, although our study provides evidence that learning bias may underlie risky behaviour, future research should confirm whether this association exists beyond the increased risk-aversion resulting from reduced negative outcome processing.

Our results also confirm previously reported associations between reward learning and performance. First, we found that increased learning from positive and negative (except in the unlucky phase) feedback was associated with lower performance across the different experimental phases. This is consistent with the notion that a slower integration of outcomes is necessary for the generalisation of probabilistic reward values (Frank et al., 2007). We speculate that increased learning from negative feedback did not remain maladaptive in the unlucky phase as the change in reward contingencies in

this condition was indicated by balloon bursts, requiring adaptation primarily based on negative feedback.

Additionally, we found that learning bias was only significantly related to performance in the unlucky phase, where increased bias was associated with reduced performance. This reflects the contradictory results whereby learning bias was found to be maladaptive by Harada (2020) but not by Lefebvre and colleagues (2017). Our results exhibit a strikingly similar pattern to those by Palminteri and colleagues (2017) in that higher learning bias was only found to be maladaptive with the introduction of increased environmental uncertainty (reversals). Palminteri and colleagues also showed that this decreased ability to flexibly adapt to changing, uncertain environments results from confirmation bias, whereby participants showed increased perseveration despite obtaining new information from negative feedback. Simulations by Caze and van der Meer (2013) similarly suggested that the (mal)adaptiveness of learning bias (either in favour of positive or negative feedback processing) depends on environmental attributes such as the rate of reward. Considering these results, it is likely that both the presence and (mal)adaptiveness of learning bias is contingent on environmental features (i.e., uncertainty), perhaps as a means to flexibly adjust one's exploration-exploitation strategy (Harada, 2020). Nevertheless, even if undue optimism has a net negative impact on learning, it may still be a "self-serving" feature of human cognition that promotes self-esteem and confidence, both of which are related to positive life outcomes (Carver et al., 2010; Weinstein et al., 1980).

Similarly to learning (Niv et al., 2012; Schultz, 1997; 2016; Frank et al., 2004; 2007; 2009), learning bias has been associated with dopaminergic and frontal structures (Lefebvre et al., 2017). Specifically, Lefebvre and colleagues (2107) found that higher bias was linked to increased reward prediction error signalling in the ventral striatum and ventromedial prefrontal cortex (vmPFC). Similarly, van den Bos et al. (2012) reported that the age-related reduction in negative feedback processing was related to increased connectivity between the striatum and medial prefrontal cortex. As in reward learning, frontal-subcortical connectivity (Moutsiana et al., 2015) as well as activity in

the striatum and vmPFC (Kuzmanovic et al., 2016) were found to underlie the optimism bias in belief updating.

Given the high degree of similarity between cortical structures of a late reward learning system (Fouragnan et al., 2015; 2018; Chapter 2 and 3) and those underlying learning/optimism bias, it is possible that the late system or interaction patterns across the early and late systems (Fouragnan et al., 2015) mediates learning bias. The latter possibility is further substantiated by the mounting evidence indicating prominent structures of the early system, such as the anterior cingulate cortex (ACC) or the thalamus (Fouragnan et al., 2015), in regulating reward learning (Behrens et al., 2007; Chakroun et al. 2020; Yu & Dayan, 2005; 2009). Accordingly, Sharot and colleagues (2007) indicated the ACC, which has strong reciprocal connections with the noradrenergic locus coeruleus (Briand et al., 2007; Joshi & Gold, 2020), in meditating the optimism bias in belief updating. Similarly, the thalamus, which was found to moderate the interaction between and early and late systems (Fouragnan et al., 2015) and plays a crucial role in avoidance learning (Kerns et al., 2004; Minamimoto et al., 2005; Seifert et al., 2011), has also been implicated in the processing of optimism bias (Kuzmanic et al., 2016). Moreover, the current study indicates that learning bias is present only under increased uncertainty, further pointing to a possible role of the early system in at least partly generating this bias. Future research could explore how learning bias may be linked to structures of the early and late systems, or their interaction.

For the first time, we provide evidence for a maladaptive learning bias in the BART that is contingent on high levels of environmental uncertainty. These results were derived from the newly developed STL(-D) model, which is the only BART model able to estimate differential learning rates. Although the STL and STL-D appear to reliably characterise behaviour, it remains to be seen whether decaying learning, as in the STL-D, or a dynamic, trial-wise learning process better characterises sequential decision making. In case the latter provides a better fit to data, it needs to be confirmed whether learning bias persists along trial-specific learning rates or disappears as in the Iowa

Gambling Task (Harada, 2020). Additionally, we found a consistent positive association between the degree of learning bias and risk-taking propensity, implying that the relative difference in learning from desirable and undesirable outcomes may generally guide risky behaviour. Future studies should also explore how learning bias may be linked to other behaviours and processes such as exploration-exploitation or confirmation bias. Finally, investigating the neural underpinnings of learning bias could shed further light on how reward learning is implemented in human frontal-subcortical networks. A particularly exciting approach includes the exploration of how the early and late reward learning systems (Chapters 2 and 3; Fouragnan et al., 2015;), or their interaction, may mediate learning bias.

Chapter 5. General discussion

5.1 Overview

Value-based decision making is an integral part of our lives and includes simple behaviours such as crossing the street when the lights turn green but also more complex problems, including which car to buy or career to pursue. In our present era, we enjoy an unprecedented supply of goods and services, which we are relentlessly compelled to choose from. Assigning and comparing the subjective value of each option and learning from decision outcomes is a complex cognitive process determined by numerous factors, including confidence, uncertainty, surprise, mood, or social influences. Understanding the behavioural and neural underpinnings of decision making and reward learning is even more crucial when considering the debilitating effects that related deficits can have on people's lives (Admon & Pizzagalli, 2015; Guitart-Masip et al., 2023; Keiflin & Janak, 2015; Strauss et al., 2014).

The current thesis is concerned with extending our understanding of the behavioural and neural correlates of learning during value-based decisions, a mechanism closely tied to all other steps of the decision making process (Fig.1-1). As discussed in Chapter 1, significant progress has been made to understand the neural underpinnings of reinforcement learning over the past several decades. A great deal of this research has focused on the role of dopamine and its networks in encoding reward predictions errors, which guide learning through indexing the difference between expected and experienced decision outcomes (Bayer & Glimcher, 2005; Fiorillo et al., 2003; Levy & Glimcher, 2012; O'Doherty et al., 2003; Schultz et al., 1997). Despite a growing body of research investigating the neural mechanisms related to key processes involved in reward learning, such as uncertainty (Behrens et al., 2007; Fiorillo et al., 2003; Soltani and Izquierdo, 2019), surprise (Fouragnan et al., 2017; 2018; Schultz et al., 2017), and outcome valence (Cools et al.,

2008; Fouragnan et al., 2015; 2018; Schultz et al., 1997), a complete spatiotemporal characterization of these processes is still lacking.

Evidence is gradually emerging about the role of non-dopaminergic neurotransmitters, such as noradrenaline (Bouret & Sara, 2005; Dayan & Yu, 2006; de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009) and serotonin (Cools et al., 2008; Homberg, 2012), as well as across-system interactions (Aston-Jones & Cohen, 2005; Briand et al., 2007) in guiding reinforcement learning. However, these endeavours are hindered by the lack of non-invasive neural recording methods with adequate spatial and temporal resolution that can measure subcortical nuclei activity containing the majority of neurotransmitter cell bodies. Moreover, due to the complex nature of reward learning, several of its subprocesses may be confounded (Ferdinand & Opitz, 2014), requiring careful experimental designs to identify the unique influence of each subprocess.

The current thesis set out to investigate the differential behavioural and neural mechanisms related to learning from positive and negative decision outcomes whilst considering uncertainty as an important modifying factor. By using the pupil as a proxy for locus-coeruleus-noradrenergic (LC-NA) activity, we studied how the LC-NA system may mediate reward learning. Additionally, we also investigated whether we preferentially learn from positive compared to negative outcomes in a paradigm with varying levels of uncertainty.

5.2 Key findings

In Chapter 2, we investigated how the early and late reward learning systems (Fouragnan et al., 2015) are linked to behaviour and may be related to the LC-NA system. We collected simultaneous electroencephalogram (EEG) and eye-tracking data during a probabilistic reward learning task with reversals, a paradigm commonly used for investigating reward learning mechanisms (Izquierdo et al., 2017; Yaple & Yu, 2019). Using multivariate, linear discriminant analysis on the single-trial, feedback-locked EEG data, we replicated the spatiotemporally distinct early and late reward leaning

components reported by Fouragnan and colleagues. In line with previous results (de Gee et al., 2021; Schneider et al., 2018; Urai et al., 2017), we found an increased phasic pupil response linked to negative compared to positive feedback. By using the pupil response as a proxy for LC-NA activity (Joshi et al., 2016; Joshi & Gold, 2020; Murphy et al., 2014; Reimer et al., 2016), we showed that this difference in the feedback-evoked pupil response was exclusively driven by increased negative feedback processing in both the early and the late systems. Furthermore, a stronger coupling between the early, but not the late, component and the pupil response was associated with reduced performance, as well as increased uncertainty and exploration propensity. We speculate that these behaviourally relevant associations suggest that the increased pupil response evoked by negative feedback likely originates in the early, rather than the late, system.

Overall, our results from Chapter 2 are consistent with the role of the LC-NA system in uncertainty processing and implementing cortical network resets (Bouret & Sara, 2005; Dayan & Yu, 2006; de Gee et al., 2017; Filipowicz et al., 2020; Sara, 2009). We propose that when internal estimates of environmental uncertainty surge in response to negative feedback, the early system, regulated by noradrenergic activity, down-regulates the late system (Fouragnan et al., 2015) by implementing a network reset. Such an interruption in reward learning structures may simultaneously increase the processing of new information and decrease the top-down impact of existing value representations. We speculate that this mechanism aids effective adaptation to changing environments by generating new, more accurate representations of external reward contingencies.

To our knowledge, the study in Chapter 3 is the first to use a neurally informed sequential sampling model to characterise post-feedback response adaptation in value-based decision making. Specifically, we utilised the early and late components established within the data set presented in Chapter 2 to investigate subsequent choice processes as a function of trial-wise EEG component amplitudes. To this end, we employed drift diffusion modelling

(Ratcliff, 1978; Ratcliff & McKoon, 2008), which resembles neural mechanisms in portraying choice as a race-to-barrier process (Kim & Shadlen, 1999; Krajbich et al., 2010; Rangel et al., 2008) and is a widely applied framework for decoupling elements of the decision process (for a review, see Ratcliff et al., 2016).

We hypothesised that if the early system implements LC-NA-induced resets in reward learning structures of the late system in response to negative outcomes, subsequent evidence accumulation would be diminished as evidence is accumulated in favour of a reversal in reward contingencies. In line with our expectations, we found that the stronger the early and late systems encoded negative feedback, as reflected by trail-wise component discriminant amplitudes, the more evidence accumulation declined over the next trial without a change in boundary separation. This drift rate reduction in response to negative outcomes was confirmed by a behavioural-only model with binary predictors (positive and negative feedback), suggesting the reliability of this effect. Furthermore, a stronger coupling between the feedback-locked pupil response and early system activity following negative outcomes was positively associated with the degree of drift rate reduction induced by the early system. This finding further implies that the drift rate reduction evoked by negative feedback is likely to originate in the early system.

Our results from Chapter 3 are also consistent with the premise that the early system is activated by noradrenergic activity, which signals increased uncertainty following negative feedback, and in turn interrupts processing in the late system. Such an interruption in reward learning structures would indeed be expected to result in reduced evidence accumulation as learned value representations become more obsolete and a reversal in reward contingencies is considered more probable.

Our final study, introduced in Chapter 4, investigated differential value learning in the Balloon Analogue Risk Task (BART; Lejuez et al., 2002), a

complex and intuitive decision making paradigm. Specifically, we were interested in whether learning bias (den Ouden et al., 2013; Frank et al., 2007; Lefebvre et al., 2017; Niv et al., 2012; Palminteri et al., 2017), characterised by preferential learning from positive compared to negative decision outcomes, characterises behaviour in a more ecologically valid task under varying levels of uncertainty. To quantify learning bias, we compared positive and negative learning rates derived from the newly developed Scaled Target Learning (STL, Zhou et al., 2021) model, and its extension with decay (STL-D). Both models were successfully fit to our data and appeared to adequately capture the learning process. Our analyses revealed for the first time a learning bias in the BART under increased levels of uncertainty. Higher learning bias was also associated with reduced performance during the task under the highest level of uncertainty, further implying the modulatory role of uncertainty in reward learning. Finally, learning bias was positively related to risk-seeking propensity independently of the level of uncertainty, suggesting that it may play a crucial role in guiding risk preferences.

5.3 Limitations and future directions

A large proportion of this thesis focused on exploring the behavioural and neural correlates of the early and late reward learning systems (Fouragnan et al., 2015). Decision making and reward learning are complex mechanisms with several subprocesses involved, including those related to uncertainty, surprise, exploration-exploitation, and feedback valence. Disentangling these often convoluted mechanisms requires careful experimental consideration. When establishing the association between reward learning components and the feedback-evoked pupil response, we explicitly accounted for the influence of surprise to avoid related confounds.

However, our reversal learning paradigm did not evenly balance feedback valence and uncertainty. As in many other similar tasks, participants often experience negative outcomes more unexpectedly as positive outcomes are increasingly expected with learning (Ferdinand & Opitz, 2014; Ferdinand et al., 2012). By balancing uncertainty across positive and negative outcomes,

future research may confirm whether the early system is indeed primarily activated by negative decision outcomes or instead reflects the degree of uncertainty. Our results in Chapter 2 indicated that the coupling between the early system and the feedback-related pupil response was linked to increased uncertainty and exploration tendency. However, we did not explicitly differentiate between expected and unexpected uncertainty (Soltani & Izquierdo, 2019; Yu & Dayan, 2005) or directed and random exploration (Warren et al., 2017; Zajkowski et al., 2017). By using specialised paradigms and computational models to disentangle exploration and uncertainty subtypes, future studies could determine whether the early system is related to a specific form of uncertainty or exploration (Warren et al., 2017; Zajkowski et al., 2017).

Due to the limited amount of research on the neural and behavioural mechanisms associated with the early and late reward learning systems (Chapter 2 and 3; Fouragnan et al. 2015), replication of the neural and behavioural correlates of these systems is crucial. Generalisability would require the reproduction of existing results with different neural measurement tools in different experimental tasks and populations. It would be particularly reassuring to directly confirm the modulatory role of the LC on early system activity via high resolution fMRI (Dahl et al., 2023; Liu et al., 2017) or single-cell recordings in animals. Moreover, pharmacological studies with NA agonists or antagonists could also provide causal evidence for the role the early system plays in mediating behaviour, such as response slowing or exploration. By implementing neurally informed modelling, such studies could also verify whether the early and late systems affect post-feedback responses in accordance with our results in Chapter 3. Finally, given the central and diverse role of the anterior cingulate cortex (ACC) in reward learning mechanisms (Behrens et al., 2007; Fouragnan et al., 2015; 2017; 2018; Soltani & Izquierdo, 2019), future research could also clarify its precise role in mediating reinforcement learning and implementing network resets.

In our final experimental chapter, we focused on differential learning in the BART under three different conditions associated with varying levels of burst

probabilities (i.e., uncertainty). Our results indicated that learning bias developed under intermediate and high levels of uncertainty and was associated with reduced task performance only under the highest level of uncertainty. The determination of condition-wise burst probabilities is arbitrary as there are countless combinations of probability curves to choose from. As such, our study was not suitable to expose an explicit cut-off level of uncertainty above which learning bias begins to develop or becomes maladaptive. Further experiments, or perhaps a meta-analysis, applying the STL(-D) model to a range of BART paradigms are crucial to determine the specific conditions under which bias emerges and influences performance.

In addition to exploring the environmental and behavioural characteristics of learning bias, it is crucial to determine the neural mechanisms that underlie this phenomenon. Optimism bias, whereby people preferentially consider positive versus negative future outcome probabilities, was found to be related to increased dopaminergic system activity (Kuzamovic et al., 2016; Palminteri et al., 2017) and reduced negative feedback processing (Palminteri et al., 2017). However, no studies to date reported the neural underpinnings of learning bias. It is intriguing to assume that learning bias may arise from reduced early system activity, which is therefore less able to down-regulate value-updating in the late system in response to negative outcomes. Unfortunately, the major visual differences between the positive and negative feedback displays in the current study did not allow for the reliable establishment of the early and late components. It remains for future studies to explore how early and late system activity and interaction may generate learning bias.

A more accurate characterisation of the neural mechanisms engaged in reward learning may also inform clinical research and practice. Several psychopathological conditions have been associated with impairments in reward learning processes, including major depressive disorder (Admon & Pizzagalli, 2015), anxiety disorder (Guitart-Masip et al., 2023), addiction (Keiflin & Janak, 2015), or schizophrenia (Strauss et al., 2014). Although the neural underpinnings of these conditions are not yet completely understood,

both the dopaminergic and noradrenergic systems have been regularly implicated in producing dysfunction (Aston-Jones & Kalivas, 2008; Itoi & Sugimoto, 2010; Yamamoto & Hornykiewicz, 2004).

Although highly speculative, early and late system interaction may provide an additional piece of the puzzle for understanding how behavioural and neural deficits related to reward learning arise in psychopathology. Indeed, there is plenty of overlap in the neural and behavioural correlates of the early and late systems on one hand, and different neuropathological conditions on the other hand, including dopaminergic and noradrenergic modulation of reward learning, uncertainty processing, and exploration-exploitation behaviour. Recently, the rate of evidence accumulation has been suggested as a transdiagnostic vulnerability factor in psychopathology (Sripada & Weigard, 2021). We found both early and late system activity to significantly influence the drift rate over the next trial, further suggesting their potential role in mediating deficits in learning and decision making. Finally, although no existing research has investigated the association between learning bias and neuropathology, this link may explain some of the observed behavioural and neural deficits observed in different conditions. If such relationships can be established, learning bias may provide an easy-to-derive measure for quantifying the severity of dysfunction in reward learning.

5.4 Conclusion

Decision making is an integral part of our daily lives. The significance and complexity of value-based decision making is highlighted by its longstanding investigation across various disciplines, including philosophy, economics, psychology, and neuroscience. Recent breakthroughs in neuroimaging and computational algorithms have provided valuable new insights into the neural processes guiding this process. Nevertheless, a complete characterisation of these mechanisms is still lacking.

The current thesis provides indirect evidence for noradrenergic modulation of reward learning in response to negative decision outcomes. Our results are consistent with the proposed role of the noradrenergic network in processing

uncertainty and promoting exploration through implementing an interruption in the cortical value processing system. Furthermore, we showed that preferential learning from positive compared to negative decision outcomes is not unique to the simple and often abstract decision tasks commonly utilised in cognitive science but persists in the BART, which more accurately mirrors real life decision scenarios.

Although the precise neural mechanisms associated with reward learning and decision making still await to be unravelled, our work provides novel insights on the valence-specific neural and behavioural characteristics associated with feedback processing, whilst highlighting the modulatory role of uncertainty and noradrenergic activity on this process.

Bibliography

Admon, R., & Pizzagalli, D. A. (2015). Dysfunctional Reward Processing in Depression. *Current opinion in psychology, 4*, 114-118. doi:10.1016/j.copsyc.2014.12.011

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics, 30*(1), 9-14. doi:10.1007/BF02480194

Aklin, W. M., Lejuez, C. W., Zvolensky, M. J., Kahler, C. W., & Gwadz, M. (2005). Evaluation of behavioral measures of risk taking propensity with inner city adolescents. *Behav Res Ther, 43*(2), 215-228. doi:10.1016/j.brat.2003.12.007

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu Rev Neurosci, 28*, 403-450. doi:10.1146/annurev.neuro.28.061604.135709

Aston-Jones, G., & Kalivas, P. W. (2008). Brain Norepinephrine Rediscovered in Addiction Research. *Biological psychiatry (1969), 63*(11), 1005-1006. doi:10.1016/j.biopsych.2008.03.016

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412. doi:10.1016/j.jml.2007.12.005

Bari, A., Theobald, D. E., Caprioli, D., Mar, A. C., Aidoo-Micah, A., Dalley, J. W., & Robbins, T. W. (2010). Serotonin Modulates Sensitivity to Reward and Negative Feedback in a Probabilistic Reversal Learning Task in Rats. *Neuropsychopharmacology, 35*(6), 1290-1301. doi:10.1038/npp.2009.233

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proc Natl Acad Sci U S A, 107*(50), 21767-21772. doi:10.1073/pnas.0908104107

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron (Cambridge, Mass.), 47*(1), 129-141. doi:10.1016/j.neuron.2005.05.020

Bayer, H. M., Lau, B., & Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *Journal of neurophysiology, 98*(3), 1428. doi:10.1152/jn.01140.2006

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat Neurosci, 10*(9), 1214-1221. doi:10.1038/nn1954

Benarroch, E. E. (2018). Locus coeruleus. *Cell and tissue research, 373*(1), 221-232. doi:10.1007/s00441-017-2649-1

Bernoulli, D. (1954). Exposition of a New Theory on the Measurement of Risk. *Econometrica, 22*(1), 23-36. doi:10.2307/1909829

Berridge, C. W., & Waterhouse, B. D. (2003). The locus coeruleus–noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. In (Vol. 42, pp. 33-84). Amsterdam: Elsevier B.V.

Bland, A. R., & Schaefer, A. (2012). Different varieties of uncertainty in human decision-making. *Front Neurosci, 6*, 85. doi:10.3389/fnins.2012.00085

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol Rev, 108*(3), 624-652. doi:10.1037/0033-295x.108.3.624

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn Sci, 8*(12), 539-546. doi:10.1016/j.tics.2004.10.003

Bouret, S., & Sara, S. J. (2005). Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. *Trends Neurosci,*

28(11), 574-582. doi:10.1016/j.tins.2005.09.002

Briand, L. A., Gritton, H., Howe, W. M., Young, D. A., & Sarter, M. (2007). Modulators in concert for cognition: modulator interactions in the prefrontal cortex. *Prog Neurobiol, 83*(2), 69-91. doi:10.1016/j.pneurobio.2007.06.007

Browning, M., Behrens, T. E., Jocham, G., O'Reillly, J. X., & Bishop, S. J. (2015). Anxious Individuals Have Difficulty Learning the Causal Statistics of Aversive Environments. *Biological Psychiatry, 77*(9), 47s-48s. Retrieved from <Go to ISI>://WOS:000352207500123

Brunello, N., Blier, P., Judd, L. L., Mendlewicz, J., Nelson, C. J., Souery, D., . . . Racagni, G. (2003). Noradrenaline in mood and anxiety disorders: basic and clinical studies. *Int Clin Psychopharmacol, 18*(4), 191-202. doi:10.1097/00004850-200307000-00001

Bush, G., Vogt, B. A., Holmes, J., Dale, A. M., Greve, D., Jenike, M. A., & Rosen, B. R. (2002). Dorsal Anterior Cingulate Cortex: A Role in Reward-Based Decision Making. *Proceedings of the National Academy of Sciences - PNAS, 99*(1), 523-528. doi:10.1073/pnas.012470999

Carver, C. S., Scheier, M. F., & Segerstrom, S. C. (2010). Optimism. *Clinical Psychology Review, 30*(7), 879-889. doi:10.1016/j.cpr.2010.01.006

Cavanagh, J. F., Figueroa, C. M., Cohen, M. X., & Frank, M. J. (2012). Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. *Cereb Cortex, 22*(11), 2575-2586. doi:10.1093/cercor/bhr332

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nat Neurosci, 14*(11), 1462-1467. doi:10.1038/nn.2925

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes.

J Exp Psychol Gen, 143(4), 1476-1488. doi:10.1037/a0035813

Cazé, R. D., & van der Meer, M. A. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics, 107*(6), 711-719. doi:10.1007/s00422-013-0571-5

Chakraborty, S., Kolling, N., Walton, M. E., & Mitchell, A. S. (2016). Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. *Elife, 5*. doi:10.7554/eLife.13588

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *Elife, 9*. doi:10.7554/eLife.51260

Chakroun, K., Wiehler, A., Wagner, B., Mathar, D., Ganzer, F., van Eimeren, T., . . . Peters, J. (2023). Dopamine regulates decision thresholds in human reinforcement learning in males. *Nat Commun, 14*(1), 5369. doi:10.1038/s41467-023-41130-y

Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A. R., & Khamassi, M. (2019). Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Sci Rep, 9*(1), 6770. doi:10.1038/s41598-019-43245-z

Clark, J. J., Hollon, N. G., & Phillips, P. E. (2012). Pavlovian valuation systems in learning and decision making. *Curr Opin Neurobiol, 22*(6), 1054-1061. doi:10.1016/j.conb.2012.06.004

Cockburn, J., Man, V., Cunningham, W. A., & O'Doherty, J. P. (2022). Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain. *Neuron, 110*(16), 2691-2702 e2698. doi:10.1016/j.neuron.2022.05.025

Cohen Hoffing, R., Karvelis, P., Rupprechter, S., Series, P., & Seitz, A. R. (2018). The Influence of Feedback on Task-Switching Performance: A Drift Diffusion Modeling Account. *Front Integr Neurosci, 12*, 1.

doi:10.3389/fnint.2018.00001

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos Trans R Soc Lond B Biol Sci, 362*(1481), 933-942. doi:10.1098/rstb.2007.2098

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature (London), 482*(7383), 85-88. doi:10.1038/nature10754

Colizoli, O., de Gee, J. W., Urai, A. E., & Donner, T. H. (2018). Task-evoked pupil responses reflect internal belief states. *Sci Rep, 8*(1), 13702. doi:10.1038/s41598-018-31985-3

Cools, R., Altamirano, L., & D'Esposito, M. (2006). Reversal learning in Parkinson's disease depends on medication status and outcome valence. *Neuropsychologia, 44*(10), 1663-1673. doi:10.1016/j.neuropsychologia.2006.03.030

Cools, R., Barker, R. A., Sahakian, B. J., & Robbins, T. W. (2001). Enhanced or Impaired Cognitive Function in Parkinson's Disease as a Function of Dopaminergic Medication and Task Demands. *Cerebral cortex (New York, N.Y. 1991), 11*(12), 1136-1143. doi:10.1093/cercor/11.12.1136

Cools, R., Barker, R. A., Sahakian, B. J., & Robbins, T. W. (2003). l-Dopa medication remediates cognitive inflexibility, but increases impulsivity in patients with Parkinson's disease. *Neuropsychologia, 41*(11), 1431-1441. doi:10.1016/S0028-3932(03)00117-9

Cools, R., Roberts, A. C., & Robbins, T. W. (2008). Serotoninergic regulation of emotional and behavioural control processes. *Trends in cognitive sciences, 12*(1), 31-40. doi:10.1016/j.tics.2007.10.011

Corr, P. J. (2004). Reinforcement sensitivity theory and personality. *Neurosci Biobehav Rev, 28*(3), 317-332. doi:10.1016/j.neubiorev.2004.01.005

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science, 319*(5867), 1264-1267. doi:10.1126/science.1150605

Dahl, M. J., Bachman, S. L., Dutt, S., Duzel, S., Bodammer, N. C., Lindenberger, U., . . . Mather, M. (2023). The integrity of dopaminergic and noradrenergic brain regions is associated with different aspects of late-life memory performance. *Nat Aging, 3*(9), 1128-1143. doi:10.1038/s43587-023-00469-z

Daniel, R., & Pollmann, S. (2010). Comparing the Neural Basis of Monetary Reward and Cognitive Feedback during Information-Integration Category Learning. *The Journal of neuroscience, 30*(1), 47-55. doi:10.1523/JNEUROSCI.2205-09.2010

Daw, N. D. (2014). Advanced Reinforcement Learning. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: decision making and the brain* (2nd ed.). Amsterdam;Boston;: Academic Press.

Daw, Nathaniel D., Gershman, Samuel J., Seymour, B., Dayan, P., & Dolan, Raymond J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron (Cambridge, Mass.), 69*(6), 1204-1215. doi:10.1016/j.neuron.2011.02.027

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks, 15*(4-6), 603-616. doi:Pii S0893-6080(02)00052-7

Doi 10.1016/S0893-6080(02)00052-7

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience, 8*(12), 1704. doi:10.1038/nn1560

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature,*

441(7095), 876-879. doi:10.1038/nature04766

Daw, N. D., & Tobler, P. N. (2014). Value Learning through Reinforcement: The Basics of Dopamine and Reinforcement Learning. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: decision making and the brain* (2nd ed.). Amsterdam;Boston;: Academic Press.

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network (Bristol), 17*(4), 335-350. doi:10.1080/09548980601004024

de, A. M. A. L., Gray, O., Al-Fatly, B., Gilmour, W., Douglas Steele, J., Kuhn, A. A., & Gilbertson, T. (2023). Pallidal neuromodulation of the explore/exploit trade-off in decision-making. *Elife, 12*. doi:10.7554/eLife.79642

de Gee, J. W., Colizoli, O., Kloosterman, N. A., Knapen, T., Nieuwenhuis, S., & Donner, T. H. (2017). Dynamic modulation of decision biases by brainstem arousal systems. *Elife, 6*. doi:10.7554/eLife.23232

de Gee, J. W., Correa, C. M. C., Weaver, M., Donner, T. H., & van Gaal, S. (2021). Pupil Dilation and the Slow Wave ERP Reflect Surprise about Choice Outcome Resulting from Intrinsic Variability in Decision Confidence. *Cereb Cortex, 31*(7), 3565-3578. doi:10.1093/cercor/bhab032

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proc Natl Acad Sci U S A, 111*(5), E618-625. doi:10.1073/pnas.1317557111

De Martino, B., Camerer, C. F., & Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences - PNAS, 107*(8), 3788-3792. doi:10.1073/pnas.0910230107

Delorme, A., Sejnowski, T., & Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage, 34*(4), 1443-1449.

doi:10.1016/j.neuroimage.2006.11.004

den Ouden, Hanneke E. M., Daw, Nathaniel D., Fernandez, G., Elshout, Joris A., Rijpkema, M., Hoogman, M., . . . Cools, R. (2013). Dissociable Effects of Dopamine and Serotonin on Reversal Learning. *Neuron (Cambridge, Mass.), 80*(4), 1090-1100. doi:10.1016/j.neuron.2013.08.030

den Ouden, H. E. M., Kok, P., & de Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in psychology, 3*, 548-548. doi:10.3389/fpsyg.2012.00548

Denison, R. N., Parker, J. A., & Carrasco, M. (2020). Modeling pupil responses to rapid sequential events. *Behav Res Methods, 52*(5), 1991-2007. doi:10.3758/s13428-020-01368-6

Devoto, P., & Flore, G. (2006). On the origin of cortical dopamine: is it a co-transmitter in noradrenergic neurons? *Curr Neuropharmacol, 4*(2), 115-125. doi:10.2174/157015906776359559

Diederen, K. M. J., & Schultz, W. (2015). Scaling prediction errors to reward variability benefits error-driven learning in humans. *Journal of neurophysiology, 114*(3), 1628. doi:10.1152/jn.00483.2015

Dimigen, O., & Ehinger, B. V. (2021). Regression-based analysis of combined EEG and eye-tracking data: Theory and applications. *J Vis, 21*(1), 3. doi:10.1167/jov.21.1.3

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012). Testing theories of post-error slowing. *Atten Percept Psychophys, 74*(2), 454-465. doi:10.3758/s13414-011-0243-2

Ehlers, M. R., & Todd, R. M. (2017). Genesis and Maintenance of Attentional Biases: The Role of the Locus Coeruleus-Noradrenaline System. *Neural Plast, 2017*, 6817349. doi:10.1155/2017/6817349

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on

attention and learning. *Nat Neurosci, 16*(8), 1146-1153. doi:10.1038/nn.3428

Élteto, N., Janacsek, K., Kóbor, A., Takács, A., Tóth-Fáber, E., & Németh, D. (2019). Do adolescents take more risks? Not when facing a novel uncertain situation. *Cognitive Development, 50*, 105-117. doi:10.1016/j.cogdev.2019.03.002

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends Cogn Sci, 15*(2), 85-93. doi:10.1016/j.tics.2010.11.004

Evers, E. A. T., Cools, R., Clark, L., Van Der Veen, F. M., Jolles, J., Sahakian, B. J., & Robbins, T. W. (2005). Serotonergic Modulation of Prefrontal Cortex during Negative Feedback in Probabilistic Reversal Learning. *Neuropsychopharmacology (New York, N.Y.), 30*(6), 1138-1147. doi:10.1038/sj.npp.1300663

Fan, C. Q., Yao, L., Zhang, J. C., Zhen, Z. L., & Wu, X. (2023). Advanced Reinforcement Learning and Its Connections with Brain Neuroscience. *Research, 6*. doi:ARTN 0064

10.34133/research.0064

Fein, G., & Chang, M. (2008). Smaller feedback ERN amplitudes during the BART are associated with a greater family history density of alcohol problems in treatment-naive alcoholics. *Drug and Alcohol Dependence, 92*(1-3), 141-148. doi:10.1016/j.drugalcdep.2007.07.017

Ferdinand, N. K., Mecklinger, A., Kray, J., & Gehring, W. J. (2012). The processing of unexpected positive response outcomes in the mediofrontal cortex. *The Journal of neuroscience, 32*(35), 12087-12092. doi:10.1523/JNEUROSCI.1410-12.2012

Ferdinand, N. K., & Opitz, B. (2014). Different aspects of performance feedback engage different brain areas: disentangling valence and expectancy in feedback processing. *Scientific reports, 4*(1), 5986-5986.

doi:10.1038/srep05986

Fievez, F., Derosiere, G., Verbruggen, F., & Duque, J. (2022). Post-error Slowing Reflects the Joint Impact of Adaptive and Maladaptive Processes During Decision Making. *Frontiers in human neuroscience, 16*, 864590-864590. doi:10.3389/fnhum.2022.864590

Filipowicz, A. L., Glaze, C. M., Kable, J. W., & Gold, J. I. (2020). Pupil diameter encodes the idiosyncratic, cognitive complexity of belief updating. *Elife, 9*. doi:10.7554/eLife.57872

Finke, J. B., Roesmann, K., Stalder, T., & Klucken, T. (2021). Pupil dilation as an index of Pavlovian conditioning. A systematic review and meta-analysis. *Neurosci Biobehav Rev, 130*, 351-368. doi:10.1016/j.neubiorev.2021.09.005

Fiorillo, C. D. (2013). Two dimensions of value: dopamine neurons represent reward but not aversiveness. *Science, 341*(6145), 546-549. doi:10.1126/science.1238699

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science (American Association for the Advancement of Science), 299*(5614), 1898-1902. doi:10.1126/science.1077349

Fiorillo, C. D., Yun, S. R., & Song, M. R. (2013). Diversity and homogeneity in responses of midbrain dopamine neurons. *J Neurosci, 33*(11), 4693-4709. doi:10.1523/JNEUROSCI.3886-12.2013

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychon Bull Rev, 26*(4), 1099-1121. doi:10.3758/s13423-018-1554-2

Fontanesi, L., Palminteri, S., & Lebreton, M. (2019). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using

diffusion decision modeling. *Cogn Affect Behav Neurosci, 19*(3), 490-502. doi:10.3758/s13415-019-00723-1

Forstmann, B. U., Anwander, A., Schafer, A., Neumann, J., Brown, S., Wagenmakers, E. J., . . . Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proc Natl Acad Sci U S A, 107*(36), 15916-15920. doi:10.1073/pnas.1004932107

Forstmann, B. U., Brass, M., Koch, I., & von Cramon, D. Y. (2006). Voluntary selection of task sets revealed by functional magnetic resonance imaging. *J Cogn Neurosci, 18*(3), 388-398. doi:10.1162/089892906775990589

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E. J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proc Natl Acad Sci U S A, 105*(45), 17538-17542. doi:10.1073/pnas.0805903105

Forstmann, B. U., Wagenmakers, E.-J., & SpringerLink. (2015). *An introduction to model-based cognitive neuroscience*. New York, NY: Springer.

Fouragnan, E., Queirazza, F., Retzler, C., Mullinger, K. J., & Philiastides, M. G. (2017). Spatiotemporal neural characterization of prediction error valence and surprise during reward learning in humans. *Sci Rep, 7*(1), 4762. doi:10.1038/s41598-017-04507-w

Fouragnan, E., Retzler, C., Mullinger, K., & Philiastides, M. G. (2015). Two spatiotemporally distinct value systems shape reward-based learning in the human brain. *Nat Commun, 6*, 8107. doi:10.1038/ncomms9107

Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Hum Brain Mapp, 39*(7), 2887-2906. doi:10.1002/hbm.24047

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat Neurosci, 12*(8), 1062-1068. doi:10.1038/nn.2342

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *J Neurosci, 35*(2), 485-494. doi:10.1523/JNEUROSCI.2036-14.2015

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America, 104*(41), 16311-16316. doi:10.1073/pnas.0706111104

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science, 306*(5703), 1940-1943. doi:10.1126/science.1102941

Franzen, L., Delis, I., De Sousa, G., Kayser, C., & Philiastides, M. G. (2020). Auditory information enhances post-sensory visual evidence during rapid multisensory decision-making. *Nat Commun, 11*(1), 5440. doi:10.1038/s41467-020-19306-7

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference* (2nd ed. Vol. 68). London;Boca Raton, FL;: Taylor & Francis.

Geana, A., Barch, D. M., Gold, J. M., Carter, C. S., MacDonald, A. W., Ragland, J. D., . . . Frank, M. J. (2022). Using Computational Modeling to Capture Schizophrenia-Specific Reinforcement Learning Differences and Their Implications on Patient Classification. *Biological Psychiatry-Cognitive Neuroscience and Neuroimaging, 7*(10), 1035-1046. doi:10.1016/j.bpsc.2021.03.017

Gelman, A. (2013). *Bayesian data analysis* (Third ed.). Boca Raton, Florida:

CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical science, 7*(4), 457-472. doi:10.1214/ss/1177011136

Gepshtein, S., Li, X., Snider, J., Plank, M., Lee, D., & Poizner, H. (2014). Dopamine function and the efficiency of human movement. *J Cogn Neurosci, 26*(3), 645-657. doi:10.1162/jocn_a_00503

Gershman, S. J. (2019). Uncertainty and Exploration. *Decision (Washington, D.C.), 6*(3), 277-286. doi:10.1037/dec0000101

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cogn Affect Behav Neurosci, 10*(2), 252-269. doi:10.3758/CABN.10.2.252

Glascher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cereb Cortex, 19*(2), 483-495. doi:10.1093/cercor/bhn098

Glimcher, P. W. (2014). Value-Based Decision Making. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: decision making and the brain* (2nd ed.). Amsterdam;Boston;: Academic Press.

Glimcher, P. W., & Fehr, E. (2014). Introduction: a Brief History of Neuroeconomics. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: decision making and the brain* (2nd ed.). Amsterdam; Boston;: Academic Press.

Goldfarb, S., Wong-Lin, K., Schwemmer, M., Leonard, N. E., & Holmes, P. (2012). Can post-error dynamics explain sequential reaction time patterns? *Front Psychol, 3*, 213. doi:10.3389/fpsyg.2012.00213

Gray, J. A. (1975). *Elements of a two-process theory of learning*. London:

Academic Press.

Guitart-Masip, M., Walsh, A., Dayan, P., & Olsson, A. (2023). Anxiety associated with perceived uncontrollable stress enhances expectations of environmental volatility and impairs reward learning. *Scientific reports, 13*(1), 18451-18451. doi:10.1038/s41598-023-45179-z

Gureckis, T., M. , & Love, B. C. (2015). Computational Reinforcement Learning. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology*. New York: Oxford University Press.

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology, 44*(6), 905-912. doi:10.1111/j.1469-8986.2007.00567.x

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J Neurosci, 26*(32), 8360-8367. doi:10.1523/JNEUROSCI.1010-06.2006

Harada, T. (2020). Learning From Success or Failure? - Positivity Biases Revisited. *Frontiers in psychology, 11*. doi:ARTN 1627

10.3389/fpsyg.2020.01627

Hassall, C. D., Holland, K., & Krigolson, O. E. (2013). What do I do now? An electroencephalographic investigation of the explore/exploit dilemma. *Neuroscience, 228*, 361-370. doi:10.1016/j.neuroscience.2012.10.040

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *The Journal of neuroscience, 31*(11), 4178-4187. doi:10.1523/JNEUROSCI.4652-10.2011

Herrmann, N., Lanctot, K. L., & Khan, L. R. (2004). The role of norepinephrine in the behavioral and psychological symptoms of dementia. *J Neuropsychiatry Clin Neurosci, 16*(3), 261-276. doi:10.1176/jnp.16.3.261

Hoeks, B., & Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior research methods, instruments, & computers, 25*(1), 16-26. doi:10.3758/BF03204445

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature neuroscience, 7*(5), 497-498. doi:10.1038/nn1238

Homberg, J. R. (2012). Serotonin and decision making processes. *Neuroscience and biobehavioral reviews, 36*(1), 218-236. doi:10.1016/j.neubiorev.2011.06.001

Hyman, J. M., Holroyd, C. B., & Seamans, J. K. (2017). A Novel Neural Prediction Error Found in Anterior Cingulate Cortex Ensembles. *Neuron (Cambridge, Mass.), 95*(2), 447-456.e443. doi:10.1016/j.neuron.2017.06.021

I.P., P. (1927). Conditioned reflexes. *Investig. Physiol. Act. Cereb. Cortex*.

Ishii, H., Ohara, S., Tobler, P. N., Tsutsui, K.-I., & Iijima, T. (2012). Inactivating anterior insular cortex reduces risk taking. *The Journal of neuroscience, 32*(45), 16031-16039. doi:10.1523/JNEUROSCI.2278-12.2012

Itoi, K., & Sugimoto, N. (2010). The Brainstem Noradrenergic Systems in Stress, Anxiety and Depression. *Journal of neuroendocrinology, 22*(5), 355-361. doi:10.1111/j.1365-2826.2010.01988.x

Izquierdo, A., Brigman, J. L., Radke, A. K., Rudebeck, P. H., & Holmes, A. (2017). The neural basis of reversal learning: An updated perspective. *Neuroscience, 345*, 12-26. doi:10.1016/j.neuroscience.2016.03.021

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory. *J Cogn Neurosci, 23*(7), 1587-1596. doi:10.1162/jocn.2010.21548

Jo, S., & Jung, M. W. (2016). Differential coding of uncertain reward in rat insular and orbitofrontal cortex. *Scientific reports, 6*(1), 24085-24085. doi:10.1038/srep24085

Joshi, S., & Gold, J. I. (2020). Pupil Size as a Window on Neural Substrates of Cognition. *Trends Cogn Sci, 24*(6), 466-480. doi:10.1016/j.tics.2020.03.005

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron, 89*(1), 221-234. doi:10.1016/j.neuron.2015.11.028

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica, 47*(2), 263-291. doi:10.2307/1914185

Keiflin, R., & Janak, Patricia H. (2015). Dopamine Prediction Errors in Reward Learning and Addiction: From Theory to Neural Circuitry. *Neuron (Cambridge, Mass.), 88*(2), 247-263. doi:10.1016/j.neuron.2015.08.037

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303*(5660), 1023-1026. doi:10.1126/science.1089910

Kim, S., Hwang, J., & Lee, D. (2008). Prefrontal Coding of Temporally Discounted Values during Inter-temporal Choice. *Neuron (Cambridge, Mass.), 59*(1), 161-172. doi:10.1016/j.neuron.2008.05.010

Kluyver, Thomas, Ragan-Kelley, Benjamin, Pérez, Fernando, Granger, Brian, Bussonnier, Matthias, Frederic, Jonathan, Kelley, Kyle, Hamrick, Jessica, Grout, Jason, Corlay, Sylvain, Ivanov, Paul, Avila, Damián, Abdalla, Safia, Willing, Carol and Jupyter development team, (2016) Jupyter Notebooks

– a publishing format for reproducible computational workflows. Loizides, Fernando and Scmidt, Birgit (eds.) In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press. pp. 87-90 . (doi:10.3233/978-1-61499-649-1-87).

Knoch, D., Gianotti, L. R. R., Pascual-Leone, A., Treyer, V., Regard, M., Hohmann, M., & Brugger, P. (2006). Disruption of Right Prefrontal Cortex by Low-Frequency Repetitive Transcranial Magnetic Stimulation Induces Risk-Taking Behavior. *The Journal of neuroscience, 26*(24), 6469-6472. doi:10.1523/JNEUROSCI.0804-06.2006

Kóbor, A., Takács, A., Janacsek, K., Németh, D., Honbolygó, F., & Csépe, V. (2015). Different strategies underlying uncertain decision making: higher executive performance is associated with enhanced feedback-related negativity. *Psychophysiology, 52*(3), 367-377. doi:10.1111/psyp.12331

Kóbor, A., Tóth-Fáber, E., Kardos, Z., Takács, Á., Éltető, N., Janacsek, K., . . . Nemeth, D. (2023). Deterministic and probabilistic regularities underlying risky choices are acquired in a changing decision context. *Scientific reports, 13*(1), 1127-1127. doi:10.1038/s41598-023-27642-z

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nat Neurosci, 13*(10), 1292-1298. doi:10.1038/nn.2635

Krishnamurthy, K., Nassar, M. R., Sarode, S., & Gold, J. I. (2017). Arousal-related adjustments of perceptual biases optimize perception in dynamic environments. *Nat Hum Behav, 1*. doi:10.1038/s41562-017-0107

Krugel, L. K., Biele, G., Mohr, P. N., Li, S. C., & Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proc Natl Acad Sci U S A, 106*(42), 17951-17956. doi:10.1073/pnas.0905191106

Kruschke, J. K. (2011). *Doing Bayesian data analysis: a tutorial with R and BUGS*. Burlington, Mass: Academic Press.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan*. Amsterdam: Academic Press.

Kuhnen, C. M., & Knutson, B. (2005). The Neural Basis of Financial Risk Taking. *Neuron (Cambridge, Mass.), 47*(5), 763-770. doi:10.1016/j.neuron.2005.08.008

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. *Neuroimage, 133*, 151-162. doi:10.1016/j.neuroimage.2016.02.014

Kuzmanovic, B., & Rigoux, L. (2017). Valence-Dependent Belief Updating: Computational Validation. *Frontiers in psychology, 8*, 1087-1087. doi:10.3389/fpsyg.2017.01087

Laeng, B., Sirois, S., & Gredeback, G. (2012). Pupillometry: A Window to the Preconscious? *Perspect Psychol Sci, 7*(1), 18-27. doi:10.1177/1745691611427305

Lanciego, J. L., Luquin, N., & Obeso, J. A. (2012). Functional neuroanatomy of the basal ganglia. *Cold Spring Harbor perspectives in medicine, 2*(12), a009621-a009621. doi:10.1101/cshperspect.a009621

Lavin, C., San Martin, R., & Rosales Jubal, E. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task. *Front Behav Neurosci, 7*, 218. doi:10.3389/fnbeh.2013.00218

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature human behaviour, 1*(4). doi:ARTN 0067

10.1038/s41562-017-0067

Lejuez, C. W., Aklin, W. M., Jones, H. A., Richards, J. B., Strong, D. R., Kahler, C. W., & Read, J. P. (2003). The Balloon Analogue Risk Task (BART) differentiates smokers and nonsmokers. *Exp Clin Psychopharmacol, 11*(1), 26-33. doi:10.1037//1064-1297.11.1.26

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., . . . Brown, R. A. (2002). Evaluation of a Behavioral Measure of Risk Taking: The Balloon Analogue Risk Task (BART). *Journal of experimental psychology. Applied, 8*(2), 75-84. doi:10.1037/1076-898X.8.2.75

Leonard, B. E. (1997). The role of noradrenaline in depression: a review. *J Psychopharmacol, 11*(4 Suppl), S39-47. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9438232

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current opinion in neurobiology, 22*(6), 1027-1038. doi:10.1016/j.conb.2012.06.001

Lin, H., Saunders, B., Hutcherson, C. A., & Inzlicht, M. (2018). Midfrontal theta and pupil dilation parametrically track subjective conflict (but also surprise) during intertemporal choice. *NeuroImage (Orlando, Fla.), 172*, 838-852. doi:10.1016/j.neuroimage.2017.10.055

Liu, K. Y., Marijatta, F., Hämmerer, D., Acosta-Cabronero, J., Düzel, E., & Howard, R. J. (2017). Magnetic resonance imaging of the human locus coeruleus: A systematic review. *Neuroscience and biobehavioral reviews, 83*, 325-355. doi:10.1016/j.neubiorev.2017.10.023

Liu, T., & Pleskac, T. J. (2011). Neural correlates of evidence accumulation in a perceptual decision task. *J Neurophysiol, 106*(5), 2383-2398. doi:10.1152/jn.00413.2011

Loued-Khenissi, L., Pfeuffer, A., Einhäuser, W., & Preuschoff, K. (2020). Anterior insula reflects surprise in value-based decision-making and perception. *NeuroImage (Orlando, Fla.), 210*, 116549-116549. doi:10.1016/j.neuroimage.2020.116549

Maness, E. B., Burk, J. A., McKenna, J. T., Schiffino, F. L., Strecker, R. E., & McCoy, J. G. (2022). Role of the locus coeruleus and basal forebrain in arousal and attention. *Brain Res Bull, 188*, 47-58.

doi:10.1016/j.brainresbull.2022.07.014

Mathot, S., & Vilotijevic, A. (2022). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behav Res Methods*. doi:10.3758/s13428-022-01957-7

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature, 459*(7248), 837-841. doi:10.1038/nature08028

Mattes, A., Porth, E., & Stahl, J. (2022). Linking neurophysiological processes of action monitoring to post-response speed-accuracy adjustments in a neuro-cognitive diffusion model. *Neuroimage, 247*, 118798. doi:10.1016/j.neuroimage.2021.118798

Matzke, D., Dolan, C. V., Logan, G. D., Brown, S. D., & Wagenmakers, E. J. (2013). Bayesian parametric estimation of stop-signal reaction time distributions. *J Exp Psychol Gen, 142*(4), 1047-1073. doi:10.1037/a0030543

Matzke, D., & Wagenmakers, E. J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: a diffusion model analysis. *Psychon Bull Rev, 16*(5), 798-817. doi:10.3758/PBR.16.5.798

Mazzoni, P., Hristova, A., & Krakauer, J. W. (2007). Why Don't We Move Faster? Parkinson's Disease, Movement Vigor, and Implicit Motivation. *The Journal of neuroscience, 27*(27), 7105-7116. doi:10.1523/JNEUROSCI.0264-07.2007

Metereau, E., & Dreher, J.-C. (2013). Cerebral correlates of salient prediction error for different rewards and punishments. *Cerebral cortex (New York, N.Y. 1991), 23*(2), 477-487. doi:10.1093/cercor/bhs037

Miletic, S., Boag, R. J., & Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia, 136*, 107261. doi:10.1016/j.neuropsychologia.2019.107261

Minamimoto, T., Hori, Y., & Kimura, M. (2005). Complementary process to response bias in the centromedian nucleus of the thalamus. *Science, 308*(5729), 1798-1801. doi:10.1126/science.1109154

Mobini, S., Body, S., Ho, M. Y., Bradshaw, C. M., Szabadi, E., Deakin, J. F. W., & Anderson, I. M. (2002). Effects of lesions of the orbitofrontal cortex on sensitivity to delayed and probabilistic reinforcement. *Psychopharmacologia, 160*(3), 290-298. doi:10.1007/s00213-001-0983-0

Monosov, I. E. (2017). Anterior cingulate is a source of valence-specific information about value and uncertainty. *Nature communications, 8*(1), 134-134. doi:10.1038/s41467-017-00072-y

Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X., & Sharot, T. (2015). Human Frontal-Subcortical Circuit and Asymmetric Belief Updating. *Journal of Neuroscience, 35*(42), 14077-14085. doi:10.1523/Jneurosci.1120-15.2015

Mulder, M. J., van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences – A model-based review. *Neuroscience, 277*, 872-884. doi:10.1016/j.neuroscience.2014.07.031

Murphy, P. R., O'Connell, R. G., O'Sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Hum Brain Mapp, 35*(8), 4140-4154. doi:10.1002/hbm.22466

Murphy, P. R., van Moort, M. L., & Nieuwenhuis, S. (2016). The Pupillary Orienting Response Predicts Adaptive Behavioral Adjustment after Errors. *PLoS One, 11*(3), e0151763. doi:10.1371/journal.pone.0151763

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat Neurosci, 15*(7), 1040-1046. doi:10.1038/nn.3130

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J Neurosci, 30*(37), 12366-12378.

doi:10.1523/JNEUROSCI.0822-10.2010

Navarro, D. J., Fuss, I.G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of mathematical psychology, 53*(4), 222-230. doi:https://doi.org/10.1016/j.jmp.2009.02.003

Nelson, A., & Mooney, R. (2016). The Basal Forebrain and Motor Cortex Provide Convergent yet Distinct Movement-Related Inputs to the Auditory Cortex. *Neuron, 90*(3), 635-648. doi:10.1016/j.neuron.2016.03.031

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of neuroscience, 35*(21), 8145-8157. doi:10.1523/JNEUROSCI.2978-14.2015

Niv, Y., Daw, N. D., & Dayan, P. (2006). Choice values. *Nature neuroscience, 9*(8), 987-988. doi:10.1038/nn0806-987

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology, 191*(3), 507-520. doi:10.1007/s00213-006-0502-4

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of neuroscience, 32*(2), 551-562. doi:10.1523/JNEUROSCI.5498-10.2012

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior, 10*(1), 5-24. doi:Doi 10.1177/10597123020101001

Nunez, M. D., Vandekerckhove, J., & Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal of mathematical psychology, 76*(Pt B), 117-130. doi:10.1016/j.jmp.2016.03.003

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature neuroscience, 4*(1), 95. doi:10.1038/82959

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron (Cambridge, Mass.), 38*(2), 329-337. doi:10.1016/S0896-6273(03)00169-7

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Ann N Y Acad Sci, 1104*, 35-53. doi:10.1196/annals.1390.022

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature, 441*(7090), 223-226. doi:10.1038/nature04676;Received28November2005;Accepted24February 2006;Publishedonline23April2006

Pajkossy, P., Szollosi, A., Demeter, G., & Racsmany, M. (2017). Tonic noradrenergic activity modulates explorative behavior and attentional set shifting: Evidence from pupillometry and gaze pattern analysis. *Psychophysiology, 54*(12), 1839-1854. doi:10.1111/psyp.12964

Palminteri, S. (2023). Choice-Confirmation Bias and Gradual Perseveration in Human Reinforcement Learning. *Behavioral Neuroscience, 137*(1), 78-88. doi:10.1037/bne0000541

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology, 13*(8), e1005684-e1005684. doi:10.1371/journal.pcbi.1005684

Palminteri, S., & Pessiglione, M. (2017). Opponent Brain Systems for Reward and Punishment Learning: Causal Evidence From Drug and Lesion Studies in Humans. *Decision Neuroscience: An Integrative Perspective*, 291-303. doi:10.1016/B978-0-12-805308-9.00023-3

Parnaudeau, S., Taylor, K., Bolkan, S. S., Ward, R. D., Balsam, P. D., & Kellendonk, C. (2015). Mediodorsal Thalamus Hypofunction Impairs Flexible Goal-Directed Behavior. *Biological psychiatry (1969), 77*(5), 445-453. doi:10.1016/j.biopsych.2014.03.020

Parra, L. C., Spence, C. D., Gerson, A. D., & Sajda, P. (2005). Recipes for the linear analysis of EEG. *Neuroimage, 28*(2), 326-341. doi:10.1016/j.neuroimage.2005.05.032

Pedersen, M. L., & Frank, M. J. (2020). Simultaneous Hierarchical Bayesian Parameter Estimation for Reinforcement Learning and Drift Diffusion Models: a Tutorial and Links to Neural Data. *Comput Brain Behav, 3*(4), 458-471. doi:10.1007/s42113-020-00084-w

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review, 24*(4), 1234-1251. doi:10.3758/s13423-016-1199-y

Pernet, C. R., Wilcox, R., & Rousselet, G. A. (2012). Robust correlation analyses: false positive and power validation using a new open source matlab toolbox. *Front Psychol, 3*, 606. doi:10.3389/fpsyg.2012.00606

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. *J Neurosci, 26*(35), 8965-8975. doi:10.1523/JNEUROSCI.1655-06.2006

Philiastides, M. G., & Sajda, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cereb Cortex, 16*(4), 509-518. doi:10.1093/cercor/bhi130

Philiastides, M. G., & Sajda, P. (2007). EEG-informed fMRI reveals spatiotemporal characteristics of perceptual decision making. *J Neurosci, 27*(48), 13082-13091. doi:10.1523/JNEUROSCI.3540-07.2007

Pleskac, T. J. (2008). Decision Making and Learning While Taking Sequential Risks. *Journal of experimental psychology. Learning, memory, and*

cognition, 34(1), 167-185. doi:10.1037/0278-7393.34.1.167

Preuschoff, K., & Bossaerts, P. (2007). Adding Prediction Risk to the Theory of Reward Learning. *Annals of the New York Academy of Sciences, 1104*(1), 135-146. doi:10.1196/annals.1390.005

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human Insula Activation Reflects Risk Prediction Errors As Well As Risk. *The Journal of neuroscience, 28*(11), 2745-2752. doi:10.1523/JNEUROSCI.4286-07.2008

Preuschoff, K., t Hart, B. M., & Einhauser, W. (2011). Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. *Front Neurosci, 5*, 115. doi:10.3389/fnins.2011.00115

Radulescu, A., Daniel, R., & Niv, Y. (2016). The Effects of Aging on the Interaction Between Reinforcement Learning and Attention. *Psychology and Aging, 31*(7), 747-757. doi:10.1037/pag0000112

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci, 9*(7), 545-556. doi:10.1038/nrn2357

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput, 20*(4), 873-922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proc Natl Acad Sci U S A, 106*(16), 6539-6544. doi:10.1073/pnas.0812589106

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological science, 9*(5), 347-356. doi:10.1111/1467-9280.00067

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychol Rev, 111*(2), 333-367. doi:10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends Cogn Sci, 20*(4), 260-281. doi:10.1016/j.tics.2016.01.007

R Core Team. (2021). "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Astria. https://R-project.org/.

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nat Commun, 7*, 13289. doi:10.1038/ncomms13289

Rescorla, R. A., Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In W. F. P. A. H. Black (Ed.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New Yoek: Appleton-Century-Crofts.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature neuroscience, 10*(12), 1615-1624. doi:10.1038/nn2013

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain: Neural correlates of RW and PH. *The European journal of neuroscience, 35*, 1190-1200. doi:10.1111/j.1460-9568.2011.07986.x

Rolls, E. T., Grabenhorst, F., & Deco, G. (2010). Decision-making, errors, and confidence in the brain. *J Neurophysiol, 104*(5), 2359-2374. doi:10.1152/jn.00571.2010

Rowe, J. B., Hughes, L., & Nimmo-Smith, I. (2010). Action selection: a race

model for selected and non-selected actions distinguishes the contribution of premotor and prefrontal areas. *Neuroimage, 51*(2), 888-896. doi:10.1016/j.neuroimage.2010.02.045

Rushworth, M. F., & Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat Neurosci, 11*(4), 389-397. doi:10.1038/nn2066

Rushworth, M. F., Behrens, T. E., Rudebeck, P. H., & Walton, M. E. (2007). Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. *Trends Cogn Sci, 11*(4), 168-176. doi:10.1016/j.tics.2007.01.004

Sajda, P., Philiastides, M. G., & Parra, L. C. (2009). Single-trial analysis of neuroimaging data: inferring neural networks underlying perceptual decision-making in the human brain. *IEEE Rev Biomed Eng, 2*, 97-109. doi:10.1109/RBME.2009.2034535

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science (American Association for the Advancement of Science), 310*(5752), 1337-1340. doi:10.1126/science.1115270

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nat Rev Neurosci, 10*(3), 211-223. doi:10.1038/nrn2573

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: the locus coeruleus mediates cognition through arousal. *Neuron, 76*(1), 130-141. doi:10.1016/j.neuron.2012.09.011

Schmid, Y., Enzler, F., Gasser, P., Grouzmann, E., Preller, K. H., Vollenweider, F. X., . . . Liechti, M. E. (2015). Acute Effects of Lysergic Acid Diethylamide in Healthy Subjects. *Biol Psychiatry, 78*(8), 544-553. doi:10.1016/j.biopsych.2014.11.015

Schmitz, F., Manske, K., Preckel, F., & Wilhelm, O. (2016). The Multiple Faces of Risk-Taking Scoring Alternatives for the Balloon-Analogue Risk Task.

European Journal of Psychological Assessment, 32(1), 17-38. doi:10.1027/1015-5759/a000335

Schneider, M., Leuchs, L., Czisch, M., Samann, P. G., & Spoormaker, V. I. (2018). Disentangling reward anticipation with simultaneous pupillometry / fMRI. *Neuroimage, 178*, 11-22. doi:10.1016/j.neuroimage.2018.04.078

Schultz, W. (1999). The Reward Signal of Midbrain Dopamine Neurons. *News Physiol Sci, 14*, 249-255. doi:10.1152/physiologyonline.1999.14.6.249

Schultz, W. (2007). Multiple dopamine functions at different time courses. *Annual review of neuroscience, 30*(1), 259-288. doi:10.1146/annurev.neuro.28.061604.135722

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience, 17*(3), 183-195. doi:10.1038/nrn.2015.26

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593-1599. doi:10.1126/science.275.5306.1593

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of statistics, 6*(2), 461-464. doi:10.1214/aos/1176344136

Seifert, S., von Cramon, D. Y., Imperati, D., Tittgemeyer, M., & Ullsperger, M. (2011). Thalamocingulate interactions in performance monitoring. *J Neurosci, 31*(9), 3375-3383. doi:10.1523/JNEUROSCI.6242-10.2011

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *J Neurosci, 27*(18), 4826-4831. doi:10.1523/JNEUROSCI.0400-07.2007

Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., & Dolan, R. (2012). Serotonin selectively modulates reward value in human decision-making. *The Journal of neuroscience, 32*(17), 5833-5842.

doi:10.1523/jneurosci.0053-12.2012

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., . . . Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature, 429*(6992), 664-667. doi:10.1038/nature02581

Shackman, A. J., Salomons, T. V., Slagter, H. A., Fox, A. S., Winter, J. J., & Davidson, R. J. (2011). The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nat Rev Neurosci, 12*(3), 154-167. doi:10.1038/nrn2994

Shadlen, M. N., & Kim, J.-N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature neuroscience, 2*(2), 176-185. doi:10.1038/5739

Sharot, T., Guitart-Masip, M., Korn, Christoph W., Chowdhury, R., & Dolan, Raymond J. (2012). How Dopamine Enhances an Optimism Bias in Humans. *Current biology, 22*(16), 1477-1481. doi:10.1016/j.cub.2012.05.053

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature neuroscience, 14*(11), 1475-U1156. doi:10.1038/nn.2949

Sharot, T., Riccardi, A. M., Raio, C. M., & Phelps, E. A. (2007). Neural mechanisms mediating optimism bias. *Nature, 450*(7166), 102-+. doi:10.1038/nature06280

Shepperd, J. A., Klein, W. M. P., Waters, E. A., & Weinstein, N. D. (2013). Taking Stock of Unrealistic Optimism. *Perspectives on Psychological Science, 8*(4), 395-411. doi:10.1177/1745691613485247

Smith, R., Taylor, S., Stewart, J. L., Guinjoan, S. M., Ironside, M., Kirlic, N., . . . Paulus, M. P. (2022). Slower Learning Rates from Negative Outcomes in Substance Use Disorder over a 1-Year Period and Their Potential Predictive Utility. *Computational psychiatry, 6*(1), 117.

doi:10.5334/cpsy.85

Sokol-Hessner, P., Camerer, C. F., & Phelps, E. A. (2013). Emotion regulation reduces loss aversion and decreases amygdala responses to losses. *Social cognitive and affective neuroscience, 8*(3), 341-350. doi:10.1093/scan/nss002

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature reviews. Neuroscience, 20*(10), 635-644. doi:10.1038/s41583-019-0180-y

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 64*, 583-616. doi:Doi 10.1111/1467-9868.00353

Sripada, C., & Weigard, A. (2021). Impaired Evidence Accumulation as a Transdiagnostic Vulnerability Factor in Psychopathology. *Front Psychiatry, 12*, 627179. doi:10.3389/fpsyt.2021.627179

Stan Developmental Team. (2019). "RStan: the R interface to Stan." R package version 2.17.5, https://mc-stan.org/.

Stock, A. K., Hoffmann, S., & Beste, C. (2016). Effects of binge drinking and hangover on response selection sub-processes-a study using EEG and drift diffusion modeling. *Addict Biol, 22*(5), 1355-1365. doi:10.1111/adb.12412

Strauss, G. P., Waltz, J. A., & Gold, J. M. (2014). A Review of Reward Processing and Motivational Impairment in Schizophrenia. *Schizophrenia bulletin, 40*(Suppl 2), S107-S116. doi:10.1093/schbul/sbt197

Sutton, R. S., Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*(88), 88:135–170.

Sutton, R. S. (1988). Learning to predict by the methods of temporal

differences. *Machine learning, 3*(1), 9-44. doi:10.1007/BF00115009

Sutton, R. S., Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. G. a. J. Moore (Ed.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (pp. 497–537). Cambridge, MA: MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction* (Second ed.). Cambridge, Massachusetts: The MIT Press.

Tan, Kelly R., Yvon, C., Turiault, M., Mirzabekov, Julie J., Doehner, J., Labouèbe, G., . . . Lüscher, C. (2012). GABA Neurons of the VTA Drive Conditioned Place Aversion. *Neuron (Cambridge, Mass.), 73*(6), 1173-1183. doi:10.1016/j.neuron.2012.02.015

The Mathworks Inc. (2018). MATLAB version 9.5.0.944444 (R2018b). Natick, Massachusetts: The Mathworks Inc. https://www.mathworks.com

Thorndike, E. L. (1898). Review of: Animal Intelligence: An Experimental Study of the Associative Processes in Animals. *Psychological Review, 5*(5), 551-553. doi:10.1037/h0067373

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive Coding of Reward Value by Dopamine Neurons. *Science (American Association for the Advancement of Science), 307*(5715), 1642-1645. doi:10.1126/science.1105370

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of neurophysiology, 95*(1), 301. doi:10.1152/jn.00762.2005

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *Journal of neurophysiology, 97*(2), 1621. doi:10.1152/jn.00745.2006

Torregrossa, M. (2019). *Neural mechanisms of addiction*. London: Academic Press.

Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., & Deisseroth, K. (2009). Phasic Firing in Dopaminergic Neurons Is Sufficient for Behavioral Conditioning. *Science (American Association for the Advancement of Science), 324*(5930), 1080-1084. doi:10.1126/science.1168878

Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: the neural drift diffusion model. *Psychol Rev, 122*(2), 312-336. doi:10.1037/a0038894

Ullsperger, M., & von Cramon, D. Y. (2001). Subprocesses of performance monitoring: a dissociation of error processing and response competition revealed by event-related fMRI and ERPs. *Neuroimage, 14*(6), 1387-1401. doi:10.1006/nimg.2001.0935

Ullsperger, M., & von Cramon, D. Y. (2003). Error Monitoring Using External Feedback: Specific Roles of the Habenular Complex, the Reward System, and the Cingulate Motor Area Revealed by Functional Magnetic Resonance Imaging. *The Journal of neuroscience, 23*(10), 4308-4314. doi:10.1523/jneurosci.23-10-04308.2003

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat Commun, 8*, 14637. doi:10.1038/ncomms14637

van den Bos, W., Cohen, M. X., Kahnt, T., & Crone, E. A. (2012). Striatum-Medial Prefrontal Cortex Connectivity Predicts Developmental Changes in Reinforcement Learning. *Cerebral Cortex, 22*(6), 1247-1255. doi:10.1093/cercor/bhr198

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of mathematical psychology, 55*(1), 94-105.

doi:10.1016/j.jmp.2010.08.010

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the Time Course of Pupillometric Data. *Trends Hear, 23*, 2331216519832483. doi:10.1177/2331216519832483

van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Van Slooten, J. C., Jahfari, S., Knapen, T., & Theeuwes, J. (2017). Individual differences in eye blink rate predict both transient and tonic pupil responses during reversal learning. *PLoS One, 12*(9), e0185665. doi:10.1371/journal.pone.0185665

Van Slooten, J. C., Jahfari, S., Knapen, T., & Theeuwes, J. (2018). How pupil responses track value-based decision-making during and after reinforcement learning. *PLoS Comput Biol, 14*(11), e1006632. doi:10.1371/journal.pcbi.1006632

Van Slooten, J. C., Jahfari, S., & Theeuwes, J. (2019). Spontaneous eye blink rate predicts individual differences in exploration and exploitation during reinforcement learning. *Sci Rep, 9*(1), 17436. doi:10.1038/s41598-019-53805-y

Vandecasteele, M., Glowinski, J., & Venance, L. (2005). Electrical synapses between dopaminergic neurons of the substantia nigra pars compacta. *The Journal of neuroscience, 25*(2), 291-298. doi:10.1523/JNEUROSCI.4167-04.2005

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: a DMAT primer. *Behav Res Methods, 40*(1), 61-72. doi:10.3758/brm.40.1.61

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychol Methods, 16*(1), 44-62. doi:10.1037/a0021765

Varazzani, C., San-Galli, A., Gilardeau, S., & Bouret, S. (2015). Noradrenaline and dopamine neurons in the reward/effort trade-off: a direct electrophysiological comparison in behaving monkeys. *J Neurosci, 35*(20), 7866-7877. doi:10.1523/JNEUROSCI.0454-15.2015

Vassena, E., Deraeve, J., & Alexander, W. H. (2020). Surprise, value and control in anterior cingulate cortex during speeded decision-making. *Nature human behaviour, 4*(4), 412-422. doi:10.1038/s41562-019-0801-5

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413-1432. doi:10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing Convergence of MCMC (with Discussion). *Bayesian analysis, 16*(2). doi:10.1214/20-BA1221

Verdonck, S., Loossens, T., & Philiastides, M. G. (2021). The Leaky Integrating Threshold and its impact on evidence accumulation models of choice response time (RT). *Psychol Rev, 128*(2), 203-221. doi:10.1037/rev0000258

Vitiello, B., Martin, A., Hill, J., Mack, C., Molchan, S., Martinez, R., . . . Sunderland, T. (1997). Cognitive and behavioral effects of cholinergic, dopaminergic, and serotonergic blockade in humans. *Neuropsychopharmacology, 16*(1), 15-24. doi:10.1016/S0893-133X(96)00134-0

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412*(6842), 43-48. doi:10.1038/35083500

Walker, A. R., Navarro, D. J., Newell, B. R., & Beesley, T. (2022). Protection from uncertainty in the exploration/exploitation trade-off. *J Exp Psychol Learn Mem Cogn, 48*(4), 547-568. doi:10.1037/xlm0000883

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling Behavior in a Clinically Diagnostic Sequential Risk-Taking Task. *Psychological Review, 112*(4), 862-880. doi:10.1037/0033-295X.112.4.862

Wang, C. A., & Munoz, D. P. (2015). A circuit for pupil orienting responses: implications for cognitive modulation of pupil size. *Curr Opin Neurobiol, 33*, 134-140. doi:10.1016/j.conb.2015.03.018

Warren, C. M., Wilson, R. C., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., Cohen, J. D., & Nieuwenhuis, S. (2017). The effect of atomoxetine on random and directed exploration in humans. *PLoS One, 12*(4), e0176034. doi:10.1371/journal.pone.0176034

Wassum, K. M., & Izquierdo, A. (2015). The basolateral amygdala in reward learning and addiction. *Neuroscience and biobehavioral reviews, 57*, 271-283. doi:10.1016/j.neubiorev.2015.08.017

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine learning, 8*(3-4), 279-292. doi:10.1007/BF00992698

Weinstein, N. D. (1980). Unrealistic Optimism About Future Life Events. *Journal of Personality and Social Psychology, 39*(5), 806-820. doi:Doi 10.1037/0022-3514.39.5.806

White, C. N., Congdon, E., Mumford, J. A., Karlsgodt, K. H., Sabb, F. W., Freimer, N. B., . . . Poldrack, R. A. (2014). Decomposing decision components in the stop-signal task: a model-based approach to individual differences in inhibitory control. *J Cogn Neurosci, 26*(8), 1601-1614. doi:10.1162/jocn_a_00567

White, C. N., Mumford, J. A., & Poldrack, R. A. (2012). Perceptual criteria in the human brain. *The Journal of neuroscience, 32*(47), 16716-16724. doi:10.1523/JNEUROSCI.1744-12.2012

White, N. M. (2011). Chapter 3 Reward: What is it? How can it be inferred from behaviour? In J. A. Gottfried (Ed.), *Neurobiology of sensation and reward*. Boca Raton: Taylor & Francis.

Wickens, J. R., Horvitz, J. C., Costa, R. M., & Killcross, S. (2007). Dopaminergic Mechanisms in Actions and Habits. *The Journal of neuroscience, 27*(31), 8181-8183. doi:10.1523/JNEUROSCI.1671-07.2007

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front Neuroinform, 7*, 14. doi:10.3389/fninf.2013.00014

Wiehler, A., Chakroun, K., & Peters, J. (2021). Attenuated Directed Exploration during Reinforcement Learning in Gambling Disorder. *J Neurosci, 41*(11), 2512-2522. doi:10.1523/JNEUROSCI.1607-20.2021

Wierda, S. M., van Rijn, H., Taatgen, N. A., & Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proc Natl Acad Sci U S A, 109*(22), 8456-8460. doi:10.1073/pnas.1201858109

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *J Exp Psychol Gen, 143*(6), 2074-2081. doi:10.1037/a0038199

Wise, R. A. (2004). Dopamine, learning and motivation. *Nat Rev Neurosci, 5*(6), 483-494. doi:10.1038/nrn1406

Yamamoto, K., & Hornykiewicz, O. (2004). Proposal for a noradrenaline hypothesis of schizophrenia. *Prog Neuropsychopharmacol Biol Psychiatry, 28*(5), 913-922. doi:10.1016/j.pnpbp.2004.05.033

Yaple, Z. A., & Yu, R. (2019). Fractionating adaptive learning: A meta-analysis of the reversal learning paradigm. *Neuroscience and biobehavioral reviews, 102*, 85-94. doi:10.1016/j.neubiorev.2019.04.006

Yau, J. O.-Y., & McNally, G. P. (2023). The Rescorla-Wagner model, prediction error, and fear learning. *Neurobiology of learning and memory, 203*, 107799-107799. doi:10.1016/j.nlm.2023.107799

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review, 12*(3), 387-402. doi:Doi 10.3758/Bf03193783

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Research Article: Using Cognitive Models to Map Relations Between Neuropsychological Disorders and Human Decision-Making Deficits. *Psychological science, 16*(12), 973-978. doi:10.1111/j.1467-9280.2005.01646.x

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol Rev, 111*(4), 931-959. doi:10.1037/0033-295x.111.4.939

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*(4), 681-692. doi:10.1016/j.neuron.2005.04.026

Zaghloul, K. A., Weidemann, C. T., Lega, B. C., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2012). Neuronal activity in the human subthalamic nucleus encodes decision conflict during action selection. *J Neurosci, 32*(7), 2453-2460. doi:10.1523/JNEUROSCI.5815-11.2012

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife, 6*. doi:10.7554/eLife.27430

Zenon, A. (2019). Eye pupil signals information gain. *Proc Biol Sci, 286*(1911), 20191593. doi:10.1098/rspb.2019.1593

Zhang, J., & Rowe, J. B. (2014). Dissociable mechanisms of speed-accuracy tradeoff during visual perceptual learning are revealed by a hierarchical drift-diffusion model. *Front Neurosci, 8*, 69. doi:10.3389/fnins.2014.00069

Zhou, R., Myung, J. I., & Pitt, M. A. (2021). The scaled target learning model: Revisiting learning in the balloon analogue risk task. *Cognitive*

psychology, 128, 101407-101407. doi:10.1016/j.cogpsych.2021.101407