Westwood, Sean (2025) *Neural characteristics of reward and punishment learning for optimising decision-making.* PhD thesis.

https://theses.gla.ac.uk/84992/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

Neural characteristics of reward and punishment learning for optimising decision-making

Sean Westwood

Submitted in fulfilment of the requirements for the

Degree of Doctor of Philosophy

School of Computing Science

College of Science and Engineering

University of Glasgow

September 2024

"No, I never go home, don't sleep, don't eat

Just do it on repeat"

- Charlotte Emma Aitchison

Abstract

Avoiding aversive or punishing outcomes is integral to the survival and function of an organism, and yet this specific dimension of learning has received comparatively little attention in comparison to its rewarding counterpart. There are a number of mechanistic theories and pathways implicated in each dimension, but as of yet there is limited consensus on a holistic model of punishment processing. They key contributions of this thesis are three-fold. First, I offer a novel insight into the role of a motivational salience signal in modulating different behaviours in rewarding and punishing contexts and make the claim that this is compatible with prominent existing mechanistic accounts of punishment learning. Second, I frame this effect with a prominent individual-difference focus, highlighting the importance of considering subject-specific sensitivities to reward and punishment when attempting to model this dichotomy. Third, I use the insights from the first two contributions to show the viability of using this paradigm as a potential means to improve task performance through a closed-loop brain-computer interface (BCI).

These contributions are made primarily from data from a reversal learning task conducted across rewarding and punishing blocks, with EEG and pupillometry as the primary measures of interest. In Chapter 2, I replicate a two-component paradigm from Philiastides et al., (2010) and Fouragnan et al. (2015) with the extension of a punishment condition. I show broad similarities in the EEG signatures, with some notable insights from pupillometry as to promising temporal signals to focus on for further investigation. In Chapter 3, I use these insights to test a motivational salience hypothesis, targeting specifically the earlier component in this two-component hypothesis. I show that individual differences in the sensitivity of this component to context can reliably track task performance. In Chapter 4, I delve further into these individual differences,

focusing more specifically on reinforcement learning parameters and psychometric personality measures to better characterise performance effects from Chapter 3. In Chapter 5, I apply all of these findings in a tentative pseudo-BCI analysis, where I retroactively estimate performance and predict whether dynamic context switching might have led to improved behavioural outcomes.

Together, these findings offer new insights into the spatiotemporal characterisation of reward-punishment differences and propose some tentative yet exciting future directions for the application of this in the performance optimisation domain.

Contents

List of Figures

Acknowledgements

I don't know many supervisors who could find the patience and composure to put up with so many deadline panics, so many indecisive pivots, so many bleary-eyed half-coherent Monday morning meetings. I've said many times that Professor Philiastides is the only reason I decided to become a postgrad, and is probably the only one who could've tempted me to consider a postdoc. He has taught me nearly everything I know, and had enough faith in me to let me cook when it mattered most. I could not be more grateful.

I don't know many partners who could survive the level of chaos I have subjected Rebecca to over the last four years, let alone continue be there for me to the extent she has. She pulled me stoically through the toughest times despite the countless nights locked away in the office, the unkempt appearance, the deteriorating sanity. Her tolerance to reward omission is unmatched (though perhaps near its limit). I promise I will never do another PhD.

I don't know many parents who could tolerate two missed graduation ceremonies in return for nearly three decades of tireless support. I promise it will be third time lucky (mum please text me when registration opens).

I don't know many colleagues who could make conference trips quite as fun as Ralitsa and Joana. I will never forget the adventures we had in San Diego and Vienna (especially those Prater rides). You've both made some challenging times a lot more enjoyable, and I couldn't ask for a better desk buddy or nemesis. I'm also grateful to Desislava and Kitti for all of the guidance when I joined the lab, and Filippo for the modelling advice and the many laughs.

I don't know many friends who would be willing to drop everything and spend their valuable evenings keeping the dreams of a delirious shell alive. It is not an understatement to say that without Christopher and Michael (and Rebecca of course), this would literally not have been possible (I ran the numbers).

To the friends who were either far away, or whom I didn't have the audacity to beg for help, I owe you no less gratitude. To Simon for the years of friendship, advice, and mutual PhD rants; to Sander for the countless hours of esoteric conversation nobody else could provide; to Tobias for the much needed solidarity throughout all teaching and research endeavours (hang in there brother); to the many others who were there for me along the way. Thank you all.

Last but by no means least I thank Professor Vinciarelli for your invaluable help in establishing the scope of the project, as well as Dr Ince and Dr Chollet for the insightful comments and words of encouragement over the years.

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.

Chapter 1

General introduction

1.1 Human Decision Making

For over a century, great effort has been made to understand how and why people make value-based decisions. This endeavour was historically explored in the context of classical economics, which has long attempted to model this facet of behaviour for a range of purposes, such as predicting market behaviour or designing economic policy (Caplin & Glimcher, 2013). Expected utility theory, based on initial ideas by Daniel Bernoulli in the mid-1700s (Modesti, 2024) and formalised in the mid-1900s (von Neumann et al., 2004), remained a highly influential model of economic decisions right up to the late 20^th century (Caplin & Glimcher, 2013). But with the growth of the fields of psychology and neuroscience, new insights into the intricacies of human behaviour began to reveal ever increasing complexity.

1.1.1 Behavioural Economics

The most famous and groundbreaking deviation from neoclassical economics came in the 1970s with the advent of a field known as behavioural economics. A landmark paper from Kahneman and Tversky (1979) demonstrated that people show robust asymmetry in choice dynamics

when faced with a prospect of monetary gain versus monetary loss. Specifically, participants consistently tend to sacrifice some amount of expected value in return for a more sure reward, and conversely tend to risk additional potential loss for a greater chance of not losing anything at all. This concept was dubbed Prospect Theory, due to its application to economic decisions (or "prospects), and is commonly characterised by the prospect theory curve Figure 1.1B, which demonstrates the asymmetry in contrast to the classic expected utility theory function Figure 1.1A.

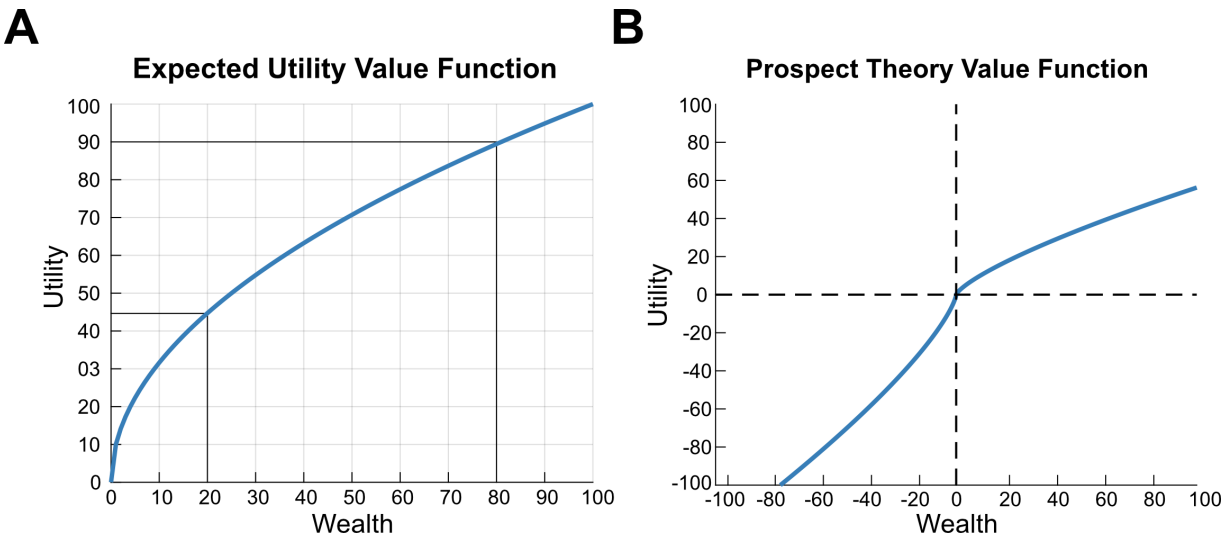

Figure 1.1: A) Expected utility theory value function. Perceived value of gains in wealth diminish as total wealth increases. B) Prospect theory value function. The value function displays a steeper weighting function for losses relative to gains, illustrating the concept of loss aversion.

This stark asymmetry was demonstrated many times by the original authors and others, quantified in a plethora of neat examples such as that displayed in Figure 1.2. Here we can see the fourfold pattern of risk attitudes, which demonstrates a clear preference for certainty in the gain domain and a clear preference for risk in the loss domain (Tversky & Kahneman, 1992). The work of Kahneman and Tversky demonstrated the complex and often irrational tendencies of humans in their choice behaviour, coinciding with increasing investigations into the complex dynamics of learning and decision-making in a value-based setting.

	Gains	Losses
Low Probability	c($100, .05) = $14	c(-$100, .05) = -$8
High Probability	c($100, .95) = $78	c(-$100, .95) = $84

Figure 1.2: Fourfold Pattern of Risk Attitudes (adapted from Tversky and Kahneman (1992)). Estimations of monetary values attributed to high- and low-probability prospects in the gain and loss domains. Here, c(X, p) shows the median certain monetary value that was treated by participants as equivalent to the uncertain prospect. Low-probability gains (top left) and high-probability losses (bottom right) depict a monetary value greater than the true Expected Value of the prospect, signifying risk-seeking attitudes. Low-probability losses (top right) and high-probability gains (bottom left) depict a monetary value lower than the true Expected Value, signifying risk-aversion.

1.1.2 Prediction error and learning

Computational models of decision-making attempt to capture internal value judgements about actions and environmental stimuli that are updated as an agent receives information from their surroundings. One of the earliest formulations of this domain originated in an associative learning context with the notion of the computational prediction error (Bush & Mosteller, 1951), which preceded the influential dopaminergic reward prediction error (Houk et al., 1994; Montague et al., 1996; Schultz et al., 1997; Schultz, 1998). Initially conceptualised in the Pavlovian sense (learning stimuli-outcome rather than action-outcome contingencies), this was proposed to reflect the difference between an expected reward V (for value) and received reward r following a stimulus s. For instance, if at timepoint t a rat were to perceive a neutral stimulus st (e.g. a light) with a V of 0, and subsequently obtain a reward r, this would be seen as a 'better than expected' outcome and thus induce a reward prediction error δ.

$$\delta_t = r_t - V_t(i) \tag{1.1}$$

The prediction error is then proposed to update the stimulus value for the next point in time V_{t+1} such that it more accurately reflects the true stimulus outcome contingencies and reduces

future prediction errors. The updating process is purportedly weighted by a free parameter known as the learning rate α, which can vary between zero and one. A higher learning rate will place more weight on new information about stimulus outcomes, and thus lead to more rapid updating of stimulus value following prediction errors.

$$V_{t+1}(i) = V_t(i) + \alpha \cdot \delta_t \tag{1.2}$$

This formulation is often described as the Rescorla-Wagner model, referring to the famous computational paper by Rescorla and Wagner (1972). However, it is worth noting that this paper proposed an extension of the Bush and Mosteller model that incorporates multiple simultaneous stimuli by summing stimulus values and calculating a net prediction error. Though an important extension that accounted for key classical conditioning effects such as blocking, the core model of Bush and Mosteller (1951) as presented in equations 1.1 and 1.2 is often misattributed to Rescorla and Wagner (Glimcher, 2011).

1.1.3 Action selection

The above model is widely employed as a simple learning rule to model both associative and instrumental learning, but in a decision-making context there must be some decision policy that enables the agent to select actions based on stimulus values. In decision-making under incomplete information it is rare that a 'greedy' policy of always choosing the highest value option is the most successful over the long term, and so the decision-maker should employ a trade-off between exploring different options to gain information and exploiting known high-value options (Gittins, 1979; Gittins & Jones, 1974). This balance is prevalent in many facets of life, such as the ubiquitous example in the decision-making literature of choosing a familiar restaurant/café/meal versus trying a something new. More formally, this can be observed in nature through fundamental evolutionary activities such as foraging (Krebs et al., 1978), hunting (Sims et al., 2008), or navigating volatile environments (Thatcher et al., 2019).

Given that optimal exploration is generally intractable (J. D. Cohen et al., 2007) there are many different heuristic models that have been proposed to more realistically reflect neural computation capabilities, generally falling under the categories of directed or random exploration (Wilson et al., 2014, 2021). Where directed exploration attempts to optimise the target of non-exploitative decisions (Averbeck, 2015; Gittins, 1979), random exploration offers a computational cheap policy that stochastically samples options in an unbiased manner (Daw et al., 2006; Thompson, 1933). There is competing evidence as to whether directed exploration models match human choice better than stochastic exploration (M. J. Frank et al., 2009; Meyer & Shi, 1995) or not (Daw et al., 2006; Payzan-LeNestour & Bossaerts, 2011), however for simple choice tasks the stochastic Softmax decision rule is often deployed:

$$p_i(t) = \frac{e^{\gamma \cdot V_i(t)}}{\sum_{j=1}^{n} e^{\gamma \cdot V_j(t)}} \tag{1.3}$$

Here, the probability p of an actor choosing option i on trial t is modelled as a sigmoid function of the value of the option V_i, the slope of which is determined by the inverse temperature parameter γ. In practice, a slope with the maximum value of 1 would reflect an entirely greedy policy where the higher value option is chosen every time, and a minimum value of 0 would indicate total randomness of choice.

1.2 Reinforcement learning in the brain

The computational modelling of behavioural learning has been highly influential, and has led to widely adopted frameworks of value-based learning and decision-making (e.g. Rangel et al. (2008). As such, many influential accounts of the neural mechanisms underpinning the learning process in animals and humans echo strongly some of the central features described in 1.1.2. The most prominent early work into this area has been developed in the reward learning paradigm, where neural responses to the prediction, presence, and omission of appetitive stimuli is tracked in key pathways using electrophysiology and brain imaging.

1.2.1 Reward prediction error

Perhaps the most influential neural theory in reinforcement learning is the dopaminergic reward prediction error (RPE) hypothesis (Schultz et al., 1997). In this landmark paper, single-unit recording of DA neurons in the ventral tegmental area of rhesus macaque monkeys initially displayed increased phasic firing to the rewarding stimulus of a drop of fruit juice. When this reward was preceded by a predictive cue, however, the magnitude of phasic firing to the reward itself was diminished, with the neurons instead responding to the cue. Importantly, when the learned cue was then followed by an unexpected reward omission, firing rates dipped below baseline at the moment of expected reward (Schultz et al., 1997). Quantifying these signals lead to the influential insight that these neurons are encoding an RPE rather than the value of a stimulus – that is, the degree to which a stimulus signals a deviation from expected reward. It was subsequently shown that this effect could be parametrically modulated by the strength of the reward contingency of the predictive cue (Fiorillo et al., 2003), making a compelling case that net phasic firing in dopaminergic VTA neurons relative to baseline following a stimulus reflects a computational RPE signal.

The dopaminergic RPE hypothesis has been extensively supported and developed since the original finding of Schultz and colleagues. Numerous electrophysiological studies have replicated this RPE signal in the mammalian VTA (e.g. Bayer and Glimcher (2005), Bayer et al. (2007), and Roesch et al. (2007). In humans, functional magnetic resonance imaging (fMRI) studies have repeatedly shown activity in the ventral striatum (vSTR) – an area that receives DA projections from VTA (Ikemoto, 2007) – that seems to reflect a similar reward prediction sensitivity (e.g. Bartra et al. (2013), Clithero and Rangel (2014), Delgado et al. (2000), McClure et al. (2003), O'Doherty (2004), O'Doherty et al. (2002, 2003), O'Doherty et al. (2004), Pagnoni et al. (2002), and Tobler et al. (2007, 2008). These projections continue into the prefrontal cortex, with the ventromedial prefrontal cortex (vmPFC; Bartra et al., 2013; Blair et al., 2006; Clithero and Rangel, 2014; Daw et al., 2006; Gläscher et al., 2009; Hampton et al., 2006) and orbitofrontal cortex (OFC; Balleine et al., 2011; Elliott et al., 2008; Gourley et al., 2016) in particular being shown to track reward expectation and value. These signals have been shown to

relate to behavioural performance on learning tasks (Pessiglione et al., 2006; Schönberg et al., 2007).

1.2.2 Striatocortical basis of explore-exploit trade-offs

An influential early framing of the implementation of explore-exploit in the brain is an opponent mechanism featuring typical dopaminergic reward pathways inhibited by top-down signals from control areas of the prefrontal cortex (PFC). This was first demonstrated using fMRI in a dynamic multi-armed bandit task with reward contingencies of each arm fluctuating throughout the task to necessitate exploration (Daw et al., 2006). The study found that typical reward-related cortical regions ventromedial and orbitofrontal PFC were sensitive to exploitative choice likelihood and reward magnitude respectively (Daw et al., 2006), in line with existing evidence of dopaminergic striatocortical pathways attributed to reward-seeking 'exploit' behaviour (Delgado et al., 2000; McClure et al., 2003; O'Doherty, 2004; O'Doherty et al., 2001, 2003; O'Doherty et al., 2004). Critically, explorative decisions were related specifically to the bilateral frontopolar cortex (Daw et al., 2006), consistent with the notion that dorsal and anterior areas of the PFC are responsible for cognitive control and promoting higher-level goals (Miller & Cohen, 2001). Further work implicated the bilateral frontopolar cortex in the tracking of environmental uncertainty, offering a more specific way in which this cortical regions may facilitate exploration (Badre et al., 2012; Cavanagh et al., 2012). These findings suggested a top-down cortical signal that promotes exploration by suppressing exploitative signals in dopaminergic reward circuits, challenging prior computational accounts of an uncertainty bonus that is incorporated more directly into the expected reward calculation (Gittins, 1979; Gittins & Jones, 1974; Kakade & Dayan, 2002).

A more recent alternative framing proposes that rather than a top-down disruptive signal from the PFC, exploration is instead a function of value-computations that are optimised for long-term gain maximisation and loss minimisation (Averbeck, 2015; Wilson et al., 2021). In this formulation, control-related frontocortical areas work in tandem with rather than in opposition to

reward networks to combine expectations of immediate and future rewards in order to determine optimal moments to explore rather than exploit (Costa & Averbeck, 2020; Costa et al., 2019; Tang et al., 2022). For example, dorsolateral PFC neurons were recently found to encode both immediate and future value in rhesus macaque monkeys (Tang et al., 2022), and a side-by-side analysis of fMRI in humans using a similar 3-armed bandit task also demonstrated a cooperative rather than opponent relationship between key PFC and motivational regions when tackling the explore-exploit challenge (Hogeveen et al., 2022).

1.2.3 Noradrenaline and adaptive gain

Regardless of whether the cortical systems in question work cooperatively or in opposition, it is clear that both control-related dorsal and anterior cortical regions, as well as value-related ventral striatocortical regions, are strongly implicated in regulating exploration. However, there does not appear to be a great deal of overlap between the areas in question and the salience-network regions associated with the early feedback-processing component highlighted in Chapter 3. As such, there are not particularly strong grounds with which to form a prediction about a relationship between exploration rates and our early salience-related EEG signal. However, in addition to striatocortical influences, the LC-driven noradrenergic pathway has been linked to explore-exploit function through a mechanism known as adaptive gain (Aston-Jones & Cohen, 2005).

This theory proposes that a double dissociation exists between tonic and phasic noradrenergic activity (Aston-Jones & Cohen, 2005; J. D. Cohen et al., 2007): a tonic (baseline) mode promotes greater general arousal and a higher degree of attentional switching; a phasic (stimulus-driven) mode promotes selective attention and more focused task performance. In essence, high tonic activity is associated with greater exploration, and high phasic activity (at stimulus onset) with greater exploitation. This is proposed to be implemented through an increase in the relative activity of neural units (individual or population firing rates) with an excitatory influence, and a corresponding decrease in the activity of units with an inhibitory influence (Figure 1.3).

Crucially, this theory is often framed in the context of optimal arousal for task performance, as described by the classic Yerkes-Dodson inverted-U model (Figure 1.3B). Overarousal leads to excessive attentional and strategic switching, whereas under arousal prevents the noticing of environmental changes and required behavioural adaptations.

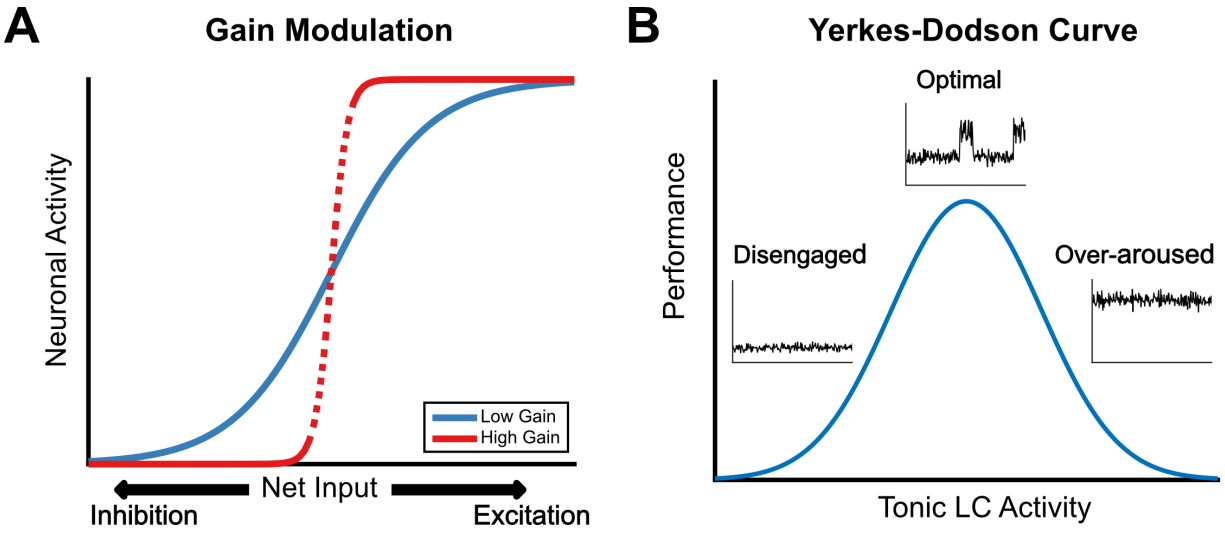

Figure 1.3: (adapted from Aston-Jones and Cohen (2005)) A) Depiction of the adaptive gain activation function. The y-axis depicts the net activity of a unit, and the x-axis depicts the degree of excitatory or inhibitory influence of a unit. Here, a unit can refer to either a single neuron or mean firing of a population of neurons). An increase in gain increases the relative activity of excitatory influences and decreases the relative activity of inhibitory influences. B) Depiction of the classic Yerkes-Dodson inverted-U model of optimal arousal for task performance. In the context of adaptive gain, excessive tonic LC activity indicates high baseline arousal and a tendency to switch too frequently between different stimuli and choices. Diminished tonic LC activity will conversely prevent identification of necessary attentional targets or strategy alterations for optimal performance.

The adaptive gain model is supported by pupillometry studies in perceptual decision making tasks that have found that ambient levels of noradrenergic arousal is related to subjective decision uncertainty (Brunyé & Gardony, 2017; Kawaguchi et al., 2018; Urai et al., 2017). Similarly, this same type of tonic arousal has been linked to the decision-maker's internal confidence in their model of the decision-making environment (de Gee et al., 2020; Nassar et al., 2012), as well as perceived volatility during value-based choice (Binetti et al., 2017; Kloosterman et al., 2015; Nassar et al., 2012). Even more directly, multiple studies have shown greater tendency to exhibit exploratory behaviour and override existing choice biases during periods of high tonic arousal (de Gee et al., 2020; Gilzenrat et al., 2010; Hayes & Petrov, 2016; Jepma & Nieuwenhuis, 2011; Krishnamurthy et al., 2017; Urai et al., 2017).

1.3 Punishment

Reward learning has been extensively studied over the past three decades, and several of the core hypotheses have enjoyed a degree of relative consensus in the field. Less certain, however, is the nature of learning from punishment. We all have the experience that an aversive stimulus as a motivator 'feels' subjectively different to a rewarding one, and perhaps even that our behavioural responses may differ as a result. But is this intuition reflective of some meaningful difference in neural mechanisms?

1.3.1 In the midbrain

Given the strong links between midbrain nuclei – particularly the VTA and substantia nigra – a natural hypothesis would be that aversive events are encoded by dips in firing in the same way that reward omission is (Mirenowicz & Schultz, 1996; Ungless et al., 2004). However, many studies failed to find a universal effect, with evidence of a mixture of positive and negative effects on firing rates in midbrain DA neurons (Coizet et al., 2006; Guarraci & Kapp, 1999; Joshua et al., 2008; Mantz et al., 1989; Schultz & Romo, 1987). These findings challenge the idea of a single continuous spectrum of midbrain DA firing for encoding the full range of reward and punishment prediction errors. Combined with the insight that there are computational issues with effectively coding a punishment prediction error through negative firing, simply due to the fact that there is limited range between baseline firing rates and the floor of zero (Bayer & Glimcher, 2005), it seems likely that some other mechanism must be involved in computing aversive information.

An influential finding in relation to this question showed evidence of two distinct subpopulations of dopaminergic neurons in the VTA and substantia nigra: one group does indeed respond positively to reward-predictive cues and negatively to aversive airpuff-predictive cues; however, the other, larger group responds positively to both cues (Matsumoto & Hikosaka, 2009). Neurons sensitive to only reward tended to be located in the ventromedial substantia nigra, whereas

those sensitive to both reward and punishment were largely located in the dorsolateral substantia nigra. Relatedly, Brischoux et al. (2009) found that neurons in the dorsal VTA were inhibited by foot shocks, whereas those located in the ventral VTA were excited by foot shocks. A subset of GABAergic VTA neurons sensitive to both delayed reward receipt and aversive stimuli has also been shown to inhibit reward-sensitive dopaminergic neurons (J. Y. Cohen et al., 2012). Taken together, it seems clear that dopaminergic midbrain neurons cannot be viewed as homogenous RPE signallers, and that more intricate dynamics are at play when punishing stimuli are involved.

1.3.2 In the subcortex

As discussed in 1.2.1, the vSTR has been extensively implicated in reward learning, largely due to its received projections from the midbrain. The heterogenous reactivity to reward and punishment found in midbrain DA neurons raises the question of whether and to what extent vSTR plays the same role in punishment learning. Early fMRI evidence indicated a differential striatal response following rewards versus punishments (Delgado et al., 2000), and subsequent studies have supported this with evidence that vSTR is uniquely activated in rewarding contexts under incomplete information (Palminteri et al., 2015). Activation to both reward and punishment has been found in the blood oxygen level dependent (BOLD) response for the whole striatum (Delgado et al., 2008, 2011), as opposed to specifically vSTR, and degeneration in the dorsal striatum (dSTR) has been shown to selectively impair punishment (but not reward) learning (Palminteri et al., 2012). This could indicate that whilst vSTR seems rather reward-specific, there could be dynamics dSTR that are more responsive to punishments.

Looking beyond the striatum, the amygdala has been extensively implicated in aversive learning. BOLD responses in this region have been repeatedly shown to have distinct reactivity to punishment (De Martino et al., 2006; Delgado et al., 2011; Metereau & Dreher, 2013; Yacubian et al., 2006), which reflects findings from primate single-cell recordings (Klavir et al., 2013). More causally, physical damage to the amygdala have been shown to specifically impair key as-

pects of punishment learning (Bechara et al., 1995; De Martino et al., 2010). It is worth noting, however, that whilst the amygdala seems integral to aversive learning, it does not appear to be responsible for a PPE signal (Delgado et al., 2008), and in some cases seems to be sensitive only to primary reinforcers (Delgado et al., 2011).

1.3.3 In the cortex

Looking finally at cortical regions, it seems that there are again distinctions in punishment learning. As discussed in 1.2.1, the vmPFC and OFC are strongly implicated in processing reward and value. In contrast, an entirely separate set of cortical regions seem to be sensitive to punishing outcomes. Most notably, BOLD activity in the anterior insula (aINS) has been specifically implicated in a PPE-like function (Combrisson et al., 2023; Gueguen et al., 2021; Kim et al., 2006; Seymour et al., 2004; Skvortsova et al., 2014), and damage in this region due to tumor has also selectively impaired punishment learning (Palminteri et al., 2012). Furthermore, in opposition to the apparent function of vSTR discussed above, Palminteri et al. (2015) found that the aINS was uniquely active in punishment contexts, mirroring the vSTR in reward contexts.

Another region that seems critical in punishment processing is the anterior cingulate cortext (ACC). Studies have found that subregions of the ACC have differential reactivity to rewards and punishments (Fujiwara et al., 2009; Monosov, 2017), and single-unit recording in primates has shown a detailed picture of the development of unsigned and signed PPE signals developing through communication between the amygdala and ACC (Klavir et al., 2013). Furthermore, the ACC has been implicated in the calculation of effort (Skvortsova et al., 2014), which is typically considered to be aversive.

1.3.4 In dopamine levels

A final source of insight into the reward-punishment dichotomy comes from research into individuals with disorders of dopamine. There are several well-studied clinical conditions that relate

strongly to the overactivity – e.g. Parkinson's Disease (PD) and schizophrenia – or underactivity – e.g. Tourette Syndrome (TS) – of DA pathways, which provide a useful paradigm to test the effects of both deficiencies and pharmacological treatments. For example, unmedicated PD patients often experience a symptom known as 'apathy', which is related to deficits in frontostriatal DA (Martínez-Horta et al., 2014; Pagonabarraga et al., 2015). Numerous studies have found that the reward learning of patients with this symptom is improved through DA agonist treatment, coinciding with a worsening of learning through punishment (Bódi et al., 2009; Frank et al., 2004; Kéri et al., 2010) – a finding that has also been found through baseline DA measurements in non-clincial populations (Cools et al., 2009). Additionally, reduced BOLD activity has been found in areas that showed specificity to punishment learning following the application of a DA agonist (Argyelan et al., 2018), and apathetic patients tend to display diminished vmPFC activity accompanied by reduced reward-sensitivity vs non-apathetic patients (Gilmour et al., 2024). The perceived cost of decision conflict – a variable that has been proposed to be inherently aversive (Cavanagh et al., 2014) – has also been shown to decrease in response to this type of medication (Cavanagh et al., 2017).

On the other end of the spectrum, PD patients who receive pharmacological intervention can be susceptible to impulsivity driven by enhanced DA levels. This can lead to behaviours such as compulsive shopping, addictive sexual behaviour, and pathological gambling (Voon et al., 2007), as well as greater dependency on dopaminergic drugs (Evans et al., 2005, 2006) and alcohol (Evans et al., 2005). Patients with this side-effect have been shown marked deficits in punishment-related learning such as non-rewarded inhibition of behaviour (Leplow et al., 2017), and an underestimation of punishing contingencies of stimuli (Piray et al., 2014). Similarly, TS is associated with excessive baseline DA levels, and a similar (albeit mirrored) dissociation to PD can be observed where the application of DA antagonists shows improved punishment learning but impaired reward learning, and vice-versa for untreated groups (Palminteri et al., 2009; Pessiglione et al., 2008). It should be noted, however, that this disparity in reward versus punishment learning is not found universally in TS (Schüller et al., 2020), nor in PD (Eisenegger et al., 2014; Jocham et al., 2011; Pessiglione et al., 2006; Rutledge et al., 2009). As such, although the balance of evidence indicates that DA plays somewhat of a reward-specific role overall, there

are likely nuances to this mechanism with regards to its role in punishment learning.

1.4 Scope

It seems clear that there exist a number of likely distinctions in the mechanisms underpinning reward and punishment learning, which seem to propagate up to both the subcortical and cortical levels. This presents an opportunity to use the high temporal acuity of electroencephalography (EEG) to further investigate this dichotomy in learning. In this vein, the present thesis has three main goals:

1. Investigate broad similarities and differences in the spatiotemporal characteristics of neural responses to reward and punishment.
2. Disentangle dynamics between these signals and behaviour at the subject-specific level.
3. Probe the potential of using insights from 1. and 2. to improve behavioural performance

In Chapter 2, I replicate the work of Fouragnan et al. (2015) on the spatiotemporal characterisation of post-feedback learning signals in a reversal learning task, with two key differences: I use pupillometry in conjunction with EEG instead of simultaneous EEG-fMRI; and I include a punishment manipulation to the reversal learning task. I show broad similarities in the EEG response to positive and negative feedback across reward and punishment contexts but identify clear context effects on the post-feedback phasic pupil signal. A particular dynamic between the pupil response and early EEG signals on punishment trials motivates the focus in the following chapter.

In Chapter 3, I make a case for a motivational salience signal captured in a weighted EEG signal designed to maximally discriminate between reward and punishment. I show that this signal is strongly predictive of individual differences in performance asymmetries across rewarding and punishing contexts, and I conservatively propose a role for the noradranergic locus

coeruleus in this effect.

In Chapter 4, I further examine the specific behavioural dynamics of the effects in Chapter 3 using a computational reinforcement learning model. I show that changes in exploration rates explain differences in performance across contexts but find no clear links between this behavioural parameter and either EEG or pupil measures. I also explore the links between these measures and psychometric scores relating to a motivational theory of personality, known as Reinforcement Sensitivity Theory.

In Chapter 5, I lay out a framework for applying the insights of chapters 1-4 in the context of brain-computer interaction for performance optimisation. I train a multinomial classifier to predict context and correctness of choice and test the plausibility of using this model to anticipate whether an individual would perform better in a reward or punishment context at a given moment in time.

I conclude by outlining some of the key questions raised throughout, including the multifaceted nature of salience and the need for attention on the individual level, and I explore the range of potential applications of the findings.

Chapter 2

Comparisons across reward and punishment in the two-component response to feedback

The first investigations of this thesis extrapolate certain findings in the reward learning domain to offer comparisons in the punishment domain. Specifically, I build on EEG and simultaneous EEG-fMRI work that has established a dual-component dynamic in neural responses to reward and reward-omission (Fouragnan et al., 2015, 2017, 2018; Philiastides et al., 2010). The main goals of this chapter are to probe the similarities and differences in these dual-component dynamics in rewarding decision-making contexts with the same signals in punishing contexts. In doing so, I establish the appetitive-aversive dichotomy as a medium for mapping individual differences in choice behaviour. More specifically, I identify particular aspects of the feedback response in the data that show potential differences across the two contexts, to determine promising avenues on which to focus subsequent investigations. Looking more broadly at the scope of the thesis, I look to lay the foundations for a viable means to optimise decision-making within neurofeedback paradigms. In addition to these longer-term goals, there are also more direct ideas relating to the nature of reward and punishment learning that I aim to test. Regarding the findings of prior studies, I replicate in the reward domain the spatiotemporal profiles of the early

and late EEG components as shown in previous work (Fouragnan et al., 2015, 2017; Philiastides et al., 2010), as well as the link between these components with the valence (but not magnitude) of model-derived prediction errors. Extending this, I show the extent to which these findings are consistent in the punishment domain and discuss this in relation to current understanding of the appetitive-aversive dichotomy in learning.

I also introduce an additional modality to the analysis – pupillometry – to provide insight into the role of noradrenergic pathways in the two-component paradigm. I compare phasic pupil responses across reward and punishment, and across positive and negative outcomes, and demonstrate relationships with EEG components and 'surprise' (defined here as model-derived unsigned prediction error). I use the findings to further motivate the target of investigation in Chapter 3 and beyond, as well as to understand the nature of any effects revealed in the punishment condition.

2.1 Background

The ability to integrate outcomes from actions into a coherent model of environmental contingencies is a vital part of the success of an organism. Much work has been done on the characterisation of the reward learning process, mapping the various mechanisms and neural pathways involved in feedback processing and value updating. Recent insights into the temporal dynamics of feedback processing have been explored with EEG (Philiastides et al., 2010) and simultaneous EEG-fMRI (Carvalheiro & Philiastides, 2023; Fouragnan et al., 2015, 2017, 2018). A key insight from these studies shows two key spatiotemporal components that distinguish responses to positive and negative feedback. However, thus far this dynamic has almost exclusively been explored in the reward domain, and it is not clear whether distinctions exist when learning in aversive contexts.

2.1.1 The two-component feedback response

A multitude of evidence has demonstrated the role of dopaminergic projections from the midbrain through the vSTR (Delgado et al., 2000; McClure et al., 2003; O'Doherty, 2004; O'Doherty et al., 2002, 2003; O'Doherty et al., 2004; Tobler et al., 2007, 2008), vmPFC (Blair et al., 2006; Daw et al., 2006; Gläscher et al., 2009; Hampton et al., 2006) and OFC (Balleine et al., 2011; Elliott et al., 2008; Gourley et al., 2016) in the processing of reward. More recently, insights from the combination of temporally precise EEG and spatially precise fMRI have shown that these regions are differentially activated in response to reward versus reward omission at around 300-320ms post feedback (Fouragnan et al., 2015, 2017). This timing is in line with prior work using only EEG that found a spatiotemporally similar post-feedback signal (Philiastides et al., 2010). This component was found to reflect an RPE valence effect rather than a magnitude effect, meaning that it tended to encode binary 'good versus bad' information about choice outcomes rather than more fine-grained information about the degree to which expectations were violated (Fouragnan et al., 2017, 2018). This signal was broadly consistent with dopaminergic reward pathways in both structure and function, and was shown to predict the degree of value updating in response to RPE (Fouragnan et al., 2017).

Notably, this methodology has also revealed an earlier component to feedback processing, emerging at the 220-230ms mark (Philiastides et al., 2010), that is associated with an initial salience and arousal response involving the ACC and aINS (Fouragnan et al., 2015). This signal was found to be primarily driven by negative outcomes, and was shown downregulate reward responses in the late component, indicating an interactive mechanism between the two distributed signals (Fouragnan et al., 2015). Also of interest is a surprise-specific signal that contained information about the unsigned RPE value. This third signal was temporally similar to the late valence-sensitive signal at 320ms, and associated more with areas involved in alertness and exploration such as the midcingulate cortex, thalamus and dorsolateral prefrontal cortex (Fouragnan et al., 2017, 2018) (Fouragnan et al., 2017, 2018).

2.1.2 Influences of noradrenaline and the locus coeruleus

More recently, a similarly designed experiment using high-resolution 7T EEG-fMRI found unique activation in the locus coeruleus (LC) associated with the early component (Carvalheiro & Philiastides, 2023). The LC is an influential brainstem nucleus that has been implicated in a wide range of processes via noradranergic projections throughout the cortex and subcortex, broadly regulating attention, arousal, and wakefulness (Sara & Bouret, 2012). In relation to the role of the two components in question, the LC has been repeatedly implicated in surprise relating to decision outcomes (de Gee et al., 2021; Filipowicz et al., 2020; Lavin et al., 2014; Preuschoff et al., 2011), and also in subsequent projections to the ACC. This is a region implicated in error detection (Carter et al., 1998), and is an important part of the early component network (Fouragnan et al., 2015), suggesting that this pathway could be involved in the early salience component of learning.

Interestingly, Carvalheiro and Philiastides (2023) did not find activation in the LC in association with the early component in a punishment learning condition, suggesting a potential differential role of this nucleus across learning contexts. Direct evidence for this notion is not abundant, but there is some indication that the LC is particularly sensitive to aversive losses (Pulcu & Browning, 2017), and there is some evidence that pupil dilation is greater following punishment than following reward (Breton-Provencher et al., 2022). It is generally accepted that noradrenergic activity from LC is the primary driver of pupil dilation from both a theoretical (Larsen & Waters, 2018; Mathôt, 2018) and experimental perspectives (de Gee et al., 2017; Joshi et al., 2016; Reimer et al., 2016). As such, pupil dilation is regularly used as a proxy for LC arousal (de Gee et al., 2014; Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011) and provides a convenient and informative measure to gain insight into this system.

2.1.3 Aims and hypotheses

Thus far, the two-component response has almost exclusively been examined in a reward context, and never alongside pupillometry. However, many of the key regions implicated in the two components are differentially associated with reward and punishment. The vmPFC, OFC, and vSTR are all heavily implicated in dopaminergic reward pathways, and concurrently with the late component of the post-feedback response (Fouragnan et al., 2015, 2017). On the other hand, areas associated with the early component have been shown to play roles in punishment learning, including the ACC (Fujiwara et al., 2009; Klavir et al., 2013; Monosov, 2017) and aINS (Combrisson et al., 2023; Gueguen et al., 2021; Kim et al., 2006; Palminteri et al., 2012; Seymour et al., 2004; Skvortsova et al., 2014).

I therefore aim to explore the spatiotemporal dynamics of the two-component feedback response using EEG and pupillometry in a reversal learning task, based on the methodology of (Fouragnan et al., 2015). Critically, I add a punishment manipulation, whereby half of the decision-making blocks have the possibility of obtaining reward or nothing, and the other half punishment or nothing. In line with previous findings, I expect to be able to discriminate between positive and negative valence outcomes from EEG signals in both the reward and punishment contexts, and to find distinct spatiotemporal profiles in line with the early and late components. Based on links between pupil and surprise, and evidence that pupil dilation is higher in response to punishment, I predict that the highest phasic pupil dilation will occur in response to punishing losses, and the lowest will be in response to rewarding wins. I also expect pupil signals across contexts to predict trial-by-trial surprise, as quantified by the absolute prediction error from a computational model. Finally, based on the relationships between arousal/error related areas such as the ACC and noradrenergic activity, I predict that the early EEG salience component will be uniquely related to the pupil response rather than the late component.

2.2 Materials and methods

2.2.1 Participants

Data was collected from 33 participants, (18 female, 15 male) with ages ranging from 18-41 (mean = 23.30 years, SD = 5.29). We excluded 6 participants from pupil analyses due to excessive missing data in the pupil recording, defined as >47% of samples missing for reward blocks and >51% for punishment blocks (based on one standard deviation from the mean). We also excluded one participant from EEG analyses due to excessive movement artifacts in the EEG signal. For combined EEG and pupil analyses this left 26 remaining participants with usable data for both EEG and pupillometry (14 female, 12 male, mean age = 23.15 years, SD age = 5.87). All participants were recruited through the University of Glasgow Subject Pool, were right-handed and had uncorrected vision. The study was approved by the College of Science and Engineering Ethics Committee at the University of Glasgow and informed consent was obtained from all participants.

2.2.2 Task and Procedure

The study used a simple probabilistic reversal learning paradigm, based largely on the design used in Fouragnan et al. (2015) with the addition of a reward-punishment manipulation (Figure 2.1). The main task consisted of 6 blocks of 80 trials, alternating between rewarding and punishing contexts. We decided to always start with a rewarding block rather than counterbalancing the order, as we wanted to maximise the feeling of earning money and then subsequently losing it to increase the subjective difference between the contexts. Each trial began with a jittered 2-3s fixation period before the decision phase, after which participants had to choose between two symbols with mirrored probabilities (70% and 30%) of a positive or negative outcome, and the same pair of symbols was used in every trial throughout the whole task. Outcomes for the symbols on a given trial were independent of each other, meaning that on a given trial it is possible for both symbols to yield the same outcome. If the participant did not respond within 1.25s

of the decision phase, they were informed that they would lose £0.50 and the message 'Please Respond Faster' was displayed. The decision phase was followed by a jittered delay period lasting between 1.5-2s, before a 0.75 display of the outcome symbol. Participants indicated their choice via left or right button press on a specialised response box (Cedrus RB-740 Response Pad, Cedrus, USA). We provided positive and negative outcomes by displaying different arrows in the centre of the screen. Specifically, in reward blocks, we used upward and neutral arrows to respectively provide positive and negative feedback, and in punishment blocks we used neutral and downwards arrows to respectively provide positive and negative feedback. To minimise pupil fluctuations due to visual properties, all arrows and fixation symbols were normalised for perceptual load and luminance using consistent pixel count and geometric structures, and transitions between fixation, decision and outcome screens were kept as subtle as possible. If a participant did not respond in time, a brief 'too slow' message appeared on the screen before the next trial, which participants were informed would carry a penalty of -£0.50 to disincentivise missed trials.

Participants aimed to ascertain once the symbol with the 70% probability of success, participants could select it repeatedly to maximise their monetary payout. However, the outcome contingencies of the symbols would switch approximately every 20 trials (+/-2), such that the 'good' option would become the 'bad' option and vice versa. The participants were told these switches would occur 'every so often' throughout each block, but both the outcome contingencies and reversal frequencies were not known to the participant. Therefore, following unexpected outcomes, participants had to infer whether this was due to inherent stochasticity in the design or a change in the underlying contingencies. This task design was chosen to provide a simple reinforcement learning paradigm with clear truth labels for correct choice and a steady degree of volatility to allow for a variety of decision-making strategies, as well as to provide consistency for comparison with similar studies such as Fouragnan et al. (2015).

Participants were paid a baseline of £10 for participation and could additionally earn between £5-20 based on task performance. This was implemented by adding £0.25 to the total reward for each 'win' outcome and subtracting £0.25 for each 'lose' outcome, while no-win and no-

loss outcomes yielded £0. The total amount won or lost was displayed after each block to keep the participant engaged with the consequences of the outcome symbols, and the experimenter reminded participants whether the upcoming block was rewarding or punishing. The average total reward was approximately £20 for a 2.5-hour session (including baseline).

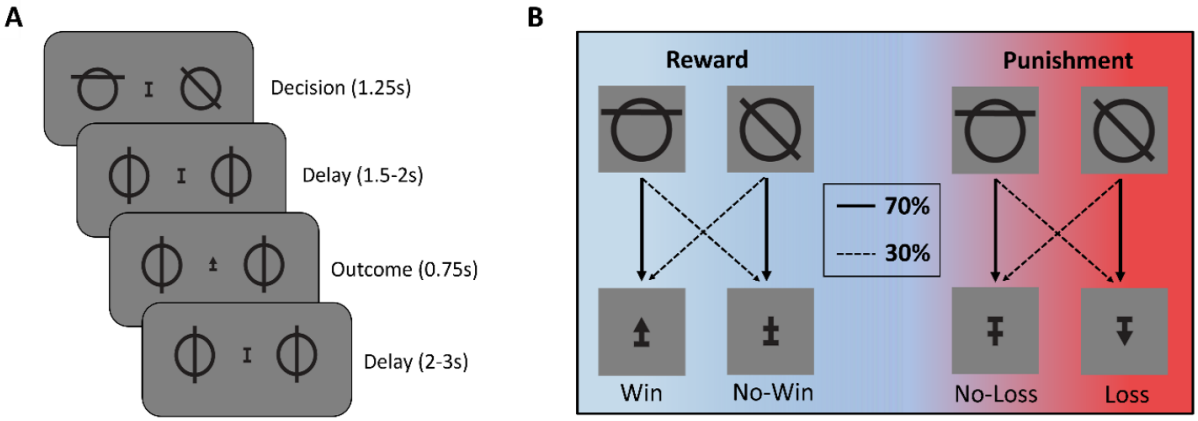

Figure 2.1: *Depiction of probabilistic reversal learning task.* **A)** Stages of a single trial. Participants choose one of two symbols with a button press for a maximum of 1.25s. If no choice was provided in this time, the message 'Please respond faster' was displayed. After a short delay, the outcome is presented in the centre of the screen. **B)** Outcome symbols and contingencies. Participants always choose between the same two symbols throughout the entire task. For a given trial, one of these symbols has a 70% chance of a positive outcome, while the other has a 30% chance. In the appetitive condition, a positive outcome is the 'win' symbol and a negative outcome is the 'no-win' symbol; in the aversive condition, a positive outcome is the 'no-loss' symbol and a negative outcome is the 'loss' symbol. These contingencies switch approximately every 20 trials during an 80-trial block.

Before attending, all participants completed a shorter online practice version of the task, which was implemented using Pavlovia, an online version of PsychoPy (Peirce et al., 2019). A minimum of 60% accuracy over 96 trials was required for participation.

2.2.3 EEG data collection and analysis

We sampled data at 1000Hz from a 64-channel EEG cap (BrainCap, BrainProducts, Germany) and accompanying amplifiers (BrainAmp, BrainProducts, Germany), using the Brain Vision Recorder software (BVR, Version 1.2.1 BrainProducts, Germany). The Ag/AgCl electrodes were positioned according to the international 10-20 system and all electrodes referenced to the

left mastoid, with a ground electrode positioned on the left mandible. All electrode impedances were kept below 20 $k\Omega$ using conductive gel. The amplifiers had a built-in hardware bandpass filter of 0.0016Hz-1000Hz. We applied a band-pass filtered to the data using a 0.5 Hz Butterworth high-pass filter to remove slow direct current drifts and a 40Hz Butterworth low-pass filter to remove higher frequencies of no interest. To remove eye-blink and -movement artifacts, participants performed an eye calibration task before the main experiment during which they were instructed to blink continuously for several seconds, and then track a cross moving horizontally and vertically whilst keeping their head still. We recorded the timing of these events and used principal component analysis (L. Parra et al., 2003) to identify linear components associated with eye-blinks and -movements, which we subsequently projected out of the broadband EEG data collected during the main task.

For each participant individually, we employed a multivariate discrimination analysis on the EEG signal, whereby an optimal set of electrode weights was estimated using a logistic regression model to maximally discriminate between positive and negative outcome trials separately for the reward and punishment contexts. This analysis was designed to replicate the two-component findings of Philiastides et al. (2010) and Fouragnan et al. (2017), with the extension of adding a punishment condition for comparison. In one analysis 'Win' trials were discriminated against 'No-Win' trials, and in the other analysis 'No-Loss' trials were discriminated against 'Loss' trials, employing a method based on L. C. Parra et al. (2005) and Sajda et al. (2007). Though positive outcomes had higher trial counts, the numerical discrepancy for positive versus negative trials was <15% for all participants and <10% for the vast majority, with the largest difference being 258 positive outcomes versus 222 negative outcomes. We applied a sliding 60ms window in 10ms increments from 100ms pre-feedback to 800ms post-feedback, and within each window data were used to train a logistic regression model, where positive outcomes (i.e. wins and no-losses) were arbitrarily mapped to positive values and negative outcomes (i.e. losses and no-wins) to negative values relative to the discriminating hyperplane. Each electrode represented one predictor variable in the model, resulting in 64 weightings w that optimally predicted context depending on the analysis. When applied to the EEG signal X, the resulting weighted amplitudes could be summed across electrodes to produce a single scalar

component amplitude Y, representing linear distance from the discriminating hyperplane:

$$Y(t) = w^T \cdot X(t) \tag{2.1}$$

To visualise the spatial representation of the resulting discriminating components, we calculated a forward model which captures the relative contribution of each sensor to the discrimination (note all topographies shown in the paper depict this forward model):

$$a = \frac{X \cdot Y}{Y^T \cdot Y} \tag{2.2}$$

Discriminator performance was quantified using the area under a receiver operating characteristic curve (AUROC) using a leave-one-out cross-validation approach. To assess the significance of these AUROC values across time, we used a permutation approach whereby a null AUROC distribution was derived from 1000 permutations of the same classifier with randomly shuffled labels for reward and punishment, and a significance threshold was set at the 99th percentile ($p < .01$). We identified early AUROC peaks for each participant separately, representing the point of individual maximum AUROC value between 170-270ms, which corresponds to the early salience-related signal outlined in the dual-component theory of feedback processing, encompassing +/-50ms from previous findings (Fouragnan et al., 2015; Philiastides et al., 2010). The 170ms also coincided with the beginning of the temporal window that exceeded the 99th percentile in the permutation test (see section 2.3.2). To avoid our early salience peaks being selected on the upward slope of a subsequent value-related peak (as found in (Fouragnan et al., 2015)), we only considered for peak selection time-points where the AUROC value was greater than that of the two preceding and two following time-points – in other words, a local maximum. Similarly, late AUROC peaks were selected using the same procedure over the subsequent 270-420ms window. This window was extended slightly longer than the early window due to the long duration of high discriminability found, in order to make sure that all peaks were true peaks rather than values on an upward slope. These subject-specific peaks were then used to

extract the corresponding Y values for use in subsequent analyses.

2.2.4 Pupillometry data collection and analysis

Pupil diameter and gaze x/y coordinates were recorded at 40hz using a screen-based eye-tracker (Tobii Pro X3-120, Tobii, Sweden). All stimuli were made with equivalent pixel counts to ensure equiluminance and were designed to minimise shape change between screens to minimise light-related pupil fluctuations.

Missing pupil data due to blinks was addressed by linearly interpolating samples within +/-100ms of blink events. We then applied a bandpass filter of 0.01-4Hz, z-scored the resulting data, and epoched each trial to -500ms/+2000ms around feedback, baseline corrected by averaging across the 500ms pre-feedback period for each subject and subtracting the result from all values. Outlier trials for each subject were identified as >3 standard deviations from the mean (averaged across trials and samples over the epoched window), or <1.5% of mean variance (variance calculated across time and averaged across trials). The latter was specifically to deal with occasional flat lines in pupil response due to errors at data collection. All outlier trials were then removed before any further analysis, averaging at 9.58 trials removed per participant.

To determine a difference in pupil response between contexts, we used a non-parametric approach based on the single-sensor time-series analysis outline by Maris and Oostenveld (2007). An independent t-test between reward and punishment contexts was conducted for each time-point across subjects, and with the non-parametric test statistic being the sum of t-values for the largest cluster of consecutive significant results ($p < .05$), which gave the window 0-1100ms post feedback. We then compared the resulting test statistic ($df = 26$, $\Sigma t = 180.01$) to the 99th percentile of 10000 permutations of test statistics from randomly allocated groups ($df = 26$, $\Sigma t = 6.30$) to determine statistical significance (Figure 2.5A).

2.2.5 EEG-informed pupil GLM analysis

To investigate the link between the pupil response and the EEG-derived discrimination components, a Generalised Linear Model (GLM) was used to estimate the full z-scored pupil time series using parametrically modulated boxcar regressors convolved with a pupil response function (de Gee et al., 2014; Denison et al., 2020). The boxcar regressors were initialised by creating a vector of zeros of the same length as the pupil time series for each predictor variable, indexing 500ms before the onset time for each corresponding regressor event within the time series, and modulating the amplitude at these indexes based on the regressor used (Figure 2.2, left).

Once the boxcar regressors were created, they were convolved with a pupil response function (PuRF) that was individualised for each subject such that it reflected their mean event-locked response parameters (Figure 2.2, centre), with the full range being 500ms pre-event to 3000ms post-event. The equation for producing the subject-specific pupil response functions (PuRFs) was obtained originally from Hoeks and Levelt (1993), where t and w are the length and width of the response function, and tmax is the point of peak amplitude:

$$h(t) = t^{w} e^{-wt/t_{\max}} \tag{2.3}$$

This was slightly modified such that a two-element vector input for w and t_{max} would produce a double gamma function to capture the subsequent dip in pupil diameter, with an additional parameter dip determining the relative amplitude of the dip to the peak.

To obtain an optimal PuRF for each participant, a grid-search approach was taken to test a series of values for w, t_{max}, and dip within the double gamma function and determine which set produced the lowest mean-squared error. Each resulting function was then visually inspected against the actual mean pupil response for prima facie goodness-of-fit, and manual adjustments were made if required.

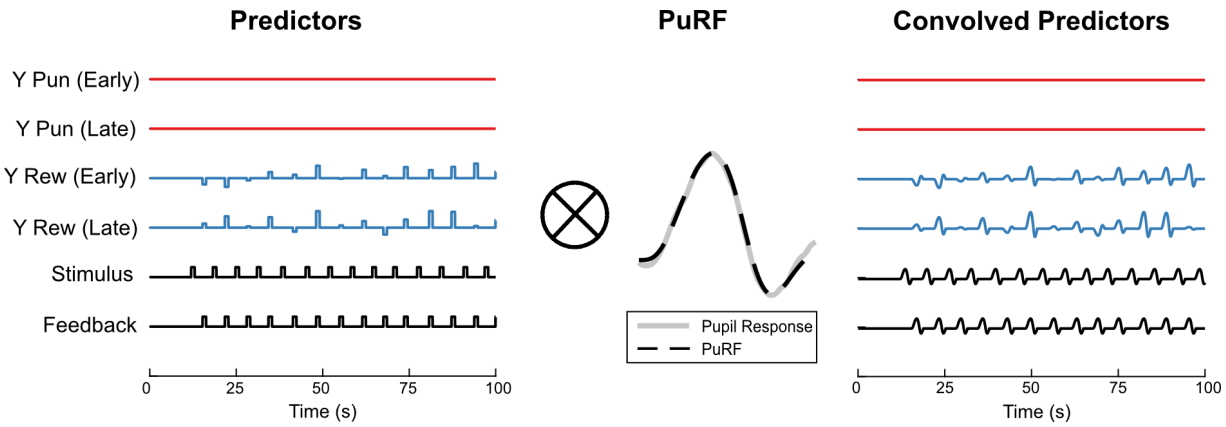

Figure 2.2: *Regressor structure for EEG-informed Pupil GLM.* **Left)**: Pupil-matched time-series of unconvolved predictor variables. X-axis indicates the first 100 seconds of the time-series from the beginning of the task. Y-axis indicates the names of the predictor variables (Top four show feedback-locked indexes modulated by early and late EEG component amplitudes for reward and punishment; bottom two show unmodulated indexes for the time of stimulus onset and feedback onset). **Centre)**: Example from one participant of the double gamma function of the pupil response function (dotted line) fitted to the feedback-locked phasic pupil response averaged over all trials (grey line). **Right)**: Predictor variables convolved with the pupil response function.

The four main regressor events of interest were the trial-by-trial early and late discrimination peaks from the separate reward and punishment models. For each of these EEG discrimination regressors, the regressor amplitude at the onset indexes were set to the corresponding discrimination amplitude (Y). Additionally, there were two control regressors included that reflected the unmodulated phasic pupil response to stimulus onset and feedback onset. For the control regressors, amplitudes were simply set to one at the onset index. Finally, a regressor was added containing the indices for every trial removed for invalid pupil data. A depiction of the final post-convolution regressors for the first 100 seconds of the task can be seen in the right-hand panel of Figure 2.2, albeit without the outlier regressor on display. Note that as the top two rows were specific to punishment trials, and the first block was a reward block, the amplitudes for these predictors remain zero.

2.2.6 Computational Modelling

We trained a model-free reinforcement learning algorithm on trial-by-trial choices for each subject. This functions by estimating for trial t prediction error δ_t from the difference in expected

value V_t and received reward r_t of choice i:

$$\delta_t = r_t - V_t(i) \tag{2.4}$$

This principle is then used to update expected value by weighting this prediction error with a learning rate parameter α. This parameter lies between 0 and 1, with a greater learning rate implying a faster updating of value expectations based on recent evidence:

$$V_{t+1}(i) = V_t(i) + \alpha \cdot \delta_t \tag{2.5}$$

To account for fluctuations in perceived environmental volatility, the learning rate parameter was also dynamically updated via the slope of the smoothed prediction error m as outlined in (Krugel et al., 2009):

$$\alpha(t) = \alpha(t-1) + f(m(t)) \cdot (1 - \alpha(t-1)), \quad \text{if} m > 0 \tag{2.6}$$

$$\alpha(t) = \alpha(t-1) + f(m(t)) \cdot \alpha(t-1), \quad \text{if} m < 0 \tag{2.7}$$

Here, $f(m(t))$ is a double sigmoid function that transforms such that $0 < m < 1$, which then scales the trial-wise dynamic learning rate. This function recruits an additional free parameter, which reduces the degree to which alpha is modulated as it increases.

Finally, choice probability for a given choice was derived according to a softmax decision rule, which adds an additional parameter for inverse temperature B (temperature being the degree of stochasticity in decisions, represented by the slope of the sigmoid):

$$p_i(t) = \frac{e^{B \cdot v_i(t)}}{\sum_{j=1}^{n} e^{B \cdot v_j(t)}} \tag{2.8}$$

2.3 Results

2.3.1 Choice behaviour is similar across reward and punishment at the group level

Subjects displayed a high level of accuracy across both conditions of the task, choosing the high-value symbol on average 70% of the time in the reward condition and 69% of the time in the punishment condition. At the group level, paired t-tests revealed no clear behavioural differences in accuracy ($df = 32$, $t = 1.20$, $p = .24$) or reaction time ($df = 32$, $t = -0.77$, $p = .45$) between the rewarding and punishing contexts (Figure 2.3A).

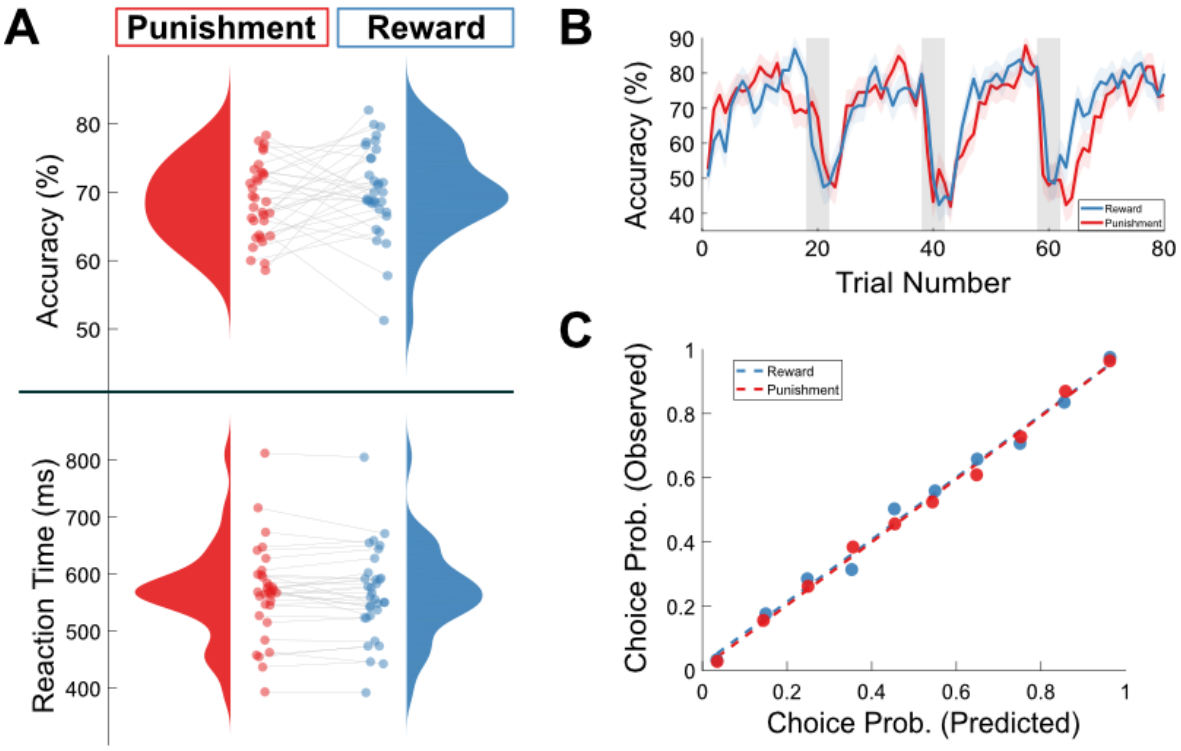

Figure 2.3: *Overview of behavioural and model results.* **A)** Comparison of choice accuracy (upper panel - percentage chosen for high-value symbol) and reaction time (lower panel – time from symbol presentation to choice in milliseconds) across reward and punishment conditions. Blue (right side of each plot) scatters show individual subject data points for reward context, while red (left side of each plot) show equivalent data for punishment context. **B)** Percentage of high-probability symbol chosen for each trial across a block, averaged across blocks and participants separately for reward (blue) and punishment (red) contexts. Shaded areas indicate trials where a reversal can occur. and pupil data from positive outcomes, and right depicts the same for negative outcomes. **C)** Reinforcement learning model performance for reward (blue) and punishment (red) trials. X-axis represents model-derived choice probabilities for a given symbol binned into deciles for each subject and averaged across subjects. Y-axis represents proportion of corresponding trials in each bin where that symbol was chosen, averaged across subjects.

Participants on average displayed the typical learning patterns we expect in a reversal learning task, with choice accuracy drastically falling following a reversal before climbing back up as the new contingencies are realised (Figure 2.3B). Observed subject choices closely matched reinforcement learning model predictions for both reward and punishment trials (Figure 2.3C; $p < .001$).

2.3.2 Broad spatiotemporal similarities across contexts in EEG discrimination

The primary analysis to be replicated from the prior dual-component literature (Fouragnan et al., 2015; Philiastides et al., 2010) is the multivariate classification of positive versus negative outcomes in a binary choice task. Specifically, for each subject and each context separately, I estimated trial-by-trial discrimination amplitudes along the feedback-locked EEG time series and quantified the discrimination performance across this period using an ROC analysis. Separability between positive and negative trials was significantly above the 0.58 significance threshold between 170-730ms in the reward context and 170-790ms in the punishment context when averaged across participants, indicating that in both cases there was a high degree of linear separability in post-feedback neural activity depending on the valence of the outcome (Figure 2.4B).

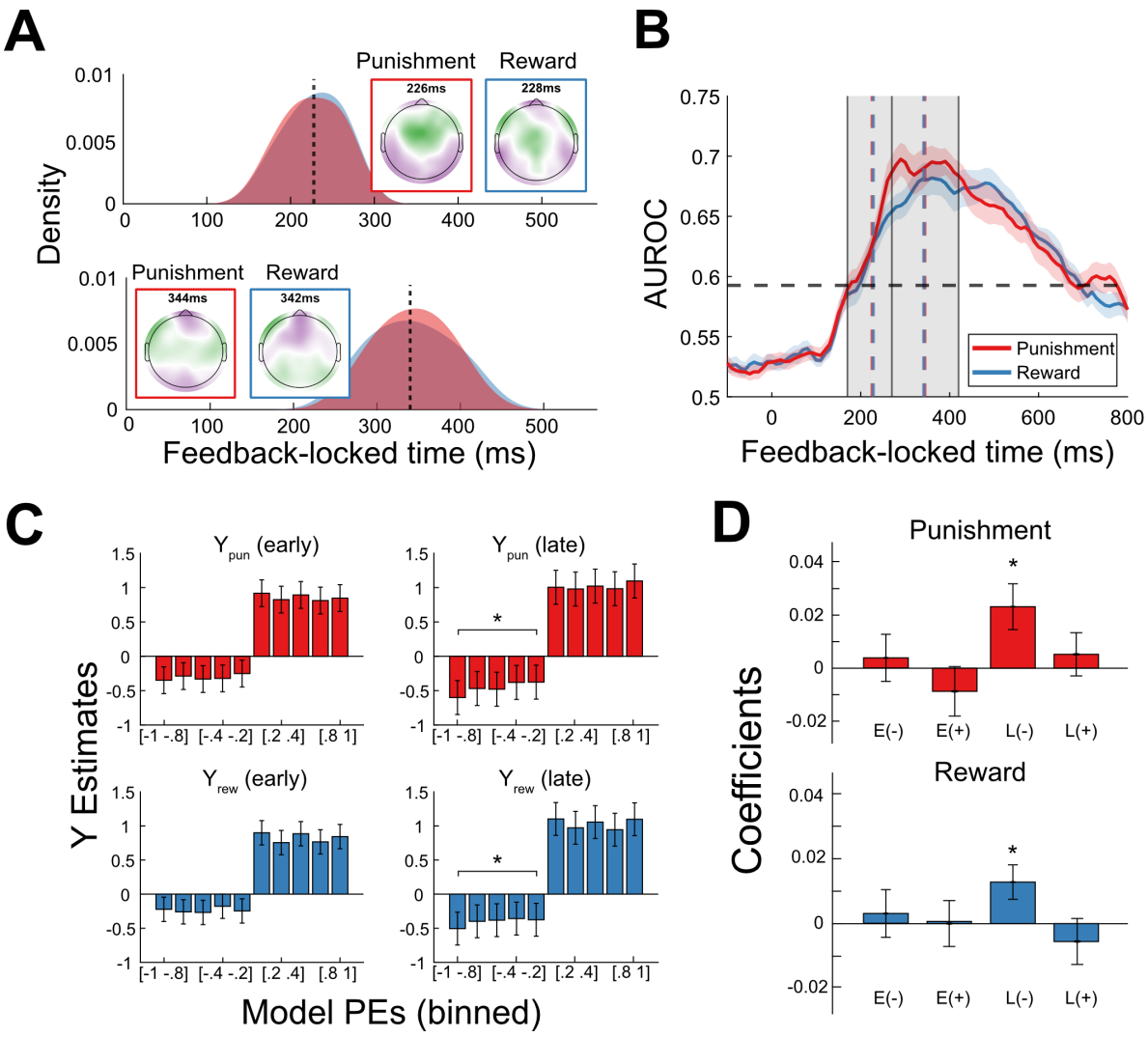

Figure 2.4: *EEG discrimination analysis results.* **A)** Distributions of subject-specific AUROC peak selections for early (top) and late (bottom) components. Red indicates punishment blocks and blue indicates reward blocks. Dotted lines show mean latency of peak AUROC averaged across subjects. Topographies (insets) show forward model of the discrimination component magnitudes. Green maps to positive Y values indicative of positive outcomes; purple mapped to negative values indicative of negative outcomes. **B)** AUROC (area under receiver operating characteristic curve) values for two separate classification models. Y-axis depicts mean feedback-locked area under AUROC for logistic regression averaged across subjects. X-axis depicts time from feedback onset in milliseconds. Shaded error bar represents standard error of the mean across subjects. Grey shaded area reflects window for peak selection, and dotted vertical lines depict average peak onset for punishment (loss vs no-loss, red) and negative (win vs no-win, blue) outcomes. Horizontal dashed line depicts p=0.01 permuted significance threshold averaged across subjects and across the two classification models. **C)** Trial-by-trial Y values from each of the punishment (top, red) and reward (bottom, blue) classifiers separated by early (left) and late (right) peaks. Y values are sorted into 10 bins based on prediction error for the corresponding trial (-1 to +1). Error bars reflect 95% confidence intervals. Significance start reflects the bins matching the significant analysis in **D**. **D)** Model coefficients across participants for punishment (top, red) and negative (bottom, blue) for multiple regressions predicting trial-wise prediction error from the early (E) and late (L) Y amplitudes. Regressions were run separately for positive and negative prediction error trials. Error bars show 95% confidence intervals. Coefficients were significantly different from zero for the late EEG component in negative-outcome trials for both punishment (top) and reward (bottom) contexts (p<.05).

AUROC performance to identify local peaks within early and late temporal windows informed by prior studies (Fouragnan et al., 2015, 2017; Philiastides et al., 2010). Each subject had a peak identified in each window, and the temporal onset of these were averaged across subjects (Figure 2.4A, density plots). Averaging the peak across subjects for the early time window revealed a mean early component peak at 226ms for the punishment context and 228ms for the reward context (Figure 2.4A, upper topographies). The same procedure yielded 344ms and 342ms peaks in the late window for reward and punishment contexts respectively (Figure 2.4A, lower topographies). The emergent scalp topographies in the early period were highly similar both temporally to those found in previous studies (Fouragnan et al., 2015; Philiastides et al., 2010), occurring within 10ms for both reward and punishment. Additionally, both contexts showed a fronto-central topographical pattern driven by positive outcomes (mapped to green colour in Figure 2.4A) that is also notably similar to the same prior studies. For the early peak, there are no obvious differences between the reward and punishment contexts in the EEG discrimination.

The timing of the late component is also highly similar across contexts, though occurring slightly later than the previous studies which report peaks in the 300-310 ms range (Fouragnan et al., 2015, 2017; Philiastides et al., 2010). When compared to these prior studies in the reward domain, the EEG discrimination in the reward context displayed an almost identical spatial shift from a positive fronto-central cluster in the early component to a negative one (mapped to green colour in Figure 2.4A). The spatial shift was particularly analogous to (Philiastides et al., 2010), which notably did not suffer from the higher impedance of the simultaneous EEG-fMRI caps used in the other studies (Fouragnan et al., 2015, 2017), and thus had a clearer topographical profile. The punishment context, however, deviated from this spatial pattern in the late component, in that this sign-flip from early-to-late in the frontocentral region was not present. Though this may be indicative of some possible difference in the relative weighting of neural response to positive and negative feedback across contexts, it is important to note that the abstract nature of the weighted signal in the discrimination output prevent any more specific interpretations. As such, the main takeaways from the raw output of the discrimination analysis should be the temporal similarities of the components with previous work and between contexts,

as well as general spatial similarities that drive the discrimination – particularly the prevalence of fronto-central electrodes around FCz.

2.3.3 EEG discrimination amplitudes track valence but not surprise

The trial-wise peak Y amplitudes displayed similar broad valence effects as those observed in Fouragnan et al. (2017): negative PEs mapped onto negative Ys, positive PEs mapped onto positive Ys, but there were no clear parametric trends of PE size tracking to Y values beyond this. Figure 2.4C shows early and late peak Y amplitudes binned into 10 groups based on the model-derived PE size for the corresponding trial, separated by context (upper = punishment, lower = reward) and timing (left = early, right = late). Bins reflected 10 intervals of width 0.2, ranging from [-1 -.8] to [.8 1], such that the leftmost 5 bars all represented negative outcome trials with surprise decreasing from left to right, and the rightmost 5 bars represented positive outcome trials with surprise increasing from left to right.

To more formally investigate the hypothesis that components from the multivariate discrimination analysis do not carry information about the level of surprise associated with an outcome, I ran four multiple linear regressions for each subject predicting trial-wise PEs derived from the computational behavioural model from the early and late discrimination amplitudes from the multivariate EEG analysis, separately for reward and punishment (figure 2.4D). In essence, this analysis roughly corresponded to the slope of the first five bars and last five bars in each of the four subplots in Figure 2.4C; if a true trend exists between signed prediction error and Y, subject-specific linear model estimates should systematically skew positive or negative.

To test whether the resulting beta coefficients from the linear models differed significantly from zero, a one-sample t-test was applied to the early and late predictors in each of the four categories. Contrary to other findings in the literature, significant effects were found for negative outcomes in the punishment late signal ($t(31) = 2.55$, $p = .016$) and the reward late signal ($t(31) = 2.29$, $p = .029$). No other comparisons for positive outcomes or for the early signals showed any effects ($p > .38$).

2.3.4 Post-feedback pupil transients are modulated by context and outcome

In the post-feedback phasic pupil response – baseline corrected and averaged across subjects - a parametric modulation can be observed whereby amplitude is higher for negative outcomes than positive outcomes, and for the punishment context versus the reward context (Figure 2.5A). To more formally compare the latter effect of greater pupil dilation in the punishment context, a non-parametric time series analysis was employed to establish a window of significant difference between the two aggregated signals (see 2.2.4). This produced a window between 0-1100ms post-feedback that showed significant context-driven differences (Figure 2.5A).

2.3.5 Phasic pupil dilation predicts surprise

Leveraging the window established from the non-parametric analysis, a scalar phasic pupil amplitude was calculated for each trial by averaging across this window. This variable was predicted by trial-by-trial surprise in a linear regression model for each subject separately, and for reward and punishment separately (Figure 2.5B). The absolute prediction error derived from the behavioural model was used to reflect surprise. As hypothesised, a one-sample t-test revealed that beta values from the subject-specific linear regressions were significantly higher than zero for both the reward ($t(25) = 3.53$, $p < .01$) and punishment contexts ($t(25) = 2.39$, $p = .026$). This indicated that in both cases, higher phasic pupil dilation – linked to noradrenergic arousal – predicted a greater level of subjective surprise.

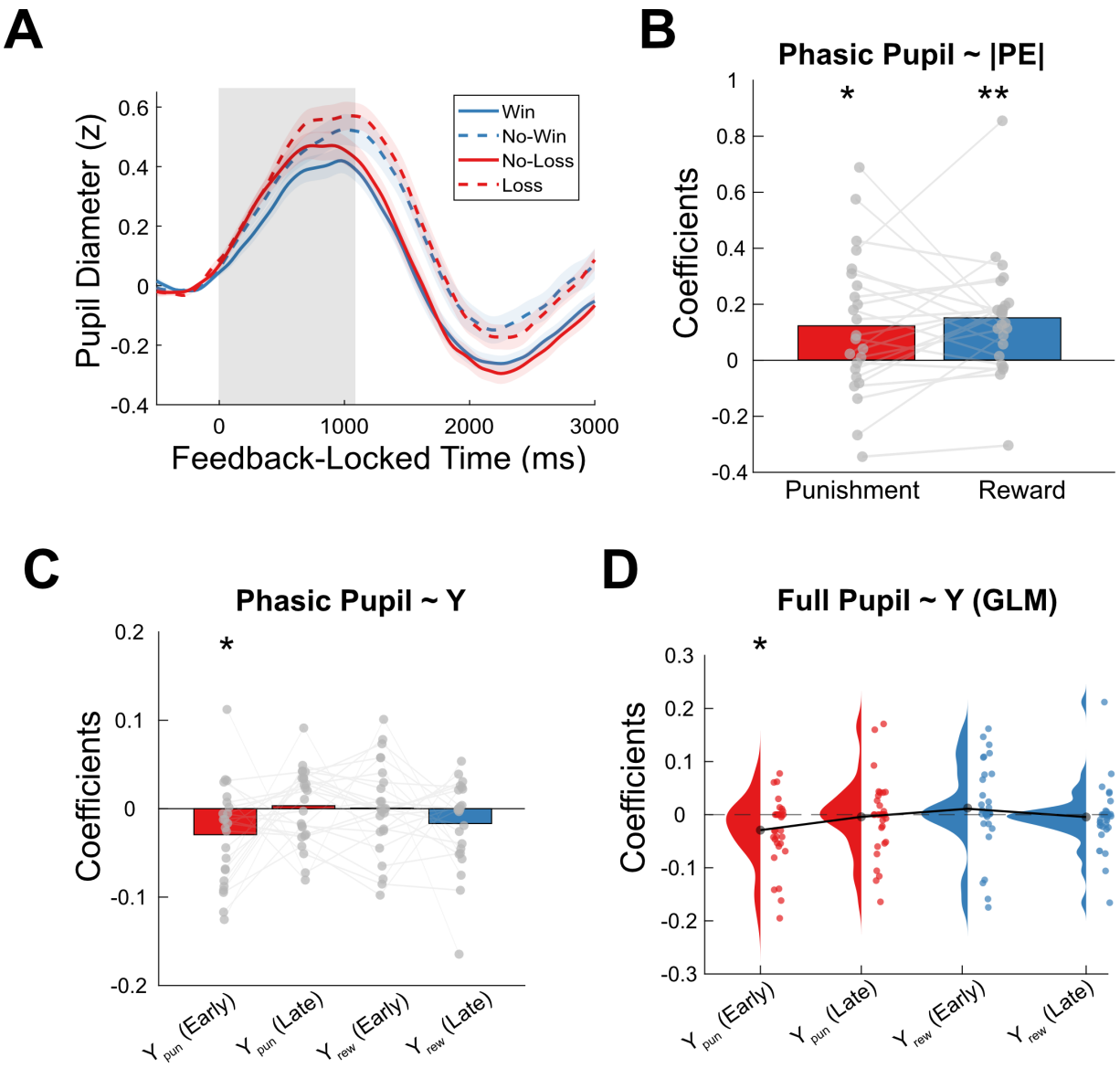

Figure 2.5: *Pupillometry results.* **A)** Post-feedback pupil response averaged across trials and participants, separated by positive (solid line) and negative (dotted line) outcomes. Red indicates punishment condition and blue indicates reward condition. X-axis represents time from feedback onset in milliseconds and Y represents z-scored pupil diameter. Shaded area indicates window of significant difference between pupil response in reward vs punishment conditions averaged across all trials, obtained from non-parametric cluster test. **B)** Subject-specific beta coefficients from two linear regressions predicting trial-by-trial phasic pupil dilation from absolute prediction error values, run separately on punishment (left, red) and reward (right, blue) trials. Coefficients were significantly different from zero in punishment ($p<.05$) and reward ($p<.01$) contexts. **C)** Subject-specific beta coefficients from two multiple linear regressions predicting trial-by-trial phasic pupil dilation from two predictors corresponding to trial-by-trial discrimination amplitudes (Y) at the moment of the early and late peaks, run separately on punishment (left two bars, red) and reward (right two bars, blue) trials. Coefficients were significantly different from zero for early punishment predictor ($p<.05$). **D)** Subject-specific beta coefficients for the first four predictors (excluding nuisance regressors) from the pupil-informed GLM analysis. Coefficients were derived from the same Y values as **C**, and were significantly different from zero for early punishment predictor ($p<.05$).

2.3.6 Phasic pupil dilation predicts early punishment EEG discrimination amplitudes

As in 2.3.3, I used early and late EEG discrimination components as the two predictors in a multiple linear regression separately for rewards and punishments and for each participant, this time predicting phasic pupil response (Figure 2.5C). Again, one-sample t-tests were used to test for significant deviations from zero in the regression coefficients across subjects. In line with expectations, the early component in the punishment context was a significant predictor ($t(25) = -2.10$, $p = .047$), but the late component was not ($t(25) = 0.39$, $p = .703$), in line with the hypothesis that the early component reflects a salience response. The negative direction of prediction was also in line with expectations, given that negative Y values were mapped to negative values, whereas the associated higher pupil amplitude increased in the positive direction. Interestingly, in the reward context, neither the late component ($t(25) = -1.68$, $p = .105$) nor the early component ($t(25) = -0.06$, $p = .951$) significantly predicted phasic pupil amplitude, suggesting that noradrenergic activity may have a unique relationship with the salience- and arousal-related early component for aversive contexts.

2.3.7 Full pupil time series is predicted by convolved EEG discrimination amplitudes for the early punishment signal only

To further explore the link between pupil diameter and Y amplitudes, I implemented a GLM analysis to predict the full pupil time series based on the methodology of de Gee et al. (2014). The four main regressor events of interest were the trial-by-trial early and late discrimination peaks from the separate reward and punishment models. For each of these EEG discrimination regressors, the regressor amplitude at the onset indexes were set to the corresponding discrimination amplitude (Y) in a vector matching the pupil time series, which was then convolved to the subject-specific PuRF. The unmodulated phasic pupil responses at the time of stimulus and feedback onset were also included as control regressors (Figure 2.5D), such that the EEG-derived

predictions should reflect a relationship with the pupil response over and above the standard impulse response to visual stimuli. An additional control regressor reflecting feedback onset on trials with problematic pupil data (see 2.2.4) was also included to account for anomalous pupil responses.

Of the four regressors of interest, only the PuRF modulated by early Ys in punishment blocks showed a significant difference from zero, as revealed by a vector contrast analysis run on the subject-specific beta coefficients ($t(25) = -2.72$, $p = .010$), similar to that seen in the analysis run on the phasic pupil response (Figure 2.5C). However, the late reward coefficients failed to show any significant deviation from zero ($t(25) = -0.44$, $p = .665$). As with the phasic pupil analysis, no significant effects emerged for the late punishment coefficients ($t(25) = -0.34$, $p = .737$), or the early reward coefficients ($t(25) = 0.81$, $p = .425$). Additionally, the vector contrast showed a significant difference between the reward and punishment contexts as a whole ($t(25) = -2.42$, $p = .022$). This difference which seems to be driven exclusively by the significant difference between the early punishment signal and the early reward signal ($t(25) = -2.34$, $p = .026$), given that no difference emerged between the late signals across context ($t(25) = X$, $p = X$). Taken with the findings from 2.3.5, this seems to implicate the early arousal-related EEG peak as the most promising candidate signal with which to explore reward versus punishment asymmetries when combined with pupillometry.

2.4 Discussion

Using a multivariate analysis to discriminate positive and negative valence outcomes from trial-by-trial EEG data in a reversal learning task, I replicated the two spatiotemporal components that serve distinct functions in feedback processing in rewarding environments (Fouragnan et al., 2015, 2017; Philiastides et al., 2010). I showed that, as in (Fouragnan et al., 2017), the trial-wise amplitudes from this analysis reflect the sign of prediction error from an outcome, but not the magnitude, suggesting that these signals capture specific valence rather than surprise. Conducting the same analyses on data obtained in a punishing context, I demonstrated that the

characteristics of these signals are largely similar when learning occurs in an aversive environment. Specifically, the timing of discrimination peaks across contexts was consistent within a couple of milliseconds, and the spatial topographies were almost identical in all cases with the exception of the lack of fronto-central sign flip in the late punishment component. Taken alongside the identical group-level accuracy and response time data between the two contexts, it can be reasonably inferred that the two-component dynamics established in previous literature are likely present in simple learning tasks independent of the aversiveness or appetitiveness of reinforcers.

Despite these similarities, I also show evidence that certain more pronounced differences across context exist in the pupil data. Clear disparities emerge in the average phasic pupil response in the 1100ms window following feedback onset, with trials in the punishment context eliciting stronger pupil transients, indicative of a pronounce noradrenergic arousal response. In relation to the EEG discrimination components, there was a notable early-late distinction in the ability of trial-by-trial Y scores to predict phasic pupil dilation, with more pronounced loss-related Ys at the early peak showing significance in the punishment context in comparison to the late peak in the reward context. Extrapolating this comparison to the full pupil time series, a GLM with Ys convolved to a pupil response function as predictors found that only the early punishment component retained significance and appeared to be driving context differences in the pupil-EEG relationship. These results seem to have implications for where differences may lie in the cortical and LC signals between reward and punishment learning, if they indeed exist.

2.4.1 Insights into the early component from the Feedback Related Negativity

To help contextualise the spatially integrated EEG signals from the multivariate discrimination output, we can look to widely studied event-related-potentials (ERPs) that show spatiotemporal and theoretical similarity with our early component. Prior work with a similar two-component EEG analysis as the present study has shown a notable link between the early salience com-

ponent and the feedback-related negativity (FRN) ERP (Philiastides et al., 2010). This tracks with subsequent EEG-fMRI analyses that found the ACC to be strongly implicated in the same early component (Fouragnan et al., 2015) – a region sensitive to error detection (Carter et al., 1998) in which the FRN is typically source localised (Walsh & Anderson, 2012). The typical temporal range appearing in FRN research is 200-300ms (M. X. Cohen et al., 2011; Holroyd & Coles, 2002; van de Vijver et al., 2011), with many studies finding peak FRN responses in the early portion of this range within 5-10ms of our discrimination peaks (Hauser et al., 2014; Philiastides et al., 2010; Talmi et al., 2013), and the primary electrode used in FRN analyses (FCz) lies directly in the centre of our frontal topographical clusters (Figure 2.4A).

Though initially proposed to reflect a direct RPE signal (Bellebaum et al., 2010; Chase et al., 2011; Holroyd & Coles, 2002), a growing body of research has challenged this view of the FRN with evidence that it better reflects a 'good versus bad' outcome valence signal that is distinct from value or surprise (Fouragnan et al., 2018; Hajcak et al., 2006; Philiastides et al., 2010; Sato et al., 2005; Toyomaki & Murohashi, 2005; Yeung & Sanfey, 2004). The distinction from surprise is also consistent with our findings that unsigned prediction errors do not significantly explain variance in our weighted EEG signal for the early component. There is also a body evidence linking FRN responses and external measures of punishment sensitivity (Balconi & Crivelli, 2010; De Pascalis et al., 2010; Massar et al., 2012; Santesso et al., 2011; Unger et al., 2012), which again implicates the early component as a potential area to target reward versus punishment comparisons in more depth. I do not suggest that my spatially weighted EEG signal is completely analogous to the FRN, which are typically reported from individual sensors of interest. However, I believe there is enough conceptual and spatiotemporal overlap to consider this a useful known signal that can motivate further investigations based on motivational environment.

2.4.2 Insights from pupillometry

The addition of pupil data to the dual-component paradigm yielded some interesting insights into the role of the LC in feedback-processing. Firstly, there were clear effects in the amplitude of post-feedback transients across positive and negative outcomes, as well as reward and punishment blocks. These dynamics are prima facie in line with expectations for two reasons. Firstly, negative outcomes are rarer and more surprising than positive outcomes, and increased phasic pupil response has been repeatedly implicated in surprise (de Gee et al., 2021; Filipowicz et al., 2020; Lavin et al., 2014; Preuschoff et al., 2011) as well as disadvantageous choice more generally (Kozunova et al., 2022). Additionally, an LC-linked increase in dynamic value updating has been shown using pupillometry for specifically loss-outcomes (Pulcu & Browning, 2017). Secondly, pupil-associated LC arousal has been shown to be greater following punishment than following reward (Breton-Provencher et al., 2022). The asymmetric findings shown in the pupil response in the present results, which seem to be echoed in the literature, are encouraging indications that further examination of the pupil signal could reveal reward-punishment dynamics.

In light of this, the unique link found between the early discrimination component in the punishment condition and both the post-feedback transients and full time-series of the pupil response again presents a compelling case to focus on this signal for further comparisons in appetitive and aversive feedback responses. Though these dynamics certainly will have some interplay with the late component, as shown by the down-regulation mechanism in (Fouragnan et al., 2015), this is likely to be difficult to find without more insight from measures such as simultaneous EEG-fMRI. As such, further investigations into the reward-punishment dynamic target the early EEG signal specifically in conjunction with pupillometry.

Chapter 3

Early salience signals predict interindividual asymmetry in decision accuracy across rewarding and punishing contexts

In Chapter 2, I demonstrated the presence of a two-component EEG response in a punishing context comparable to that found both in this work and others (Fouragnan et al., 2015, 2017; Philiastides et al., 2010). There were notable distinctions, however, in the LC-driven phasic pupil response across reward and punishment, as well as across positive and negative outcomes. Furthermore, I showed links between pupil arousal and specific elements of the weighted EEG signal, with the early punishment signal in particular appearing to explain the majority of asymmetry in the pupil-EEG relationship across contexts. Additionally, there are several attractive theoretical links between the LC and the areas implicated in this early signal in EEG-fMRI work, which are discussed more thoroughly in this chapter. This makes the early EEG component a compelling lens through which to further investigate the reward-punishment dichotomy in learning.

To pursue this, I adapted the methodology of Chapter 2 to discriminate directly between the rewarding and punishing contexts within outcomes of equivalent valence. The goal was to identify a motivational salience signal temporally aligned with the early component, and to see whether this could reliably predict accuracy asymmetry across contexts. I also investigated the effects of phasic pupil arousal on this signal to gain insight into the role of the locus coeruleus.

3.1 Background

The classical account of instrumental learning dictates that actions leading to favourable outcomes will be reinforced, whilst actions leading to unfavourable outcomes will be diminished (Skinner, 1938; Thorndike, 1911). This basic principle of reinforcing behaviour has typically been understood through the reward prediction error (RPE) hypothesis, whereby the difference between expected and received outcomes is computed by phasic firing of midbrain dopamine (DA) neurons (Bayer & Glimcher, 2005; Glimcher, 2011; Schultz et al., 1997). The dopaminergic RPE in this framework acts as a 'teaching signal' that updates an internal value representation for a given stimulus following an associated outcome (Hollerman & Schultz, 1998), enabling the actor to better select for rewarding behaviours.

3.1.1 Salience within the dual-component framework

In a more recent perspective on the classical RPE hypothesis, it has been proposed that an early unselective salience signal precedes the later RPE and value-updating response independent of feedback valence or value (Schultz, 2016). The concept of salience is broadly defined as the degree of bottom-up attention attracted by a stimulus (Bordalo et al., 2012, 2022), which can incorporate a variety of factors such as sensory intensity, novelty, surprise, and relevance to motivational goals. With respect to the temporal dynamics of the dopaminergic RPE signal, there is evidence that adjusting different aspects of salience causes changes in the early response to stimulus presentation regardless of reward contingencies. For instance, the early activation of

dopaminergic neurons has been shown to be diminished by reduced visual intensity (Tobler et al., 2003) and reduced novelty through repeated exposure (Schultz, 1998). Similarly, dopaminergic neurons show substantial activation to non-rewarding stimuli only in the context of a reward-rich environment (Kobayashi & Schultz, 2014), implying that the degree of potential goal-relevance (motivational salience) also contributes to the salience response.

In human neuroimaging, a similar two-component (i.e. early/late) response has been observed with electroencephalography (EEG) during reinforcement learning (Philiastides et al., 2010). Subsequent EEG work with simultaneous functional magnetic resonance imagining (fMRI) showed that the early component of feedback processing was related to regions including the anterior insula (aINS) and anterior cingulate cortex (ACC) (Fouragnan et al., 2015), which are key areas within the so-called salience network (Seeley, 2019). The late component, on the other hand, involved areas traditionally implicated in reward and value processing, such as the vSTR (Bartra et al., 2013; Clithero & Rangel, 2014; O'Doherty et al., 2004; Pagnoni et al., 2002) and vmPFC (Bartra et al., 2013; Clithero & Rangel, 2014; Gläscher et al., 2009). Furthermore, it was found that this later value signal was downregulated by the early salience signal (Fouragnan et al., 2015), indicating a modulatory effect of outcome salience on value processing and raising clear parallels to the midbrain dynamics outlined by Schultz (2016).

A key aspect of learning that the two-component hypothesis may help to illuminate is the nature of learning in rewarding versus punishing contexts. This is due to the idea that individuals can have differing responses to these environmental conditions depending on their sensitivity to the goal of gaining reward versus the goal of avoiding punishment (McNaughton & Corr, 2008), which would alter the motivational salience of feedback in each of these contexts and perhaps explain individual asymmetries in learning. Evidence from human neuroimaging has shown that certain regions such as the locus coeruleus (LC), aINS and vSTR show particularly distinct activation patterns in rewarding versus punishing contexts (Carvalheiro & Philiastides, 2023; Palminteri et al., 2015), indicating the potential for highly variable individual dynamics in response to different types of reinforcer.

3.1.2 Motivational salience is compatible with prominent theories of punishment learning

A prominent mechanistic account of punishment learning is that the RPE mechanism incorporates aversive feedback as a negative signal via the suppression of dopaminergic firing, similar to the unexpected omission of reward (Mirenowicz & Schultz, 1996; Ungless et al., 2004). If this account is accurate, a modulatory salience component could plausibly act via the habenula, which has been directly implicated in the processing of motivational salience (Bromberg-Martin et al., 2010a, 2010b; Danna et al., 2013; Fakhoury & Domínguez López, 2014; Hikosaka, 2010), and seems influential for encoding aversive events and driving avoidance behaviour (Hennigan et al., 2015; Lawson et al., 2014; Lecca et al., 2017; Mondoloni et al., 2022). Importantly, the habenula has an inhibitory projection to dopaminergic activity in the ventral tegmental area (VTA) and substantia nigra (Christoph et al., 1986; Hikosaka, 2010; Matsumoto & Hikosaka, 2007), suggesting compatibility between the early salience hypothesis and this shared-mechanism account of punishment learning.

However, some prominent findings have shown that distinct subpopulations of DA neurons in the midbrain show phasic excitation to aversive stimuli rather than inhibition (Brischoux et al., 2009; J. Y. Cohen et al., 2012; Matsumoto & Hikosaka, 2009). Additionally, certain studies have found no effects of pharmacological DA agents on punishment learning, despite significant concurrent effects on reward learning (Eisenegger et al., 2014; Jocham et al., 2011; Pessiglione et al., 2006; Rutledge et al., 2009), which could suggest that punishment learning depends on a specific punishment prediction error (PPE) signal from separate non-DA system (Palminteri & Pessiglione, 2017). If this is the case, an early salience signal as presented by Fouragnan et al. (2015) is also compatible with many of the regions shown to exhibit distinct activation to aversive feedback during learning, including aINS (Combrisson et al., 2023; Gueguen et al., 2021; Kim et al., 2006; Palminteri et al., 2015; Seymour et al., 2004; Skvortsova et al., 2014), ACC (Fujiwara et al., 2009; Klavir et al., 2013; Monosov, 2017), and amygdala (De Martino et al., 2006; Delgado et al., 2011; Klavir et al., 2013; Metereau & Dreher, 2013; Yacubian et al., 2006). Crucially, though it is not yet clear exactly how the reward-punishment dichotomy

is processed in the brain, the most prominent accounts that have been proposed thus far seem to be mechanistically compatible with an early salience signal that modulates subsequent value processing, making this a plausible avenue for investigation. As such, possible salience effects can be examined from an agnostic position as to the core mechanism of reward and punishment encoding.

3.1.3 Aims and hypotheses

In this work, I aimed to investigate the extent to which I can differentiate interindividual learning propensities across the two contexts from neural and physiological measures. Specifically, I exploited an early salience electrophysiological (EEG) component, appearing at around 220ms post-feedback (Philiastides et al., 2010), that has previously been shown to emerge following reward omissions, with a subsequent downstream influence on a separate value processing stage (Fouragnan et al., 2015, 2017, 2018). This relationship is consistent with dual-component dynamics observed in midbrain DA neurons, where an early salience response to feedback modulates a later value-related signal (Schultz, 2016). This could point to a general salience mechanism, compatible with any of the main theories of reward and punishment encoding, that forms a crucial initial stage of reinforcement learning in the brain and explains a degree of individual variability in behavioural responses.

Adapting the paradigm of Fouragnan et al. (2015) to include distinct rewarding and punishing contexts in a reversal learning task, I first aimed to identify EEG post-feedback responses that are linearly separable across the two contexts independently for both positive and negative outcomes, leveraging the high temporal resolution to isolate the early salience-related component. This approach allows us to have a direct valence comparison without any confounding effects of outcome sign, such that the unique distinguishing factor in each comparison is whether the outcomes are relevant to a reward- or punishment-related outcome. Subsequently, I investigated whether these representations are consistent with the early salience signals reported in previous studies and test the extent to which they explain interindividual asymmetries in behaviour across

contexts Since there is evidence that the LC has both distinct contextual dynamics across reward and punishment as well as functional connectivity to key salience areas (Carvalheiro & Philiastides, 2023), and this nucleus is known to drive phasic pupil dilation (Larsen & Waters, 2018; Mathôt, 2018), I also used phasic pupil dilation as an indirect proxy measure to test how differences in LC-driven noradrenergic activations relate to EEG-derived salience representations and whether they further explain subject-specific behavioural changes across contexts.

3.2 Materials and methods

Refer to Chapter 2 for detailed description of the task, participants, EEG and pupil methodology, and computational model.

3.2.1 Subject-specific context sensitivity

My main aim was to test whether differences in neural or pupil signals between contexts can predict corresponding behavioural asymmetries across participants. Going forward, these comparisons will be referred to with the Δ prefix, which in all cases indicates the punishment condition subtracted from the reward condition for a given measure. The primary behavioural measure of context sensitivity is Δaccuracy, which is simply the proportion of correct choices attained in the punishment context subtracted from the proportion of correct choices attained in the reward context. As such, as positive value for Δaccuracy indicates greater average accuracy in the reward context. A correct choice refers to trials where the symbol with higher probability of reward or punishment-omission was chosen.

Given that the EEG-derived Y measurement reflects the distance from the discriminating hyperplane towards either the rewarding or punishing context, ΔY is designed to show the average asymmetry in neural signals across contexts. For positive and negative outcomes separately, ΔY for an individual participant is calculated by subtracting the absolute mean Y magnitude

for punishment condition trials from the absolute mean Y magnitude for reward condition trials (Figure 3.1A). For example, a ΔY value greater than 0 for positive outcomes would indicate that on average, for an individual participant, the neural signal induced by reward more pronounced and distinct than the neural signal induced by punishment omission.

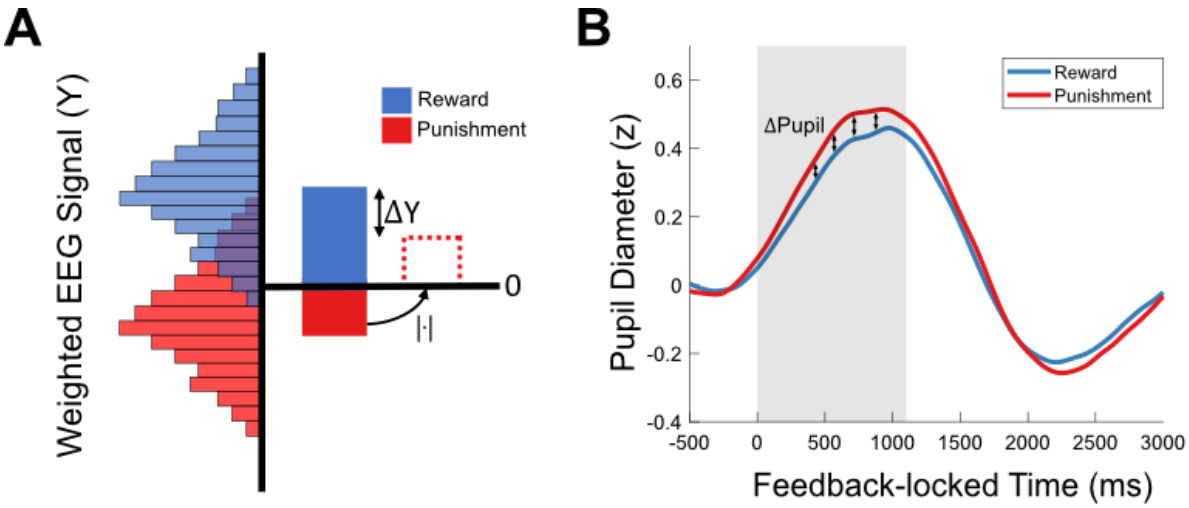

Figure 3.1: **A)**Depiction of ΔY measure for a hypothetical participant. Histograms show trial-by-trial distribution of weighted EEG (Y) values from the multivariate discrimination (as defined in Equation 2.1) across reward (blue, upper) and punishment (red, lower) trials. For the positive outcome model, reward trials reflect wins and punishment trials reflect non-losses. For the negative outcome model, reward trials reflect no-wins and punishment trials reflect losses. Solid bars show mean Y value averaged across trials. Dotted red bar shows conversion of mean punishment Y to absolute value for subtraction from the mean reward Y. **B)** Depiction of Δpupil measure for an example participant. Blue and red lines show mean z-scored pupil response post-feedback for reward and punishment trials respectively. Single Δpupil measure is computed by averaging over the shaded area – which highlights the window of significance from the non-parametric clustering test – for each condition and subtracting the punishment value from the reward value. This procedure is carried out separately for positive and negative outcome trials.

I leveraged the non-parametric window of significance (0-1100ms as outlined in the previous chapter) to calculate a Δpupil score, where mean pupil amplitude across the window in the punishment context was subtracted from the reward context for each participant. I chose to average across the window rather than select a single value at the peak as the non-parametric analysis demonstrated that many of the between-context differences are not accounted for by differences at the peak alone. As with the ΔY above, Δpupil was computed separately for positive and negative outcome trials to avoid possible confounding effects of outcome (e.g. signals associated with error detection) on the pupil diameter. A positive value for Δpupil would indicate that a participant exhibited greater phasic dilation in response to outcomes in the reward context

compared to the punishment context. Taking the difference score here isolates context-driven dilation effects by subtracting out common outcome-related arousal responses.

Together, these Δ scores allow us to quantify the extent to which context-dependent differences in EEG and pupil signals track context-dependent asymmetries in task performance. I therefore leveraged these scores to test the second hypotheses using simple linear regression, specifically that across reward and punishment contexts, differences in LC-driven pupil dilation and in discriminating EEG signals will predict subsequent asymmetries in behavioural accuracy.

3.2.2 Mediation analysis

The final hypothesis proposes that task performance is influenced by a salience signal visible in EEG data, which is in turn downstream of LC activation that drives pupil dilation. Because of the sequential nature of this hypothesis, a mediation analysis was used to determine whether the neural processes behind the ΔY value facilitate a relationship between LC-driven Δpupil and subsequent Δaccuracy. The goal of the mediation analysis is to identify whether the relationship between a predictor variable (Δpupil) and an outcome variable (Δaccuracy) can be explained by a mediator variable (ΔY).

Typically for a mediation effect to be considered plausible here there are three preconditions: 1) the predictor variable (Δpupil) should significantly predict the outcome variable (Δaccuracy) in a simple linear regression; 2) the predictor variable (Δpupil) should significantly predict the mediator variable (ΔY) in a simple linear regression; and 3) the mediator variable (ΔY) should significantly predict the outcome variable (Δaccuracy) (Baron & Kenny, 1986; Shrout & Bolger, 2002). In some cases, condition 1) can be considered non-essential, such as where the effects in 2) and 3) have opposite directions (MacKinnon, 2000). The mediation effect itself reflects the difference in predictive strength (the beta coefficient) of Δpupil on Δaccuracy in the simple regression model versus in the multiple regression model that includes ΔY (VanderWeele, 2016). For positive and negative outcomes separately, I used the M3 toolbox for Matlab (Wager et al., 2008) to establish the preconditions and significance of the mediation effect using a 10,000

sample bootstrap test on the resulting statistic (Wager et al., 2008).

3.3 Results

3.3.1 Distinct EEG and pupil responses to reward and punishment capture more than surprise

Pupil diameter post-feedback followed a typical impulse response profile for all contexts and outcomes, however these factors parametrically affected deviation from pre-feedback baseline. Negative outcomes elicited a greater dilation than positive outcomes, as did punishing contexts compared to rewarding contexts (Figure 3.2A). The non-parametric cluster test (Maris & Oostenveld, 2007) revealed significant differences for each of these comparisons across the 0-1100ms window. This is depicted by the shaded area in Figure 3.2A, which contains the negative Δpupil signal (reward – punishment) separately for positive and negative outcomes. Taken alongside the EEG findings, this supports the first hypothesis that salience-related signals will be significantly different across reward and punishment contexts.

To investigate whether any group differences emerged at the neural level, two single-trial multivariate discriminant analyses were used on EEG data locked to the time of decision feedback to separate the reward and punishment contexts; one trained on trials where the outcome was positive, the other negative. Separability between reward and punishment context was significantly greater than .5 between 170-530ms for positive-outcome trials and 170-500ms for negative-outcome trials, determined by AUROC values that exceeded the significance threshold of 0.58 from a 1000-sample permutation test ($p < .05$) (Figure 3.2C). A window of interest was set at 170-270ms to isolate the early salience component from a later value updating component, based on timings from previous studies (Fouragnan et al., 2015; Philiastides et al., 2010). At the individual level, a subject-specific discrimination peak was taken as the highest out of all AUROC values greater than the preceding and following two AUROC values within the speci-

fied window of interest. Averaged across participants, this yielded a component peak at 221ms for positive outcomes, and 230ms for negative outcomes (Figure 3.2B). The scalp topographies averaged across subjects at these moments reflected a similar fronto-central cluster to that observed in previous early components (Fouragnan et al., 2015, 2017; Philiastides et al., 2010), and were highly comparable across the two discrimination analyses trained separately on positive and negative outcomes (Fig. 3.2B; insets).

To test whether the context-driven EEG discrimination component was reflective of surprise (as the outcome-driven component was in Chapter 2), I used a linear regression to predict the trial-wise discrimination component amplitudes (Ys) from unsigned prediction error derived from the computational reinforcement learning model (Figure 3.2D). A one-sample t-test showed that for both positive ($t(31) = -0.743$, $p = .462$) and negative outcomes ($t(31) = -0.603$, $p = .551$), subject-specific model coefficients were not statistically significant from zero, indicating that the EEG component amplitude contains information other than pure surprise at an outcome.

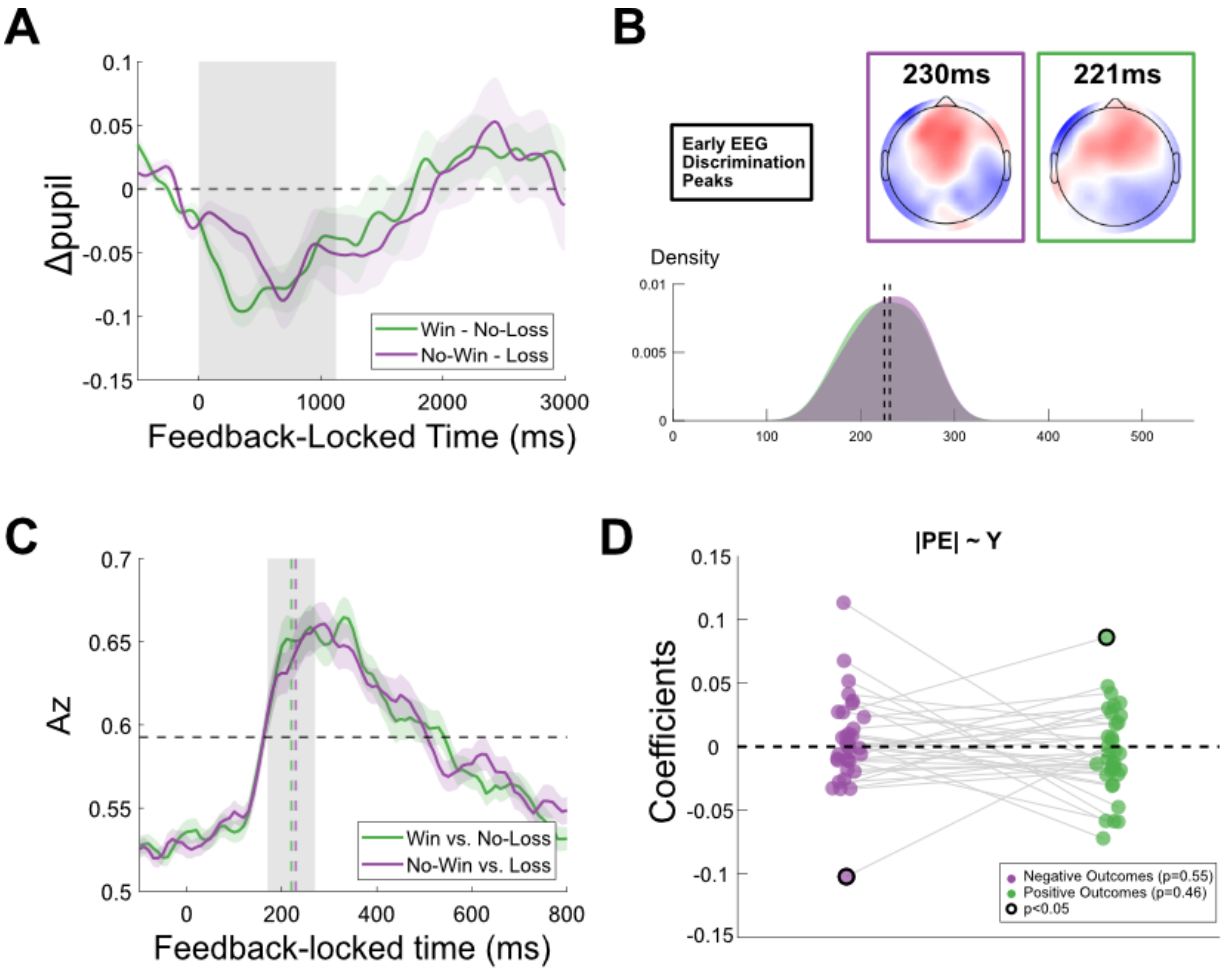

Figure 3.2: **A)** Difference score (reward – punishment) of the post-feedback pupil signal averaged across participants separately for positive outcomes (win – no-loss, green) and negative outcomes (no-win – loss, purple). Shaded area indicates window of significant difference between pupil response in reward vs punishment conditions averaged across all trials, obtained from non-parametric cluster test. **B)** Distributions of subject-specific AUROC peak selections for early (top) and late (bottom) components. Green indicates positive outcome trials and purple indicates negative outcome trials. Dotted lines show mean latency of peak AUROC averaged across subjects. Scalp topographies show average forward model from subject-specific early peaks – conditioned were arbitrarily mapped as negative (red) for punishment and positive (blue) for reward. **C)** AUROC (area under receiver operating characteristic curve) values and scalp topographies for two separate classification models. Y-axis depicts mean feedback-locked area under AUROC for logistic regression averaged across subjects. X-axis depicts time from feedback onset in milliseconds. Shaded error bar represents standard error of the mean across subjects. Grey shaded area reflects window for peak selection, and dotted vertical lines depict average peak onset for positive (win vs no-loss, green) and negative (no-win vs loss, purple) outcomes. Horizontal dashed line depicts p=0.01 permuted significance threshold averaged across subjects and across the two classification models. **D)** Beta coefficients for individual participants from a linear model predicting trial-by-trial Y amplitudes from unsigned prediction error from the reinforcement learning model. Purple dots (left) show coefficients from negative outcome trials only, and green dots (right) show coefficients from positive outcome trials only. Black outline indicates the beta coefficient value for that subject was significant.

3.3.2 Accuracy changes across contexts are tracked by EEG and pupil metrics

Despite similarities in behaviour across reward and punishment contexts at the group level, there was significant inter-individual variability in accuracy (Figure 2.3A), and clear differences in neural and physiological signals emerged. To address the second hypothesis and understand whether individual dynamics in accuracy were predicted by changes in EEG and pupil signals across contexts, I used simple linear regression to predict the individual Δaccuracy values across participants using the other Δ measures outlined in the methods section.

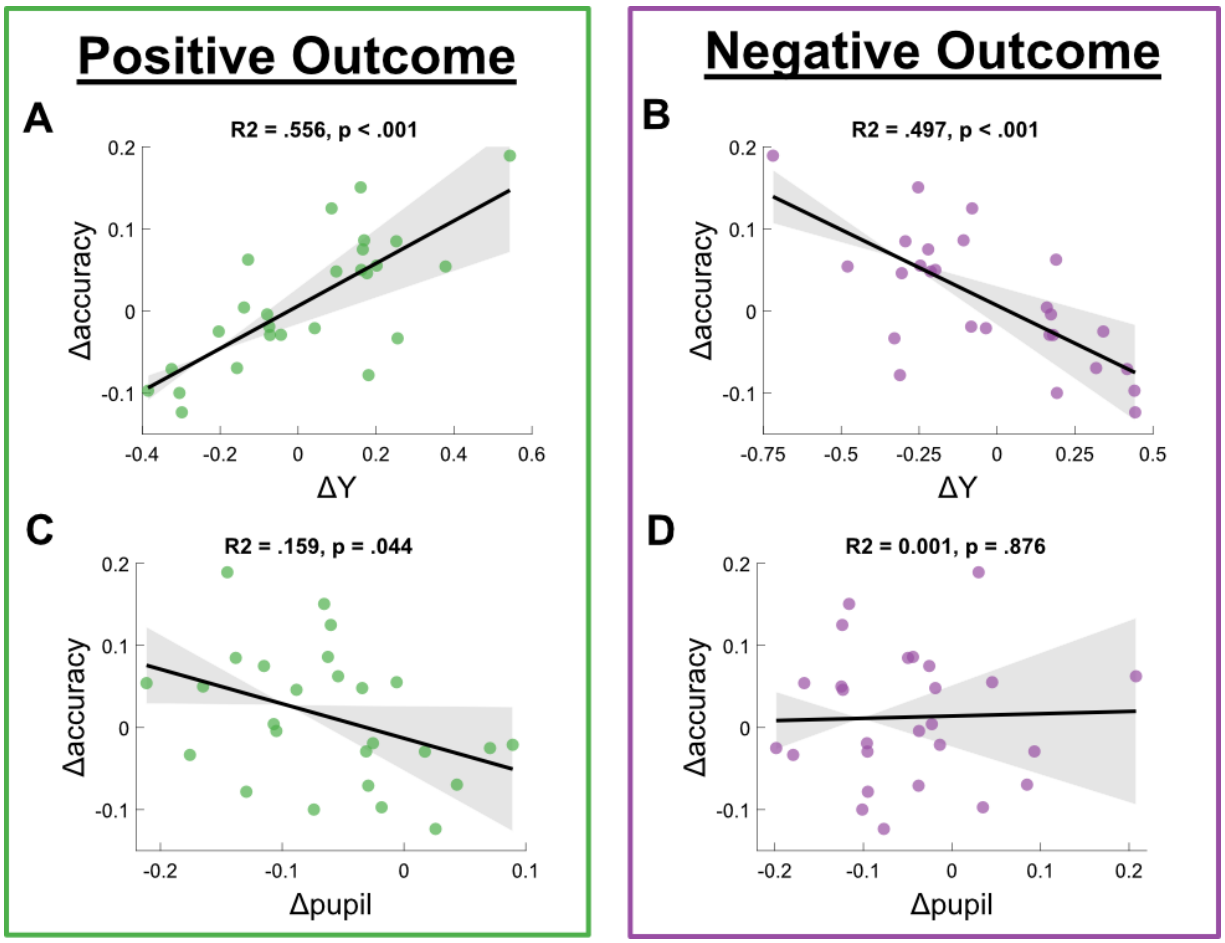

Figure 3.3: **A & B)** Δaccuracy linearly predicted by ΔY across subjects. Shaded error bars indicate 95% confidence intervals for the estimate. Δ-accuracy is the same measure calculated across all trials for all plots, whereas ΔY is separated by classification model trained on positive-outcome (left, green) and negative-outcome (right, purple) trials. Positive value on X axis indicates that EEG data for reward condition is on average further from the discriminating hyperplane than EEG data for punishment condition in a given participant, and vice versa. Positive value on the Y axis indicates higher proportion of correct choices in reward condition versus punishment condition for a given participant. C & D) Equivalent plots with Δpupil (reward – punishment) depicted on the X axis rather than EEG components. Again, Δaccuracy is identical across both plots, whereas Δ-pupil is separated by outcome type. .

I found that Δaccuracy was strongly positively predicted by ΔY for positive outcomes (Figure 3.3A; R2 = .556, F(1,24) = 30.039, p < .001) and negatively predicted for negative outcomes (Figure 3.3B; R2 = .497, F(1,24) = 23.671, p < .001). In each case, a discrimination component driven primarily by the polarised outcome (rewarding win or punishing loss) tends to bias accuracy in favour of the same context (e.g. more pronounced response to reward over punishment-omission predicts higher accuracy in reward condition over punishment condition and vice versa). I also found that as Δpupil increases, Δaccuracy significantly decreases for

positive-outcome trials (Figure 3.3C; R2 = .156, F(1,24) = 0.437, p = .046), but not for negative-outcome trials (Figure 3.3D; R2 = .001, F(1,24) = 0.028, p = .868). For positive outcomes, this suggests that relatively greater phasic arousal in response to wins reduces relative accuracy in the reward condition compared to the punishment condition, and vice versa. These findings show that, in line with the second aim it is possible to predict behavioural changes across context from EEG and pupil signals. It should be noted that the significant pupil result does not survive a Bonferroni correction for multiple comparisons, which lowers the alpha to .0125, demanding a level of caution for interpretation. All other results are unaffected.

3.3.3 EEG discrimination component mediates pupil effects on accuracy for positive outcomes only

The final hypothesis proposed that the salience-related EEG component would be related to pupil dilation, and that this relationship might offer further explanatory power in relation to behavioural changes across context. Given that pupil dilation is used here as a proxy for early LC arousal signals in the brainstem, and the projections that exist from LC to the regions associated with the early salience component in the EEG (Joshi & Gold, 2022), I believe that the EEG component may reflect a downstream cortical salience representation of which is influenced by LC activation and subsequently drives behaviour. As such, given that both signals influence behaviour for positive outcomes, I believe that the EEG signals may be mediating an effect of LC arousal on behaviour. To reiterate the precondition checks, 1) the predictor variable (Δpupil) should significantly predict the outcome variable (Δaccuracy) in a simple linear regression; 2) the predictor variable (Δpupil) should significantly predict the mediator variable (ΔY) in a simple linear regression; and 3) the mediator variable (ΔY) should significantly predict the outcome variable (Δaccuracy) (Baron & Kenny, 1986; Shrout & Bolger, 2002).

As with Δaccuracy, Δpupil was found to significantly predict ΔY for positive-outcome trials (Figure 3.4A; R2 = .367, F(1,24) = 13.584, p = .001), but not for negative-outcome trials (Figure 3.4B; R2 = .056, F(1,24) = 1.414, p = .246), so the mediation analysis was only conducted for

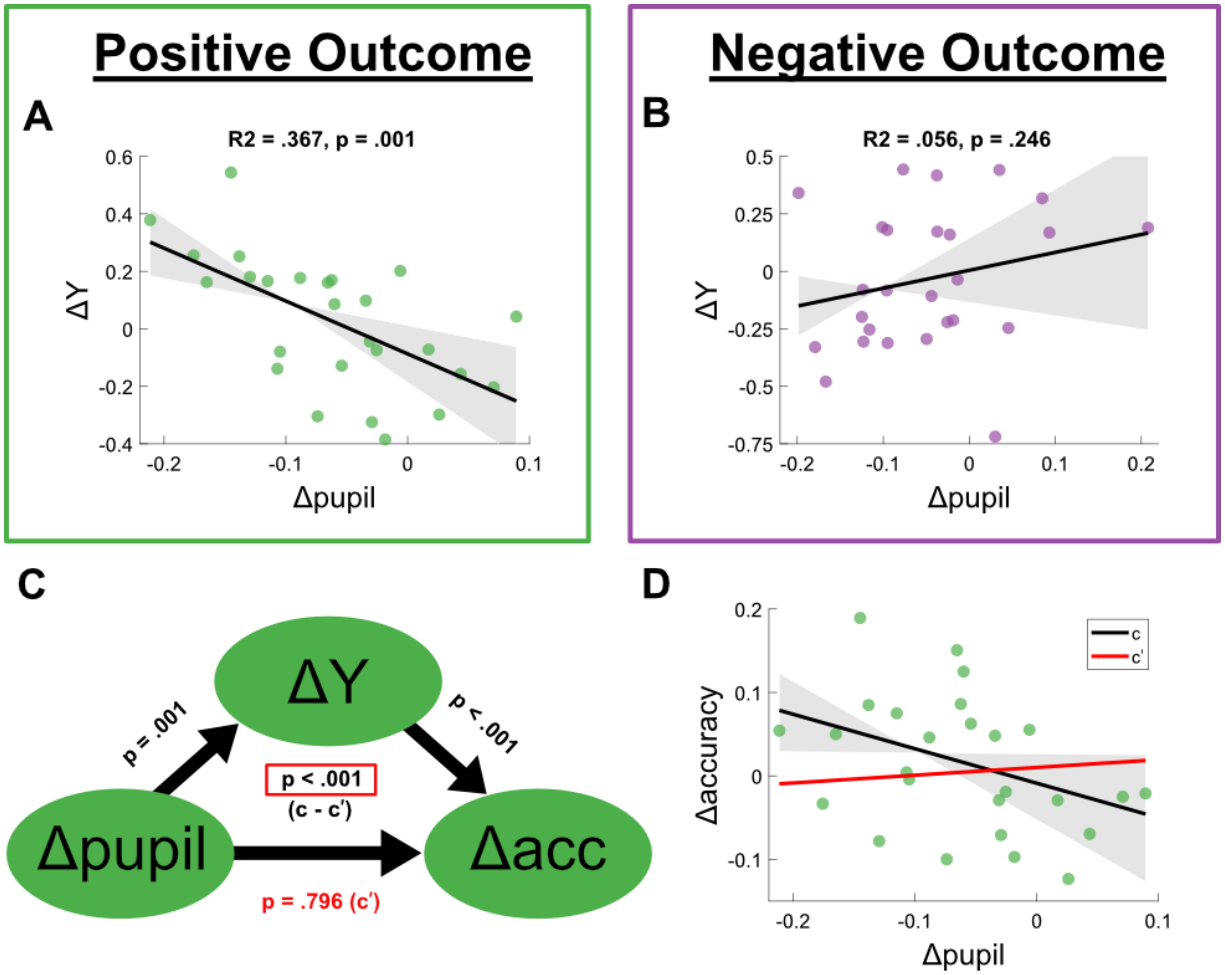

Figure 3.4: **A & B)** ΔY predicted by Δpupil across subjects. As in Figure 3.3, **A** shows a significant prediction of ΔY from Δpupil for positive-outcome trials, whereas **B)** shows no significant relationship between the two for negative-outcome trials. **C)** Mediation analysis (for positive outcomes only) showing the effect of Δpupil on Accuracy with ΔY as a mediating variable. P-values indicate as follows: Left – linear prediction of ΔY by Δpupil; Right – linear prediction of Accuracy by ΔY; Bottom – direct effect of pupil change on accuracy change when ΔY is included as a predictor in a multivariate regression (c'; direct effect). Middle – permutation test of comparison of model coefficient for Δpupil predicting Accuracy when ΔY is included as a predictor (c'; direct effect) versus not (c; total effect). **D)** Depiction of the two coefficient lines c and c' from the mediation analysis. The black line indicates the slope of the effect of Δpupil on accuracy in a simple linear regression, as depicted fully in Figure 3.3C ($\beta = 0.201$, p = .046). The red line indicates the slope of the same effect in a model where ΔY is included as an additional predictor ($\beta = 0.018$, p = .796).

positive outcomes. The final bootstrapped comparison between the coefficient of Δpupil for predicting Δaccuracy with (c) and without (c') ΔY included as a predictor was highly significant ($p<.001$, Figures 3.4C & 3.4D), indicating that changes in pupil-related arousal signals following positive outcomes may influence accuracy changes via distinct cortical activity across reward and punishment contexts.

3.4 Discussion

In this chapter I aimed to determine whether feedback in a punishing context elicits a distinct salience-related signal when compared to a rewarding context. I showed through multivariate discrimination analysis that EEG signals in response to punishment are highly separable from reward omission, and likewise for punishment avoidance and reward. By isolating an EEG signal that temporally coincides with a typical salience component of feedback processing (Fouragnan et al., 2015; Philiastides et al., 2010), I find distinct associations between mean discrimination amplitude and broad performance asymmetries across context. The phasic pupil responses to feedback were significantly amplified in the punishing context compared to the rewarding context, the magnitude of which also predicted performance differences, with a significant mediation effect of the EEG signal on this relationship. These findings suggest firstly that an initial salience response to feedback – possibly originating in the noradrenergic system in the brainstem - is modulated by an aversive context, and secondly that the degree to which this occurs has a significant direct effect on overall decision accuracy.

3.4.1 Motivational salience in relation to reward and punishment

Central to the hypothesis that individual differences in reinforcement sensitivity drive performance differences across contexts containing rewarding versus punishing reinforcers, I propose that the corresponding motivational asymmetry produces systematic differences in the motivational salience response to feedback. It been shown that the mere possibility of receiving

a rewarding outcome in a given environment can provoke a motivational salience response in dopaminergic regions to completely neutral stimuli (Kobayashi & Schultz, 2014). Accordingly, altered motivational responses in the presence of potential rewards or punishments may lead to behavioural changes, such as a shift in exploration tendency (D. Blanchard et al., 2001; J. Blanchard et al., 1998) or startle response (Aluja et al., 2015). Such behavioural shifts may affect task performance, bringing the agent closer to or further from optimal action, consistent with the strong link found between context differences in the EEG signal and overall choice accuracy.

Consistent with the proposed role of a motivational salience response in differentiating reward and punishment learning, the early component is strongly linked with the aINS and amygdala in both rewarding (Carvalheiro & Philiastides, 2023; Fouragnan et al., 2015, 2017) and punishing contexts (Carvalheiro & Philiastides, 2023). Though active in both contexts, these two regions have been implicated repeatedly in a specific capacity within punishment learning. Activity in the aINS has been directly related to a computational PPE-like function (Kim et al., 2006; Seymour et al., 2004; Skvortsova et al., 2014), while damage to the amygdala is known to inhibit salience processing of arousing stimuli (e.g. Anderson and Phelps, 2001) as well as punishment learning (Bechara et al., 1995; De Martino et al., 2010). The specific role in punishment learning combined with the presence in the early component of reward learning suggest that these regions could house motivational salience signals that are asymmetrically sensitive to appetitive and aversive reinforcers.

It is also worth returning to links between the FRN and the early feedback component established in 2.4.2. Although the early signal in Chapter 2 is inherently different to that shown in the current chapter due to the spatial weights being derived from a different discrimination dimension, there are doubtless commonalities due to the highly coherent spatial and temporal profiles. The FRN is purported to be generated primarily in the ACC (Walsh & Anderson, 2012), a region also central to the cortical salience network and shown to be related to early discrimination components (Fouragnan et al., 2015). Importantly, the FRN has also been characterised explicitly as a motivational salience signal common across rewarding and aversive stimuli (Mason et al., 2016; Talmi et al., 2013), a view consistent with findings that an active rather than

passive learning enhances FRN magnitudes, implying that motivational relevance is a key element of the signal (Itagaki & Katayama, 2008; Marco-Pallarés et al., 2010; Martin & Potts, 2011; Yeung et al., 2005). This is further evidence that motivational salience is a likely driver of early component amplitudes that are sensitive to the distinction between appetitive and aversive outcomes from actions.

3.4.2 Insights from pupillometry into theories of punishment

I showed that differences in pupil dilation in response to rewards versus punishment-omissions seem to strongly predict the corresponding differences in weighted EEG signal (Figure 3.4A), and moderately track accuracy asymmetries (Figure 3.3C). Given that noradrenergic LC activation is known to drive phasic pupil dilation (Larsen & Waters, 2018; Mathôt, 2018), I interpret these signals conservatively as an indirect proxy for activity in this nucleus. The LC has noradrenergic projections to both the amygdala (Buffalari & Grace, 2007; McCall et al., 2017) and the ACC (Carvalheiro & Philiastides, 2023; Chandler & Waterhouse, 2012; Hamner et al., 1999; Joshi & Gold, 2022; Koga et al., 2020), and these projections are implicated in alertness and attention (Sara, 2009; Sara & Bouret, 2012), which I use as the basis for a possible early arousal signal propagating from the LC to influence salience processing from outcomes. Though it has been shown that cortical signals which occur after the early EEG component, such as the P3 ERP, can exhibit a relationship with the phasic pupil response, these signals are generally believed to be co-generated alongside pupil dilation by noradrenergic LC signals (Chang et al., 2024; Menicucci et al., 2024; Nieuwenhuis, 2011; Nieuwenhuis et al., 2005). This also accounts for cases where the P3 and pupil dilations were found to be uncorrelated (de Gee et al., 2021; Kamp & Donchin, 2015; LoTemplio et al., 2021), and I believe that these findings are in line with the proposed mediation pathway from LC to cortex to behaviour that I propose in the results.

It important to note that the pupil effects from the data were not present for negative outcomes – the reward-omission versus punishment comparisons. Since negative outcomes were

less frequent (therefore more surprising) and provoked a much larger pupil response on average (Figure 3.2A), I speculate that this is due to a ceiling effect of pupil diameter, whereby more subtle changes across context are less detectable as the pupil nears maximum dilation. Recent work has shown that LC activity differs significantly between positive and negative outcomes in a rewarding context but not in a punishing context (Carvalheiro & Philiastides, 2023), which seems consistent with the idea that LC activity is higher across the board in a punishing context and perhaps therefore less differentiable, as indicated by the broadly higher dilation I observe. However, this hypothesis has not been directly tested and remains conservative.

In addition to areas in the cortical salience network, the LC also projects to the habenula (Purvis et al., 2018; Root et al., 2015), which has been directly implicated in the processing of motivational salience (Bromberg-Martin et al., 2010a, 2010b; Danna et al., 2013; Fakhoury & Domínguez López, 2014; Hikosaka, 2010) as well as aversive stimuli (Hennigan et al., 2015; Lawson et al., 2014; Lecca et al., 2017; Mondoloni et al., 2022). This is relevant to one of the main hypotheses of outcome encoding in punishment learning - that punishment is encoded with firing dips in midbrain dopaminergic neurons (Matsumoto & Hikosaka, 2009) - as it provides a possible route from early outcome-driven arousal signals in the LC to the encoding of reward and punishment in the VTA and substantia nigra via inhibitory signals from the habenula (Christoph et al., 1986; Hikosaka, 2010; Matsumoto & Hikosaka, 2007). This provides further support for the plausibility of the proposed mediation pathway, although I reiterate that this would require further research to test.

Chapter 4

Individual dynamics across reward and punishment in behavioural patterns and personality

In thesis thus far, I have focused primarily on broad comparisons of neural and pupil dynamics across rewarding and punishing contexts (Chapter 1), and I have established a general relationship between these signals and contextual asymmetry in task performance in value-based decision-making (Chapter 2). I now aim to explore the nature of the behavioural differences at a more descriptive level, diving deeper into specific behavioural and psychometric measures that underpin choice accuracy.

The investigation in Chapter 4 has two main avenues. First, I employ computational behavioural modelling to examine the extent to which behavioural accuracy asymmetries are explained by corresponding differences in specific reinforcement learning parameters. I then examine whether questionnaire scores on a relevant reinforcement-based theory of personality can predict behavioural or physiological measures.

4.1 Background

Defining broad principles of economic and value-based decision-making is an endeavour popularised with the advent of behavioural economics in the late 1970s. The landmark paper from Kahneman and Tversky (1979) demonstrated that people show robust asymmetry in choice dynamics when faced with a prospect of monetary gain versus monetary loss. Specifically, participants consistently tend to be more risk-averse in rewarding versus punishing contexts (Figure 4.1 A), and tend to overweight low probabilities and underweight high probabilities (Figure 4.1 B). A plethora of behavioural biases and heuristics have since been evidenced, such as the tendency for stimulus values to be biased by the range of possible outcomes within the context they were learned (Bavard & Palminteri, 2023; Bavard et al., 2018, 2021).

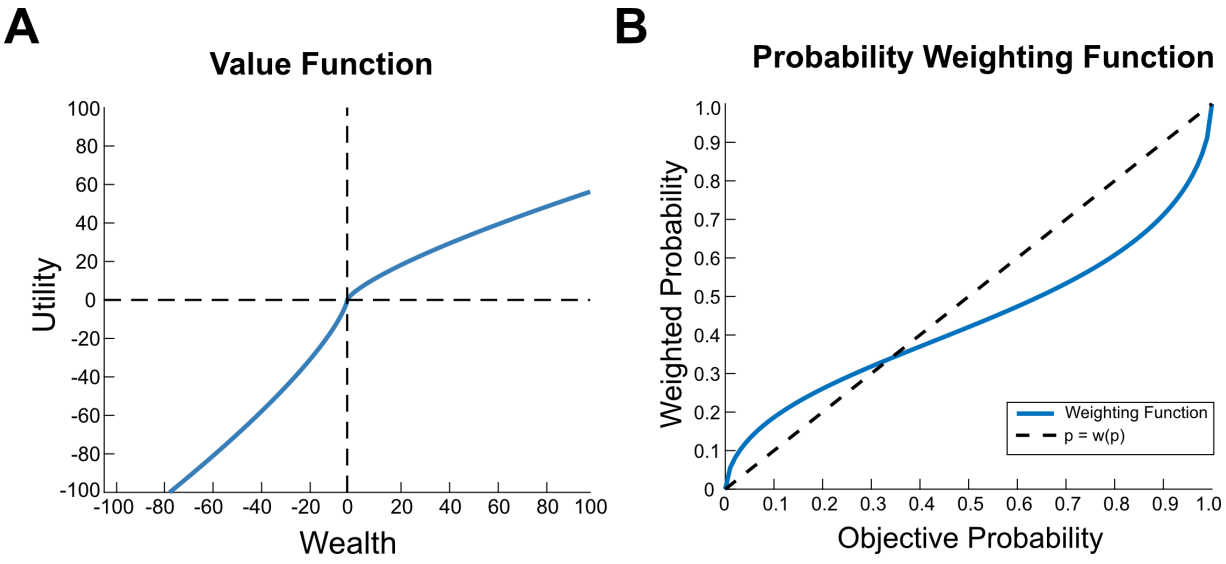

Figure 4.1: *Depiction of key prospect theory principles.* **A)** Prospect theory value function. The value function displays a steeper weighting function for losses relative to gains, illustrating the concept of loss aversion. B) Prospect theory weighting function. The probability weighting functions shows how people tend to over-weigh low probabilities and under-weigh med-high probabilities. The asymmetry in the inverse S curve also demonstrates a tendency for complementary probabilities to sum to less than one.

However, despite the presence of general principles that seem to generally apply to human decision-making, there are clear individual differences in choice behaviour across research and everyday life. To take just one example, the presence of acute stress has been shown to increase (Galván & McGlennen, 2012; Starcke et al., 2008), decrease (Clark et al., 2012), or have no effect on risk-taking in economic decisions (Sokol-Hessner et al., 2016). The stress-risk interac-

tion has therefore been deemed multi-faceted and sensitive to a range of individual differences (Porcelli & Delgado, 2017; Starcke & Brand, 2012, 2016), including gender (Daughters et al., 2013; Lighthall et al., 2009; Preston et al., 2007; van den Bos et al., 2009, 2014) and personality metrics (Carvalheiro et al., 2022; Lauriola & Weller, 2018; Lauriola et al., 2014). This highlights the importance of probing individual differences that may well contain important dynamics that are masked at the group level.

4.1.1 Individual differences through the lens of reinforcement learning

Perhaps the most notable way in which researchers have attempted to address variability across actors in learning tasks is through computational models of decision making. One particularly well-documented feature of such models is the need to solve the problem of the explore-exploit trade-off. Given some amount of volatility in an environment, the actor must occasionally sample perceived lower-value options to test whether underlying action values have changed enough to affect the optimal behavioural strategy. Though many models solve this problem through targeted sampling of non-exploitative options (e.g. Averbeck, 2015; Gittins, 1979, random exploration is often favoured due to its computational efficiency (Daw et al., 2006; Thompson, 1933). As such, the stochastic softmax decision rule provides an effective and simple solution in uncomplicated tasks, particularly when the set of possible actions is small or binary (see Equation 1.3). The key parameter in this sigmoid-shaped function is the central slope, known as the inverse temperature, which dictates how much an actor's decision-making is dictated by perceived value, and how much is due to stochastic exploration. This parameter can range from zero to one, where one represents a greedy policy where the higher value option is chosen every time, and zero represents completely random choice.

Within the reinforcement learning framework employed in the present work, there are two main free parameters that characterise an individual's decision-making behaviour: a learning rate parameter α from the Bush and Mosteller value updating policy which reflects speed of value updating; and an inverse temperature (or 'slope') parameter Υ from the Softmax decision

rule which reflects the degree of exploitation versus exploration in choice selection. Though the particular computational implementation of the explore-exploit trade-off is highly varied in the literature (Wilson et al., 2021), this aspect of learning has received extensive attention from a behavioural and neurobiological perspective – more so than the learning rate - and as such will be the focus of the reinforcement learning aspect of the current chapter.

4.1.2 Explore-exploit in the brain

There have been multiple accounts of how the brain solves the exploration problem over the past two decades. Early accounts proposed an opponent mechanism in the cortex, with a key role for the bilateral frontopolar cortex in inhibiting reward-seeking tendencies in dopaminergic striatocortical pathways (Badre et al., 2012; Cavanagh et al., 2012; Daw et al., 2006). An alternative view later emerged suggesting that rather than an opponent mechanism, frontal areas of the cortex worked cooperatively with reward pathways to motivate exploration by incorporating future potential rewards into value calculations to bias behaviour away from suboptimal greedy policies (Averbeck, 2015; Costa & Averbeck, 2020; Costa et al., 2019; Tang et al., 2022; Wilson et al., 2021). Indeed, modern evidence seems to trend more towards a cooperative rather than opponent coupling (Hogeveen et al., 2022; Tang et al., 2022), though it is not a settled question. In either case, some interaction between frontal control areas and ventral value-related areas seems to be well-supported.

Another perspective on the neural systems involved in exploration focuses on the noradrenergic LC pathway, specifically in relation to its role in regulating arousal. The adaptive gain model (Aston-Jones & Cohen, 2005) outlines a Yerkes-Dodson like relationship between baseline LC activity and adaptive task engagement, whereby excessively high or low tonic firing disrupts phasic activity needed for optimal performance. Specifically, over-arousal is proposed to lead to high attentional switching, whereas under-arousal prevents necessary acknowledgement of changing environmental features. A range of evidence has since supported this model, such as findings linking higher tonic LC arousal to increased switching behaviour and exploration (de

Gee et al., 2020; Gilzenrat et al., 2010; Hayes & Petrov, 2016; Jepma & Nieuwenhuis, 2011; Krishnamurthy et al., 2017; Urai et al., 2017).

Though this kind of tonic arousal is believed to work in opposition to a phasic mode of activation in the noradrenergic system (Aston-Jones & Cohen, 2005), there is an important distinction to make between phasic responses to task-relevant stimuli and surprise-driven responses to environmental feedback post-decision. This is evident in findings that have shown an increase in post-feedback phasic dilation following exploratory decisions that were preceded by advantageous outcomes (Kozunova et al., 2022). These post-feedback arousal responses have also been linked to computational surprise signals following decision outcomes (de Gee et al., 2021; Filipowicz et al., 2020; Lavin et al., 2014; Preuschoff et al., 2011), and in turn are purported to drive plasticity in performance-related areas such as the ACC – implicated in process such as error detection (Carter et al., 1998) – which in turn communicates back to LC to promote exploration of alternative strategies through the tonic mode (Aston-Jones & Cohen, 2005).

Given this link between the LC and an important salience-network region in the ACC (see also Carvalheiro and Philiastides, 2023; Chandler and Waterhouse, 2012; Hamner et al., 1999; Joshi and Gold, 2022; Koga et al., 2020), as well as other regions implicated in the early arousal such as the amygdala (Buffalari & Grace, 2007; McCall et al., 2017), it is plausible that subject-specific pupil and EEG dynamics shown in Chapter 3 bear some relationship to behavioural explore-exploit tendencies, which may underpin some of the accuracy effects observed. Furthermore, as shown in Figure 3.3, these dynamics can have direct explanatory links with regards to task performance if differences in arousal are present.

4.1.3 Altered reinforcement learning in clinical populations

There have been important developments in our understanding of the pathways and mechanisms involved in implementing key reinforcement learning parameters such as reward prediction error, surprise, and exploration. However, mapping the variation in these components across the spectrum of human individual difference – as well as subsequent effects on behaviour and life

outcomes – remains a difficult challenge. One approach that has seen progress in this goal is the study of known neurological disorders with mechanistic links to reinforcement learning processes, with a prominent example of this being Parkinson's disease (PD). PD typically induces neuronal loss in dopaminergic centres, and as such is often treated with DA-enhancing drugs (Samii et al., 2004), which can sometimes produce a side effect known as impulsive compulsive behaviours (ICB). Such behaviours can encompass compulsive shopping, addictive sexual behaviour, and pathological gambling (Voon et al., 2007), as well as increased dependency on dopaminergic drugs (Evans et al., 2005; Evans et al., 2006) and other substances such as alcohol (Evans et al., 2005). More broadly, ICB presentation has been associated with behavioural metrics of novelty seeking and enhanced exploration tendencies (Djamshidian et al., 2011), which has been partially attributed to uncertainty in establishing contingencies between actions and rewards (Averbeck et al., 2013).

In untreated PD, the associated dopaminergic impairment can often lead to a symptom known as motivational apathy, attributed to deficits in frontostriatal DA (Martínez-Horta et al., 2014; Pagonabarraga et al., 2015). A key finding within motivational research in PD has been that apathetic patients show improved reward learning when medicated with DA agonists at the detriment of punishment learning, and vice versa when unmedicated (Bódi et al., 2009; M. J. Frank et al., 2004; Kéri et al., 2010). Similarly, apathetic PD patients showed greatly reduced reward-sensitivity and task performance, accompanied by diminished firing in the ventromedial PFC, when compared with non-apathetic patients (Gilmour et al., 2024). Furthermore, mirrored results can be seen in patients with Tourette syndrome, a condition where dopaminergic activity is excessively high and treated with DA antagonists (Leckman, 2002). In such cases, the exact opposite reward-punishment dissociation can be observed where the treatment group shows improved punishment learning but impaired reward learning and vice-versa (Palminteri et al., 2009; Pessiglione et al., 2008).

The reward-punishment dichotomy observed in these cases has been linked to two separate dopaminergic mechanisms within motivational pathways in the basal ganglia characterised as 'go' and 'no-go' pathways (M. Frank, 2006). These pathways are proposed to be implemented

by two different types of dopamine receptor. D1 receptors send inhibitory projections to the internal global pallidus, which disinhibits the thalamus and promotes motor readiness and reward-seeking 'go' behaviour. Conversely, D2 receptors send inhibitory projections to the external global pallidus, which disinhibits the internal area and has the opposite effect on motor function. Critically, dopaminergic activity driven by the basal ganglia are proposed to regulate the balance between these two types of receptor, with increases in dopamine promoting excitation in D1 receptors and dips in dopamine favouring activity in D2 receptors (M. Frank, 2006)). In the context of individual differences, this offers a straightforward mechanism by which atypical learning occurs in cases like PD and Tourette syndrome, where altered DA function could drive impulsivity if increased or provoke apathy if decreased. However, when applied to motivation-based accounts of personality, this mechanistic perspective offers insight into learning differences across the full range of human phenotype rather than just the outer limits.

4.1.4 Reinforcement sensitivity theory as a paradigm for reward-punishment asymmetry

Reinforcement Sensitivity Theory (RST) is a theory of personality that proposes a dual-dimension model of individual differences based on sensitivity to appetitive and aversive stimuli and goals (Corr, 2004; Gray, 1981; McNaughton & Corr, 2008). This theory proposes that at the fundamental level, human behaviour is largely built upon innate sensitivity to different kinds of reinforcers, which manifest in distinct approach and avoidance behavioural systems (McNaughton & Corr, 2008). The approach system (Behavioural Activation System; BAS) largely overlaps with reward- and motivation-related dopaminergic pathways including VTA and the vSTR (Depue & Collins, 1999), whereas the avoidance (Behavioural Inhibition System; BIS) system involves the amygdala and ACC amongst other arousal-related regions (Corr, 2004). This advance-retreat dichotomy echoes the highly replicated finding that rewards are more associated with a 'go' response of behavioural invigoration (McNaughton & Gray, 2000), and punishments are conversely associated with a 'no-go' response of behavioural suppression.

Later, the theory was revised to incorporate an additional Fight, Flight or Freeze (FFFS) mechanism following evidence that a double-dissociation existed between cautious approach and fear response (Corr, 2004; McNaughton & Corr, 2008; McNaughton & Gray, 2000), as demonstrated with panicolytic and anxiolytic drugs in mice (D. Blanchard et al., 2001, 2003; J. Blanchard et al., 1998). This further helped to isolate the motivational components of the theory, with clear parallels visible between the RST BIS/BAS and the D1/D2 dopaminergic pathways. More explicit links can be seen in a variety of studies that have examined RST personality metrics through the lens of a Go-NoGo task paradigm, with consistent relationships found between the BIS/BAS dimensions and external measures such as N2 ERP response to positive or negative reinforcement (Hewig et al., 2005; Leue et al., 2012) or motivational conflict (Leue et al., 2009). Given links between the BIS and areas involved in our early salience-related EEG signal from Chapter 3, and the particular relevance to reward and punishment sensitivity, exploring psychometric RST data may offer insight into the characteristics of participants who show strong asymmetries in performance accuracy across rewarding and punishing contexts.

4.1.5 Aims and hypotheses

The aim of the current chapter is primarily to explore the specific behavioural dynamics that underpin the accuracy asymmetry effects displayed in Chapter 3. The explore-exploit literature – particularly in relation to adaptive gain and the optimal arousal model – presents the most likely candidate for explaining the link between our EEG and pupil difference scores and accuracy asymmetry across reward and punishment. Specifically, given the highly simple nature and relatively low volatility of the task environment, I hypothesise that maladaptive decision making will be linked to excessive tonic arousal leading to overly high exploration and stochasticity in choice selection, as reflected in the computational inverse temperature parameter. Furthermore, I predict that cross-context differences in EEG across wins and losses would be correlated with inverse temperature in the same direction as they were with accuracy in Chapter 3 (Figure 3.3 A & 3.3 B; positively for positive outcomes and negatively for negative outcomes). Also following the findings of Chapter 3 (Figure 3.3 C & 3.3 D), I predict that context differences in the surprise-

related post-feedback pupil dilation will correlate negatively with inverse temperature following positive outcomes but will not reveal any relationship following negative outcomes.

Regarding tonic pupillometry measures, I predict that pre-stimulus baseline pupil dilation will be significantly higher on trials containing an 'exploratory decision' than on trials containing an 'exploitative decision', as defined by whether the symbol with the highest model-estimated value was chosen or not. I also predict that differences in this tonic pupil arousal measure across rewarding and punishing blocks will correlate with corresponding differences in the inverse temperature parameter.

Finally, for the psychometric RST measures, the analysis approach is a largely exploratory examination of the relationships between subscale scores and measures from EEG, pupillometry, and computational reinforcement learning. Though I refrain from making any strong predictions, there are certain clear theoretical links between, for example, BAS scores and more frequent exploitation or BIS scores and higher pupil dilation following negative outcomes that may be more likely than others to show a relationship. Any findings here will be addressed conservatively in the discussion.

4.2 Methods

Refer to Chapter 2 for detailed description of the task, participants, EEG and pupil methodology, and computational model. Refer to Chapter 3 for more detailed description of difference (delta) metrics.

4.2.1 Tonic pupil dilation and delta scores

In addition to the phasic pupil response (described in Chapter 2) and corresponding Δpupil difference score (described in Chapter 3), I calculated pre-stimulus baseline arousal on each trial as a measure of trial-by-trial tonic arousal. This was simply taken as the mean amplitude across

the 500ms preceding stimulus onset on each trial, comprised of 20 samples at the 40Hz frequency. This procedure was based approximately on several other studies that have specifically investigated the effects of tonic pupil arousal (e.g. Gilzenrat et al., 2010).

For tonic pupil differences scores (in line with the phasic Δpupil difference scores used in Chapter 3), I employed the same procedure by simply calculating the average tonic pupil dilation score for each participant across reward and punishment trials separately, then subtracting the reward score from the punishment score to produce a single Δtonic score. Similarly, a difference score for exploration rates was calculated as the inverse temperature (or slope) parameter from the punishment-trial trained computational model subtracted from the equivalent parameter from the reward-trial trained model, producing a single Δslope score for each participant.

4.2.2 RST-PQ

The Reinforcement Sensitivity Theory Personality Questionnaire (RST-PQ) was developed in 2016 as a means to more comprehensively capture the three main subsystems (BIS, BAS & FFFS) of the revised RST (Corr & Cooper, 2016). Specifically, it addresses several issues that appeared in previous questionnaires, such as the inclusion of just a single BAS subscale which seemed to fail to capture certain important delineations such as reward sensitivity versus impulsivity (Dawe et al., 2004; Quilty & Oakman, 2004; Smillie & Jackson, 2006; Smillie, Jackson, & Dalgleish, 2006; Smillie, Pickering, & Jackson, 2006).

The scale itself consists of 65 items scored on a scale of 1 to 4, with items phrased as a series of statements accompanied by the question “How accurately does each statement describe you?”. The available responses are “Not at all”, “Slightly”, “Moderately”, and “Highly”. Items are grouped into the BIS, BAS and FFFS subscales, with BIS containing 23 items (e.g. “I am often preoccupied with unpleasant thoughts”), BAS containing 32 items (e.g. “I regularly try new activities just to see if I enjoy them”), and FFFS containing 10 items (e.g. “There are some things I simply cannot go near”). Scores for each subscale were determined by simply summing across all items pertaining to the subscale in question.

Additionally, the BAS items are broken down into 4 further subscales, as extracted from exploratory factor analysis (Corr & Cooper, 2016) guided by the theoretical subdivisions proposed by Carver and White (1994). These groupings are Reward Interest (7 items), Goal-Drive Persistence (7 items), Reward Reactivity (10 items), and Impulsivity (8 items). However, for the analysis in the current chapter, we will only be using a single overarching BAS score rather than examining BAS subscales individually.

4.3 Results

4.3.1 Accuracy asymmetry is predicted by exploration rate but not learning rate

To explore the nature of the context effects on choice accuracy, we compared differences across context in the free parameters of a reinforcement learning model with accuracy asymmetry. The learning rate reflects the weight applied to new information, and the slope (inverse temperature) reflects the degree of stochasticity or exploration in choice behaviour. As with accuracy and reaction time (Figure 2.3A), paired t-tests revealed no significant group-level differences across context for either slope (df = 32, t = 0.13, p = .90) or learning rate (df = 32, t = 0.34, p = .73), although a high degree of inter-individual variability was present regarding which context produced a higher value (Figure 4.2 A & B).

As with our other measures, we computed a delta value for each by subtracting the value estimated from a model trained on punishment blocks from that of a model trained on reward blocks. Using a robust correlation (bendcorr: https://github.com/CPernet/Robust-Correlations/blob/v2/bendcorr.m), we found that Δslope was significantly correlated with Δaccuracy (r(31) = .464, p = .006), but Δlrate was not (r(31) = -.268, p = .131). This result suggests that reduction in accuracy going from one context to another tended to be driven by an increase in exploration and lower stability in symbol selection, in line with my hypothesis.

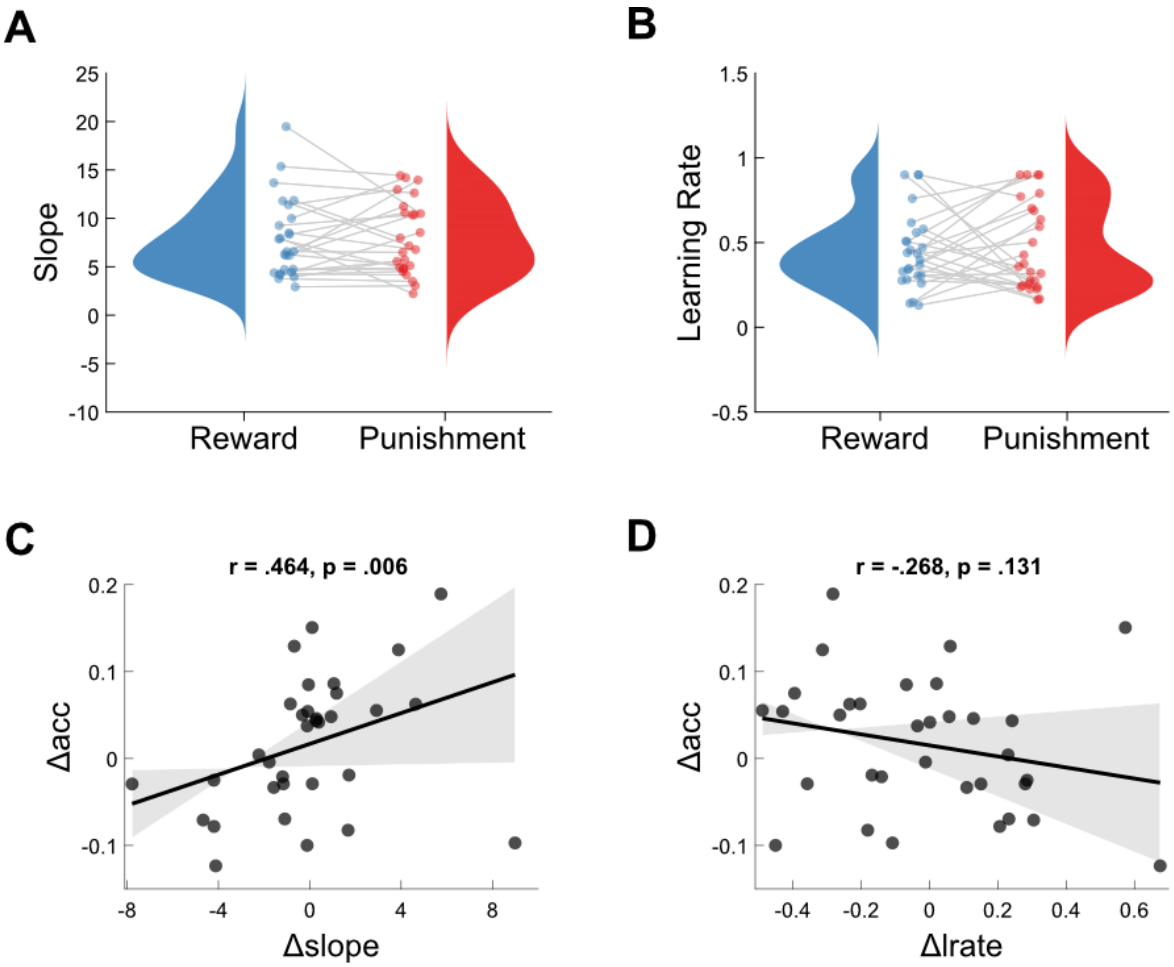

Figure 4.2: *Comparisons of reinforcement learning parameters across context.* **A) & B)** Cross-context comparisons of **A)** model-derived slope of the SoftMax sigmoid (i.e. inverse temperature and **B)** learning rate for individual subjects. **C) & D)** Δaccuracy linearly predicted by **C)** Δslope and **D)** Δlrate across subjects. Shaded error bars indicate 95% confidence intervals for the estimate. Positive value on either axis indicates that parameter was greater in the reward context than the punishment context.

4.3.2 Pupil and EEG signals are not significantly correlated with exploration rate

Following the significant relationship between slope and accuracy, I explored further this parameter in relation to the EEG and pupil delta measures. I examined the relationship of Δslope with Δpupil and ΔY for positive and negative outcomes separately, following a similar analysis strategy as in Chapter 3 except using robust correlation instead of a linear model.

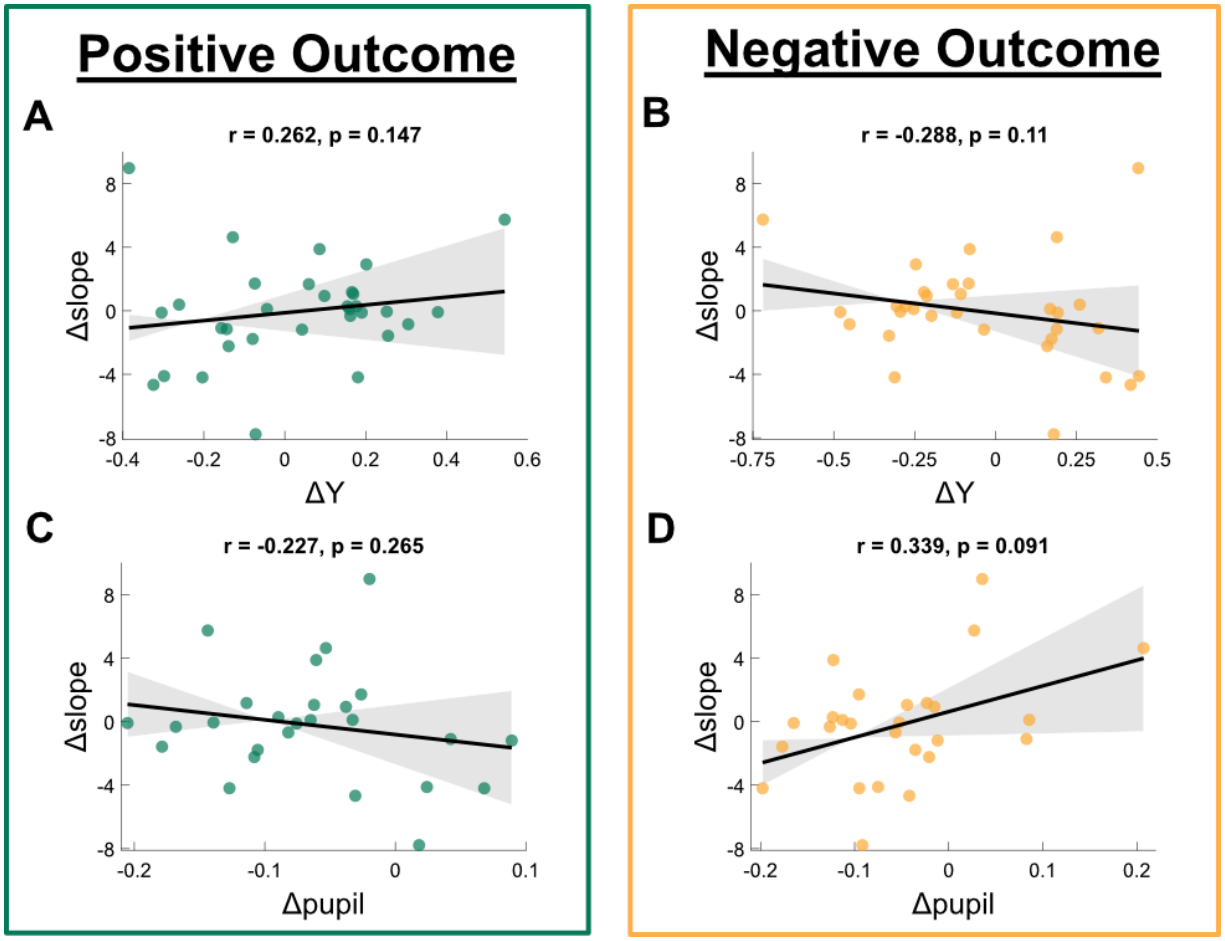

Figure 4.3: *Comparisons between Δslope and physiological measures.* **A & B)** ΔY correlated with Δslope across subjects. Shaded error bars indicate 95% confidence intervals for the estimate. ΔY is separated by classification model trained on positive-outcome (left, green) and negative-outcome (right, orange) trials. Δslope is identical across both plots. Positive value on X axis indicates that EEG data for reward condition is on average further from the discriminating hyperplane than EEG data for punishment condition in a given participant. Positive value on the Y axis indicates lower choice stochasticity in reward condition versus punishment condition for a given participant. **C & D)** Equivalent plots with Δpupil depicted on the X axis rather than ΔY. Positive value on the X axis indicates greater average post-feedback phasic pupil response for reward trials versus punishment trials.

Though trend-level relationships were visible, there were no significant correlations between

Δslope and ΔY for positive (r(31) = -.262, p = .147) or negative (r(31) = -.288, p = .110) outcomes (Figure 4.3 A & B). This was also the case for Δpupil for both positive (r(31) = -.227, p = .265) or negative (r(31) = -.339, p = .091) outcomes (Figure 4.3 C & D). This suggests that whilst it is not implausible that the context sensitivity signals from the EEG and pupil analyses have some effect on choice stochasticity and rate of value updating, this is not enough to explain the strong accuracy asymmetry effects that we see in Chapter 3.

4.3.3 Tonic pupil arousal tracks trial-wise exploration but not context differences

In addition to the surprise-related post-feedback pupil signals analysed in 4.3.3, adaptive gain theory also makes specific predictions about tonic levels of noradrenergic arousal and exploration. I first tested whether pre-decision baseline levels of pupil dilation (as a proxy for tonic arousal) were higher on trials where an exploratory choice was made rather than an exploitative one. Here, exploratory decisions were defined as decisions where the model-derived symbol value for the chosen symbol was lower than the unchosen symbol, whereas exploitative decisions chose the symbol with the higher model-derived value. A paired-ttest calculated within participants revealed a highly significant difference between these two types of choice (df = 32, t = 4.16, p < .001), as shown in Figure 4.4 A.

To test whether broad differences in tonic pupil arousal averaged across reward and punishment trials varied with differences in exploration rates, a robust correlation was used. Unlike with the phasic pupil responses in Figure 4.4B, this was conducted across all trials within each context rather than separating by positive and negative outcomes, as in this case there is no confound of outcome type since the pupil data comes from the pre-decision phase. Contrary to expectations, the robust correlation revealed no relationship between these two measures (r(31) = -.009, p = .965).

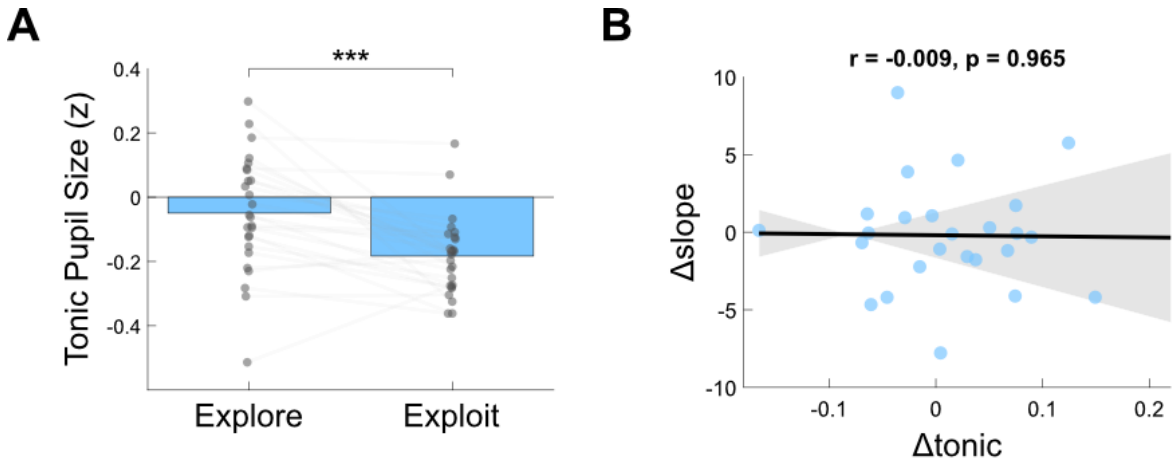

Figure 4.4: *Tonic pupil dilation in relation to exploration tendencies.* **A)** Average subject-specific pre-decision pupil diameter (in z scores) preceding decisions where the participant selected a perceived lower-value option (explore, left) verses decisions where the participant selected a perceived higher-value option (exploit, right). Triple asterisks indicate a p-value of <.001 in a paired samples t-test. **B)** Difference in average tonic pupil dilation on reward trials minus punishment trials (Δtonic) correlated with Δslope across subjects. Shaded error bars indicate 95% confidence intervals for the estimate. Δtonic grouped across both positive and negative outcomes. Positive value on X axis indicates that pre-choice baseline pupil diameter was on average higher in the reward context versus punishment context.

4.3.4 RST metrics are mostly uncorrelated with physiological and behavioural measures

In a series of exploratory investigations, I computed a robust correlation between participant scores on the BIS and BAS subscales of the RST-PQ with the main physiological and behavioural delta measures that yielded results from Chapter 3 and the current chapter. On the behavioural side, the only significant correlation that emerged was between BAS scores and exploration rates (r(31) = -.35, p = .044), as depicted in Figure 4.5 (BAS x Δslope). This is in the direction that would be theoretically predicted, as individuals with higher reward sensitivity would be expected to make relatively more exploitative decisions in the reward condition compared to the punishment condition.

For post-feedback ΔY and Δpupil measures, all correlations were non-significant with the exception of BAS scores and Δpupil following negative outcomes (Figure 4.5). This is also a relationship that is theoretically in line with RST, as higher reward sensitivity should predict relatively greater arousal in response to a reward omission versus punishment. It is important to

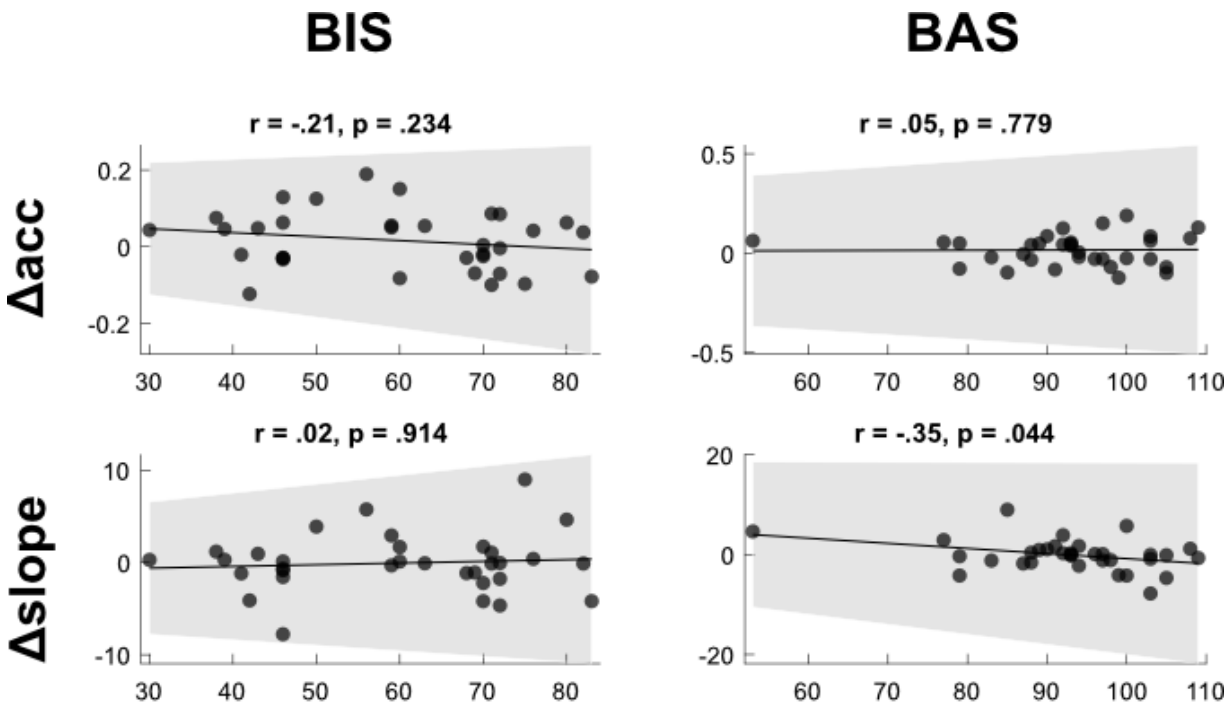

Figure 4.5: *Correlations between behavioural ΔY scores and RST dimensions.* Columns indicate X-axis scores for BIS (left) and BAS (right) subscales of the RST-PQ. Rows depict Y-axis scores from Δaccuracy and Δslope measures (as described in Figure 4.2).

note, however, that for both significant correlations the p value is not far below .05, and as such does not survive any Bonferroni multiple comparison correction.

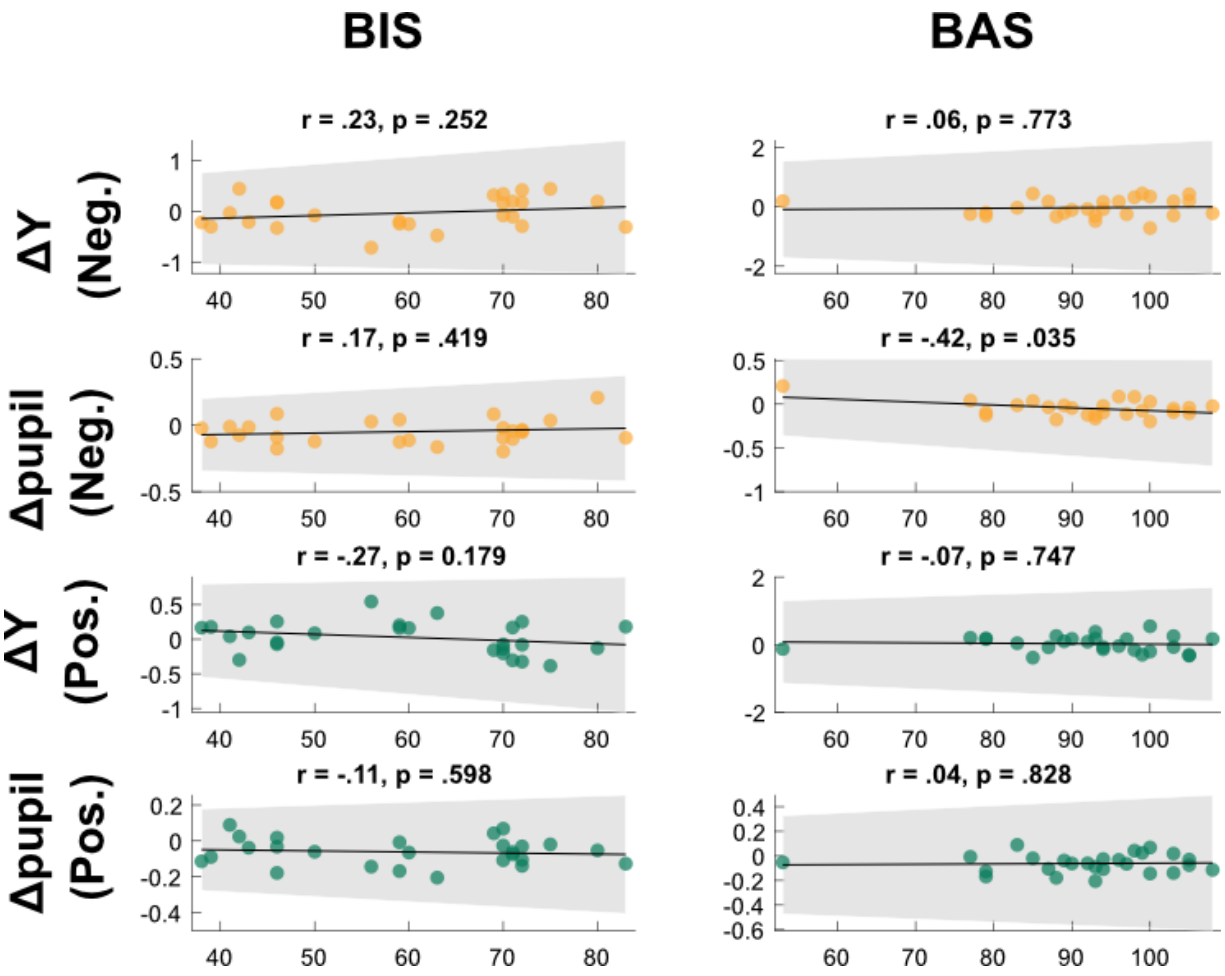

Figure 4.6: *Linear regressions predicting physiological Δ scores from RST dimensions.* Columns indicate X-axis scores for BIS (left) and BAS (right) subscales of the RST-PQ. Rows depict Y-axis scores from post-feedback ΔY and Δpupil measures (as described in Figure 4.3) for positive (green, lower half) and negative (orange, upper half) outcomes separately.

4.4 Discussion

Following the broad accuracy asymmetry findings of Chapter 3, this chapter aimed to probe deeper into the behavioural, physiological, and psychometric measures that may illuminate the factors that underpin performance changes. Using a computational reinforcement learning model, I captured the degree of exploration in the participants choices (inverse temperature, i.e. slope of the sigmoid in a Softmax learning rule), and the readiness with which they updated their beliefs about the symbol values (learning rate) with a methodology based on Fouragnan et al. (2017). The model was trained on trial-by-trial choices for each subject separately for

reward and punishment contexts, producing separate estimates for the two free parameters for each context, with the goal of capturing systematic differences in strategy or outcome processing across the conditions. Both models fit the behavioural data with a high degree of accuracy, as depicted in Figure 2.3 in Chapter 2.

Of the two free parameters, only the inverse temperature was significantly predictive of performance asymmetry. The positive correlation shown in Figure 4.2C suggests that greater exploration rates (shallower slope i.e. lower inverse temperature) correspond to lower accuracy, suggesting that a more stable and exploitative decision-making strategy is beneficial in this task. In relation to the context comparisons of Chapter 3, this can be interpreted speculatively to mean that an internal response to feedback that increases exploration and choice stochasticity in a given context, possibly driven by a modulated arousal response, is detrimental to performance. However, conclusive evidence of links between exploration asymmetry and EEG or pupil measures was not forthcoming, with only directional but non-significant trends emerging, suggesting that the full picture of the link between internal responses to feedback and task performance is not yet clear.

Baseline pupil dilation, characterised by the pre-stimulus average pupil diameter and indicative of the tonic mode of noradrenergic activation, was significantly higher on trials containing an exploratory decision than on trials containing an exploitative decision. This suggests that tonic LC activity does drive increased attentional switching and thus greater exploration. However, these effects were unable to account for aggregate differences in exploration across context.

4.4.1 The missing link between neural response and task performance

As with unsigned prediction errors (or surprise) in Chapter 3, there were no significant relationships across subjects between the weighted EEG signal and the inverse temperature (slope) and learning rate parameters from the reinforcement learning model (Figures 4.3 A-D). This is somewhat surprising, as Δslope – which can be conceptualised as the rate of stochasticity in choice behaviour – was highly predictive of accuracy asymmetries (Figure 4.2C), as was the EEG com-

ponent itself (Figure 3.3 A & 3.3 B). Additionally, the exploration-implicated tonic mode of LC activation is purported to be driven by perceived long-term task utility (Aston-Jones & Cohen, 2005; Jepma & Nieuwenhuis, 2011; Nieuwenhuis, 2011), which would suggest that an aversive learning environment such as the punishment blocks in the present task would invoke greater exploration due to an inherently negative task utility. One possible explanation for this could be that the binary reversal learning task is simple enough that certain participants are able to focus on a simple choice strategy regardless of subjective responses to decision outcomes. This could also explain the group-level similarities in choice accuracy, and why only when examining individual differences did notable context effects emerge. This is speculative, however, and would have to be tested by employing a similar paradigm in a more dynamic or intuition-based task, for example by utilising multi-armed choice tasks and/or drifting symbol values rather than reversing, as implemented by Jepma and Nieuwenhuis (2011)).

The findings seem to suggest that although increases in EEG amplitude from one context to another weakly track corresponding increases in choice randomness and exploration at a non-significant trend level, there seem to be other processes contained in the weighted EEG signal that reduce accuracy in a non-systematic manner. It is important to note that the direction of these context effects are heterogenous across subjects, in the sense that some subjects experience accuracy reduction in the punishment context, while others see an enhancement (Figure 2.3 A). This could imply a myriad of possible interactions between a context-related motivational salience response and downstream behavioural effects such as a helpful enhancement of memory and focus (e.g. Sutherland and Mather, 2015, 2018), or an unhelpful dysregulated arousal response to non-salient or neutral motivational stimuli that could be exacerbated, for example, in cases of anxiety or schizophrenia (Neumann et al., 2021).

Accordingly, the link between neural differences during reward versus punishment processing and subsequent behaviour may have important applications in clinical settings. For instance, disorders characterised by elevated DA in fronto-striatal regions typically predict deficits in punishment learning and behavioural inhibition, including schizophrenia (Moustafa et al., 2015) and Tourette's syndrome (Palminteri et al., 2012). Conversely, patients with Parkinson's disease

and Major Depressive Disorder can experience severe motivational apathy, largely attributed to a deficit in fronto-striatal DA (Pagonabarraga et al., 2015) and blunted RPE responses in the striatum and amygdala (Queirazza et al., 2019), respectively. These individuals also show significantly reduced distinction in neural response to outcome valence (e.g. gains versus losses), characterised by changes in FRN amplitudes (Martínez-Horta et al., 2014). Here, I have identified a similar neural signature predicting inter-individual behavioural performance across outcome types – in the absence of group-level trends – which could be used as a proxy for a more targeted diagnostic stratification and a more individualised treatment planning.

4.4.2 A possible role for LC-driven cortical and pupillary signals

Alternatively, it is also possible that cortical dynamics relating to exploration may be present later in the post-feedback response than the early EEG signal targeted here. Certain known cortical signals, particularly the P3, have been long believed to be linked to LC activity following Nieuwenhuis et al. (2005). Furthermore, the timing of this signal – canonically reported as between 300-400ms (Nieuwenhuis et al., 2005; Sutton et al., 1965) – is notably compatible with the timing of the late discrimination components identified in Chapter 2. With a couple of exceptions (Chang et al., 2024; Menicucci et al., 2024), a strictly direct link between the P3 and pupil dilation has generally opposed in recent years (de Gee et al., 2021; Hong et al., 2014; Kamp & Donchin, 2015; LoTemplio et al., 2021; Murphy et al., 2011). The prevailing account is therefore that pupil dilations and the P3 are co-generated from LC-driven activity rather than a causal role of the P3 (Menicucci et al., 2024; Nieuwenhuis et al., 2005).

This is not to discount the case made for the link between LC activity, pupil dilation and the early component that has been made in the preceding chapters; the 150-200ms taken for LC activity to project to cortical structures (Laeng et al., 2012) largely coincides with the initial emergence of the early component, and the exclusively relationships with the early rather than late signals found in chapter 2 should not be overlooked. It is also true that the early component has been shown to downregulate the late component (Fouragnan et al., 2015), and so it is

not impossible that effects relating to exploration could be found through formulating hypotheses around longer EEG latencies. Indeed, this would not be incompatible with the hypotheses targeted in Chapters 3 and 4 of this thesis relating to the early signal.

In line with many prior findings in the literature (de Gee et al., 2020; Gilzenrat et al., 2010; Hayes & Petrov, 2016; Jepma & Nieuwenhuis, 2011; Krishnamurthy et al., 2017; Urai et al., 2017), the results showed that tonic baseline pupil dilation was significantly higher preceding decisions where the participant made an exploratory choice (Figure 4.4A). Specifically, this result echoes directly a key finding from Jepma and Nieuwenhuis (2011) that shows a highly similar disparity in the pupil signal when averaged across a comparable pre-stimulus window. However, Jepma and Nieuwenhuis (2011) also found that individual differences in baseline diameter predicted an individual's tendency to explore, whereas in the results of this chapter differences in average tonic firing across trials in reward and punishment blocks was unable to account for corresponding differences in broad exploration rates (Figure 4.4B). The lack of effect here is somewhat surprising given the strong effect on the trial-wise analysis and the findings of the studies mentioned above. It also fails to match the finding that pre-task tonic pupil dilation predicts risk-taking behaviour specifically for aversive losses, and not for reward-omissions (Yechiam & Telpaz, 2011). Again, it is possible that the simplicity of the task employed here restricts the possible exploration dynamics that could occur, and it would be valuable to test this further in a more complex decision-making environment.

4.4.3 Possibilities and challenges of RST

Finally, it is necessary to evaluate the fairly underwhelming findings of the RST analysis presented in section 4.3.5 with a critical eye. Of the battery of correlations performed, two passed the uncorrected .05 significance threshold: a negative relationship between Δslope and BAS scores; and a negative relationship between Δpupil and BAS scores following negative outcomes. To interpret these at face value, the first finding is indeed compatible with the prediction that RST would compel: a higher degree of reward seeking should of course result in lower

exploration. The second finding is slightly less intuitive and can be summarised as: individuals with higher reward seeking propensities have higher phasic pupil responses to losses than to no-wins. In the loss domain, it seems more likely that BIS scores would be relevant to an arousal response to negative outcomes given the links to areas such as the amygdala and ACC (Corr, 2004). Regardless, although these results may be useful in generating hypotheses for further investigation, neither passes even one level of Bonferroni correction, and so should be treated with a high degree of scepticism.

The inclusion of the RST in the data collection was an exploratory decision driven by the intriguing theoretical links between its subscales and the neural pathways implicated in reinforcement learning and by extensions the early and late signals in the two-component model. Specifically, the association between the BAS and reward-related regions such as VTA and vSTR (Depue & Collins, 1999), and the BIS with the amygdala and ACC (Corr, 2004). However, in relating an abstract psychometric measure directly to neural or physiological measures, effect sizes are unlikely to be very large. As such, to achieve appropriate power to reasonably investigate specific relationships, a more targeted a-priori plan is warranted. To provide some context for this type of experiment using RST and EEG, the standard range of sample sizes on display in published research is typically at least 40-50 (De Pascalis et al., 2010, 2017a, 2017b, 2019), and can be well over 100 (De Pascalis et al., 2018). Indeed, significant effects have been found with sample sizes in this range for BIS scores in relation to FRN amplitudes following monetary loss and gain (De Pascalis et al., 2010). Therefore, to adequately explore this aspect of the EEG and pupil signals in relation to RST components, more targeted hypotheses are required to reduce multiple comparison issues, and greater sample sizes should be used to facilitate this.

Chapter 5

Predicting context-dependent performance for neurofeedback paradigms

Chapters 2-4 of the thesis have targeted differences in learning across rewarding and punishing contexts to reveal key similarities and differences in neural signatures, pupil dynamics, and behavioural markers. In doing so, I have identified a key salience-related signal that seems to differentiate reward and punishment learning in a way that predicts behavioural dynamics, and provided evidence for the potential role of the noradrenergic system in contributing to this signal by means of pupillometry. Importantly, the degree to which task performance depended on rewarding or punishing environment was highly variable across individuals, and specific individual asymmetries were reliably predicted by EEG signals and, to a lesser extent, pupil dilation.

The final experimental chapter will explore the potential application of these findings in the context of a closed-loop brain-computer interface (BCI) system. I present a general framework to conceptualise the challenge of improving task performance through modelling individual responses to different environmental factors. I then provide a pseudo-example of this approach in practice using data collected in a non-interactive task to demonstrate how the techniques could be applied and tested in a closed-loop BCI setting. The results of this will then be discussed in

the context of similar BCI endeavours, with attention paid to methodological room for improvement.

5.1 Background

The goal of using personalised data from an individual to improve performance outcomes has been tackled from a range of perspectives using a range of measures and techniques. In neuroscience, extensive research has shown that direct data from the brain can aid everything from relaxation to golf putting to the regeneration of motor function in paralysis. On the other end of the spectrum, computer science has produced many interesting methodological insights with regards to pattern recognition in large sets of behavioural data for the purpose of building individualised action profiles, identifying strengths and weaknesses, and in some cases developing personalised experiences in the form of targeted tutorials and guidance. The relative advantages of each of these approaches offers an enticing prospect for an integrative approach to BCI.

5.1.1 Neurofeedback and BCI for performance enhancement

Neurofeedback – also known as biofeedback – is the practice of using direct insight from neural measurements to improve behavioural performance. Early attempts demonstrated that people have the ability to regulate their alpha brain activity when provided with live feedback about their EEG spectral power profiles (Kamiya, 1968, 1969). Since its advent, a number of well-established paradigms have been developed and employed in a variety of use cases. EEG-based frequency-power neurofeedback is the most commonly employed method, where real-time information about brain activity delivered through audio or visual cues is used to modulate frequency and amplitude of brain activity (Da Silva & De Souza, 2021; Marzbani et al., 2016). This is normally achieved following a training protocol where a participant will become familiarised with how the feedback responds to their own metacognitive efforts, known as neurofeedback training (NFT), which is then utilised in a task environment once a baseline level of control has

been acquired (J. H. Gruzelier et al., 2014).

A popular method in this domain has encouraged augmented activity the Sensory Motor Rhythm (SMR) band of activity in the range of 12-15Hz, sometimes also known as the beta protocol, whilst suppressing other frequency ranges. This approach is based on the idea that the beta band of brainwaves over sensorimeter regions on the scalp are related to focus and attention, and as such has often been used for improving outcomes in attention-deficit hyperactivity disorders (Egner & Gruzelier, 2004; Heinrich et al., 2007; Lubar et al., 1995), as well as more general attentional performance (e.g. Egner and Gruzelier (2001) and Mikicin (2015, 2021)). Indeed, a systematic review found the SMR to be by far the most prevalent paradigm for attention-related objectives (Da Silva & De Souza, 2021).

A different approach has focused on maximising the relative ratio of theta to alpha (A/T) amplitudes for the purpose of inducing relaxation and promoting creativity, typically implemented in an eyes-closed setting with relaxing sounds used as indicators of desired brainwaves (J. Gruzelier, 2009). The A/T paradigm has been used to enhance outcomes in creative domains as musical or dance performance (Egner & Gruzelier, 2003; J. Gruzelier, 2009; J. H. Gruzelier, 2014; Raymond et al., 2005). Though this is by no means an exhaustive account of the vast array of different EEG-neurofeedback approaches, these popular protocols have been repeatedly shown to have efficacy in a wide range of settings such as sports performance (Arns et al., 2008; Cheng et al., 2015; Gong et al., 2021; Ring et al., 2015; Rostami et al., 2012; Xiang et al., 2018), memory (Escolano et al., 2011; Lecomte & Juhel, 2011), and epilepsy (Sterman & Friar, 1972; Sterman et al., 1974).

A branch of research related to NFT, known as brain-computer interface (BCI), takes a slightly different approach to augmenting human functioning. Whereas NFT presents a direct representation of some target signal in the brain for autoregulation, BCI bypasses this metacognitive component by directly effecting some event external to the individual based on neural signals. Traditionally, this paradigm has focused on getting around the need for physical action in motor-impaired individuals (Nicolas-Alonso & Gomez-Gil, 2012; Wolpaw, 2013), such as those with forms of paralysis, by controlling an external device such as a robotic limb (Robin-

son et al., 2021) or computer mouse (Citi et al., 2008; McFarland et al., 2008). However, the BCI paradigm can also refer more broadly to methods of optimising behaviour in non-clinical populations without neurofeedback, such as the integration of neural signals from groups of people to aid decision-making in collaborative BCI tasks (Poli et al., 2014; Valeriani et al., 2015; Wang & Jung, 2011; Yuan et al., 2013). The BCI framework therefore presents a fairly broad paradigm through which to view EEG-informed attempts to manipulate task performance.

5.1.2 Predictive models for performance optimisation

Human behaviour in decision making tasks has understandably been a focus within psychology and neuroscience, but there have also been notable efforts from the field of computer science to predict individual actions. In non-cooperative games such as Poker, opponent modelling has been a key focus when building game-playing AI, as it is possible to improve performance with exploitative strategy based on opponent play (Berger, 2007; Billings et al., 1998; Davidson et al., 2000; Xu & Chen, 2021)). Early examples first used neural networks to identify which features of play were most relevant to prediction, which were then implemented in a simple table-based system for estimating median hand strength (Davidson et al., 2000). Modern approaches are more sophisticated, first classifying an opponent on the axes of loose-tight and passive-aggressive, then employing population-based evolution to develop a range of effective counter-strategies that can be flexibly employed to match opponent behaviour in an exploitation phase (Xu & Chen, 2021).

A variation of this research focuses on player modelling to identify weaknesses for the purpose of aiding improvement. A variation of the Chess engine Maia was used to predict the most likely move for players of a certain skill level (McIlroy-Young et al., 2020), which has promising implications for targeting coaching to common weaknesses throughout Chess progression. This was subsequently enhanced to offer even more specificity, modelling individual players to achieve a player identification accuracy of 98% (McIlroy-Young et al., 2022). In addition, the authors were able to develop unique ‘blunder profiles’ that characterised the kinds of mistakes

that a player was susceptible to make – a highly useful tool for delivering personalised feedback and practice (McIlroy-Young et al., 2021).

5.1.3 A hybrid approach

Both the neurofeedback and machine learning angles have shown unique successes and very different types of insight into behavioural optimisation. In the specific context of human performance, the neurofeedback literature has a great deal more depth and overall practical success. Nonetheless, there are certainly elements from the computer science literature that could offer some interesting and novel adaptations.

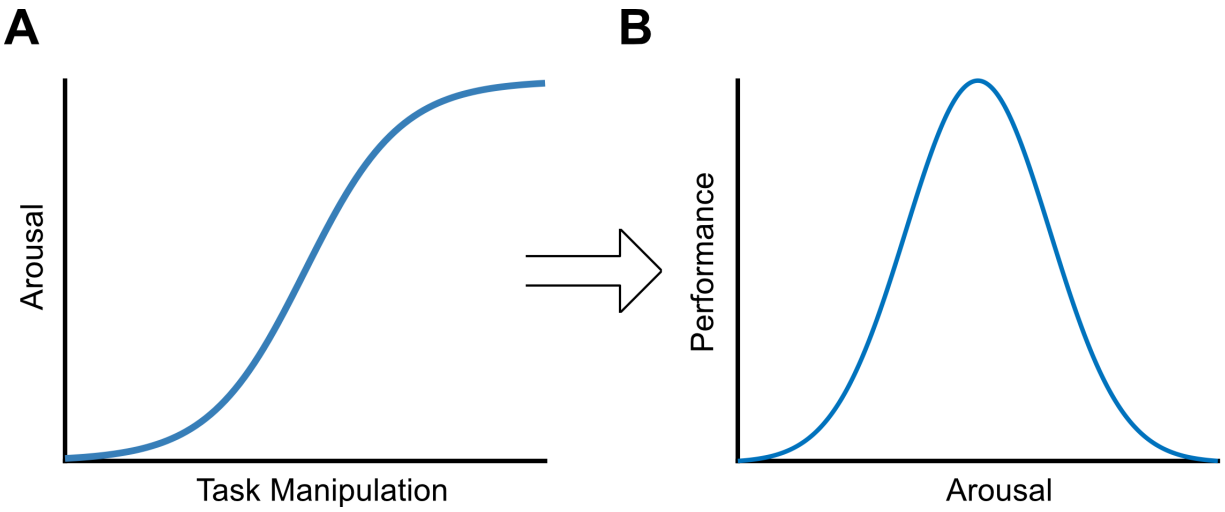

Figure 5.1: *Theoretical representation of functional relationship between manipulation, physiological process, and task performance.* **A)** Arousal theoretically increases as a sigmoid function of manipulation (e.g. stressor) strength, and in turn **B)** task performance theoretically is maximised at the peak of the Yerkes-Dodson inverted U.

Figure 5.1 demonstrates a typical unidimensional optimisation case based on neurofeedback studies such as Faller et al. (2019), who achieved significant performance gain on a flight simulation task using a heartbeat sound to manage arousal levels alongside real-time EEG. Using this arousal scenario as a hypothetical example for a given participant, performance exists as a function of arousal (Yerkes & Dodson, 1908), and arousal exists as a function of 'peripheral' stress manipulation. Therefore, to optimise performance, a stress manipulation is required such that:

$$M = argmax f(g(m)) \tag{5.1}$$

Where M is the optimal task manipulation, $g(m)$ is arousal as a function of task manipulation and $f(g(m))$ is task performance as a function of arousal. Though simple, this framing offers some flexibility to help conceptualise how this can be extrapolated to borrow from the strengths of other approaches to tackling this challenge.

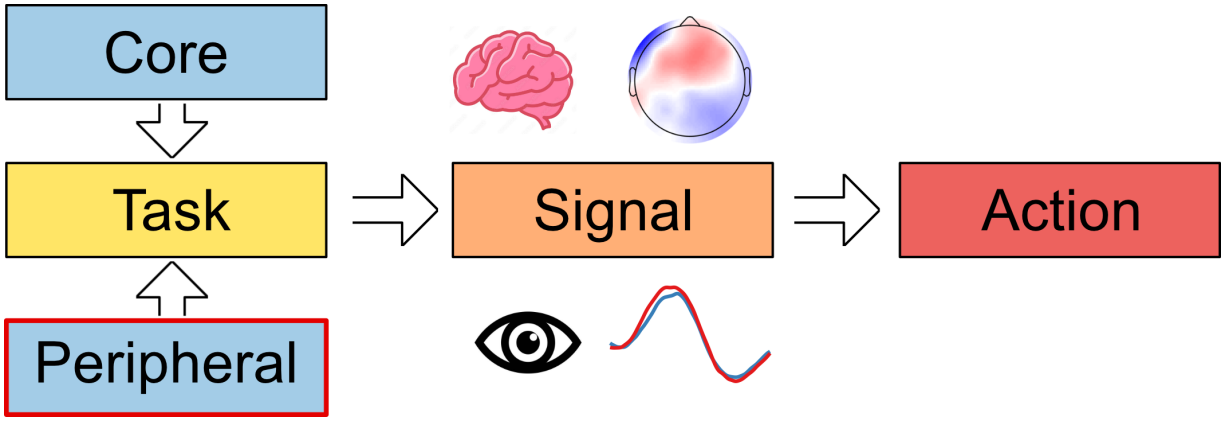

Figure 5.2: *Flow chart of task features influencing action via neural response.* A task or environment is comprised of core and peripheral features (left) that elicit a neural response (centre) in a given state. This neural response is then responsible for action selection (right). Whilst core features are exclusively desirable when building an optimal agent, peripheral features are also relevant to human behaviour.

The neurofeedback approach has shown that in many cases, this goal can be achieved by simply providing some meta-insight into an individual's position on the right-hand function. On the other hand, the behavioural data-driven approach is limited in an important respect: it cannot account for the intermediate physiological activity that bridges psychological state and subsequent action. In other words, it skips the 'signal' step highlighted in Figure 5.2. This means that, whilst the general relationship between task and action may be modelled, there is no way to account for moment-to-moment changes in internal signals that ultimately are the causal drive of an eventual action. Certain recent endeavours have made some steps towards unifying these different ideas, for example Faller et al. (2019) used a statistical classification approach to identify instances of maladaptive oscillations associated with over- or under-correction when navigating a course, which then manipulated the heartbeat sound. In this way, the feedback was not based on the typical power-frequency read-outs and had some degree of machine learning involved in

the performance optimisation. However, this still operated in the arousal optimisation paradigm typical of approaches like SMR NFT, and there is an opportunity to widen the scope further.

One possible framework to combine the two approaches is outlined in Figure 5.2. If we were to develop an artificial intelligence agent to perform a task, it would optimise exclusively for the features of a task that are directly relevant to performing it well, such as environmental uncertainty or desired outcomes. Such features can be thought of as **core features**, which are necessary to navigate a task successfully. Humans, on the other hand, must perform a huge range of functions, and must also be vigilant of external threats to safety. Take a video game example, such as Multiplayer Online Battle Arena (MOBA), where you are engaged in a long-term battle in a highly dynamic and uncertain environment. There may be factors such as the stress of gathering resources in a risky area, or aversive loss of being ambushed and killed, that are actually irrelevant to an optimal game-playing strategy. And yet such factors will still influence decisions due to presumed evolutionary advantages, such as the high cost of being ambushed while gathering resources in real life. Features of a task that relate to these factors, but not to optimal gameplay, can be thought of as **peripheral features**. Thus, human physiological and neural signals are functions of both core and peripheral features of a task state, and action is a function of these signals.

There seems to be an opportunity here, therefore, to bridge the gap between these two approaches to apply BCI to a decision-making paradigm using insights about environmental effects on brain and behaviour rather than standard neurofeedback. If we extrapolate Figure 5.2 and Equation 5.1 such that M becomes a set of peripheral features optimised over a multidimensional parameter space, the set of optimisation effects is greatly expanded and allows for manipulations such as rewarding versus punishing feedback to be explored.

5.1.4 Aims and hypotheses

The aim of the current chapter is to demonstrate a proof-of-concept for a real-time closed-loop analysis strategy to optimise decision-making performance. Specifically, I train a multimodal

classifier to estimate whether a rewarding or punishing environment would be most conducive to an optimal action. In cases where novel methods with BCI or neurofeedback are developed for task performance (e.g. Faller et al., 2019), it often preceded by theoretical groundwork to show that the manipulation-signal-action relationship is robust enough for a closed-loop experiment to work (Saproo et al., 2016).

As such, this chapter will apply the classifier retroactively to existing data to test whether it has the capability to successfully discriminate context and correctness based on EEG and pupillometry data, and whether this can plausibly be applied in a way that improves decision accuracy. There are therefore several criteria that I aim to meet: 1) a level of multinomial classification accuracy satisfactorily above chance level; 2) a level of subject-specific model-predicted choice accuracy that is not significantly different to their actual choice accuracy; 3) a hypothetical context-switching rule that produces estimated counterfactual choice accuracy that is significantly higher than both predicted accuracy with the true context and actual accuracy. Goals 2) and 3) represent an attempt at what could be thought of as "pseudo-BCI", whereby I attempt to make some inference about the counterfactual performance of the participants in the context that they did not actually experience.

5.2 Materials and methods

Refer to Chapter 2 for detailed description of the task, participants, EEG and pupil data collection and processing, and computational model. For consistency across all variations of the classification model, data was only trained on the subset of 24 participants for whom both EEG and pupil data was usable (see Chapter 1 for details). Furthermore, for each participants, trials with excessively noisy pupil data were removed for all variations of the model (see Chapter 1 for details). The average number of remaining trials across these 24 participants was X.

5.2.1 Multinomial classification architecture

In order to combine the multimodality of the EEG, pupil and response time (RT) data into a single classifier, a two-layer multinomial logistic regression was employed based on the methodology of (Shih et al., 2016). In this architecture, the first layer is trained on analysis windows from stimulus-, choice-, and feedback-locked data in the EEG and pupil time series to classify the four-way combination of context (reward/punishment) and choice (correct/incorrect) on a given trial (Figure 5.3). This produces a matrix of coefficients for each class and input dimension (e.g. EEG channels), which is then used to produce a weighted signal (or classification 'score') for each class on each trial. These scores are then collated into a single matrix (alongside the lower dimensional RT) to be used as the input to the second layer, which again estimates a matrix of coefficients from which probabilities are calculated using the SoftMax function (Figure 5.3).

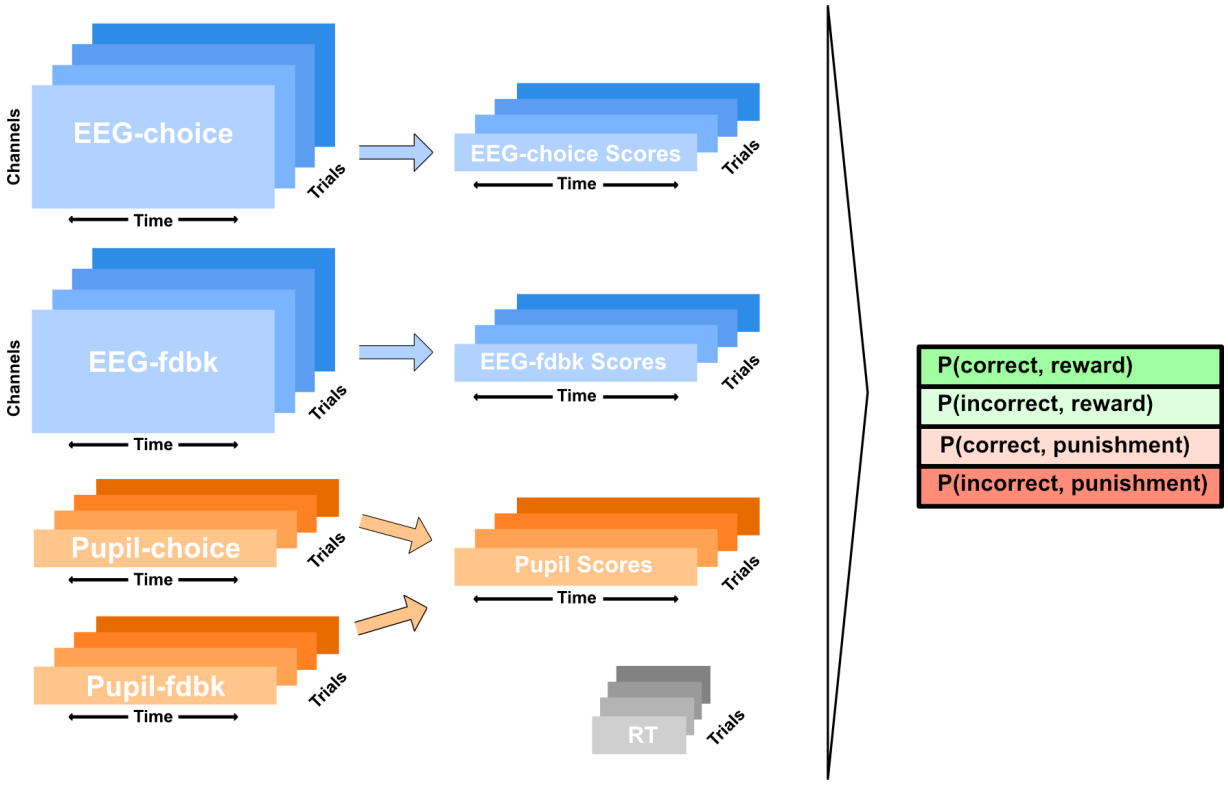

Figure 5.3: *Diagram of the 2-step multinomial regression architecture (adapted from Shih et al. (2016).* A two-layer architecture reduces the dimensionality of each input modality to a single weighted score for each time window on each trial. **Left)** Layer 1 contains the raw inputs for the choice- and feedback-locked EEG and pupil data, with input dimensions on the x-axis, time-windows on the y-axis, and trials on the z-axis. **Centre)** Layer 2 contains the unidimensional weighted outputs from Layer 1, plus raw response times in seconds. **Right)** shows the final 4-way classification output, with estimated class probabilities for each combination of context and choice accuracy.

5.2.2 EEG and pupil input data

For the first layer of classification, the data from EEG and pupillometry was applied in a series of temporal windows locked to relevant events in the course of the task. In each case, the input from the temporal window was a matrix where each row was an observation and each column was an input dimension. In both cases, the total number of observations N consisted of a multitude of samples for each trial extracted from a particular time window, such that N was equal to the product of the number of trials Nt and the size of the window in samples Ns. Accordingly, the accompanying vector of truth labels denoting which class a given trial belonged to was augmented, such that each trial label was replicated to match the number of samples, elongating the truth vector to length N.

The EEG input data was split into two broad groups: feedback-locked data and stimulus-locked data. The feedback-locked data was split into 9 non-overlapping windows of 75ms, ranging from 100ms pre-feedback to 575ms post-feedback (Figure 5.4). For this data, inputs were from trial i were used to predict context and accuracy on the subsequent trial i+1, meaning that there was no prediction for the very first trial of each block. The stimulus-locked data was in practice a combination of stimulus-locked and choice-locked windows, in order to establish some consistency relative to stimulus and choice onset given the variability in reaction time. The stimulus-locked data was split into 6 non-overlapping windows of 75ms, ranging from 100ms pre-feedback to 300ms post-feedback, and the choice-locked data was split into 9 windows ranging from 300ms pre-choice to 300ms post-choice (Figure 5.4).

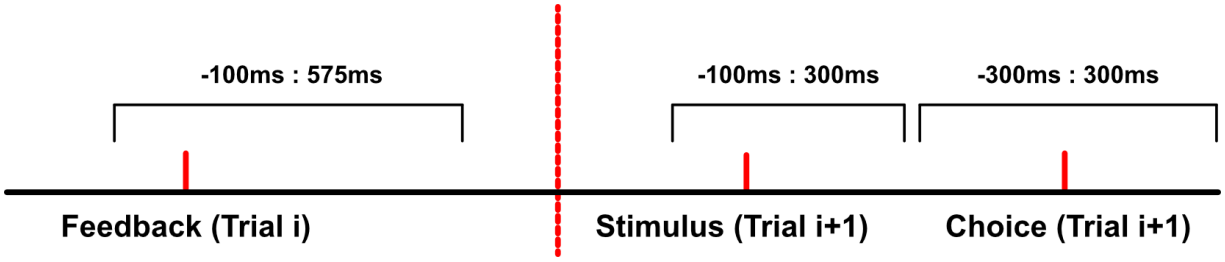

Figure 5.4: *Event-locked timings of EEG feedback and choice data for predicting trial i+1.* Feedback windows range from 100ms before to 575ms after feedback. For consistency in timing around impactful events, choice-locked data was comprised of -100ms pre-stimulus to 300ms post-stimulus and -300ms pre-choice to 300ms post-choice.

The pupil data was similarly segmented around stimulus and feedback, with some key differences. For stimulus-locked data, there was a single pre-stimulus window trained on the pupil data (normalised across the full pupil time series) from 500ms pre-stimulus to stimulus onset. There were then 8 overlapping windows of 500ms in length, with a new window onset every 250ms from stimulus onset to 1500ms post-stimulus. The post-stimulus data was baseline corrected using the 500ms of pre-stimulus data, such that it reflected the phasic deviation in response to the stimulus rather than the regular normalised diameter. The same approach was taken for feedback-locked pupil data, with one 500ms pre-feedback window and 11 baseline-corrected post-feedback windows of 500ms occurring every 250ms from feedback onset to 2500ms post-feedback. As with the EEG data, the feedback-locked windows were used to predict the class for the subsequent trial, whereas stimulus-locked windows were used to predict the class for the current trial.

5.2.3 Multinomial Logistic Regression Model

At both layers of the classification model, coefficients were estimated using the glmnet algorithm for multinomial logistic regression (Friedman et al., 2010; Tay et al., 2023). To estimate the optimal weighting across N observations for each of K classes with p input dimensions, the algorithm minimises the following loss function for multinomial outcomes using cyclical coordinate descent, which optimises over each parameter individually whilst holding the others constant, cycling through them until convergence:

$$\ell\left(\{\beta_{0k},\beta_k\}_{k=1}^{K}\right) = -\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{ik}(\beta_{0k}+x_i^T\beta_k) - \log\left(\sum_{\ell=1}^{K} e^{\beta_{0\ell}+x_i^T\beta_\ell}\right)\right) \tag{5.2}$$

Here glmnet optimises the β coefficients, where β_{0k} is the intercept term for class k, while β_k is the p-length vector of coefficients for each predictor variable for class k. The term yik is a binary indicator that is set to 1 if the observation i belongs to class k and 0 if not, while xTi is the transposed feature vector for observation i.

5.2.4 L2 regularisation

The glmnet algorithm employs an elastic net approach to regularisation during model fitting, whereby predictors are regularised using a combination of L1 (lasso) and L2 (ridge) regularisation that is balanced using a mixing parameter α. Ridge regularisation tends to shrink coefficients with high collinearity towards each other, whilst lasso favours one coefficient in particular and discards the others (Friedman et al., 2010). The elastic net penalty is given by:

$$\lambda \left[(1-\alpha) \frac{\|\beta\|_F^2}{2} + \alpha \sum_{j=1}^{p} \|\beta_j\|_1 \right] \tag{5.3}$$

Here, the Frobenius norm $\|\beta\|_F^2$ for coefficient matrix β reflects the L2 penalty, whilst the L1 penalty is given by $\|\beta\|_1$, with α controlling the relative mix of the two. The overall regularisation strength is set by λ, which sits between 0 and 1 (with 0 indicating no penalty applied).

There are different costs and benefits for selecting an α closer to 0 for L2 regularisation or 1 for L1 regularisation, and often it can be useful to compare different balances to find the best solution for the data. However, for consistency with the approach of Shih et al. (2016), and to save on the considerable training time required to optimise the first layer repeatedly, an α of 0 was selected in all cases to render the elastic net equivalent to standard L2 regularisation. When applied to the loss function, this equated to:

$$\ell\left(\{\beta_{0k}, \beta_k\}_{k=1}^{K}\right) = -\frac{1}{N} \sum_{i=1}^{N} \left(\sum_{k=1}^{K} y_{ik} (\beta_{0k} + x_i^T \beta_k) - \log \left(\sum_{\ell=1}^{K} e^{\beta_{0\ell} + x_i^T \beta_\ell} \right) \right) + \lambda \left[(1-\alpha) \frac{\|\beta\|_F^2}{2} \right] \tag{5.4}$$

The ideal value for λ was selected using a k-fold cross validation approach each time the glmfit function was called, whereby the fitting procedure was run k+1 times; the first time was used to obtain a sequence of lambda values to cycle through on each fold, and the remaining times computed the fit on the omitted fold by training on the rest of the data. A value of 10 was

selected for k as the recommended default for the algorithm. The error is accumulated across folds, and the mean and standard deviation are calculated from this value. The value of λ that gives the minimum mean cross validated error is selected.

It should be noted that a λ of 0 (i.e. no regularisation) was used for one-dimensional predictors, including pupil data at layer one and the response-time-only model at layer two.

5.2.5 Cross validation

To address the issue of overfitting in the training procedure, 5-fold cross validation was used for every window at layer one and every input combination at layer two. The folds were created using a stratified approach to account for imbalances in the frequencies of class labels, and in every case the classification scores were produced for each fold by taking the dot product of the in-fold test data and the model coefficients optimised on the out-of-fold training data. As such, all scores and resultant probabilities were produced from unseen data, minimising any risk of overfitting.

5.2.6 Class weighting and scaling

Due to the fact that participants typically show choice accuracy in the region of 60-70%, there is an inherent imbalance in the number of trials within the two 'correct' classes versus the two 'incorrect' classes. As such, when trained with equal weighting, the model has a tendency to greatly inflate probabilities for the more frequent classes. To combat this, a simple inverse frequency weighting strategy was used, which assigned a weight to each class proportional to the number of times it occurred in the truth labels, such that less frequent classes were weighted higher and the combined weights summed to 1. By inputting the corresponding weights w for each trial i into the glmnet function, the impact of each class on the loss function (before regularisation) is modified, such that it now pays more attention to underrepresented classes:

$$\ell(\{\beta_{0k},\beta_k\}_{k=1}^{K}) = -\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} w_k y_{ik}(\beta_{0k}+x_i^T\beta_k) - w_k \log\left(\sum_{\ell=1}^{K}\exp(\beta_{0\ell}+x_i^T\beta_\ell)\right)\right) \quad (5.5)$$

To check whether the weighting worked as intended, the proportion of predicted correct choices was compared to the true proportion of correct choices. A pure inverse weighting tended to overcorrect the predictions in this regard, and so a scaling factor was applied to reduce the impact of the reweighting. A simple grid-search approach found that a scaling factor of just .99 was enough to bring the predicted choice accuracy up to realistic levels, which was implemented by multiplying all inverse frequency weights by .99 and re-normalising such that they summed to 1. This weighting solution was applied to all levels of the model to obtain the final results.

5.2.7 Projected task performance

In order to simulate the range of choice accuracy we might expect to see if manipulating the task context based on the classifier probabilities, I calculated a series of projected proportions of correct choice based on different sequences of reward and punishment context. The five variations of this shown in Figure 5.5B were as follows:

1. 'Optimal' shows the proportion of predicted correct choice if the participant was automatically placed in the context with the highest relative probability of correct choice over incorrect choice.

2. 'Enhanced' shows the proportion of predicted correct choice if the context was regularly adjusted based on a switching rule designed for a BCI setting. Here, this would cause a context switch if the relative probability of correct choice over the previous 5 trials was higher for opposite context than for the current context. For example, if the participant was in the reward context, and the relative chance of correct choice from the model was on average higher for the punishment context over the prior 5 trials, then the context would

switch.

3. 'Predicted' shows the proportion of predicted correct choice based on the context that the participant actually experienced for each trial.

4. 'Actual' shows the true behavioural performance achieved by the participant, expressed in proportion of correct decisions.

5. 'Floor' is similar to 'Optimal', except it automatically places the participant in the context with the lowest relative probability of correct choice over incorrect choice.

In each case (with the exception of 'Actual'), I simply summed the number of times the model had a higher probability of correct choice for the context provided (e.g. P(correct, reward) > P(incorrect, reward)), dividing by the number of trials.

5.3 Results

5.3.1 EEG data provides the best classification accuracy

To discover the optimal combination of input data for classification, 9 different versions of the 2nd layer of the model were trained, similar to Shih et al. (2016). The classification accuracy is defined as the proportion of trials where the model correctly predicts the class label based on the highest estimated probability from layer 2 (Figure 5.5A, white bars). The top performing model was that trained on only EEG data, combining the epochs relating to the choice period and feedback period, yielding an average correct classification rate of 41.31%. The lowest performing model contained only RT data, classifying correctly just 32.03% of the time.

There are two notable jumps in performance due to the addition of predictors. The inclusion of EEG broadly increases classification accuracy, and more specifically including EEG from the time of feedback presentation is particularly valuable for classification. To formally test these

insights, I ran two vector contrast analyses on the subject-specific classification accuracy values. Firstly, to compare the added benefit of including EEG data, I ran a contrast of RT, Pupil, and Pupil+RT against EEG+RT, EEG+Pupil and EEG+Pupil+RT, weighting the first three as -1, the second three as +1, and the remaining three as 0. The result was highly significant ($t(25) = -6.83$, $p < .001$), indicating that the addition of EEG greatly improves model performances. Secondly, to compare the choice-related EEG model to the feedback-related EEG model, I ran a simple binary contrast setting EEGchoice to -1 and EEGstim to 1, and all others to 0. This was again significant ($t(25) = -8.91$, $p < .01$), demonstrating that EEG signals at the time of feedback seem to carry the most useful information for this classification problem.

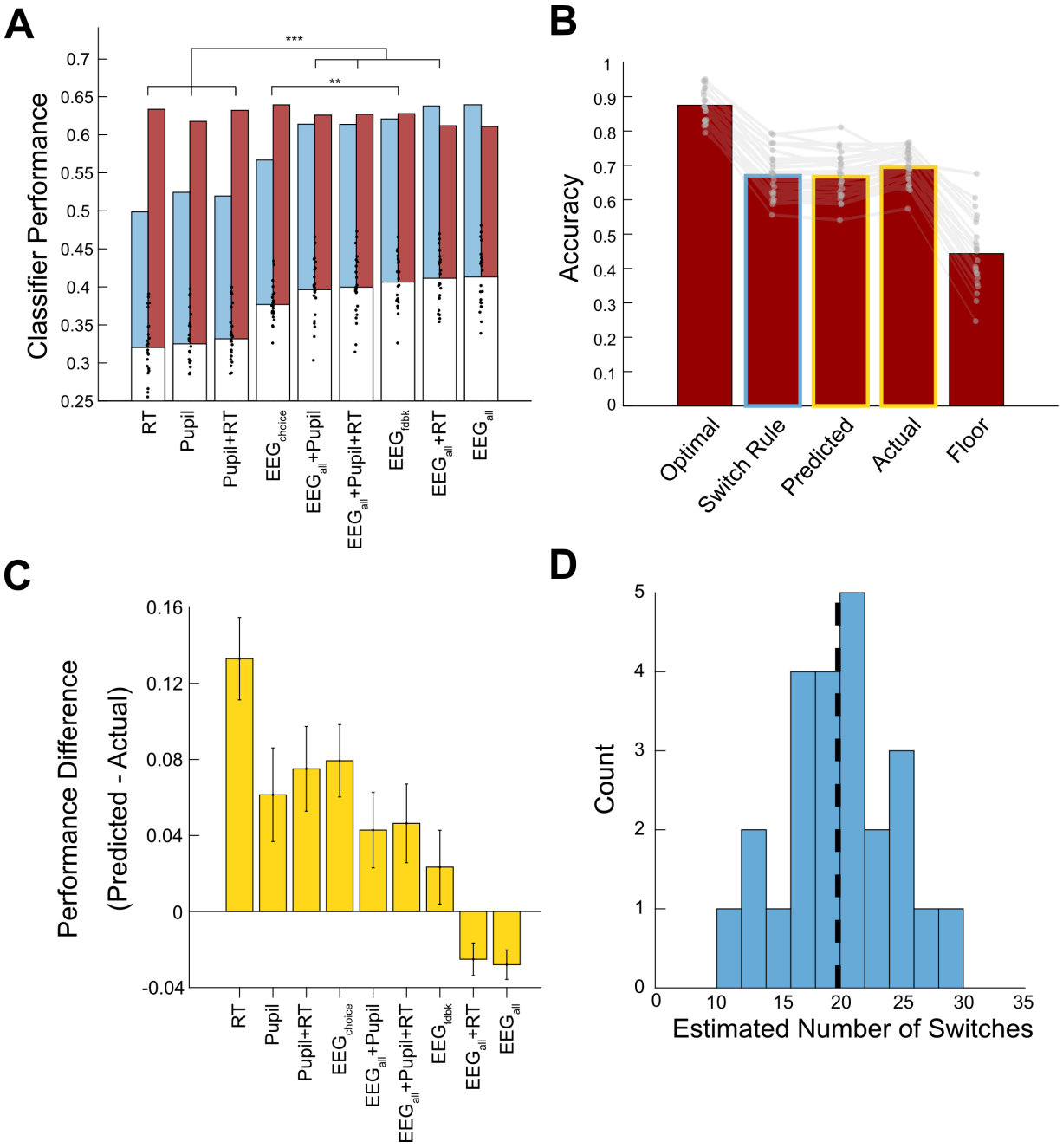

Figure 5.5: *Multinomial classifier performance metrics.* **A)** Proportion of correct classification for nine different input combinations. White bars indicate overall performance averaged across subjects i.e. proportion of correctly identified class. Dots indicate overall performance for each subject. Blue bars (left) indicate proportion of correctly identified context, defined as the context of the class with the highest estimated probability aligning with the true context. Red bars (right) indicate proportion of correctly identified choice accuracy, defined the same as context but for the correct/incorrect estimation. **B)** Model-estimated behavioural accuracy from the winning EEG model for each variation outlined in 5.2.7. Bars indicate mean predicted accuracy averaged across subjects; dots indicate individual subject estimates. **C)** Difference between predicted and actual performance accuracy (relating to yellow outlined bars in **B**). **D)** Histogram showing the distribution of estimated number of total switches across all trials based on the hypothetical switch rule outlined in 5.2.7 (relating to blue outlined bar in **B**). Dotted line indicates the mean across subjects.

Given the two-by-two classification, I also calculated the performance of predicting the correct context (regardless of outcome) and the correct outcome (regardless of context). In each case this was based on the context or outcome of the class with the highest assigned probability (Figure 5.5A, coloured bars; left-hand blue bars depict context prediction; right-hand red bars depict choice accuracy prediction). It is worth noting that there are negligible differences in these values when calculating by combining the probabilities of the two subclasses (e.g. p(correct, punishment) + p(correct, reward)). Generally, all models classified choice accuracy well, with the lowest performance coming from the EEG model with 61.09% and the highest coming from the choice-only EEG model with 63.94%. As such, the overall accuracy was largely dictated by the ability of the model to discriminate between the reward and punishment contexts, with the lowest performance here coming from the RT model at a below-chance 49.86% and the highest performance coming from the full EEG model at 63.94%.

5.3.2 Projected task performance indicates BCI potential

Using the winning EEG_{all} model, I then created a series of projections for model-predicted task performance under a variety of conditions. The goal was to simulate a variety of different series of reward/punishment contexts and project how well a participant would perform based on whether the model estimates a higher probability for a correct or incorrect choice within the context provided. As shown in the 3rd and 4th bars of Figure 5.5B, the model estimates for choice accuracy for the true context were similar at 67.70% to that actually achieved by the participant at 69.49% (although statistically different; $t(25) = -3.43$, $p < .01$). Comparisons with the other models on this metric can be seen in Figure 5.5C, showing that the best performing models tended to more closely reflect real behavioural patterns. Furthermore, as we would expect, the theoretical optimal context (as described in 5.2.6) on each trial produces a much better estimated performance than any other variation at 87.5%, and similarly the theoretical worst context produces significantly worse estimated performance at 44.34% (Figure 5.5B, 1st and 5th bars respectively).

Finally, the proposed BCI switching rule (outline in 5.2.6) shows reasonable heuristic value in the rate at which it would hypothetically trigger switches, averaging at approximately 20 total switches throughout the task at a rate of 1 every 24 trials. However, compared to the other metrics, the estimated 70.00% task accuracy gives a very small, non-significant improvement of performance over predicted accuracy given the true context ($t(24) = -0.74$, $p = .463$), it gives a significantly lower performance than actual accuracy ($t(24) = -2.61$, $p = .016$). This indicates that the results from this model are unable to provide evidence in the direction that a targeted switching rule based on live neural data would improve behavioural performance on a value-based decision-making task (Figure 5.5B, 2nd-4th bars).

5.4 Discussion

Despite the challenge of decoding four classes rather than just two, the multimodal classifier demonstrated a reasonable degree of accuracy across a number of different input combinations, with the best performing configurations exceeding 40% on average across subjects (Figure 5.5A). Furthermore, when breaking down by the binary subdivisions of context (reward versus punishment trials) and choice accuracy (correct versus incorrect trials), classifier performance exceeded 63% for the best combinations (Figure 5.5B). Though this is lower than that achieved by Shih et al. (2016), the metrics are generally encouraging given that a) the task was not designed with this type of analysis in mind, and b) the four-way classification presented a notably more challenging case to solve statistically. Furthermore, the ability of the model to produce intuitive predictions for task performance and potential enhancements with BCI helps to motivate further research in this vein. However, it should be noted that the model was unable to provide direct support for the efficacy of a targeted switching rule from this data.

5.4.1 Implications of predictive qualities of EEG, RT and pupil

Though many of our input combinations yielded good results, the predictive performance of certain variables warrants some review. Firstly, the feedback-locked EEG response period was by far the most valuable signal in the four-way classification, showing particular success in the reward-punishment aspect relative to other measures such as pupil and RT. This is both expected and encouraging in light of the links to task accuracy seen in Chapter 3, and further validates the idea that post-feedback dynamics contain meaningful and useful information about disparities in reward and punishment processing.

As the predictor with the lowest dimensionality and least rich information, RT was not necessarily expected to be a strong predictor in its own right. However, the fact that its addition did not offer much improvement in combination with other models, and that it was not present in the winning model, was somewhat surprising. RT featured in the best performing models on the visual task in Shih et al. (2016), in line with the idea that greater RT is a reflection of higher uncertainty (McDougle & Collins, 2021; Richer & Beatty, 1987) and lower confidence (Zylberberg et al., 2016), variables that are directly related to choice accuracy. Similarly, pupil dilation has been robustly linked to surprise (de Gee et al., 2021; Filipowicz et al., 2020; Preuschoff et al., 2011)and uncertainty (Brunyé & Gardony, 2017; Kawaguchi et al., 2018; Nassar et al., 2012; Urai et al., 2017). However, despite the fact that both of these predictors had poor overall classification accuracy absent any EEG signals, they did provide some of the highest performance for the correct-incorrect dimension, which is in line with the above parallels to parameters relevant to choice accuracy. In this light, it is possible that the pupil and RT signals are relevant for this aspect of choice behaviour but less so for reward-punishment distinctions – at least relative to EEG.

5.4.2 Evaluation of the multinomial classifier

It is also possible that the contributions from RT and pupillometry were limited by the methodology used in the classification approach. The choice of multinomial classifier was motivated by two main factors: firstly, it bore a degree of similarity to the linear methods employed in the preceding chapters; secondly, it had been directly implemented in prior work to predict task accuracy in a research group with multiple successful BCI endeavours (Shih et al., 2016). This combination offered an appealing starting point for the proof-of-concept aim of this chapter as a means to evaluate whether the prediction required for a BCI paradigm would be at all plausible. Though the results presented here do provide support for plausibility, there are a number of ways in which the classification methodology could be potentially developed and improved for more effective neurofeedback.

Firstly, the use of a linear classifier places certain assumptions on the data that might not necessarily hold. In a four-way classification, it is entirely possible that the optimal hyperplanes for separating classes within multidimensional data are curved rather than linear – something the multinomial logistic regression employed in my model would be unable to account for. The most simple way to test this would be to replicate the process using identical architecture with a non-linear model in the place of the multinomial regression, such as a neural network or non-linear support vector machine. This would offer the opportunity to compare model performance in the most direct way possible, offering insight as to whether the relationships between pupillometry, response time, EEG signals and task context and performance are in fact nonlinear.

Taking this idea further, there are a multitude of more sophisticated techniques available from machine learning literature that can leverage a variety of features of multimodal data to maximally extract useful information. For instance, there are modern deep learning architectures that allow for effective processing of high-dimensional multimodal data from neural and physiological measures for use cases such as sleep-stage classification (Chambon et al., 2018), the identification of epileptic seizures (Samiee et al., 2015), or emotion categorisation (Zhang et al., 2018). In the latter case, a deep multimodal neurophysiological transformer has recently

demonstrated comparable or superior performance to the best available competitors on fewer samples for the classification of affective valence and arousal (Koorathota et al., 2022). Briefly, this model takes advantage of the ability of the transformer architecture to map long-term temporal and contextual dependencies with the groundbreaking self-attention mechanism (Vaswani et al., 2017). While on the cutting edge end of neurophysiological classification, this example demonstrates the increasingly sophisticated techniques for extracting maximal value from available data.

5.4.3 Limitations of pseudo-BCI

In addition to improvements available in the classification approach, there are also notable methodological limitations of the retrospective analysis used that significantly restrict the strength of conclusion that can be drawn. The primary goal of the analysis was to demonstrate a reasonable level of predictive capability; however a secondary "pseudo-BCI" goal was to make a case that the output probabilities plausibly map onto counterfactual performance of participants in the non-experienced context on each trial, such that some rough inferences can be made about the viability of the experimental design for a live BCI implementation. This yielded mixed results: whilst the hypothetical switch rates seemed to fall in a reasonable range, and the predicted accuracy was slightly better in the switch-rule condition than the true context condition, it was still significantly lower than the actual behavioural accuracy of the participants. It is worth noting that this seemed to reflect a trend of predicted accuracy generally underestimating in the winning model, and I believe the results are reasonable enough so as not to rule out further pursuit in a closed-loop BCI experiment.

For a more affirmative case to be made, however, there are issues of causal understanding that are not adequately mapped out in the present example. In fields such as the healthcare (Bica et al., 2020) and marketing (Bottou et al., 2013), counterfactual prediction has received a great deal of focus due to the value of understanding how outcomes might have been affected given different conditions for risk identification and strategy development. In order to make effective

inferences about counterfactual outcomes, though, robust causal understandings of the relationships between variables is required (Hartford et al., 2017; Prosperi et al., 2020). To borrow an example from Hartford and colleagues (2017), imagine an airline company trying to determine how much to raise prices based on demand from customers. Without a causal understanding of the effects of variables such as seasonal interest in different locations, school holidays, current events etc., a purely data-driven approach could easily come to incorrect conclusions about optimal pricing. Returning to the case of the pseudo-BCI in the current chapter, the results should accordingly be taken with a high degree of caution, and require either a greater degree of understanding about the causal relationships between the neural and pupil signals with respect to reward, punishment, and task performance, or alternatively a true BCI test where counterfactual inference is not needed.

Chapter 6

General Discussion

Exploring motivational sensitivity to reward and punishment is a fascinating endeavour, as it is something that each of us has a personally emotive connection to. Every one of us has experienced the strong internal motivation associated with some especially rewarding goal, whether that be pursuing a love interest or winning a sports competition. Equally, we all have been faced with the strong avoidance drive associated with scary or stressful situations, such as defending a PhD thesis. Furthermore, there is a rich variety of individual experience in relation to these two types of reinforcer that provide us with great diversity of interaction and specialisation. From a performance perspective, this presents both a great challenge in managing the environment to get the best out of someone, but also an opportunity to let people shine in the roles that suit them the best. Consider the dynamic, unpredictable flare of the attacking winger in comparison to the steady, no-nonsense centre-back on a football pitch. Each will have their unique sensitivities to risk and reward, and in both cases this is necessary to carry out their role effectively.

This thesis examined in detail the individual differences in respective sensitivity to reward and punishment, using insights from EEG and pupillometry. It explored the relationship between these signals and individual differences in behavioural propensities and personality features. And finally, it touched on the possibility of leveraging information about this gathered from neural signals to optimally cater a decision-making environment to suit the needs of the

individual. The main insights gleaned from this work will hopefully set the scene for a number of logical next steps.

6.1 Further insights into punishment learning

One obvious area that calls for more targeted investigation is the late component of the putative two-component EEG response. Due to compelling evidence highlighted in Chapter 1 that linked the early component to the punishment-sensitive pupil response, and the hypotheses that were subsequently generated, the analyses in Chapters 3 and 4 were targeted on this aspect of the post-feedback signal. However, as demonstrated in Fouragnan et al. (2015), the two signals are mechanistically coupled, meaning that we should not rule out the possibility that further investigation into punishment effects in the late signal would be fruitful. The amygdala was one of the areas found to be associated with this late signal in EEG-fMRI work (Fouragnan et al., 2015, 2017), which has been purported to play an important role in punishment learning (Bechara et al., 1995; De Martino et al., 2006, 2010; Delgado et al., 2008, 2011; Metereau & Dreher, 2013; Yacubian et al., 2006). Further research targeting specific hypothesis in relation to this signal may therefore be valuable.

An additional limitation of the reversal learning paradigm employed in this thesis is the constant level of volatility in the environment. Volatility refers to the degree to which underlying outcome contingencies change over time, while stochasticity is the degree of uncertainty in the outcome contingencies. For example, in the task used here, an increase in volatility would translate to an increase in the frequency of reversals, whereas an increase in stochasticity would mean a set of outcome contingencies closer to 50/50. These two environmental parameters have independent consequences for behavioural strategy, and have been shown to likely be estimated in parallel by actors during learning (Piray & Daw, 2021). Furthermore, estimation of task uncertainty (Brunyé & Gardony, 2017; Kawaguchi et al., 2018; Urai et al., 2017) and volatility (Binetti et al., 2017; Kloosterman et al., 2015; Nassar et al., 2012) have both been shown to have significant interplay with pupil dynamics in reinforcement learning. By only examining a single

volatility and stochasticity level, there were perhaps meaningful effects in relation to this that were missed. Future research would benefit from using a range of decision-making tasks that allow for these sorts of dynamics to reveal new comparisons in reward- and punishment-related signals.

Finally, although some effort was made to investigate individual differences in relation to reinforcement sensitivity, there could more subtle interactive effects that are difficult to pick up on without more targeted investigation. For example, there is a large body of evidence that suggests that acute stress is an influential factor in regulating important decision-making variables such as risk taking (Porcelli & Delgado, 2017; Starcke & Brand, 2016) and cortico-striatal dopamine release (Nagano-Saito et al., 2013; Vaessen et al., 2015; van Ruitenbeek et al., 2021). Furthermore, stress responses in a learning context are almost certainly heavily influenced by personality factors, for example in the degree to which risk taking is altered (Lauriola & Weller, 2018; Lauriola et al., 2014). There are two approaches that could be taken to help elucidate factors such as these. Firstly, simple measures associated with stress responses could be added, such as the galvanic skin response or heart rate, without causing much additional cost or difficulty to data collection. This could allow for additional analyses into the interactions between stress response, reward and punishment, and personality factors. Secondly, specific task designs could be employed to artificially induce stress responses for a more causal analysis, such as the cold pressor test.

6.2 Closing the loop

The four-way classification model presented in Chapter 5 offers a promising and novel approach to the behaviour optimisation challenge, integrating perspectives from neuroscience and computer science to reframe traditional neurofeedback and BCI approaches. The model output was able to classify the four-way combination of reward, punishment, correct and incorrect trials with moderately good accuracy, performing best when trained on choice- and feedback-locked EEG signals without RT and pupil data included. Importantly, model-predicted choice behaviour

and accuracy was reasonably close to that displayed behaviourally by participants, meaning that the approach was not ruled out as implausible for application in a live BCI setting. However, the blocked nature of the reversal-learning task from which the data was obtained was not a very realistic representation of the dynamic context-switching that would occur in a true BCI variation. To address this, an intermediate step using the same methodology in a more dynamic task would be a useful improvement on the existing blocked structure of Chapter 5. Two changes to the paradigm that offer this whilst maintaining a good degree of conceptual consistency with the reversal learning task could be: a) the implementation of dynamic contingency drift via markov random walk as opposed to the binary reversals; b) the expansion from a two-armed to multi-armed bandit. These adaptations have been used successfully to investigate relevant effects such as the relationship between pupil dilation and exploration (Jepma & Nieuwenhuis, 2011), and would certainly bring things closer to the eventual BCI goal. Thirdly, I believe the BCI approach would have the most interesting application in a truly ecologically valid task, to see whether the principles explored in reversal tasks or multi-armed bandits generalise to realistic use-cases. There are certain preconditions that present a challenge for finding a suitable testing ground:

a) a known optimal decision strategy to use as ground truth when training the model b) symmetric play for reward and punishment scenarios c) an increased degree of realism and complexity d) parallels to the original initial risky choice paradigm

An option that meets all of these criteria is the popular casino card game known as BlackJack or 21, a binary choice task which has the benefit of being solved (Baldwin et al., 1956) and so has a readily available ground truth for correct decision. There have already been several attempts to use machine learning to emulate realistic BlackJack playing strategies (Kakvi, 2009; Kendall & Smith, 2003; Schiller & Gobet, 2012; Srinivasaiah et al., 2024), one electrophysiological study that examines ERPs in response to feedback during play (Hewig et al., 2007), and one investigation into trait impulsivity in relation to behavioural patterns in the game (Webster & Crysel, 2012). Thus far, however, there have been no attempts to combine these approaches for either modelling or performance optimisation. As such, this presents an appealing medium to advance the ecological validity of the proposed framework.

6.3 Beyond reward and punishment

In Chapter 5, I laid out a framework for integrating physiological signals to model how the relationships between certain task-irrelevant environmental parameters affect an individual's success on a target outcome. The framework represents this as an optimisation problem. This is an attempt to open the door to a range of machine learning approaches that could add unique value to traditional neurofeedback and BCI methodologies prevalent in neuroscience.

Though models used in computer science literature are not based on known mechanisms in the brain, this does not necessarily mean that the prediction-focused approach is not compatible with desires for interpretation. Neuroscience is far from fully understanding decision making on a mechanistic level, and whilst we may have valuable insights into specific elements such as the role of the reward prediction error (Schultz, 2016), most everyday tasks are complex enough to exceed the explanatory power of theory. There are cases where simple generalisability-focused machine learning algorithms such as SARSA outperform models that are specifically designed to emulate human processes (Schiller & Gobet, 2012).

As an applied example, a recent paper leveraged machine learning to narrow down successful features of popular neuroscientific models of human choice under risk (Peterson et al., 2021). Using a hierarchical neural network with different groups of model assumptions at each level (e.g. models that incorporate alternative option value versus models that do not), they identified which kinds of decisions yielded the most improvement when different parameters were added. For instance, dominated gambles (i.e. where all outcomes of one prospect are better than the other) yielded the highest improvement in prediction when adding context-related parameters, suggesting that people may use different decision strategies for different types of gamble. From here, they built a new 'mixture of theories' model that incorporated principles from a range of other models and showed that this outperformed the predictive capacity of all other traditional models. In short, applying deep learning to the existing 'plausible parameter space' yielded new insights into the underlying processes in human risky choice. This type of approach could potentially be applied to the case of narrowing down and identifying environmental variables

– such as rewarding or punishing feedback – that possess a meaningful relationship to task performance. This could then open the door for a multivariate expansion of the BCI paradigm introduced in Chapter 5 that incorporates more than one parameter to manipulate; for instance, it could be possible to model the interaction between state arousal and reinforcement sensitivity to add an additional dimension to the optimisation process.

6.4 Conclusions

Punishment is an integral yet sometimes overlooked aspect of reinforcement learning, and there are still many mysteries that remain unexplored. In this thesis, I have provided evidence that an early salience response post-feedback may play a key role in modulating the individual response dynamics to appetitive and aversive feedback. I have made a case that this mechanism is compatible with existing accounts of reward and punishment learning, and I have demonstrated a meaningful link between this signal and measures of behavioural performance on a reinforcement learning task. I believe that this perspective on reward and punishment dynamics provides a novel and interesting addition to the current understanding of the topic.

I have also presented a framework through which these insights could potentially be leveraged to improve task performance in a targeted manner. Though in the very early stages of implementation, I believe I have laid the very early foundations for what could prove to be a valuable lens through which to apply existing BCI and neurofeedback techniques. It is my hope that this work can be built upon to realise new and exciting applications of reinforcement learning theory in settings that are meaningful to life in the real world.

References

Aluja, A., Blanch, A., Blanco, E., & Balada, F. (2015). Affective modulation of the startle reflex and the reinforcement sensitivity theory of personality: The role of sensitivity to reward. *Physiology & Behavior*, *138*, 332–339. https://doi.org/10.1016/j.physbeh.2014.09.009

Anderson, A. K., & Phelps, E. A. (2001). Lesions of the human amygdala impair enhanced perception of emotionally salient events [Number: 6835 Publisher: Nature Publishing Group]. *Nature*, *411*(6835), 305–309. https://doi.org/10.1038/35077083

Argyelan, M., Herzallah, M., Sako, W., DeLucia, I., Sarpal, D., Vo, A., Fitzpatrick, T., Moustafa, A. A., Eidelberg, D., & Gluck, M. (2018). Dopamine modulates striatal response to reward and punishment in patients with parkinson's disease: A pharmacological challenge fMRI study. *NeuroReport*, *29*(7), 532. https://doi.org/10.1097/WNR.0000000000000970

Arns, M., Kleinnijenhuis, M., Fallahpour, K., & Breteler, R. (2008). Golf performance enhancement and real-life neurofeedback training using personalized event-locked EEG profiles [Publisher: Routledge _eprint: https://doi.org/10.1080/10874200802149656]. *Journal of Neurotherapy*, *11*(4), 11–18. https://doi.org/10.1080/10874200802149656

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Averbeck, B. B. (2015). Theory of choice in bandit, information sampling and foraging tasks [Publisher: Public Library of Science]. *PLOS Computational Biology*, *11*(3), e1004164. https://doi.org/10.1371/journal.pcbi.1004164

Averbeck, B. B., Djamshidian, A., O'Sullivan, S. S., Housden, C. R., Roiser, J. P., & Lees, A. J. (2013). Uncertainty about mapping future actions into rewards may underlie performance on multiple measures of impulsivity in behavioral addiction: Evidence from parkinson's disease. *Behavioral neuroscience*, *127*(2), 245–255. https://doi.org/10.1037/a0032079

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration [Publisher: Elsevier]. *Neuron*, *73*(3), 595–607. https://doi.org/10.1016/j.neuron.2011.12.025

Balconi, M., & Crivelli, D. (2010). FRN and p300 ERP effect modulation in response to feedback sensitivity: The contribution of punishment-reward system (BIS/BAS) and behaviour identification of action. *Neuroscience Research*, *66*(2), 162–172. https://doi.org/10.1016/j.neures.2009.10.011

Balleine, B. W., Leung, B. K., & Ostlund, S. B. (2011). The orbitofrontal cortex, predicted value, and choice [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.2011.06270.x]. *Annals of the New York Academy of Sciences*, *1239*(1), 43–50. https://doi.org/10.1111/j.1749-6632.2011.06270.x

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., & Palminteri, S. (2018). Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences [Publisher: Nature Publishing Group]. *Nature Communications*, *9*(1), 4503. https://doi.org/10.1038/s41467-018-06781-2

Bavard, S., & Palminteri, S. (2023). The functional form of value normalization in human reinforcement learning (T. Kahnt, M. J. Frank, & N. Garrett, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *12*, e83891. https://doi.org/10.7554/eLife.83891

Bavard, S., Rustichini, A., & Palminteri, S. (2021). Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning [Publisher: American Association for the Advancement of Science]. *Science Advances*, *7*(14). https://doi.org/10.1126/SCIADV.ABE0340

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal [Publisher: Elsevier]. *Neuron*, *47*(1), 129–141.

Bayer, H. M., Lau, B., & Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate [Publisher: American Physiological Society]. *Journal of Neurophysiology*, *98*(3), 1428–1439. https://doi.org/10.1152/jn.01140.2006

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., & Damasio, A. R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans [Publisher: American Association for the Advancement of Science]. *Science*, *269*(5227), 1115–1118. https://doi.org/10.1126/science.7652558

Bellebaum, C., Polezzi, D., & Daum, I. (2010). It is less than you expected: The feedback-related negativity reflects violations of reward magnitude expectations. *Neuropsychologia*, *48*(11), 3343–3350. https://doi.org/10.1016/j.neuropsychologia.2010.07.023

Berger, U. (2007). Brown's original fictitious play. *Journal of Economic Theory*, *135*(1), 572–578. https://doi.org/10.1016/j.jet.2005.12.010

Bica, I., Jordon, J., & van der Schaar, M. (2020). Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, *33*, 16434–16445. Retrieved September 19, 2024, from https://proceedings.neurips.cc/paper_files/paper/2020/hash/bea5955b308361a1b07bc55042e25e54-Abstract.html

Billings, D., Papp, D., Schaeffer, J., & Szafron, D. (1998). Opponent modeling in poker.

Binetti, N., Harrison, C., Mareschal, I., & Johnston, A. (2017). Pupil response hazard rates predict perceived gaze durations [Publisher: Nature Publishing Group]. *Scientific Reports*, *7*(1), 3969. https://doi.org/10.1038/s41598-017-04249-9

Blair, K., Marsh, A. A., Morton, J., Vythilingam, M., Jones, M., Mondillo, K., Pine, D. C., Drevets, W. C., & Blair, J. R. (2006). Choosing the lesser of two evils, the better of two goods: Specifying the roles of ventromedial prefrontal cortex and dorsal anterior cin-

gulate in object choice [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *26*(44), 11379–11386. https://doi.org/10.1523/JNEUROSCI.1640-06.2006

Blanchard, D., Griebel, G., & Blanchard, R. (2001). Mouse defensive behaviors: Pharmacological and behavioral assays for anxiety and panic. *Neuroscience & Biobehavioral Reviews*, *25*(3). https://doi.org/https://doi.org/10.1016/S0149-7634(01)00009-4

Blanchard, D., Griebel, G., & Blanchard, R. (2003). The mouse defense test battery: Pharmacological and behavioral assays for anxiety and panic. *Elsevier*. Retrieved June 29, 2022, from https://www.sciencedirect.com/science/article/pii/S0014299903012767

Blanchard, J., Hebert, A., Ferrari, P., Palanza, P., Figueira, R., Blanchard, D., & Parmigiani, S. (1998). Defensive behaviors in wild and laboratory (swiss) mice: The mouse defense test battery. *Physiology & Behavior*, *65*(2). https://doi.org/https://doi.org/10.1016/S0031-9384(98)00012-2

Bódi, N., Kéri, S., Nagy, H., Moustafa, A., Myers, C. E., Daw, N., Dibó, G., Takáts, A., Bereczki, D., & Gluck, M. A. (2009). Reward-learning and the novelty-seeking personality: A between-and within-subjects study of the effects of dopamine agonists on young parkinson's patients*. *A JOURNAL OF NEUROLOGY*. https://doi.org/10.1093/brain/awp094

Bordalo, P., Gennaioli, N., & Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics*, *127*(3), 1243–1285. https://doi.org/10.1093/qje/qjs018

Bordalo, P., Gennaioli, N., & Shleifer, A. (2022). Salience [_eprint: https://doi.org/10.1146/annurev-economics-051520-011616]. *Annual Review of Economics*, *14*(1), 521–544. https://doi.org/10.1146/annurev-economics-051520-011616

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., & Snelson, E. (2013, July 27). Counterfactual reasoning and learning systems. https://doi.org/10.48550/arXiv.1209.2355

Breton-Provencher, V., Drummond, G. T., Feng, J., Li, Y., & Sur, M. (2022). Spatiotemporal dynamics of noradrenaline during learned behaviour [Publisher: Nature Publishing Group]. *Nature*, *606*(7915), 732–738. https://doi.org/10.1038/s41586-022-04782-2

Brischoux, F., Chakraborty, S., Brierley, D. I., & Ungless, M. A. (2009). Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli [Publisher: Proceedings of the

National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *106*(12), 4894–4899. https://doi.org/10.1073/pnas.0811507106

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010a). Distinct tonic and phasic anticipatory activity in lateral habenula and dopamine neurons. *Neuron*, *67*(1), 144–155. https://doi.org/10.1016/j.neuron.2010.06.016

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010b). Dopamine in motivational control: Rewarding, aversive, and alerting. *Neuron*, *68*(5), 815–834. https://doi.org/10.1016/j.neuron.2010.11.022

Brunyé, T. T., & Gardony, A. L. (2017). Eye tracking measures of uncertainty during perceptual decision making [Publisher: Elsevier B.V.]. *International Journal of Psychophysiology*, *120*, 60–68. https://doi.org/10.1016/j.ijpsycho.2017.07.008

Buffalari, D. M., & Grace, A. A. (2007). Noradrenergic modulation of basolateral amygdala neuronal activity: Opposing influences of -2 and receptor activation [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *27*(45), 12358–12366. https://doi.org/10.1523/JNEUROSCI.2007-07.2007

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning [Place: US Publisher: American Psychological Association]. *Psychological Review*, *58*(5), 313–323. https://doi.org/10.1037/h0054388

Caplin, A., & Glimcher, P. W. (2013, September). 1) basic methods from neoclassical economics. In *Neuroeconomics: Decision making and the brain: Second edition* (pp. 3–17). Elsevier Inc. https://doi.org/10.1016/B978-0-12-416008-8.00001-2

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance [Publisher: American Association for the Advancement of Science]. *Science*, *280*(5364), 747–749. https://doi.org/10.1126/science.280.5364.747

Carvalheiro, J., Conceição, V. A., Mesquita, A., & Seara-Cardoso, A. (2022). Psychopathic traits and reinforcement learning under acute stress [Publisher: John Wiley & Sons, Ltd]. *Journal of Personality*, *90*(3), 393–404. https://doi.org/10.1111/JOPY.12673

Carvalheiro, J., & Philiastides, M. G. (2023). Distinct spatiotemporal brainstem pathways of outcome valence during reward- and punishment-based learning [Publisher: Elsevier]. *Cell Reports*, *42*(12). https://doi.org/10.1016/j.celrep.2023.113589

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *67*(2), 319–333. https://doi.org/10.1037/0022-3514.67.2.319

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

Cavanagh, J. F., Masters, S. E., Bath, K., & Frank, M. J. (2014). Conflict acts as an implicit cost in reinforcement learning [Publisher: Nature Publishing Group]. *Nature Communications*, *5*(1), 5394. https://doi.org/10.1038/ncomms6394

Cavanagh, J. F., Mueller, A. A., Brown, D. R., Janowich, J. R., Story-Remer, J. H., Wegele, A., & Richardson, S. P. (2017). Cognitive states influence dopamine-driven aberrant learning in parkinson's disease. *Cortex*, *90*, 115–124. https://doi.org/10.1016/j.cortex.2017.02.021

Chambon, S., Galtier, M. N., Arnal, P. J., Wainrib, G., & Gramfort, A. (2018). A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series [Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *26*(4), 758–769. https://doi.org/10.1109/TNSRE.2018.2813138

Chandler, D., & Waterhouse, B. (2012). Evidence for broad versus segregated projections from cholinergic and noradrenergic nuclei to functionally and anatomically discrete subregions of prefrontal cortex. *Frontiers in Behavioral Neuroscience*, *6*. Retrieved September 19, 2023, from https://www.frontiersin.org/articles/10.3389/fnbeh.2012.00020

Chang, Y.-H., Chen, H.-J., Barquero, C., Tsai, H. J., Liang, W.-K., Hsu, C.-H., Muggleton, N. G., & Wang, C.-A. (2024). Linking tonic and phasic pupil responses to p300 amplitude in an emotional face-word stroop task [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.14 *Psychophysiology*, *61*(4), e14479. https://doi.org/10.1111/psyp.14479

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience*, *23*(4), 936–946. https://doi.org/10.1162/jocn.2010.21456

Cheng, M.-Y., Huang, C.-J., Chang, Y.-K., Koester, D., Schack, T., & Hung, T.-M. (2015). Sensorimotor rhythm neurofeedback enhances golf putting performance [Section: Journal of Sport and Exercise Psychology]. https://doi.org/10.1123/jsep.2015-0166

Christoph, G. R., Leonzio, R. J., & Wilcox, K. S. (1986). Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *6*(3), 613–619. https://doi.org/10.1523/JNEUROSCI.06-03-00613.1986

Citi, L., Poli, R., Cinel, C., & Sepulveda, F. (2008). P300-based BCI mouse with genetically-optimized analogue control [Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *16*(1), 51–61. https://doi.org/10.1109/TNSRE.2007.913184

Clark, L., Li, R., Wright, C. M., Rome, F., Fairchild, G., Dunn, B. D., & Aitken, M. R. F. (2012). Risk-avoidant decision making increased by threat of electric shock [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.2012.01454.x]. *Psychophysiology*, *49*(10), 1436–1443. https://doi.org/10.1111/j.1469-8986.2012.01454.x

Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, *9*(9), 1289–1302. https://doi.org/10.1093/scan/nst106

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 [Publisher: Nature Publishing Group]. *Nature*, *482*(7383), 85–88. https://doi.org/10.1038/nature10754

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 [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

Cohen, M. X., Wilmes, K. A., & van de Vijver, I. (2011). Cortical electrophysiological network dynamics of feedback learning. *Trends in Cognitive Sciences*, *15*(12), 558–566. https://doi.org/10.1016/j.tics.2011.10.004

Coizet, V., Dommett, E. J., Redgrave, P., & Overton, P. G. (2006). Nociceptive responses of midbrain dopaminergic neurones are modulated by the superior colliculus in the rat. *Neuroscience*, *139*(4), 1479–1493. https://doi.org/10.1016/j.neuroscience.2006.01.030

Combrisson, E., Basanisi, R., Gueguen, M. C. M., Rheims, S., Kahane, P., Bastin, J., & Brovelli, A. (2023). Neural interactions in the human frontal cortex dissociate reward and punishment learning [Publisher: eLife Sciences Publications Limited]. *eLife*, *12*. https://doi.org/10.7554/eLife.92938.1

Cools, R., Frank, M. J., Gibbs, S. E., Miyakawa, A., Jagust, W., & D'Esposito, M. (2009). Striatal dopamine predicts outcome-specific reversal learning and its sensitivity to dopaminergic drug administration [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *29*(5), 1538–1543. https://doi.org/10.1523/JNEUROSCI.4467-08.2009

Corr, P. J. (2004). Reinforcement sensitivity theory and personality [Publisher: Pergamon]. *Neuroscience & Biobehavioral Reviews*, *28*(3), 317–332. https://doi.org/10.1016/J.NEUBIOREV.2004.01.005

Corr, P. J., & Cooper, A. J. (2016). The reinforcement sensitivity theory of personality questionnaire (RST-PQ): Development and validation [Publisher: American Psychological Association Inc.]. *Psychological Assessment*, *28*(11), 1427–1440. https://doi.org/10.1037/PAS0000273

Costa, V. D., & Averbeck, B. B. (2020). Primate orbitofrontal cortex codes information relevant for managing explore–exploit tradeoffs [Publisher: Society for Neuroscience Section: Research Articles]. *Journal of Neuroscience*, *40*(12), 2553–2561. https://doi.org/10.1523/JNEUROSCI.2355-19.2020

Costa, V. D., Mitz, A. R., & Averbeck, B. B. (2019). Subcortical substrates of explore-exploit decisions in primates [Publisher: Elsevier]. *Neuron*, *103*(3), 533–545.e5. https://doi.org/10.1016/j.neuron.2019.05.017

Da Silva, J. C., & De Souza, M. L. (2021). Neurofeedback training for cognitive performance improvement in healthy subjects: A systematic review [Publisher: Educational Publishing Foundation]. *Psychology & Neuroscience*, *14*(3), 262–279. https://doi.org/10.1037/pne0000261

Danna, C., Shepard, P., & Elmer, G. (2013). The habenula governs the attribution of incentive salience to reward predictive cues. *Frontiers in Human Neuroscience*, *7*. Retrieved October 22, 2023, from https://www.frontiersin.org/articles/10.3389/fnhum.2013.00781

Daughters, S. B., Gorka, S. M., Matusiewicz, A., & Anderson, K. (2013). Gender specific effect of psychological stress and cortisol reactivity on adolescent risk taking. *Journal of Abnormal Child Psychology*, *41*(5), 749–758. https://doi.org/10.1007/s10802-013-9713-4

Davidson, A., Billings, D., Schaeffer, J., & Szafron, D. (2000). Improved opponent modeling in poker.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans [Publisher: Nature Publishing Group]. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Dawe, S., Gullo, M. J., & Loxton, N. J. (2004). Reward drive and rash impulsiveness as dimensions of impulsivity: Implications for substance misuse. *Addictive Behaviors*, *29*(7), 1389–1405. https://doi.org/10.1016/j.addbeh.2004.06.004

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 [Publisher: eLife Sciences Publications Ltd]. *eLife*, *6*. https://doi.org/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. *Cerebral Cortex*, *31*(7), 3565–3578. https://doi.org/10.1093/cercor/bhab032

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *111*(5), E618–E625. https://doi.org/10.1073/pnas.1317557111

de Gee, J. W., Tsetsos, K., Schwabe, L., Urai, A. E., McCormick, D., McGinley, M. J., & Donner, T. H. (2020). Pupil-linked phasic arousal predicts a reduction of choice bias across species and decision domains [Publisher: eLife Sciences Publications Ltd]. *eLife*, *9*, 1–25. https://doi.org/10.7554/eLife.54014

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

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain [Publisher: American Association for the Advancement of Science]. *Science*, *313*(5787), 684–687. https://doi.org/10.1126/science.1128356

De Pascalis, V., Fracasso, F., & Corr, P. J. (2017a). The behavioral approach system and augmenting/reducing in auditory event-related potentials during emotional visual stimulation. *Biological Psychology*, *123*, 310–323. https://doi.org/10.1016/j.biopsycho.2016.10.015

De Pascalis, V., Fracasso, F., & Corr, P. J. (2017b). Personality and augmenting/reducing (a/r) in auditory event-related potentials (ERPs) during emotional visual stimulation [Publisher: Nature Publishing Group]. *Scientific Reports*, *7*(1), 41588. https://doi.org/10.1038/srep41588

De Pascalis, V., Scacchia, P., Sommer, K., & Checcucci, C. (2019). Psychopathy traits and reinforcement sensitivity theory: Prepulse inhibition and ERP responses. *Biological Psychology*, *148*, 107771. https://doi.org/10.1016/j.biopsycho.2019.107771

De Pascalis, V., Sommer, K., & Scacchia, P. (2018). Resting frontal asymmetry and reward sensitivity theory motivational traits [Publisher: Nature Publishing Group]. *Scientific Reports*, *8*(1), 13154. https://doi.org/10.1038/s41598-018-31404-7

De Pascalis, V., Varriale, V., & D'Antuono, L. (2010). Event-related components of the punishment and reward sensitivity. *Clinical Neurophysiology*, *121*(1), 60–76. https://doi.org/10.1016/j.clinph.2009.10.004

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., & Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum [Publisher: American

Physiological Society]. *Journal of Neurophysiology*, *84*(6), 3072–3077. https://doi.org/10.1152/jn.2000.84.6.3072

Delgado, M. R., Jou, R. L., & Phelps, E. A. (2011). Neural systems underlying aversive conditioning in humans with primary and secondary reinforcers [Publisher: Frontiers]. *Frontiers in Neuroscience*, *5*. https://doi.org/10.3389/fnins.2011.00071

Delgado, M. R., Li, J., Schiller, D., & Phelps, E. A. (2008). The role of the striatum in aversive learning and aversive prediction errors [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3787–3800. https://doi.org/10.1098/rstb.2008.0161

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

Depue, R. A., & Collins, P. F. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion [Publisher: Cambridge University Press]. *Behavioral and Brain Sciences*, *22*(3), 491–517. https://doi.org/10.1017/S0140525X99002046

Djamshidian, A., O'Sullivan, S. S., Wittmann, B. C., Lees, A. J., & Averbeck, B. B. (2011). Novelty seeking behaviour in parkinson's disease. *Neuropsychologia*, *49*(9), 2483–2488. https://doi.org/10.1016/j.neuropsychologia.2011.04.026

Egner, T., & Gruzelier, J. H. (2004). EEG biofeedback of low beta band components: Frequency-specific effects on variables of attention and event-related brain potentials. *Clinical Neurophysiology*, *115*(1), 131–139. https://doi.org/10.1016/S1388-2457(03)00353-5

Egner, T., & Gruzelier, J. H. (2001). Learned self-regulation of EEG frequency components affects attention and event-related brain potentials in humans. *NeuroReport*, *12*(18), 4155. Retrieved September 19, 2024, from https://journals.lww.com/neuroreport/fulltext/2001/12210/learned_self_regulation_of_eeg_frequency.58.aspx

Egner, T., & Gruzelier, J. H. (2003). Ecological validity of neurofeedback: Modulation of slow wave EEG enhances musical performance. *NeuroReport*, *14*(9), 1221. Retrieved September 18, 2024, from https://journals.lww.com/neuroreport/abstract/2003/07010/ecological_validity_of_neurofeedback__modulation.6.aspx

Eisenegger, C., Naef, M., Linssen, A., Clark, L., Gandamaneni, P. K., Müller, U., & Robbins, T. W. (2014). Role of dopamine d2 receptors in human reinforcement learning [Number: 10 Publisher: Nature Publishing Group]. *Neuropsychopharmacology*, *39*(10), 2366–2375. https://doi.org/10.1038/npp.2014.84

Elliott, R., Agnew, Z., & Deakin, J. F. W. (2008). Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans [_eprint: https://onlinelibrary.wiley.com/9568.2008.06202.x]. *European Journal of Neuroscience*, *27*(9), 2213–2218. https://doi.org/10.1111/j.1460-9568.2008.06202.x

Escolano, C., Aguilar, M., & Minguez, J. (2011). EEG-based upper alpha neurofeedback training improves working memory performance [ISSN: 1558-4615]. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2327–2330. https://doi.org/10.1109/IEMBS.2011.6090651

Evans, A. H., Lawrence, A. D., Potts, J., Appel, S., & Lees, A. J. (2005). Factors influencing susceptibility to compulsive dopaminergic drug use in parkinson disease [Publisher: Wolters Kluwer]. *Neurology*, *65*(10), 1570–1574. https://doi.org/10.1212/01.wnl.0000184487.72289.f0

Evans, A. H., Pavese, N., Lawrence, A. D., Tai, Y. F., Appel, S., Doder, M., Brooks, D. J., Lees, A. J., & Piccini, P. (2006). Compulsive drug use linked to sensitized ventral striatal dopamine transmission [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.20822]. *Annals of Neurology*, *59*(5), 852–858. https://doi.org/10.1002/ana.20822

Fakhoury, M., & Domínguez López, S. (2014). The role of habenula in motivation and reward. *Advances in Neuroscience*, *2014*, 1–6. https://doi.org/10.1155/2014/862048

Faller, J., Cummings, J., Saproo, S., & Sajda, P. (2019). Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task [Publisher: National Academy of Sciences]. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(13), 6482–6490. https://doi.org/10.1073/pnas.1817207116

Filipowicz, A. L. S., Glaze, C. M., Kable, J. W., Gold, J. I., & Filipowicz, A. (2020). Pupil diameter encodes the idiosyncratic, cognitive complexity of belief updating. https://doi.org/10.1101/736140

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons [Publisher: American Association for the Advancement of Science]. *Science*, *299*(5614), 1898–1902. https://doi.org/10.1126/science.1077349

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 [Publisher: Nature Publishing Group]. *Scientific Reports*, *7*(1), 1–18. https://doi.org/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 [Publisher: Nature Publishing Group]. *Nature Communications*, *6*(1), 1–11. https://doi.org/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 [Publisher: John Wiley & Sons, Ltd]. *Human Brain Mapping*, *39*(7), 2887–2906. https://doi.org/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 [Publisher: Nature Publishing Group]. *Nature Neuroscience*, *12*(8), 1062–1068. https://doi.org/10.1038/nn.2342

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism [Publisher: American Association for the Advancement of Science]. *Science*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Frank, M. (2006). Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Elsevier*. Retrieved July 22, 2022, from https://www.sciencedirect.com/science/article/pii/S089360800600150X?casa_token=uUHAndWS3A0AAAAA:5DTnqcCkspYMofogLVid6YwNCya_Y1MAzGQJpbRW3DCPLZ93Zk4VWC8lCBjNaXdIiV1Tg9g

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22. https://doi.org/10.18637/jss.v033.i01

Fujiwara, J., Tobler, P. N., Taira, M., Iijima, T., & Tsutsui, K.-I. (2009). Segregated and integrated coding of reward and punishment in the cingulate cortex [Publisher: American Physiological Society]. *Journal of Neurophysiology*, *101*(6), 3284–3293. https://doi.org/10.1152/jn.90909.2008

Galván, A., & McGlennen, K. M. (2012). Daily stress increases risky decision-making in adolescents: A preliminary study [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/dev.20602]. *Developmental Psychobiology*, *54*(4), 433–440. https://doi.org/10.1002/dev.20602

Gilmour, W., Mackenzie, G., Feile, M., Tayler-Grint, L., Suveges, S., Macfarlane, J. A., Macleod, A. D., Marshall, V., Grunwald, I. Q., Steele, J. D., & Gilbertson, T. (2024). Impaired value-based decision-making in parkinson's disease apathy. *Brain*, *147*(4), 1362–1376. https://doi.org/10.1093/brain/awae025

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. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 252–269. https://doi.org/10.3758/CABN.10.2.252

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 148–164. https://doi.org/10.1111/j.2517-6161.1979.tb01068.x

Gittins, J., & Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. *Progress in Statistics*, 241–266. Retrieved August 26, 2024, from https://cir.nii.ac.jp/crid/1574231875963679616

Gläscher, 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 [Publisher: Oxford University Press]. *Cerebral Cortex*, *19*(2), 483–495. https://doi.org/10.1093/cercor/bhn098

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis [Publisher: Proceedings of the National Academy of

Sciences]. *Proceedings of the National Academy of Sciences*, *108*, 15647–15654. https://doi.org/10.1073/pnas.1014269108

Gong, A., Gu, F., Nan, W., Qu, Y., Jiang, C., & Fu, Y. (2021). A review of neurofeedback training for improving sport performance from the perspective of user experience [Publisher: Frontiers]. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.638369

Gourley, S. L., Zimmermann, K. S., Allen, A. G., & Taylor, J. R. (2016). The medial orbitofrontal cortex regulates sensitivity to outcome value [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *36*(16), 4600–4613. https://doi.org/10.1523/JNEUROSCI.4253-15.2016

Gray, J. A. (1981). A critique of eysenck's theory of personality [Publisher: Springer, Berlin, Heidelberg]. *A Model for Personality*, 246–276. https://doi.org/10.1007/978-3-642-67783-0_8

Gruzelier, J. H., Foks, M., Steffert, T., Chen, M. J. .-., & Ros, T. (2014). Beneficial outcome from EEG-neurofeedback on creative music performance, attention and well-being in school children. *Biological Psychology*, *95*, 86–95. https://doi.org/10.1016/j.biopsycho.2013.04.005

Gruzelier, J. (2009). A theory of alpha/theta neurofeedback, creative performance enhancement, long distance functional connectivity and psychological integration. *Cognitive Processing*, *10*(1), 101–109. https://doi.org/10.1007/s10339-008-0248-5

Gruzelier, J. H. (2014). EEG-neurofeedback for optimising performance. II: Creativity, the performing arts and ecological validity. *Neuroscience & Biobehavioral Reviews*, *44*, 142–158. https://doi.org/10.1016/j.neubiorev.2013.11.004

Guarraci, F. A., & Kapp, B. S. (1999). An electrophysiological characterization of ventral tegmental area dopaminergic neurons during differential pavlovian fear conditioning in the awake rabbit. *Behavioural Brain Research*, *99*(2), 169–179. https://doi.org/10.1016/S0166-4328(98)00102-8

Gueguen, M. C. M., Lopez-Persem, A., Billeke, P., Lachaux, J.-P., Rheims, S., Kahane, P., Minotti, L., David, O., Pessiglione, M., & Bastin, J. (2021). Anatomical dissociation of intracerebral signals for reward and punishment prediction errors in humans [Num-

ber: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *12*(1), 3344. https://doi.org/10.1038/s41467-021-23704-w

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*(2), 148–154. https://doi.org/10.1016/j.biopsycho.2005.04.001

Hamner, M. B., Lorberbaum, J. P., & George, M. S. (1999). Potential role of the anterior cingulate cortex in PTSD: Review and hypothesis [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1 6394%281999%299%3A1%3C1%3A%3AAID-DA1%3E3.0.CO%3B2-4]. *Depression and Anxiety*, *9*(1), 1–14. https://doi.org/10.1002/(SICI)1520-6394(1999)9:1<1::AID-DA1>3.0.CO;2-4

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 [Publisher: Society for Neuroscience]. *Journal of Neuroscience*, *26*(32), 8360–8367. https://doi.org/10.1523/JNEUROSCI.1010-06.2006

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction [ISSN: 2640-3498]. *Proceedings of the 34th International Conference on Machine Learning*, 1414–1423. Retrieved September 19, 2024, from https://proceedings.mlr.press/v70/hartford17a.html

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage*, *84*, 159–168. https://doi.org/10.1016/j.neuroimage.2013.08.028

Hayes, T. R., & Petrov, A. A. (2016). Pupil diameter tracks the exploration–exploitation tradeoff during analogical reasoning and explains individual differences in fluid intelligence. *Journal of Cognitive Neuroscience*, *28*(2), 308–318. https://doi.org/10.1162/jocn_a_00895

Heinrich, H., Gevensleben, H., & Strehl, U. (2007). Annotation: Neurofeedback – train your brain to train behaviour [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-7610.2006.01665.x]. *Journal of Child Psychology and Psychiatry*, *48*(1), 3–16. https://doi.org/10.1111/j.1469-7610.2006.01665.x

Hennigan, K., D'Ardenne, K., & McClure, S. M. (2015). Distinct midbrain and habenula pathways are involved in processing aversive events in humans [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *35*(1), 198–208. https://doi.org/10.1523/JNEUROSCI.0927-14.2015

Hewig, J., Hagemann, D., Seifert, J., Naumann, E., & Bartussek, D. (2005). The relationship of cortical activity and personality in a reinforced go-nogo paradigm [Publisher: Hogrefe Publishing]. *Journal of Individual Differences*, *26*(2), 86–99. https://doi.org/10.1027/1614-0001.26.2.86

Hewig, J., Trippe, R., Hecht, H., Coles, M. G., Holroyd, C. B., & Miltner, W. H. (2007). Decision-making in blackjack: An electrophysiological analysis. *Cerebral Cortex*, *17*(4), 865–877. https://doi.org/10.1093/cercor/bhk040

Hikosaka, O. (2010). The habenula: From stress evasion to value-based decision-making [Number: 7 Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *11*(7), 503–513. https://doi.org/10.1038/nrn2866

Hoeks, B., & Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis, 16–26.

Hogeveen, J., Mullins, T. S., Romero, J. D., Eversole, E., Rogge-Obando, K., Mayer, A. R., & Costa, V. D. (2022). The neurocomputational bases of explore-exploit decision-making [Publisher: Elsevier]. *Neuron*, *110*(11), 1869–1879.e5. https://doi.org/10.1016/j.neuron.2022.03.014

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning [Number: 4 Publisher: Nature Publishing Group]. *Nature Neuroscience*, *1*(4), 304–309. https://doi.org/10.1038/1124

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity [Place: US Publisher: American Psychological Association]. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Hong, L., Walz, J. M., & Sajda, P. (2014). Your eyes give you away: Prestimulus changes in pupil diameter correlate with poststimulus task-related EEG dynamics [Publisher: Public

Library of Science]. *PLOS ONE*, *9*(3), e91321. https://doi.org/10.1371/journal.pone.0091321

Houk, J. C., Adams, J. L., & Barto, A. G. (1994). A model of how the basal ganglia generate and use neural signals that predict reinforcement. Retrieved August 24, 2024, from https://direct.mit.edu/books/edited-volume/chapter-pdf/2303165/9780262275774_cao.pdf

Ikemoto, S. (2007). Dopamine reward circuitry: Two projection systems from the ventral midbrain to the nucleus accumbens–olfactory tubercle complex. *Brain Research Reviews*, *56*(1), 27–78. https://doi.org/10.1016/j.brainresrev.2007.05.004

Itagaki, S., & Katayama, J. (2008). Self-relevant criteria determine the evaluation of outcomes induced by others. *NeuroReport*, *19*(3), 383. https://doi.org/10.1097/WNR.0b013e3282f556e8

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration–exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Jocham, G., Klein, T., Neuroscience, M. U. .-. J. o., & 2011, u. (2011). Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *Soc Neuroscience*. https://doi.org/10.1523/JNEUROSCI.3904-10.2011

Joshi, S., & Gold, J. I. (2022). Context-dependent relationships between locus coeruleus firing patterns and coordinated neural activity in the anterior cingulate cortex (E. L. Rich, T. Moore, & T. H. Donner, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *11*, e63490. https://doi.org/10.7554/eLife.63490

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 [Publisher: Elsevier]. *Neuron*, *89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Joshua, M., Adler, A., Mitelman, R., Vaadia, E., & Bergman, H. (2008). Midbrain dopaminergic neurons and striatal cholinergic interneurons encode the difference between reward and aversive events at different epochs of probabilistic classical conditioning trials [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *28*(45), 11673–11684. https://doi.org/10.1523/JNEUROSCI.3839-08.2008

Kahneman & Tversky. (1979). Prospect theory: An analysis of decision under risk | handbook of the fundamentals of financial decision making. Retrieved December 24, 2022, from https://www.worldscientific.com/doi/abs/10.1142/9789814417358_0006

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, *15*(4), 549–559. https://doi.org/10.1016/S0893-6080(02)00048-5

Kakvi, S. A. (2009). Reinforcement learning for blackjack. In S. Natkin & J. Dupire (Eds.), *Entertainment computing – ICEC 2009* (pp. 300–301). Springer. https://doi.org/10.1007/978-3-642-04052-8_43

Kamiya, J. (1968). Conscious control of brain waves. *Psychology Today*, *1*, 56–60.

Kamiya, J. (1969). Operant control of the EEG alpha rhythm and some of its reported effects on consciousness [Publisher: Wiley]. *Altered states of consciousness*, 519–529. Retrieved September 18, 2024, from https://cir.nii.ac.jp/crid/1574231874062667008

Kamp, S.-M., & Donchin, E. (2015). ERP and pupil responses to deviance in an oddball paradigm [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.12378]. *Psychophysiology*, *52*(4), 460–471. https://doi.org/10.1111/psyp.12378

Kawaguchi, K., Clery, S., Pourriahi, P., Seillier, L., Haefner, R. M., & Nienborg, H. (2018). Differentiating between models of perceptual decision making using pupil size inferred confidence [Publisher: Society for Neuroscience Section: Research Articles]. *Journal of Neuroscience*, *38*(41), 8874–8888. https://doi.org/10.1523/JNEUROSCI.0735-18.2018

Kendall, G., & Smith, C. (2003). The evolution of blackjack strategies. *The 2003 Congress on Evolutionary Computation, 2003. CEC '03.*, *4*, 2474–2481 Vol.4. https://doi.org/10.1109/CEC.2003.1299399

Kéri, S., Moustafa, A. A., Myers, C. E., Benedek, G., & Gluck, M. A. (2010). -synuclein gene duplication impairs reward learning [Publisher: Proceedings of the National Academy of Sciences ISBN: 17:11911194]. *Proceedings of the National Academy of Sciences*, *107*(36), 15992–15994. https://doi.org/10.1073/PNAS.1006068107

Kim, H., Shimojo, S., & O'Doherty, J. P. (2006). Is avoiding an aversive outcome rewarding? neural substrates of avoidance learning in the human brain [Publisher: Public Library of Science]. *PLOS Biology*, *4*(8), e233. https://doi.org/10.1371/journal.pbio.0040233

Klavir, O., Genud-Gabai, R., & Paz, R. (2013). Functional connectivity between amygdala and cingulate cortex for adaptive aversive learning [Publisher: Cell Press]. *Neuron*, *80*(5), 1290–1300. https://doi.org/10.1016/J.NEURON.2013.09.035

Kloosterman, N. A., Meindertsma, T., van Loon, A. M., Lamme, V. A. F., Bonneh, Y. S., & Donner, T. H. (2015). Pupil size tracks perceptual content and surprise [Publisher: Blackwell Publishing Ltd]. *European Journal of Neuroscience*, *41*(8), 1068–1078. https://doi.org/10.1111/ejn.12859

Kobayashi, S., & Schultz, W. (2014). Reward contexts extend dopamine signals to unrewarded stimuli [Publisher: Elsevier]. *Current Biology*, *24*(1), 56–62. https://doi.org/10.1016/j.cub.2013.10.061

Koga, K., Yamada, A., Song, Q., Li, X.-H., Chen, Q.-Y., Liu, R.-H., Ge, J., Zhan, C., Furue, H., Zhuo, M., & Chen, T. (2020). Ascending noradrenergic excitation from the locus coeruleus to the anterior cingulate cortex. *Molecular Brain*, *13*(1), 49. https://doi.org/10.1186/s13041-020-00586-5

Koorathota, S., Khan, Z., Lapborisuth, P., & Sajda, P. (2022). Multimodal neurophysiological transformer for emotion recognition [ISSN: 2694-0604]. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3563–3567. https://doi.org/10.1109/EMBC48229.2022.9871421

Kozunova, G. L., Sayfulina, K. E., Prokofyev, A. O., Medvedev, V. A., Rytikova, A. M., Stroganova, T. A., & Chernyshev, B. V. (2022). Pupil dilation and response slowing distinguish deliberate explorative choices in the probabilistic learning task. *Cognitive, Affective, & Behavioral Neuroscience*, *22*(5), 1108–1129. https://doi.org/10.3758/s13415-022-00996-z

Krebs, J. R., Kacelnik, A., & Taylor, P. (1978). Test of optimal sampling by foraging great tits [Publisher: Nature Publishing Group]. *Nature*, *275*(5675), 27–31. https://doi.org/10.1038/275027a0

Krishnamurthy, K., Nassar, M. R., Sarode, S., & Gold, J. I. (2017). Arousal-related adjustments of perceptual biases optimize perception in dynamic environments [Publisher: Nature Publishing Group]. *Nature Human Behaviour*, *1*(6), 1–11. https://doi.org/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 [Publisher: National Academy of Sciences]. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(42), 17951–17956. https://doi.org/10.1073/pnas.0905191106

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Larsen, R. S., & Waters, J. (2018). Neuromodulatory correlates of pupil dilation [Publisher: Frontiers]. *Frontiers in Neural Circuits*, *12*. https://doi.org/10.3389/fncir.2018.00021

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2014). Individual differences in risky decision making: A meta-analysis of sensation seeking and impulsivity with the balloon analogue risk task [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.1784]. *Journal of Behavioral Decision Making*, *27*(1), 20–36. https://doi.org/10.1002/bdm.1784

Lauriola, M., & Weller, J. (2018). Personality and risk: Beyond daredevils— risk taking from a temperament perspective. In M. Raue, E. Lermer, & B. Streicher (Eds.), *Psychological perspectives on risk and risk analysis: Theory, models, and applications* (pp. 3–36). Springer International Publishing. https://doi.org/10.1007/978-3-319-92478-6_1

Lavin, C., San Martín, R., & Rosales Jubal, E. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task [Publisher: Frontiers]. *Frontiers in Behavioral Neuroscience*, *7*. https://doi.org/10.3389/fnbeh.2013.00218

Lawson, R. P., Seymour, B., Loh, E., Lutti, A., Dolan, R. J., Dayan, P., Weiskopf, N., & Roiser, J. P. (2014). The habenula encodes negative motivational value associated with primary punishment in humans [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *111*(32), 11858–11863. https://doi.org/10.1073/pnas.1323586111

Lecca, S., Meye, F. J., Trusel, M., Tchenio, A., Harris, J., Schwarz, M. K., Burdakov, D., Georges, F., & Mameli, M. (2017). Aversive stimuli drive hypothalamus-to-habenula

excitation to promote escape behavior (O. Manzoni, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *6*, e30697. https://doi.org/10.7554/eLife.30697

Leckman, J. F. (2002). Tourette's syndrome [Publisher: Elsevier]. *The Lancet*, *360*(9345), 1577–1586. https://doi.org/10.1016/S0140-6736(02)11526-1

Lecomte, G., & Juhel, J. (2011). The effects of neurofeedback training on memory performance in elderly subjects [Number: 08 Publisher: Scientific Research Publishing]. *Psychology*, *02*(8), 846. https://doi.org/10.4236/psych.2011.28129

Leplow, B., Sepke, M., Schönfeld, R., Pohl, J., Oelsner, H., Latzko, L., & Ebersbach, G. (2017). Impaired learning of punishments in parkinson's disease with and without impulse control disorder. *Journal of Neural Transmission*, *124*(2), 217–225. https://doi.org/10.1007/s00702-016-1648-9

Leue, A., Chavanon, M.-L., Wacker, J., & Stemmler, G. (2009). On the differentiation of n2 components in an appetitive choice task: Evidence for the revised reinforcement sensitivity theory [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.2009.00872.x]. *Psychophysiology*, *46*(6), 1244–1257. https://doi.org/10.1111/j.1469-8986.2009.00872.x

Leue, A., Lange, S., & Beauducel, A. (2012). Reinforcement sensitivity and conflict processing [Publisher: Hogrefe Publishing]. *Journal of Individual Differences*, *33*(3), 160–168. https://doi.org/10.1027/1614-0001/a000096

Lighthall, N. R., Mather, M., & Gorlick, M. A. (2009). Acute stress increases sex differences in risk seeking in the balloon analogue risk task [Publisher: Public Library of Science]. *PLOS ONE*, *4*(7), e6002. https://doi.org/10.1371/journal.pone.0006002

LoTemplio, S., Silcox, J., Federmeier, K. D., & Payne, B. R. (2021). Inter- and intra-individual coupling between pupillary, electrophysiological, and behavioral responses in a visual oddball task [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.13758]. *Psychophysiology*, *58*(4), e13758. https://doi.org/10.1111/psyp.13758

Lubar, J. F., Swartwood, M. O., Swartwood, J. N., & O'Donnell, P. H. (1995). Evaluation of the effectiveness of EEG neurofeedback training for ADHD in a clinical setting as measured by changes in t.o.v.a. scores, behavioral ratings, and WISC-r performance. *Biofeedback and Self-regulation*, *20*(1), 83–99. https://doi.org/10.1007/BF01712768

MacKinnon, D. P. (2000). Contrasts in multiple mediator models [Num Pages: 20]. In *Multivariate applications in substance use research*. Psychology Press.

Mantz, J., Thierry, A. M., & Glowinski, J. (1989). Effect of noxious tail pinch on the discharge rate of mesocortical and mesolimbic dopamine neurons: Selective activation of the mesocortical system. *Brain Research*, *476*(2), 377–381. https://doi.org/10.1016/0006-8993(89)91263-8

Marco-Pallarés, J., Krämer, U. M., Strehl, S., Schröder, A., & Münte, T. F. (2010). When decisions of others matter to me: An electrophysiological analysis. *BMC Neuroscience*, *11*(1), 86. https://doi.org/10.1186/1471-2202-11-86

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data [Publisher: Elsevier]. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/J.JNEUMETH.2007.03.024

Martin, L. E., & Potts, G. F. (2011). Medial frontal event-related potentials and reward prediction: Do responses matter? *Brain and Cognition*, *77*(1), 128–134. https://doi.org/10.1016/j.bandc.2011.04.001

Martínez-Horta, S., Riba, J., Bobadilla, R. F. d., Pagonabarraga, J., Pascual-Sedano, B., Antonijoan, R. M., Romero, S., Mañanas, M. À., García-Sanchez, C., & Kulisevsky, J. (2014). Apathy in parkinson's disease: Neurophysiological evidence of impaired incentive processing [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *34*(17), 5918–5926. https://doi.org/10.1523/JNEUROSCI.0251-14.2014

Marzbani, H., Marateb, H. R., & Mansourian, M. (2016). Neurofeedback: A comprehensive review on system design, methodology and clinical applications. *Basic and Clinical Neuroscience*, *7*(2), 143–158. https://doi.org/10.15412/J.BCN.03070208

Mason, L., Trujillo-Barreto, N. J., Bentall, R. P., & El-Deredy, W. (2016). Attentional bias predicts increased reward salience and risk taking in bipolar disorder. *Biological Psychiatry*, *79*(4), 311–319. https://doi.org/10.1016/j.biopsych.2015.03.014

Massar, S. A. A., Rossi, V., Schutter, D. J. L. G., & Kenemans, J. L. (2012). Baseline EEG theta/beta ratio and punishment sensitivity as biomarkers for feedback-related negativity (FRN) and risk-taking. *Clinical Neurophysiology*, *123*(10), 1958–1965. https://doi.org/10.1016/j.clinph.2012.03.005

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, *1*(1), 16. https://doi.org/10.5334/joc.18

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons [Number: 7148 Publisher: Nature Publishing Group]. *Nature*, *447*(7148), 1111–1115. https://doi.org/10.1038/nature05860

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals [Publisher: Nature Publishing Group]. *Nature*, *459*(7248), 837–841. https://doi.org/10.1038/nature08028

McCall, J. G., Siuda, E. R., Bhatti, D. L., Lawson, L. A., McElligott, Z. A., Stuber, G. D., & Bruchas, M. R. (2017). Locus coeruleus to basolateral amygdala noradrenergic projections promote anxiety-like behavior (L. Luo, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *6*, e18247. https://doi.org/10.7554/eLife.18247

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum [Publisher: Elsevier]. *Neuron*, *38*(2), 339–346. https://doi.org/10.1016/S0896-6273(03)00154-5

McDougle, S. D., & Collins, A. G. E. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, *28*(1), 20–39. https://doi.org/10.3758/s13423-020-01774-z

McFarland, D. J., Krusienski, D. J., Sarnacki, W. A., & Wolpaw, J. R. (2008). Emulation of computer mouse control with a noninvasive brain–computer interface. *Journal of Neural Engineering*, *5*(2), 101. https://doi.org/10.1088/1741-2560/5/2/001

McIlroy-Young, R., Sen, S., Kleinberg, J., & Anderson, A. (2020). Aligning superhuman AI with human behavior: Chess as a model system. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1677–1687. https://doi.org/10.1145/3394486.3403219

McIlroy-Young, R., Wang, R., Sen, S., Kleinberg, J., & Anderson, A. (2022). Learning models of individual behavior in chess. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1253–1263. https://doi.org/10.1145/3534678.3539367

McIlroy-Young, R., Wang, Y., Sen, S., Kleinberg, J., & Anderson, A. (2021). Detecting individual decision-making style: Exploring behavioral stylometry in chess. *Advances in Neural Information Processing Systems*, *34*, 24482–24497. Retrieved December 31, 2022, from https://proceedings.neurips.cc/paper/2021/hash/ccf8111910291ba472b385e9c5f59099-Abstract.html

McNaughton, N., & Corr, P. J. (2008). The neuropsychology of fear and anxiety: A foundation for reinforcement sensitivity theory [Publisher: Cambridge University Press ISBN: 9780511819384]. *The Reinforcement Sensitivity Theory of Personality*, 44–94. https://doi.org/10.1017/CBO9780511819384.003

McNaughton, N., & Gray, J. A. (2000). Anxiolytic action on the behavioural inhibition system implies multiple types of arousal contribute to anxiety [Publisher: Elsevier]. *Journal of affective disorders*, *61*(3), 161–176.

Menicucci, D., Animali, S., Malloggi, E., Gemignani, A., Bonanni, E., Fornai, F., Giorgi, F. S., & Binda, P. (2024). Correlated p300b and phasic pupil-dilation responses to motivationally significant stimuli [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.14550]. *Psychophysiology*, *61*(6), e14550. https://doi.org/10.1111/psyp.14550

Metereau, E., & Dreher, J.-C. (2013). Cerebral correlates of salient prediction error for different rewards and punishments. *Cerebral Cortex*, *23*(2), 477–487. https://doi.org/10.1093/cercor/bhs037

Meyer, R. J., & Shi, Y. (1995). Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem [Publisher: INFORMS]. *Management Science*, *41*(5), 817–834. https://doi.org/10.1287/mnsc.41.5.817

Mikicin, M. (2015). The autotelic involvement of attention induced by EEG neurofeedback training improves the performance of an athlete's mind. *Biomedical Human Kinetics*, *7*(1). https://doi.org/10.1515/bhk-2015-0010

Mikicin, M. (2021). Psychological evaluation of attention indices and directed visual perception using neurofeedback training. *Advances in Cognitive Psychology*, *17*(3), 230–238. https://doi.org/10.5709/acp-0332-9

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function [Publisher: Annual Reviews]. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Mirenowicz, J., & Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli [Publisher: Nature Publishing Group]. *Nature*, *379*(6564), 449–451. https://doi.org/10.1038/379449a0

Modesti, P. (2024). On specimen theoriae novae de mensura sortis of daniel bernoulli. *Decisions in Economics and Finance*. https://doi.org/10.1007/s10203-024-00471-z

Mondoloni, S., Mameli, M., & Congiu, M. (2022). Reward and aversion encoding in the lateral habenula for innate and learned behaviours [Number: 1 Publisher: Nature Publishing Group]. *Translational Psychiatry*, *12*(1), 1–8. https://doi.org/10.1038/s41398-021-01774-0

Monosov, I. E. (2017). Anterior cingulate is a source of valence-specific information about value and uncertainty [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *8*(1), 134. https://doi.org/10.1038/s41467-017-00072-y

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *16*(5), 1936–1947. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996

Moustafa, A. A., Kéri, S., Somlai, Z., Balsdon, T., Frydecka, D., Misiak, B., & White, C. (2015). Drift diffusion model of reward and punishment learning in schizophrenia: Modeling and experimental data [Publisher: Elsevier]. *Behavioural Brain Research*, *291*, 147–154. https://doi.org/10.1016/J.BBR.2015.05.024

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'connell, R. G. (2011). Pupillometry and p3 index the locus coeruleus–noradrenergic arousal function in humans [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.2011.01226.x]. *Psychophysiology*, *48*(11), 1532–1543. https://doi.org/10.1111/j.1469-8986.2011.01226.x

Nagano-Saito, A., Dagher, A., Booij, L., Gravel, P., Welfeld, K., Casey, K. F., Leyton, M., & Benkelfat, C. (2013). Stress-induced dopamine release in human medial prefrontal

cortex—18f-fallypride/PET study in healthy volunteers [_eprint: https://onlinelibrary.wiley.com/d *Synapse*, *67*(12), 821–830. https://doi.org/10.1002/syn.21700

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 [Publisher: Nature Publishing Group]. *Nature Neuroscience*, *15*(7), 1040–1046. https://doi.org/10.1038/nn.3130

Neumann, S. R., Glue, P., & Linscott, R. J. (2021). Aberrant salience and reward processing: A comparison of measures in schizophrenia and anxiety. *Psychological Medicine*, *51*(9), 1507–1515. https://doi.org/10.1017/S0033291720000264

Nicolas-Alonso, L. F., & Gomez-Gil, J. (2012). Brain computer interfaces, a review [Number: 2 Publisher: Molecular Diversity Preservation International]. *Sensors*, *12*(2), 1211–1279. https://doi.org/10.3390/s120201211

Nieuwenhuis, S. (2011). Learning, the p3, and the locus coeruleus-norepinephrine system. https://doi.org/10.7551/mitpress/8791.003.0016

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the p3, and the locus coeruleus–norepinephrine system [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *131*(4), 510–532. https://doi.org/10.1037/0033-2909.131.4.510

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, *14*(6), 769–776. https://doi.org/10.1016/j.conb.2004.10.016

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 [Publisher: Elsevier]. *Neuron*, *38*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning [Publisher: American Association for the Advancement of Science]. *Science*, *304*(5669), 452–454. https://doi.org/10.1126/science.1094285

O'Doherty, J. P., Deichmann, R., Critchley, H. D., & Dolan, R. J. (2002). Neural responses during anticipation of a primary taste reward [Publisher: Elsevier]. *Neuron*, *33*(5), 815–826. https://doi.org/10.1016/S0896-6273(02)00603-7

O'Doherty, J. P., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex [Publisher: Nature Publishing Group]. *Nature Neuroscience*, *4*(1), 95–102. https://doi.org/10.1038/82959

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction [Number: 2 Publisher: Nature Publishing Group]. *Nature Neuroscience*, *5*(2), 97–98. https://doi.org/10.1038/nn802

Pagonabarraga, J., Kulisevsky, J., Strafella, A. P., & Krack, P. (2015). Apathy in parkinson's disease: Clinical features, neural substrates, diagnosis, and treatment. *The Lancet Neurology*, *14*(5), 518–531. https://doi.org/10.1016/S1474-4422(15)00019-8

Palminteri, S., & Pessiglione, M. (2017). Opponent brain systems for reward and punishment learning: Causal evidence from drug and lesion studies in humans. In *Decision neuroscience: An integrative approach* (pp. 291–303). Retrieved February 2, 2024, from https://www.sciencedirect.com/science/article/pii/B9780128053089000233

Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., Czernecki, V., Karachi, C., Capelle, L., Durr, A., & Pessiglione, M. (2012). Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron*, *76*(5), 998–1009. https://doi.org/10.1016/j.neuron.2012.10.017

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning [Publisher: Nature Publishing Group]. *Nature Communications 2015 6:1*, *6*(1), 1–14. https://doi.org/10.1038/ncomms9096

Palminteri, S., Lebreton, M., Worbe, Y., Grabli, D., Hartmann, A., & Pessiglione, M. (2009). Pharmacological modulation of subliminal learning in parkinson's and tourette's syndromes. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(45), 19179–19184. https://doi.org/10.1073/PNAS.0904035106

Parra, L., Spence, C., Gerson, A., & Sajda, P. (2003). Response error correction-a demonstration of improved human-machine performance using real-time EEG monitoring [Conference

Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *11*(2), 173–177. https://doi.org/10.1109/TNSRE.2003.814446

Parra, L. C., Spence, C. D., Gerson, A. D., & Sajda, P. (2005). Recipes for the linear analysis of EEG [Publisher: Academic Press]. *NeuroImage*, *28*(2), 326–341. https://doi.org/10.1016/j.neuroimage.2005.05.032

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings [Publisher: Public Library of Science]. *PLOS Computational Biology*, *7*(1), e1001048. https://doi.org/10.1371/journal.pcbi.1001048

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pessiglione, M., Seymour, B., Flandin, G., & Dolan, R. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *nature.com*. Retrieved July 26, 2022, from https://www.nature.com/articles/nature05051

Pessiglione, M., Petrovic, P., Daunizeau, J., Palminteri, S., Dolan, R. J., & Frith, C. D. (2008). Subliminal instrumental conditioning demonstrated in the human brain. *Neuron*, *59*(4), 561–567. https://doi.org/10.1016/J.NEURON.2008.07.005

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making [Publisher: American Association for the Advancement of Science]. *Science*, *372*(6547), 1209–1214. https://doi.org/10.1126/science.abe2629

Philiastides, M. G., Biele, G., Vavatzanidis, N., Kazzer, P., & Heekeren, H. R. (2010). Temporal dynamics of prediction error processing during reward-based decision making. *NeuroImage*, *53*(1), 221–232. https://doi.org/10.1016/j.neuroimage.2010.05.052

Piray, P., & Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility [Publisher: Nature Publishing Group]. *Nature Communications 2021 12:1*, *12*(1), 1–16. https://doi.org/10.1038/s41467-021-26731-9

Piray, P., Zeighami, Y., Bahrami, F., Eissa, A. M., Hewedi, D. H., & Moustafa, A. A. (2014). Impulse control disorders in parkinson's disease are associated with dysfunction in stimulus valuation but not action valuation [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *34*(23), 7814–7824. https://doi.org/10.1523/JNEUROSCI.4063-13.2014

Poli, R., Valeriani, D., & Cinel, C. (2014). Collaborative brain-computer interface for aiding decision-making [Publisher: Public Library of Science]. *PLOS ONE*, *9*(7), e102693. https://doi.org/10.1371/journal.pone.0102693

Porcelli, A. J., & Delgado, M. R. (2017). Stress and decision making: Effects on valuation, learning, and risk-taking. *Current Opinion in Behavioral Sciences*, *14*, 33–39. https://doi.org/10.1016/j.cobeha.2016.11.015

Preston, S. D., Buchanan, T. W., Stansfield, R. B., & Bechara, A. (2007). Effects of anticipatory stress on decision making in a gambling task [Place: US Publisher: American Psychological Association]. *Behavioral Neuroscience*, *121*(2), 257–263. https://doi.org/10.1037/0735-7044.121.2.257

Preuschoff, K., 't Hart, B. M., & Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making [Publisher: Frontiers]. *Frontiers in Neuroscience*, *5*. https://doi.org/10.3389/fnins.2011.00115

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare [Publisher: Springer Nature]. *Nature Machine Intelligence*, *2*(7), 369–375. https://doi.org/10.1038/s42256-020-0197-y

Pulcu, E., & Browning, M. (2017). Affective bias as a rational response to the statistics of rewards and punishments (M. J. Frank, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *6*, e27879. https://doi.org/10.7554/eLife.27879

Purvis, E. M., Klein, A. K., & Ettenberg, A. (2018). Lateral habenular norepinephrine contributes to states of arousal and anxiety in male rats. *Behavioural Brain Research*, *347*, 108–115. https://doi.org/10.1016/j.bbr.2018.03.012

Queirazza, F., Fouragnan, E., Steele, J. D., Cavanagh, J., & Philiastides, M. G. (2019). SUPPLEMENTARY MATERIALS - neural correlates of weighted reward prediction error

during reinforcement learning classify response to cognitive behavioral therapy in depression [Publisher: American Association for the Advancement of Science]. *Science Advances*, *5*(7). https://doi.org/10.1126/SCIADV.AAV4962

Quilty, L. C., & Oakman, J. M. (2004). The assessment of behavioural activation—the relationship between impulsivity and behavioural activation. *Personality and Individual Differences*, *37*(2), 429–442. https://doi.org/10.1016/j.paid.2003.09.014

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making [Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Raymond, J., Sajid, I., Parkinson, L. A., & Gruzelier, J. H. (2005). Biofeedback and dance performance: A preliminary investigation. *Applied Psychophysiology and Biofeedback*, *30*(1), 65–73. https://doi.org/10.1007/s10484-005-2175-x

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 [Publisher: Nature Publishing Group]. *Nature Communications*, *7*(1), 13289. https://doi.org/10.1038/ncomms13289

Richer, F., & Beatty, J. (1987). Contrasting effects of response uncertainty on the task-evoked pupillary response and reaction time [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.14 8986.1987.tb00291.x]. *Psychophysiology*, *24*(3), 258–262. https://doi.org/10.1111/j.1469-8986.1987.tb00291.x

Ring, C., Cooke, A., Kavussanu, M., McIntyre, D., & Masters, R. (2015). Investigating the efficacy of neurofeedback training for expediting expertise and excellence in sport. *Psychology of Sport and Exercise*, *16*, 118–127. https://doi.org/10.1016/j.psychsport.2014.08.005

Robinson, N., Mane, R., Chouhan, T., & Guan, C. (2021). Emerging trends in BCI-robotics for motor control and rehabilitation. *Current Opinion in Biomedical Engineering*, *20*, 100354. https://doi.org/10.1016/j.cobme.2021.100354

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards [Publisher: Nature

Publishing Group]. *Nature Neuroscience*, *10*(12), 1615–1624. https://doi.org/10.1038/nn2013

Root, D. H., Hoffman, A. F., Good, C. H., Zhang, S., Gigante, E., Lupica, C. R., & Morales, M. (2015). Norepinephrine activates dopamine d4 receptors in the rat lateral habenula [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *35*(8), 3460–3469. https://doi.org/10.1523/JNEUROSCI.4525-13.2015

Rostami, R., Sadeghi, H., Karami, K. A., Abadi, M. N., & Salamati, P. (2012). The effects of neurofeedback on the improvement of rifle shooters' performance [Publisher: Routledge _eprint: https://doi.org/10.1080/10874208.2012.730388]. *Journal of Neurotherapy*, *16*(4), 264–269. https://doi.org/10.1080/10874208.2012.730388

Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. A., & Glimcher, P. W. (2009). Dopaminergic drugs modulate learning rates and perseveration in parkinson's patients in a dynamic foraging task. *Soc Neuroscience*. https://doi.org/10.1523/JNEUROSCI.3524-09.2009

Sajda, P., Goldman, R. I., Philiastides, M. G., Gerson, A. D., & Brown, T. R. (2007). A system for single-trial analysis of simultaneously acquired EEG and fMRI [ISSN: 1948-3554]. *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, 287–290. https://doi.org/10.1109/CNE.2007.369667

Samiee, K., Kovács, P., & Gabbouj, M. (2015). Epileptic seizure classification of EEG time-series using rational discrete short-time fourier transform [Conference Name: IEEE Transactions on Biomedical Engineering]. *IEEE Transactions on Biomedical Engineering*, *62*(2), 541–552. https://doi.org/10.1109/TBME.2014.2360101

Samii, A., Nutt, J. G., & Ransom, B. R. (2004). Parkinson's disease [Publisher: Elsevier]. *The Lancet*, *363*(9423), 1783–1793. https://doi.org/10.1016/S0140-6736(04)16305-8

Santesso, D. L., Dzyundzyak, A., & Segalowitz, S. J. (2011). Age, sex and individual differences in punishment sensitivity: Factors influencing the feedback-related negativity [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.2011.01229.x]. *Psychophysiology*, *48*(11), 1481–1489. https://doi.org/10.1111/j.1469-8986.2011.01229.x

Saproo, S., Shih, V., Jangraw, D. C., & Sajda, P. (2016). Neural mechanisms underlying catastrophic failure in human–machine interaction during aerial navigation [Publisher: IOP Publishing]. *Journal of Neural Engineering*, *13*(6), 066005. https://doi.org/10.1088/1741-2560/13/6/066005

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition [Number: 3 Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *10*(3), 211–223. https://doi.org/10.1038/nrn2573

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: The locus coeruleus mediates cognition through arousal [Publisher: Cell Press]. *Neuron*, *76*(1), 130–141. https://doi.org/10.1016/J.NEURON.2012.09.011

Sato, A., Yasuda, A., Ohira, H., Miyawaki, K., Nishikawa, M., Kumano, H., & Kuboki, T. (2005). Effects of value and reward magnitude on feedback negativity and p300. *NeuroReport*, *16*(4), 407. Retrieved August 4, 2024, from https://journals.lww.com/neuroreport/fulltext/2005/03150/Artifact_Correction_of_the_Ongoing_EEG_Using.00020.aspx

Schiller, M. R. G., & Gobet, F. R. (2012). A comparison between cognitive and AI models of blackjack strategy learning. In B. Glimm & A. Krüger (Eds.), *KI 2012: Advances in artificial intelligence* (pp. 143–155). Springer. https://doi.org/10.1007/978-3-642-33347-7_13

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *27*(47), 12860–12867. https://doi.org/10.1523/JNEUROSCI.2496-07.2007

Schüller, T., Fischer, A. G., Gruendler, T. O. J., Baldermann, J. C., Huys, D., Ullsperger, M., & Kuhn, J. (2020). Decreased transfer of value to action in tourette syndrome. *Cortex*, *126*, 39–48. https://doi.org/10.1016/j.cortex.2019.12.027

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

Schultz, W., & Romo, R. (1987). Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey [Publisher: American Physiolog-

ical Society]. *Journal of Neurophysiology*, *57*(1), 201–217. https://doi.org/10.1152/jn.1987.57.1.201

Schultz, W. (1998). Predictive reward signal of dopamine neurons [Publisher: American Physiological Society]. *Journal of Neurophysiology*, *80*(1), 1–27. https://doi.org/10.1152/jn.1998.80.1.1

Schultz, W. (2016). Dopamine reward prediction-error signalling: A two-component response [Publisher: NIH Public Access]. *Nature reviews. Neuroscience*, *17*(3), 183. https://doi.org/10.1038/NRN.2015.26

Seeley, W. W. (2019). The salience network: A neural system for perceiving and responding to homeostatic demands [Publisher: Society for Neuroscience Section: Progressions]. *Journal of Neuroscience*, *39*(50), 9878–9882. https://doi.org/10.1523/JNEUROSCI.1138-17.2019

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., & Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans [Number: 6992 Publisher: Nature Publishing Group]. *Nature*, *429*(6992), 664–667. https://doi.org/10.1038/nature02581

Shih, V., Zhang, L., Kothe, C., Makeig, S., & Sajda, P. (2016). Predicting decision accuracy and certainty in complex brain-machine interactions. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 004076–004081. https://doi.org/10.1109/SMC.2016.7844870

Shrout, P., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological methods*, *7*, 422–45. https://doi.org/10.1037/1082-989X.7.4.422

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J. A., Pitchford, J. W., James, A., Ahmed, M. Z., Brierley, A. S., Hindell, M. A., Morritt, D., Musyl, M. K., Righton, D., Shepard, E. L. C., Wearmouth, V. J., Wilson, R. P., Witt, M. J., & Metcalfe, J. D. (2008). Scaling laws of marine predator search behaviour [Publisher: Nature Publishing Group]. *Nature*, *451*(7182), 1098–1102. https://doi.org/10.1038/nature06518

Skinner, B. F. (1938). The behavior of organisms. new york: Appleton-century-crofts. *American Psychologist*, *221*, 233.

Skvortsova, V., Palminteri, S., & Pessiglione, M. (2014). Learning to minimize efforts versus maximizing rewards: Computational principles and neural correlates. *Journal of Neuroscience*, *34*(47), 15621–15630. https://doi.org/10.1523/JNEUROSCI.1350-14.2014

Smillie, L. D., & Jackson, C. J. (2006). Functional impulsivity and reinforcement sensitivity theory [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.2005.00369.x]. *Journal of Personality*, *74*(1), 47–84. https://doi.org/10.1111/j.1467-6494.2005.00369.x

Smillie, L. D., Jackson, C. J., & Dalgleish, L. I. (2006). Conceptual distinctions among carver and white's (1994) BAS scales: A reward-reactivity versus trait impulsivity perspective. *Personality and Individual Differences*, *40*(5), 1039–1050. https://doi.org/10.1016/j.paid.2005.10.012

Smillie, L. D., Pickering, A. D., & Jackson, C. J. (2006). The new reinforcement sensitivity theory: Implications for personality measurement [Publisher: SAGE Publications Inc]. *Personality and Social Psychology Review*, *10*(4), 320–335. https://doi.org/10.1207/s15327957pspr1004_3

Sokol-Hessner, P., Raio, C. M., Gottesman, S. P., Lackovic, S. F., & Phelps, E. A. (2016). Acute stress does not affect risky monetary decision-making. *Neurobiology of Stress*, *5*, 19–25. https://doi.org/10.1016/j.ynstr.2016.10.003

Srinivasaiah, R., Biju, V. G., Jankatti, S. K., Channegowda, R. H., & Jinachandra, N. S. (2024). Reinforcement learning strategies using monte-carlo to solve the blackjack problem. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(1), 904. https://doi.org/10.11591/ijece.v14i1.pp904-910

Starcke, K., & Brand, M. (2012). Decision making under stress: A selective review [Publisher: Pergamon]. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1228–1248. https://doi.org/10.1016/j.neubiorev.2012.02.003

Starcke, K., & Brand, M. (2016). Effects of stress on decisions under uncertainty: A meta-analysis [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *142*, 909–933. https://doi.org/10.1037/bul0000060

Starcke, K., Wolf, O. T., Markowitsch, H. J., & Brand, M. (2008). Anticipatory stress influences decision making under explicit risk conditions [Place: US Publisher: American Psychological Association]. *Behavioral Neuroscience*, *122*(6), 1352–1360. https://doi.org/10.1037/a0013281

Sterman, M. B., & Friar, L. (1972). Suppression of seizures in an epileptic following sensorimotor EEG feedback training. *Electroencephalography and Clinical Neurophysiology*, *33*(1), 89–95. https://doi.org/10.1016/0013-4694(72)90028-4

Sterman, M. B., Macdonald, L. R., & Stone, R. K. (1974). Biofeedback training of the sensorimotor electroencephalogram rhythm in man: Effects on epilepsy [_eprint: https://onlinelibrary.wil 1157.1974.tb04016.x]. *Epilepsia*, *15*(3), 395–416. https://doi.org/10.1111/j.1528-1157.1974.tb04016.x

Sutherland, M. R., & Mather, M. (2015). Negative arousal increases the effects of stimulus salience in older adults [Publisher: Routledge]. *Experimental Aging Research*. Retrieved August 4, 2024, from https://www.tandfonline.com/doi/abs/10.1080/0361073X.2015.1021644

Sutherland, M. R., & Mather, M. (2018). Arousal (but not valence) amplifies the impact of salience [Publisher: Routledge _eprint: https://doi.org/10.1080/02699931.2017.1330189]. *Cognition and Emotion*, *32*(3), 616–622. https://doi.org/10.1080/02699931.2017.1330189

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty [Publisher: American Association for the Advancement of Science]. *Science*, *150*(3700), 1187–1188. https://doi.org/10.1126/science.150.3700.1187

Talmi, D., Atkinson, R., & El-Deredy, W. (2013). The feedback-related negativity signals salience prediction errors, not reward prediction errors [Publisher: Society for Neuroscience Section: Brief Communications]. *Journal of Neuroscience*, *33*(19), 8264–8269. https://doi.org/10.1523/JNEUROSCI.5695-12.2013

Tang, H., Costa, V. D., Bartolo, R., & Averbeck, B. B. (2022). Differential coding of goals and actions in ventral and dorsal corticostriatal circuits during goal-directed behavior [Publisher: Elsevier]. *Cell Reports*, *38*(1). https://doi.org/10.1016/j.celrep.2021.110198

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, *106*, 1–31. https://doi.org/10.18637/jss.v106.i01

Thatcher, H. R., Downs, C. T., & Koyama, N. F. (2019). Anthropogenic influences on the time budgets of urban vervet monkeys. *Landscape and Urban Planning*, *181*, 38–44. https://doi.org/10.1016/j.landurbplan.2018.09.014

Thompson, W. R. (1933). ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES. *Biometrika*, *25*(3), 285–294. https://doi.org/10.1093/biomet/25.3-4.285

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies* [Google-Books-ID: Go8XozILUJYC] Transaction Publishers.

Tobler, P. N., Christopoulos, G. I., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2008). Neuronal distortions of reward probability without choice [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *28*(45), 11703–11711. https://doi.org/10.1523/JNEUROSCI.2870-08.2008

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm [Publisher: Society for Neuroscience Section: Behavioral/Systems/Cognitive]. *Journal of Neuroscience*, *23*(32), 10402–10410. https://doi.org/10.1523/JNEUROSCI.23-32-10402.2003

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), 1621–1632. https://doi.org/10.1152/jn.00745.2006

Toyomaki, A., & Murohashi, H. (2005). Discrepancy between feedback negativity and subjective evaluation in gambling. *NeuroReport*, *16*(16), 1865. https://doi.org/10.1097/01.wnr.0000185962.96217.36

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Unger, K., Heintz, S., & Kray, J. (2012). Punishment sensitivity modulates the processing of negative feedback but not error-induced learning [Publisher: Frontiers]. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00186

Ungless, M. A., Magill, P. J., & Bolam, J. P. (2004). Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli [Publisher: American Association for the Advancement of Science]. *Science*, *303*(5666), 2040–2042. https://doi.org/10.1126/science.1093360

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias [Publisher: Nature Publishing Group]. *Nature Communications*, *8*(1), 1–11. https://doi.org/10.1038/ncomms14637

Vaessen, T., Hernaus, D., Myin-Germeys, I., & van Amelsvoort, T. (2015). The dopaminergic response to acute stress in health and psychopathology: A systematic review. *Neuroscience & Biobehavioral Reviews*, *56*, 241–251. https://doi.org/10.1016/j.neubiorev.2015.07.008

Valeriani, D., Poli, R., & Cinel, C. (2015). A collaborative brain-computer interface to improve human performance in a visual search task [ISSN: 1948-3554]. *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, 218–223. https://doi.org/10.1109/NER.2015.7146599

van Ruitenbeek, P., Quaedflieg, C. W., Hernaus, D., Hartogsveld, B., & Smeets, T. (2021). Dopaminergic and noradrenergic modulation of stress-induced alterations in brain activation associated with goal-directed behaviour [Publisher: SAGE Publications Ltd STM]. *Journal of Psychopharmacology*, *35*(12), 1449–1463. https://doi.org/10.1177/02698811211044679

van de Vijver, I., Ridderinkhof, K. R., & Cohen, M. X. (2011). Frontal oscillatory dynamics predict feedback learning and action adjustment [Publisher: MIT Press]. *Journal of Cognitive Neuroscience*, *23*(12), 4106–4121. https://doi.org/10.1162/jocn_a_00110

van den Bos, R., Harteveld, M., & Stoop, H. (2009). Stress and decision-making in humans: Performance is related to cortisol reactivity, albeit differently in men and women. *Psychoneuroendocrinology*, *34*(10), 1449–1458. https://doi.org/10.1016/j.psyneuen.2009.04.016

van den Bos, R., Taris, R., Scheppink, B., de Haan, L., & Verster, J. (2014). Salivary cortisol and alpha-amylase levels during an assessment procedure correlate differently with risk-taking measures in male and female police recruits [Publisher: Frontiers]. *Frontiers in Behavioral Neuroscience*, *7*. https://doi.org/10.3389/fnbeh.2013.00219

VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide [Publisher: Annual Reviews]. *Annual Review of Public Health*, *37*, 17–32. https://doi.org/10.1146/annurev-publhealth-032315-021402

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5999–6009.

von Neumann, J., Morgenstern, O., Kuhn, H. W., & Rubinstein, A. (2004). *Theory of games and economic behavior: 60th anniversary commemorative edition*. Princeton University Press. Retrieved September 20, 2024, from http://ebookcentral.proquest.com/lib/gla/detail.action?docID=1092486

Voon, V., Thomsen, T., Miyasaki, J. M., de Souza, M., Shafro, A., Fox, S. H., Duff-Canning, S., Lang, A. E., & Zurowski, M. (2007). Factors associated with dopaminergic drug–related pathological gambling in parkinson disease. *Archives of Neurology*, *64*(2), 212–216. https://doi.org/10.1001/archneur.64.2.212

Wager, T. D., Davidson, M. L., Hughes, B. L., Lindquist, M. A., & Ochsner, K. N. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron*, *59*(6), 1037–1050. https://doi.org/10.1016/j.neuron.2008.09.006

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews*, *36*(8), 1870–1884. https://doi.org/10.1016/j.neubiorev.2012.05.008

Wang, Y., & Jung, T.-P. (2011). A collaborative brain-computer interface for improving human performance [Publisher: Public Library of Science]. *PLOS ONE*, *6*(5), e20422. https://doi.org/10.1371/journal.pone.0020422

Webster, G. D., & Crysel, L. C. (2012). "hit me, maybe, one more time": Brief measures of impulsivity and sensation seeking and their prediction of blackjack bets and sexual promis-

cuity. *Journal of Research in Personality*, *46*(5), 591–598. https://doi.org/10.1016/j.jrp.2012.07.001

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56. https://doi.org/10.1016/j.cobeha.2020.10.001

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. *Journal of experimental psychology. General*, *143*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wolpaw, J. R. (2013, January 1). Chapter 6 - brain–computer interfaces. In M. P. Barnes & D. C. Good (Eds.), *Handbook of clinical neurology* (pp. 67–74, Vol. 110). Elsevier. https://doi.org/10.1016/B978-0-444-52901-5.00006-X

Xiang, M.-Q., Hou, X.-H., Liao, B.-G., Liao, J.-W., & Hu, M. (2018). The effect of neurofeedback training for sport performance in athletes: A meta-analysis. *Psychology of Sport and Exercise*, *36*, 114–122. https://doi.org/10.1016/j.psychsport.2018.02.004

Xu, J., & Chen, S. (2021). A neuroevolutionary approach for opponent modeling and exploitation in no-limit texas hold'em poker [ISSN: 2688-0938]. *2021 China Automation Congress (CAC)*, 2270–2275. https://doi.org/10.1109/CAC53003.2021.9727922

Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D. F., & Büchel, C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *26*(37), 9530–9537. https://doi.org/10.1523/JNEUROSCI.2915-06.2006

Yechiam, E., & Telpaz, A. (2011). To take risk is to face loss: A tonic pupillometry study [Publisher: Frontiers]. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00344

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), 459–482. https://doi.org/10.1002/cne.920180503

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex*, *15*(5), 535–544. https://doi.org/10.1093/cercor/bhh153

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain [Publisher: Society for Neuroscience]. *Journal of Neuroscience*, *24*(28), 6258–6264. https://doi.org/10.1523/JNEUROSCI.4537-03.2004

Yuan, P., Wang, Y., Gao, X., Jung, T.-P., & Gao, S. (2013). A collaborative brain-computer interface for accelerating human decision making. In C. Stephanidis & M. Antona (Eds.), *Universal access in human-computer interaction. design methods, tools, and interaction techniques for eInclusion* (pp. 672–681). Springer. https://doi.org/10.1007/978-3-642-39188-0_72

Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2018). Learning affective features with a hybrid deep model for audio–visual emotion recognition [Conference Name: IEEE Transactions on Circuits and Systems for Video Technology]. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 3030–3043. https://doi.org/10.1109/TCSVT.2017.2719043

Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision (M. J. Frank, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *5*, e17688. https://doi.org/10.7554/eLife.17688