The Deconstruction of Reinforcement Learning in Human Substance Use Disorder

Tsen Vei Lim

Department of Psychiatry

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Trinity Hall November 2021

Preface

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Some data reported in this thesis have been published in the following form:

Research reported in Chapter 3 has been published as:

Lim, T. V., Cardinal, R. N., Bullmore, E. T., Robbins, T. W., & Ersche, K. D. (2021). Impaired Learning From Negative Feedback in Stimulant Use Disorder: Dopaminergic Modulation. *International Journal of Neuropsychopharmacology*, *24*(11), 867–878. https://doi.org/10.1093/ijnp/pyab041

Research reported in Chapter 5 is a reanalysis of Ersche et al (2016), and has been published as:

Lim, T. V., Cardinal, R. N., Savulich, G., Jones, P. S., Moustafa, A. A., Robbins, T. W., & Ersche, K. D. (2019). Impairments in reinforcement learning do not explain enhanced habit formation in cocaine use disorder. *Psychopharmacology*, *236*(8), 2359–2371. https://doi.org/10.1007/s00213-019-05330-z

Part of the data presented in Chapter 6 has been reported in:

Ersche, K. D., **Lim, T. V.**, Murley, A. G., Rua, C., Vaghi, M. M., White, T. L., Williams, G. B., & Robbins, T. W. (2021). Reduced Glutamate Turnover in the Putamen Is Linked with Automatic Habits in Human Cocaine Addiction. *Biological Psychiatry*, *89*(10), 970–979. https://doi.org/10.1016/j.biopsych.2020.12.009

I have also co-authored the following papers on the validation of the Creature of Habit Scale – a self-report instrument on habitual tendencies – which was used in this thesis:

Ersche, K. D., **Lim, T.-V.**, Ward, L. H. E., Robbins, T. W., & Stochl, J. (2017). Creature of Habit: A self-report measure of habitual routines and automatic tendencies in everyday life. *Personality and Individual Differences*, *116*, 73–85. https://doi.org/10.1016/j.paid.2017.04.024

Ersche, K. D., Ward, L. H. E., **Lim, T.-V.**, Lumsden, R. J., Sawiak, S. J., Robbins, T. W., & Stochl, J. (2019). Impulsivity and compulsivity are differentially associated with automaticity and routine on the Creature of Habit Scale. *Personality and Individual Differences*, *150*, 109493. https://doi.org/10.1016/j.paid.2019.07.003

Thesis summary

The Deconstruction of Reinforcement Learning in Human Substance Use Disorder

Tsen Vei Lim

Individuals diagnosed with substance use disorder (SUD) often behave in ways detrimental to their own interest and well-being. The mechanisms behind such maladaptive behaviour in human SUD remain unclear, but can be explained by disruptions to reinforcement learning processes that under normal circumstances shape behaviour adaptively. This perspective has led to two different, but not mutually exclusive, hypotheses: (1) reinforcement learning is impaired in drug-addicted individuals, as they are unable to learn from the consequences of their actions, and (2) learned behaviour in drug-addicted individuals reflects an imbalance between two regulatory systems: the goal-directed and the habit system. Recently, trial-by-trial computational modelling lends itself as a promising tool to deconstruct latent cognitive processes that underpin learning, which can provide mechanistic insights into these impairments. Thus, with multiple learning paradigms, the objectives of this thesis are two-fold: (1) to characterise the cognitive profile related to impaired reinforcement learning and its supporting processes in SUD with computational modelling; (2) to clarify the relationship between impaired reinforcement learning and habit learning in SUD.

The first part of the thesis describes the computational analyses of task performance in probabilistic reinforcement learning. These analyses identified in two independent cohorts of stimulant-addicted individuals a selectively reduced learning rate from punishment, suggesting that their behaviour may be less amenable to negative feedback. In one of these cohorts, participants underwent pharmacological manipulations with dopamine $D_{2/3}$ receptor agents, which found that both dopamine $D_{2/3}$ receptor antagonist (400mg amisulpride) and agonist (0.5 mg pramipexole) differentially modulated behaviour in stimulant-addicted individuals: while both dopamine agents impaired performance in control participants, they ameliorated learning from negative feedback in stimulant-addicted individuals – confirming the link between aberrant learning and dopamine dysfunction in SUD. Next, I investigated the integrity of declarative and non-declarative memory systems in cocaine use disorder patients with a category learning task, as these systems are thought to complement reinforcement learning. I found that cocaine use disorder patients showed clear deficits in both declarative and non-

declarative memory. Analyses of their response strategies revealed that these patients were more likely than control participants to adopt a simple but suboptimal memorisation strategy during learning, as opposed to a more complex integrative strategy, which supports the notion of an aberrant engagement of memory systems during reinforcement learning.

Given that SUD is associated with enhanced habit formation, I then tested the hypothesis that reinforcement learning impairments exacerbate subsequent habit formation in cocaine use disorder, by reanalysing prior data on an appetitive instrumental learning task with computational methods. Contrary to the hypothesis, I found that impaired reinforcement learning in cocaine use disorder, in the form of a reduced learning rate, is insufficient to account for enhanced habit formation in these patients, suggesting other modulatory factors at play. I subsequently addressed the question of whether patients with cocaine use disorder have insight into their behavioural tendencies by using self-report questionnaires. These data revealed evidence for a predilection for automatic habits and reduced goal-directed actions in their daily lives. Finally, I expanded my work by measuring instrumental learning in a community sample of individuals recruited online who consume alcohol hazardously (as measured with the Alcohol Use Disorder Identification Test questionnaire) – but not formally diagnosed with alcohol use disorder. I tested this with a novel task paradigm which measures goal-directed and habitual responses in a conflict situation, but did not find any evidence for an impaired goal-directed or augmented habitual control associated with harmful alcohol use.

Jointly, the study of reinforcement learning with multiple paradigms refined our understanding of maladaptive behaviours in severe SUD, which may be characterised by the attenuated effects of negative feedback on behaviour, as well as aberrant non-declarative and declarative memory systems. Impaired reinforcement learning, however, cannot fully account for habit predominance associated with SUD. Instead, this predominance might be modulated differentially by different drugs of abuse, drug use severity and individual differences in habitual tendencies.

Acknowledgements

I would like to take this opportunity to express my gratitude towards several individuals who supported me throughout this journey.

First and foremost, I would like to thank my primary supervisor, Professor Karen Ersche, for her tireless guidance and support towards my PhD training. She has been extremely generous with her time whenever I needed advice or feedback on my work. I am inspired by her ingenuity in research, and have benefitted tremendously from her practical advices on how to conduct research, and tips of the trade on becoming a good scientist. Her unwavering support and faith in my abilities have enabled me to complete this thesis.

Next, I am grateful to Dr Rudolf Cardinal, who introduced me to the world of computational modelling, and provided me with expert advice and support for all things computational whenever I need them. I find his insights into modelling and how best to optimise cognitive modelling utterly fascinating, and have always enjoyed our discussions.

Special thanks to these individuals who have supported my research in various ways: Mr Simon Jones for his help in analysing the diffusion tensor imaging data reported in Chapter 5; Dr Amy Bland, an expert in *PsychoPy*, for her help in translating the behavioural tasks for online data collection in Chapter 7; Professor Trevor Robbins, for his illuminating expert insights on the meaning and wider implications of some of my behavioural findings.

The research projects undertaken in this thesis were largely collaborative in nature, and completed with the Addiction Research Cambridge group, led by Professor Karen Ersche. I would like to thank Eva Groot and Ibtisam Abdi, who I have worked alongside with to collect the data that forms a substantial part of this thesis. Special thanks to Dr Roderick Lumsden for providing technical support and for programming some of the behavioural tasks that were integral to this thesis. Thanks also to other former members of the team: Dr Chun Meng, Rachel Rodrigues, Jeremy Guild, and Joanna Vamvakopoulou. I miss our lunch talks and our times working together in the lab.

To the friends I have made along the way: Peter Zhukovsky, Yin Jou Khong, Samantha Sie, Qin Kane Toh, and others, thank you for enriching my time here in Cambridge. I am also grateful to Athina Aruldass, Aleya Marzuki, Joshua Khoo and Caspar Geißler, not only for their friendships, but also for their insightful comments and discussions on earlier drafts of my chapters, which improved the writing of this thesis greatly.

I am indebted to my dear uncle, Sunny Lee, for without his support, this PhD would not have been possible. Thank you Mum and Dad, and my siblings, Jze Ling and Yu Her. Thank you Trinity Hall, for providing pastoral, and in particular hardship support during the covid-19 pandemic.

Finally, I would like to thank my partner, Jie Yi Ng, whose love and support I can always count on.

Table of Contents

List of Figures

List of Tables

List of Supplementary Figures

List of Supplementary Tables

Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
AIC	Akaike Information Criterion
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
A-O	Action-Outcome
AUD	Alcohol Use Disorder
AUDIT	Alcohol Use Disorder Identification Test
BIC	Bayesian Information Criterion
BIS-11	Barratt Impulsiveness Scale
COHS	Creature of Habit Scale
CUD	Cocaine Use Disorder
DASS-21	Depression, Anxiety and Stress Subscale (21-item)
DSM-5	Diagnostic and Statistical Manual (5^{th} edition)
DSM-IV-TR	Diagnostic and Statistical Manual (4^{th} edition) Text revision
DWI	Diffusion Weighted Imaging
FA	Fractional anisotropy
fMRI	Functional Magnetic Resonance Imaging
HDI	Highest Density Interval
HLAT	Habit learning at test
HSCQ	Habitual Self Control Questionnaire
IQ	Intelligence Quotient
M	Mean value
MINI	Mini International Neuropsychiatric Inventory
MLE	Maximum Likelihood Estimation

MRI	Magnetic Resonance Imaging
NART	National Adult's Reading Test
NHLAT	No habit learning at test
OCD	Obsessive-Compulsive Disorder
OCDUS	Obsessive-Compulsive Drug Use Scale
OFC	Orbitofrontal cortex
PIT	Pavlovian-to-Instrumental Transfer
RDoC	Research Domain Criteria
RL	Reinforcement Learning
ROI	Region of Interest
SCID	Structured Clinical Interview for the DSM-IV
SCR	Skin Conductance Response
SD	Standard deviation
SEM	Standard error to the mean
S-R	Stimulus-Response
SSRI	Selective serotonin reuptake inhibitor
SUD	Substance use disorder
StimUD	Stimulant use disorder
UNODC	United Nations Office for Drugs and Crime
vmPFC	Ventromedial prefrontal cortex

Chapter 1: General Introduction

1.1 Substance Use Disorder and reinforcement learning

"But here's the rub of addiction. By its nature, people afflicted are unable to do what, from the outside, appears to be a simple solution—don't drink. Don't use drugs."

(Sheff, 2008, p.184)

This heart-wrenching quote from David Sheff's *Beautiful Boy*, a first-person account of his son's battle with alcohol and methamphetamine addiction, describes the reality of addicted users that many cannot fathom. Substance use disorder (SUD henceforth for the chapter), a condition that affects approximately 35 million people worldwide (UNODC, 2021), is primarily characterised by maladaptive patterns of drug use that span across domains of impaired control, social impairments, risky use and pharmacological tolerance (American Psychiatric Association, 2013). For these patients, drug use dominates their lives, to the point where they risk ill health, familial breakups, neglecting school or work responsibilities, in favour of scoring and using drugs. In some cases, even prior near-death experiences (e.g. by overdose) have little effect in deterring them from drugs in the future. These pathological behaviour patterns are widely recognised as a consequence of disrupted psychological and neurobiological processes that subserve adaptive behaviour (Volkow et al., 2016). Thus, the main impetus for cognitive research in SUD has been to elucidate the nature of these disruptions. Exact causes for such maladaptive behaviours are likely multifactorial; over the years, many psychological theories have been proposed to understand maladaptive behaviour, including, among others, aberrant incentive salience (Berridge & Robinson, 2016; T. E. Robinson & Berridge, 1993), disrupted self-regulation (Baler & Volkow, 2006; Baumeister, 2003), dysregulated opponent processes (Koob et al., 1989; Koob & Le Moal, 2008; Solomon & Corbit, 1974), and impaired interoceptive awareness (Goldstein et al., 2009; Paulus et al., 2009). Although these theories have made great strides in advancing our knowledge of addictive behaviours, the cognitive characteristics of maladaptive behaviours beyond drug-taking in SUD remain elusive. This thesis will focus on an emerging yet understudied perspective of maladaptive behaviour in human SUD, which is conceived in terms of aberrant reinforcement learning.

Actions are guided by their consequences. This innate tendency to learn "by carrot or by stick" is known as reinforcement learning, and explains how humans maintain adaptive and functional behaviours to serve our best interests (Niv, 2009). For example, we learn to revisit restaurants where we had good dining experiences, or we stop drinking alcohol when feeling uncomfortable or nauseated. Indeed, many people are driven to consume psychoactive drugs because of their reinforcing properties. In a subset of these people, however, drug-taking persists even when its effects are highly detrimental. For example, someone with alcohol use disorder, even with aversive visceral sensations (e.g. vomiting and nausea), persists with drinking, and drinks again in the future. Decades of research identified that chronic use of drugs like alcohol or cocaine targets brain systems implicated in reward and motivation (Volkow et al., 2016). Thus, maladaptive behaviours characteristic of drug addiction are thought to reflect aberrant reinforcement learning (Hyman, 2005; Maia & Frank, 2011). Impaired reinforcement learning processes, which go against self-preservation instincts, pose serious challenges to addiction recovery, yet their nature is complex and only partially understood.

Broadly, this thesis sets out to study reinforcement learning in SUD. This first chapter serves as an introductory chapter to outline the state of the research into the learning impairments in human SUD. I begin by describing the psychological and neural processes that underpin learning under normal circumstances, including reinforcement learning and instrumental learning processes. This is followed by a review of the extant literature on the learning mechanisms affected in SUD or by chronic drug use. In closing, this chapter identifies the outstanding questions, and delineate the aims and hypotheses tested in this thesis.

1.2 Learning mechanisms in health

The study of learning functions has emerged as an influential framework to explain the mechanisms of functional behaviour. In this section, I will outline two learning processes that greatly shaped our recent understanding of adaptive behaviour: reinforcement learning and instrumental learning. I acknowledge that these two learning systems have substantial overlaps with one another (e.g. instrumental actions are initially learnt through reinforcement). However, I elect to discuss each process in turn, as the former relates more specifically to feedback learning, whereas the latter involves the regulation of learned actions.

1.2.1 Learning from feedback: reinforcement learning

Imagine that you are gambling on an electronic slot machine with two levers. Whilst playing, you notice that picking the right lever pays out more often, whereas the left lever frequently loses your money. Any sensible person would naturally bias their choices towards the right lever. This phenomenon can be explained by reinforcement learning (RL): a theoretical model of how humans use past consequences to better guide future behaviour (Niv, 2009). The goal of reinforcement learning is to make predictions on which action maximises one's benefits or minimises unfavourable outcomes, thereby facilitating functional and adaptive behaviours (Niv, 2009; Sutton & Barto, 1998).

Reinforcement learning is an example of an instrumental behaviour that encompasses two elements: associative learning and reinforcing feedback. Associative learning refers to the learning of the contingency between specific events or actions, and specific outcomes (Shanks, 1995), such as knowing the right lever is linked with winning more money. These learned contingencies are then reinforced by the outcome (positive or negative), as predicted by the law of effect (Thorndike, 1911). For example, a mother praises her son when he finishes his homework. This positive feedback then strengthens the likelihood of the son completing his homework in the future. Actions could also be reinforced in a negative manner, such that the outcome reduces the likelihood of negative consequences, e.g. one may choose to get the annual flu shot to avoid experiencing the negative health consequences from a previous flu.

1.2.2.1 Psychological and neural substrates of reinforcement learning

There are several components that aid optimal reinforcement learning, but more broadly reinforcement learning concerns the learning and update of subjective values of each available action, and selecting actions that maximises that value (Sutton & Barto, 1998). These subjective values are intimately linked with the human orbitofrontal cortex (OFC) and its animal homologue (Kringelbach, 2005; O'Doherty, 2004; Rolls, 2004; Schoenbaum et al., 2011), and lesions to these regions impair subjects' ability to adjust actions based on incentive values (Gallagher et al., 1999; Izquierdo et al., 2004; Jones & Mishkin, 1972). Considerable research suggests that the learning and update of this value depend on the neurotransmitter

dopamine. One notable role of dopamine in supporting reinforcement learning is to signal prediction errors – the discrepancies between expectations and actual outcomes (Glimcher, 2011; McClure et al., 2003; Montague et al., 2004). Seminal work by Schultz and colleagues (1997) identified in non-human primates that when presented with an unexpected reward, midbrain dopaminergic neurons sharply increase in activity. However, when this reward becomes predicted by a discriminative cue, the onset of this phasic signal changes from reward receipt to the predictive cue – signalling reward prediction instead of reward. Conversely, when the reward was expected but did not materialise (reward omission), there was a sharp dip in dopaminergic neuronal activity. These important observations led to the proposal of a dopaminergic learning signal that guides reward prediction and may underpin reinforced behaviour. Motivated by this hypothesis, subsequent works have supported the relationship between dopamine prediction errors and feedback learning (Frank et al., 2004; Pessiglione et al., 2006; Steinberg et al., 2013). In humans, these prediction error signals are also predominantly localised in regions with rich dopamine innervations, including the midbrain, striatum (ventral and dorsal), OFC, and the anterior insula (Bayer & Glimcher, 2005; D'Ardenne et al., 2008; Jensen et al., 2007; Menon et al., 2007; Pagnoni et al., 2002; Seymour et al., 2007).

It has been suggested that the learning of value differs with opposing valences, mostly from neural evidence of distinct regions associated with appetitive and aversive learning: learning from positive outcomes is linked with the medial OFC and the ventral striatum, whereas learning from negative feedback involves the lateral OFC, anterior insula, anterior cingulate cortex, the periaqueductal grey, and the lateral habenula (Elliott, Dolan, et al., 2000; Elliott, Friston, et al., 2000; Hennigan et al., 2015; Jensen et al., 2007; Lawson et al., 2014; Palminteri et al., 2012; Roy et al., 2014). However, since positive or negative outcomes are highly dependent on contexts and reference points (Kim et al., 2006), it is difficult to delineate a precise neuroanatomical distinction between these systems (Pessiglione & Delgado, 2015). An alternate neuro-computational model suggests that the dopaminergic receptors in the dorsal striatum play a role in determining approach and avoidance behaviour, as different dopamine neuronal populations have differential sensitivity to phasic and tonic levels of dopamine (Frank, 2005; Frank & O'Reilly, 2006). Specifically, D1 receptors are excitatory and are thought to be sensitive to phasic dopamine bursts related to positive (reward) prediction errors. By contrast, D2 receptors are inhibitory by nature and sensitive to dips in tonic dopamine levels, which is

linked with negative (aversive) prediction errors. Because these two receptor subtypes have been associated with positive and negative prediction errors, it has also been argued that D1 and D2 receptors differentially underpin learning from positive and negative feedback respectively (Cox et al., 2015; Hikida et al., 2010; Surmeier et al., 2007). There is also growing evidence for the opponency of dopamine and serotonin, such that these neurotransmitter systems modulate appetitive and aversive behaviours respectively (Boureau & Dayan, 2011; Cools et al., 2011; Daw et al., 2002). It is likely that these views overlap significantly with one another, but nevertheless, they allude to the notion that different valences are subserved by different neurobiological systems.

In addition to value signals, another important component of reinforcement learning involves the action selection process. Logically, selecting the option with the highest value would be the default way to maximise gains. However, in an uncertain environment, agents usually have to evaluate their choices based on the information at hand (Sutton & Barto, 1998). In such situations, humans could either *exploit* their knowledge and stick with choices with the highest value, or *explore* other choices to gather more information. This is known as the exploration/exploitation trade-off (also known as stochasticity or reinforcement sensitivity) in the reinforcement learning literature, and this process has been modelled experimentally in a probabilistic learning task (Daw et al., 2006). Daw and colleagues (2006) simulated this process by having participants complete a reinforcement learning task with probabilities that change over time. Their data identified that the tendency to exploit values is associated with activations in the medial OFC – unsurprisingly a region implicated in value representation (O'Doherty, 2004). By contrast, exploratory activity involves the frontopolar cortex and the intraparietal sulcus (Daw et al., 2006). Interestingly, there is some evidence that the brain also signals the uncertainty levels for the available choices, notably in the anterior cingulate cortex (Behrens et al., 2007; Brown & Braver, 2005). These results support the notion that value and probabilistic information are integrated during the action selection process. However, although action selection is often driven by action values and its probability, this is not always the case. For example, an action could be selected simply because of familiarity, irrespective of reinforcement history – a process known as stickiness (also known as perseverative tendency; Christakou et al., 2013), and this process has been linked with compulsive disorders like substance use disorder and OCD (Kanen et al., 2019). Whether this process plays out differently in psychopathologies is an active and ongoing research endeavour.

Reinforcement learning is likely not an isolated process. A parallel and overlapping line of research has identified two dissociable memory systems that are involved in acquiring new knowledge – namely declarative and non-declarative memory (Packard & Knowlton, 2002; Seger & Miller, 2010). On one hand, declarative memory facilitates the rapid learning of factual knowledge or simple rules (e.g. at a traffic light, red means stop, green means go). This system is flexible and conscious, and depends on the hippocampus (Squire & Zola-Morgan, 1991). On the other hand, non-declarative memory supports the incremental learning of complex associations through trial-and-error. Compared to declarative memory, the non-declarative memory system is slow and implicit, but detects regular patterns in a changing environment (e.g. cloudy days and strong winds most likely predict an incoming storm); this system is thought to be striatal-dependent (Poldrack et al., 1999; Shohamy et al., 2008). Both declarative and non-declarative memory support reinforcement learning by mediating the acquisition of associative knowledge, albeit through different routes (Seger & Miller, 2010). Converging evidence suggests that these systems are dissociable (Knowlton et al., 1996; Poldrack et al., 2001; Shohamy, Myers, Grossman, et al., 2004). A double dissociation is demonstrated in one such study with a probabilistic category learning task designed to test these systems (Knowlton et al., 1996). Patients with Parkinson's disease, characterised by deficits in striatal dopamine, showed impairments in non-declarative, but not declarative memory, as reflected by poor trial-and-error learning but preserved task knowledge. This is in stark contrast to amnesia patients (who typically show hippocampal damage), who were able to learn by trial-and-error (non-declarative memory), but were unable to retain any factual knowledge about the task. The dissociable nature of these systems thereby allows learning to occur via different routes, and makes it possible for one system to compensate for the other in neuropathological conditions (Gluck et al., 2002; Shohamy, Myers, Onlaor, et al., 2004).

1.2.2.2 Experimental paradigms used to study reinforcement learning

Experimental paradigms that assess reinforcement learning typically involve the use of corrective feedback to enable participants to learn from their choices. Common tasks that probe this ability include the probabilistic reinforcement learning task (O'Doherty et al., 2004), the reversal learning task (Murphy et al., 2003), the Iowa Gambling Task (Bechara et al., 1994, 1997), and the Weather Prediction Task (Knowlton et al., 1996).

The probabilistic reinforcement learning task (sometimes also known as the n-arm bandit task) is a straightforward way to assess reinforcement learning (O'Doherty et al., 2004; Pessiglione et al., 2006). Participants are presented with several choices, and must learn by trial-and-error to select choices that more often lead to rewarding feedback, or more likely avoid punishing feedback. Feedback could take the form of positive (e.g. winning 10 points) or negative feedback (e.g. losing 10 points). In most cases, feedback is probabilistic (e.g. choosing A most likely leads to reward, but not always), but feedback can also be deterministic, depending on the process being assessed (e.g. to ensure participant reaches a certain criterion). Learning is then measured by task accuracy (e.g. rate of optimal choices) over time, or by inferring its latent variables with computational models (discussed later). Typically, these tasks have been used to assay reward or punishment learning, which are hypothesised to be disrupted in some neuropsychiatric disorders (Heinz et al., 2016).

Sometimes, adaptive behaviour depends on one's flexibility in switching behaviour when previously rewarded choices are no longer beneficial. This can be modelled in an extension to the reinforcement learning task, known as the probabilistic reversal learning task (Gallagher et al., 1999; Jentsch et al., 2002). Similar to reinforcement learning, this task uses corrective feedback to inform participants the outcome of their choices. However, instead of a stable contingency, this task switches learned contingencies from time to time, such that a previously rewarded choice becomes punished. The rationale behind this is to ascertain if participants switch their choices according to feedback, or stick to a previously rewarded choice despite negative feedback. The latter is thought to signify cognitive inflexibility in face of changes – a hallmark of compulsivity. There could be many reasons for such inflexible behaviour. One possibility is impaired learning from feedback (Fineberg et al., 2014), but this has yet to be conclusively shown. This inflexibility could also in part reflect dysregulated top-down behavioural control over instrumental actions (Vandaele & Janak, 2018).

Real-life decisions usually require integrating potential costs and benefits, which also depends on feedback learning. This is modelled in a task known as the Iowa Gambling Task (Bechara et al., 1994). In this task, participants are required to draw cards from four deck of cards to

maximise their point gains. For each card drawn, participants receive varying magnitudes of monetary rewards, but on some card draws, these rewards are accompanied with penalties. For example, a participant could win $100 but lose $150 on a single trial. Unbeknownst to the participant, two of the card decks (termed risky decks) have higher monetary gains, but incur larger costs, which ultimately results in a net loss; the other two card decks (termed safe decks) yield smaller monetary gains, but would incur less penalty, and choosing these would eventually result in a net gain. Advantageous decision-making is reflected by the number of draws made from the safe decks against the risky decks. Further, it has been hypothesised that autonomic signals elicited in anticipation of penalties help guide decisions, and this process recruits the ventromedial prefrontal cortex (vmPFC). Bechara and colleagues identified that patients with vmPFC lesions consistently chose the risky deck over the safe decks, and did not elicit any anticipatory skin conductance response (SCR) (Bechara et al., 1994, 1997).

Another way to gain insight into the reinforcement learning system is by probing the declarative and non-declarative memory systems that support it. This is commonly done with a probabilistic category learning task known as the Weather Prediction Task (Gluck & Bower, 1988; Knowlton et al., 1996). Participants learn by trial-and-error to categorise card combinations into one of two categories (sunshine or rain). On each trial, the card combination consists of either one, two or three cards out of four possible unique tarot cards. Each of the four unique cards vary in their probability of producing sun, which gives an overall probability when combined. Task performance (% optimal responses made) is traditionally viewed as a measure of non-declarative learning. However, given that the two systems operate concurrently (Packard & Knowlton, 2002; Poldrack & Foerde, 2008), prior research has shown that it is possible to approach this task using various strategies: explicit verbalisable rule-based strategies (declarative) or implicit non-verbal learning by trial-and-error (non-declarative) (Gluck et al., 2002). Moreover, researchers have created a variant of the task, such that the task can only be learnt via declarative processes (explicit memorisation) instead of feedback (Poldrack et al., 2001; Shohamy, Myers, Grossman, et al., 2004), thus making it possible to dissociate between the two systems more cleanly. Combining analyses of both performance and strategy for the weather prediction task can provide insight into the processes that facilitate the learning of reinforced behaviour.

1.2.2 Goal-directed versus habit learning: dual-process theory of instrumental learning

Control over learned actions, such as those acquired through reinforcement learning, is increasingly conceived in instrumental learning terms (Balleine & Dickinson, 1998; Dickinson, 1985). This theory predicts that learned behaviours are regulated by two dissociable processes: a goal-directed system and a goal-independent habit system. The goal-directed system regulates actions that are motivated by a desired outcome. Psychologically, actions are deemed goal-directed if: (1) actions are directed towards achieving a certain outcome, (2) the outcome itself is desirable (Balleine & Dickinson, 1998; de Wit & Dickinson, 2009). An example of a goal-directed behaviour is being motivated to study hard in order to get good grades in an upcoming exam. In associative terms, goal-directed actions (e.g. studying hard) are mediated by action-outcome contingencies (e.g. studying hard *causes* good grades). This form of behaviour relies on knowledge between actions and outcomes, and therefore is characterised as prospective and flexible, but computationally costly as it requires conscious deliberation. Because of its sensitivity to reinforcement, I argue that reinforcement learning, at least under conditions of minimal training, falls largely within the remit of goal-directed learning.

However, when actions become repetitive in a predictable environment, the brain automates actions into habits, so that they become reflexive and easier to perform (Wood & Neal, 2007). In psychological terms, habits no longer depend on intentions, outcomes, or action-outcome contingency (Dickinson, 1985). Rather, behaviour is stimulus-bound i.e. elicited by the presence of a conditioned cue that was previously predictive of a prior reward (Adams & Dickinson, 1981). Habits are adaptive such that they allow us to perform routine behaviours without the need for conscious deliberation, thereby maximising efficiency (Wood et al., 2002). However, they lack flexibility because, unlike goal-directed actions, behaviours controlled by the habit system are not modified by outcomes (Dickinson, 1985). Clearest examples of these are action slips – actions performed unintentionally. For instance, we unwittingly turn on the light switch upon entering our bedroom during daytime; or entering the kitchen prompts us to open the fridge in the absence of any intention to look for food. Environmental cues trigger habitual responses. The goal-directed and the habit systems are thought to work in parallel to regulate instrumental actions (Balleine & O'Doherty, 2010). As such, adaptive behaviour has been hypothesised to arise from the flexible switching between the two systems, depending on situational demands.

1.2.2.1 Psychological and neural substrates of goal-directed and habit learning

Actions are jointly regulated by the goal-directed and habit systems in the brain, but under different circumstances, one system can dominate over another to control instrumental actions (Balleine & O'Doherty, 2010). A common method to bias actions towards the habit system is with extensive repetition. Adams and Dickinson (1981) found that after extensive repetition of an appetitive behaviour, rodents are no longer sensitive to changes in outcome value, which is indicative of habitual control over behaviour. Subsequent investigations characterise the neural substrates underpinning goal-directed to habit learning. There is widespread agreement that the goal-directed and habit systems are subserved by dissociable brain systems. On one hand, the goal-directed system is underpinned by fronto-striatal systems closely linked with value computation, such as the anterior caudate, ventral striatum and the medial OFC (Balleine & O'Doherty, 2010). On the other hand, the habit system is linked with cortico-striatal systems implicated in motor responses, including the putamen and the premotor cortex (Knowlton & Patterson, 2016). Early studies in rats over-trained with appetitive food-seeking behaviours have identified, through brain lesion procedures, that inactivating the dorsomedial striatum (homologue to the human anterior caudate) causes behaviour to be habitual and insensitive to changes in outcome or contingency (Yin, Knowlton, et al., 2005; Yin, Ostlund, et al., 2005). By contrast, inactivating the dorsolateral striatum (homologue to human putamen) enabled habitual rats to regain sensitivity to outcomes and contingency (Yin et al., 2006). This reciprocal relationship is replicated in human studies. For example, overtraining an operant response increases the relative engagement of the putamen (Tricomi et al., 2009). By contrast, the medial OFC and the anterior caudate nucleus were sensitive to both changes in outcome value (Valentin et al., 2007) and action-outcome contingency (Tanaka et al., 2008). Further, the inter-individual differences within the cortico-striatal connectivity have also been demonstrated to underlie variations in habit formation (de Wit et al., 2012).

As goal-directed and habit systems are viewed to regulate behaviour in parallel, there is an increase in interest in the mechanisms that arbitrate the balance between the two systems. Converging evidence suggests that the lateral prefrontal cortex, which has been implicated in goal-directed planning, plays a role in this balance (Bogdanov et al., 2018; Lee et al., 2014). Lee and colleagues (2014) showed that switching flexibly between computational proxies of

goal-directed and habitual behaviours (i.e. model-based and model-free behaviour) during a volatile learning task corresponds to the activity in the inferolateral prefrontal cortex, but this result was only correlational at that time. Subsequently, Bogdanov and colleagues (2018) sought to test this hypothesis by applying a theta-burst stimulation to transiently inhibit the inferolateral prefrontal cortex before completing an instrumental learning task (the slips of action task). They found that participants' task performance was more biased towards habitual responding when the inferolateral prefrontal cortex was inhibited relative to control conditions, suggesting a causal role for this region in switching between the control systems depending on the situational demands. Therefore, understanding the balance between goal-directed and habitual responding does not only involve each individual system, but also the arbitration between the two.

1.2.2.2 Experimental paradigms of goal-directed and habit systems

Conventional experimental paradigms of habits in animal studies have taken advantage of the independence of habits from goals. Essentially, habits are viewed as the reciprocal of goal-directed actions, and have been operationally defined as the absence of goal-directed behaviour. Two classical tests of habits have developed from this: outcome devaluation, which manipulates the value of the outcome (Adams & Dickinson, 1981); and contingency degradation, which alters the action-outcome contingency (Dickinson & Balleine, 1994; Hammond, 1980). In both task paradigms, instrumental responses are learnt via overtraining in the presence of a predictive cue, but they differ in terms of subsequent phases. In outcome devaluation paradigms, once a learned response is established, the reinforcer is then devalued. This can be done in multiple ways, such as inducing satiety or inducing sickness, with the goal to reduce the desirability or value of the outcomes so that they no longer should motivate behaviour. If humans (or animals) continue to respond despite devaluation, their actions are thought to be no longer guided by the outcomes, but instead by the cue, which is a feature of a habit. By contrast, in contingency degradation paradigms, once the initial action-outcome contingencies are established, they are gradually degraded, so that the causal relationship between actions and outcomes no longer exist. If the instrumental response continues despite degradation, that action is said to be controlled by the habit system. Both outcome devaluation and contingency degradation were successful paradigms in identifying habits in animals, and have been translated for human experiments (de Wit et al., 2007, 2012; Vaghi et al., 2019).

An alternative model conceived instrumental processes of goal-directed and habitual control in computational terms (Daw, 2014; Daw et al., 2011; Dayan & Daw, 2008). On the one hand, goal-directed control is thought to involve an internal model of the environment as it typically tracks values and action-outcome contingencies during action selection. Therefore, it is known as a model-based process that involves forward prospective simulations of all possible outcomes to maximise values. On the other hand, habit learning may only involve the stamping-in of stimulus-response associations through reinforcement and is largely guided by prior rewards. As such, the habit system is deemed a model-free system guided only by prior rewards (Dolan & Dayan, 2013). These processes could be tested in the two-step decision-making task (Daw et al., 2011; Gläscher et al., 2010). In this task, participants needed to pick a stimulus that they think might maximise their rewards in a two-stage decision process. Based on their choice selections on each stage, this task purportedly enables the dissociation between actions that are guided either by prior experiences of rewards (model-free) or the knowledge of the transition between stages (model-based).

1.2.3 Summary

In brief, reinforcement learning is concerned with the use of past reinforcement to guide future behaviour to maximise potential benefits and avoid aversive states. This process involves multiple latent processes such as value learning, reward prediction and action selection, most of which implicates the dopaminergic system. Supporting memory systems such as declarative and non-declarative processes can also facilitate reinforcement learning. Learned behaviours as such are thought to be regulated by the goal-directed system – which is sensitive to outcomes; or, upon extensive repetition, by the habit system that is insensitive to changes in outcomes or action-outcome contingency. Both reinforcement learning and instrumental learning theories are overlapping but distinct frameworks that can jointly be used to understand the psychological basis of maladaptive behaviour observed in SUD.

1.3 Learning impairments in substance use disorder

Chronic drug use induces neuro-adaptive changes that alter brain circuits (Volkow et al., 2004). In particular, the potent reinforcing effects of addictive drugs are attributed to its interactions

with brain reward and learning pathways, which these drugs alter over extended use (Hyman, 2005; Hyman et al., 2006; Koob & Volkow, 2010). Individuals addicted to stimulant drugs (e.g. cocaine, amphetamines, methamphetamine) and alcohol show notably reduced striatal dopamine receptors, which speaks for a downregulation of dopamine transmission (Martinez et al., 2004; Volkow et al., 1993, 1996). Hence, it is conceivable that dopamine-dependent processes such as learning and motivation are affected in substance use disorder. Our knowledge that dopaminergic pathways crucially underpin learned and motivated behaviours (Wickens et al., 2007) led to two prevailing hypotheses within the addiction literature: (1) reinforcement learning is impaired in SUD; (2) SUD patients show increased habitual control over behaviour. This section discusses the available evidence relevant to these hypotheses, with a special focus on cocaine and alcohol.

1.3.1 Altered feedback learning in human cocaine and alcohol use disorder

Deficits in optimising actions with reinforcing feedback is thought to be one pathway in which pathological drug use is sustained (Maia & Frank, 2011). This section discusses several strands of evidence that provide support for this hypothesis. Evidence presented here is largely derived from behavioural and neuroimaging experiments that incorporate the use of corrective feedback to adjust ongoing behaviour, such as gambling tasks, reinforcement and reversal learning tasks. The synthesis of this evidence would lead to the conclusion that despite the clear impairments of feedback learning in cocaine and alcohol use disorder, the underlying cognitive profile in these patients is less well understood.

Adaptive actions in real life are guided by their consequences, and the probabilistic reinforcement learning task is an appropriate means to investigate this process. Converging evidence show that chronic users of methamphetamine (Harlé et al., 2015), cocaine (Morie et al., 2016; Strickland et al., 2016) and alcohol (Jokisch et al., 2014; Rustemeier et al., 2012) learned slower than their healthy counterparts in response to monetary gains, consistent with the notion that reinforcing feedback guides actions less well. This deficit putatively impacts drug use initiation and subsequent relapse. For instance, cocaine and methamphetamine users who showed attenuated striatal activations during reinforcement learning were more likely to relapse within a 12-month period (Stewart et al., 2014a, 2014b). Moreover, the use of negative feedback to learn avoidance behaviour, although less studied, also seems to be impaired;

studies found that negative feedback such as electric shocks, symbolic errors, disgust cues or monetary losses were not effective in altering behaviours in patients addicted to alcohol or cocaine (Ersche et al., 2014, 2016; Hester et al., 2013; Thompson et al., 2012). However, poor reinforcement learning can result from a multitude of factors, and the specific nature of such impairments is less clear.

Deficits in reinforcement learning also hamper one's ability to adjust behaviour flexibly according to situational demands, which results in the abnormal persistence of behaviour. This is known as perseveration and is viewed as a behavioural marker for compulsivity (Figee et al., 2016). Perseverative tendencies and cognitive inflexibility are measured in reversal learning tasks, which require the flexible update of incentive values acquired from feedback. Studies in stimulant- (Ersche et al., 2008; Ersche, Roiser, Abbott, et al., 2011) and alcohol-addicted individuals (Vanes et al., 2014) have shown behavioural signatures of perseveration, such that initially learned contingencies do not flexibly adapt to the changed feedback. Deficits in reversal learning in patients could also be traced to impairments in latent processes of learning, such that patients are more likely to repeat prior choices irrespective of incentive values (Kanen et al., 2019). Moreover, this deficit has been linked to reduced grey matter in the orbitofrontal cortex in alcohol use disorder patients (Moreno-López et al., 2015), further corroborating the role of OFC in signalling incentive values, and its dysfunction in addicted individuals.

The decision to continue to use drugs by SUD patients is thought to be largely driven by immediate positive reinforcement at the expense of long term losses. Therefore, these patients are thought to exhibit decision-making deficits that have been characterised as a "myopia for future consequences" (Bechara, 2005). This hypothesis is tested in an experimental task that requires the real-time incorporation of positive and negative prospects of a decision, such as the Iowa Gambling Task. Individuals addicted to cocaine and alcohol were more likely to select cards that led to immediate large rewards, even if their choices are accompanied by large losses (resulting in a net loss over time), revealing a form of risky decision-making (Bechara et al., 2001; Verdejo-Garcia et al., 2007). By contrast, their healthy counterparts were more likely to select the advantageous deck with smaller gains, but also smaller losses (resulting in a net gain over time). A subsequent study revealed that cocaine-dependent users showed a blunted anticipatory SCR, suggesting that the unpredictable losses did not affect their decisions – a

behavioural profile that is similar to those with ventromedial prefrontal cortex (vmPFC) lesions (Bechara & Damasio, 2002). The involvement of the vmPFC is further confirmed in a subsequent study where cocaine users also show abnormal vmPFC activation when completing this task (Bolla et al., 2003). To confirm the specificity of the anticipatory SCR, Bechara and colleagues (2002) conducted a subsequent study with a reversed design – participants instead receive immediate punishing feedback on their choices, but occasionally receive a delayed reward. The results were opposite to that of the standard design. Cocaine users showed intact anticipatory SCR for delayed reward and unimpaired performance, thereby confirming that their choices are driven by reward at the expense of punishment (Bechara et al., 2002). Alcohol users share a similar behaviour profile – they too, tend to select riskier options and lack anticipatory SCR to unpredictable punishments (Bechara et al., 2001; Brevers et al., 2014; Loeber et al., 2009). Collectively, results from these decision-making studies characterised SUD impairments as a hypersensitivity to reward, but hyposensitivity to future losses, which may mirror their drug use in real life.

The prevailing view is that reinforcement learning impairments in addicted individuals are related to dysfunctions in dopamine-related processes (Kalivas & O'Brien, 2008; Wise & Robble, 2020). A candidate mechanism for that is striatal prediction error signals (Keiflin & Janak, 2015). Several studies on stimulant-addicted individuals have found blunted striatal prediction error signals towards unexpected outcomes (Parvaz et al., 2015; Rose et al., 2014; Tanabe et al., 2013). In particular, this was specific to unexpected negative feedback, and correlates with poor learning (Parvaz et al., 2015; Tanabe et al., 2013). Although studies on alcohol use disorder patients did not identify abnormalities in striatal prediction errors per se, further analyses revealed aberrant functional connectivity between the striatum and the prefrontal cortex, which was presumed to underpin action selection (Deserno et al., 2015; Park et al., 2010). Another candidate mechanism that supports reinforcement learning is the role of dopamine D_2 receptors, which is consistently shown to be downregulated in chronic alcohol and cocaine users (Heinz, 2002; Volkow et al., 1993, 1996). It has been hypothesised that dopamine D_2 receptors are more sensitive to negative prediction errors, which facilitate learning from negative feedback (Frank & Hutchison, 2009; Hikida et al., 2010). Indeed, individuals with genetically-determined lower dopamine D_2 receptors have problems incorporating negative feedback into their actions (Jocham et al., 2009; Klein et al., 2007). Thus, it is possible that addicted individuals are impaired in learning from negative feedback,

which could perpetuate their pathological drug taking patterns, as the effects of negative feedback, which should deter behaviour, is dampened. Whether this really is the case remains to be determined. However, in addition to negative reinforcement learning, reductions in dopamine D_2 receptors in rats have been shown to predict trait impulsivity and subsequent cocaine reinforcement in cocaine-naïve rats (Dalley et al., 2007), which suggests another possible mechanism for low dopamine D_2 receptor density to elevate addiction risk (Gleich et al., 2021; Noble et al., 1993). It is likely that dopamine underpins several dissociable reinforcement learning processes (Frank et al., 2007), but these have not yet been investigated in the context of SUD.

It has also been suggested that non-declarative memory, a system supporting reinforcement learning, is also impaired, although evidence for this is limited and equivocal. The non-declarative system in these studies were tested with the Weather Prediction Task. Although a study found preserved non-declarative memory in cocaine-dependent users (Vadhan et al., 2008), other studies have found clear deficits in short and long term cocaine users (Kumar et al., 2019; Vadhan et al., 2014). In particular, deficits in non-declarative memory were only present after heavy alcohol and cannabis use were statistically controlled for in the cocaine users (Vadhan et al., 2014). However, it is noteworthy that these studies have only analysed summary score performance for the task but did not identify their learning strategies or assessed declarative memory. Hence, the nature of this profile remains unclear.

To summarise, extant evidence largely supports the notion of altered learning from feedback in cocaine and alcohol users, which could have implications for maladaptive behaviour frequently reported in patients. Possible mechanisms for this impairment include dysfunctions to dopaminergic processes implicated in prediction error signalling and learning from negative feedback, and possibly memory systems associated with learning. However, the nature of these deficits have not been conclusively determined in human drug users.

1.3.2 Habit predominance in substance use disorder

An alternative but not mutually exclusive hypothesis that explains maladaptive behaviour in SUD concerns the dysregulated regulatory control over learned actions. It has been

hypothesised that the balance between goal-directed and habitual behaviours is disrupted in SUD, heavily biased towards the latter (Everitt & Robbins, 2005, 2016). Since actions controlled by the habit system are automatic by nature and insensitive to consequences, this is thought to be one way in which bad habits (e.g. drug use) persist despite adverse consequences. Most animal studies test this hypothesis with the outcome devaluation procedure. These studies have shown that whilst healthy rodents reduced responding after outcome devaluation, those treated with stimulants or alcohol did not show this decrement, which is indicative of habitual control (Corbit et al., 2012; Corbit, Chieng, et al., 2014; Dickinson et al., 2002; Hopf et al., 2010; Lesscher et al., 2010; Miles et al., 2003; A. Nelson & Killcross, 2006). It is noteworthy that instrumental response only becomes habitual after prolonged training – cocaine and alcohol-treated rats with moderate training do not demonstrate this habit bias (Corbit et al., 2012; Zapata et al., 2010). Nevertheless, current evidence suggests that extended cocaine and alcohol exposure seems to facilitate habit formation, as treated rats formed habits quicker than their non-treated counterparts (Mangieri et al., 2012; Nordquist et al., 2007). Neurally, this transition might be underpinned by a quicker devolution of control from the dorsomedial striatum to the dorsolateral striatum (Belin & Everitt, 2008; Corbit et al., 2012). Indeed, inactivating the dorsolateral striatum in alcohol-exposed rats that received extensive practice renders initially habitual behaviours sensitive to outcomes again, further corroborating the enhanced habitual control over their actions (Corbit et al., 2012). However, the psychological and neural mechanisms that underlie this shift is yet unclear.

In humans, evidence for this process is limited and equivocal. Human studies mostly use outcome devaluation tasks, but adopt different devaluation techniques, such as instructed devaluation (e.g. participants are told that outcome A is no longer valuable) in the slips of actions task (de Wit et al., 2007); or taste aversion techniques that render certain outcomes undesirable (van Timmeren et al., 2020). Although the majority of available studies in human cocaine and alcohol use disorder contend that learned behaviour becomes habitual quicker in these patients (Ersche et al., 2016; Sjoerds et al., 2013), this is not always the case (van Timmeren et al., 2020). An alternative paradigm often used to study habits computationally is the two-step task. This task approximates goal-directed and habits as computational accounts of model-based and model-free control respectively. Extant evidence shows that alcohol users (Sebold et al., 2014) and methamphetamine users (Voon et al., 2015) have reduced model-based, but comparable model-free behaviour. In other words, these studies tell us that the goal-

directed system, rather than the habit system is impaired in these drug users. However, subsequent studies with larger and more heterogeneous samples did not replicate this finding (Nebe et al., 2018; Sebold et al., 2017). Interestingly, one study found that alcohol use disorder patients who relapsed had lower neural signatures of model-based control (Sebold et al., 2017), which echoed earlier findings in reinforcement learning in methamphetamine users (Stewart et al., 2014b), suggesting that the ability to exert goal-directed control might be related to protracted abstinence during recovery.

In summary, whilst the evidence for a habit predominance over instrumental action is compelling in animal studies, the findings in humans are equivocal. Moreover, the exact nature of this disrupted balance, whether this is related to an impaired goal-directed system, an augmented habit system, or poor regulation between the two, remains unknown.

1.4 Outstanding questions and the potential of computational learning models

It is clear that reinforcement learning is impaired in individuals with SUD, and this might be related to their impaired ability to respond appropriately to reinforcing feedback. However, there remain gaps in our knowledge pertaining to these areas. Clarifications to the psychological processes might be clinically informative:

(1) What are the underlying computational features that underpin such impairments? On a cognitive level, which components of reinforcement learning (e.g. value learning, action selection) contribute to the behavioural profile in SUD patients? A potential mechanism discussed earlier implicates deficits in dopamine neurotransmission, but this link has not been concretely established. I address these questions in Chapter 3.

(2) The differences in the declarative and non-declarative memory systems that support reinforcement learning are unclear in SUD. Do SUD patients rely on a different strategy during learning from feedback? This is addressed in Chapter 4.

In the context of behavioural control, SUD is linked with a predominance of the habit system over the goal-directed system (Everitt & Robbins, 2016). Consequently, conscious control over well-learned actions, such as drug use, might have devolved to a system insensitive to the consequences of such actions. One possible way this occurs, among several, is that the goal-

directed system fails to exert control when actions are rendered meaningless (Vandaele & Janak, 2018), but this has not been concretely shown in humans, which raises several outstanding queries:

(3) Does the initial goal-directed learning of appetitive behaviours affect subsequent habit formation? This question formed the basis of Chapter 5.

(4) Does enhanced habitual control seen in behavioural tasks also mean patients are more habitual in their daily lives? I explore this question in Chapter 6 with self-reported questionnaires.

(5) Do impaired reinforcement learning and habit predominance manifest during initial stages of harmful substance use? Although this habit predominance associated with SUD is clear within animal models of addiction, limited research has looked into the early stages of alcohol use disorder. Even in severe alcohol use disorder, findings in humans have been equivocal (Sebold et al., 2014; Sjoerds et al., 2013; van Timmeren et al., 2020), which could be attributed to limitations in existing task paradigms. I use a novel behavioural task to study these processes in a population characterised by harmful alcohol consumption in Chapter 7.

Reinforcement learning is increasingly studied using computational techniques. These techniques interface between our knowledge of computer science, neuroscience and psychology, providing an instrumental tool for researchers to deconstruct the learning process into its constituent components. Leveraging mathematical frameworks, computational learning models offer a mechanistic insight into the reinforcement learning process by providing a quantitative means to measure psychological processes (Huys et al., 2021). This offers the advantage of exploring latent variables quantitatively that were previously not accessible through summary scores alone. The advent of reinforcement learning algorithms proved fruitful in identifying behavioural and neural signatures that underpin learning. For example, Stout and colleagues (2004) applied an expectancy valence model to computationally deconstruct behavioural performance of a small sample of cocaine-dependent users on the Iowa Gambling Task. They identified a reduced loss weightage parameter in the cocaine user group, suggesting that cocaine-dependent users were less likely to take monetary losses into account during decision-making. A similar observation was made in abstinent heroin-dependent users, who also showed a reduced loss aversion parameter (Ahn et al., 2014). Both studies highlight

the utility of computational models in identifying latent processes that could account for behavioural performances. The use of modelling was also able to pinpoint selective mechanistic disruptions in compulsive disorders based on common experimental paradigms such as reversal learning tasks (Kanen et al., 2019). Therefore, computational methods would be useful in elucidating reinforcement learning impairments in SUD that are still elusive.

1.5 Thesis objectives and outline

The objectives of this thesis are two-fold: First, I sought to characterise the cognitive characteristics and extent of impairments in reinforcement learning and its supporting memory processes in mild and severe substance use disorder. Second, I aim to clarify how impairments to the goal-directed system, as a function of reinforcement learning, impact habit predominance in the context of substance use disorder. Before I report the behavioural findings, I first contextualise, in Chapter 2, my experimental findings by providing an overview of the samples tested, experimental paradigms used and the general statistical approach used to analyse behavioural data.

Chapter 3 tested the hypothesis that impaired reinforcement learning in severe SUD is due to disruptions to dopamine-dependent learning processes, such as the impact of feedback on behaviour and the tendency to pursue reward values. It reports behavioural findings from two independent patient samples of stimulant use disorder who completed a probabilistic reinforcement learning task that assessed learning from reward and punishment separately. To probe the neurochemical substrates of reinforcement learning impairments in these patients, patients in one study were subjected to pharmacological drug challenges that target the dopamine $D_{2/3}$ receptors with selective antagonist (400 mg amisulpride) and agonist (0.5 mg pramipexole) in a randomised, placebo-controlled, parallel, crossover trial design. I predicted that dopamine receptor blockade by amisulpride would further impair task performance, whilst the dopaminergic receptor agonist, pramipexole, would ameliorate learning performance in SUD.

Chapter 4 explored the notion that memory processes known to support reinforcement learning, namely declarative and non-declarative memory, are disrupted in cocaine use disorder. These

processes were measured in a well-known probabilistic category learning task, the weather prediction task (Knowlton et al., 1996). Participants completed two variants of this task, which required learning either from trial-and-error (non-declarative) or explicit memorisation (declarative). These learning methods are suggested to underpin striatal-dependent and hippocampal-dependent memory systems respectively. I also analysed participants' response strategy to determine how participants learn during this task. As these patients have been characterised with striatal deficits, I predicted that these patients rely less on strategies that require the striatum such as feedback learning, and more on hippocampal-based strategies such as direct memorisation.

Chapter 5 applied computational reinforcement learning algorithms to decompose the appetitive goal-directed learning performance in a published dataset and investigate its links with habit formation in cocaine use disorder. Whilst the view of a habit bias in addiction is widely discussed, the goal-directed system, which co-regulates instrumental behaviour with the habit system, has received comparatively less focus in human drug addiction. I tested the hypothesis that deficits in appetitive goal-directed learning, specifically in computational learning parameters, contribute to habit preponderance in patients with cocaine use disorder. Specifically, I predicted that impairments in computational parameters that explain appetitive goal-directed learning should predict the habit bias scores during the outcome devaluation phase.

Chapter 6 continued to interrogate the habit theory of addiction by measuring goal-directed and habit systems via self-reported questionnaires. If, as the theory posits, the habit system predominates behaviour in substance use disorder, this should manifest in either an increase of habitual tendencies, a reduction in goal-directed actions, or both, in daily lives of patients with substance use disorder. I measured habitual and goal-directed tendencies with the Creature of Habit Scale (Ersche et al., 2017) and the Habitual Self Control Questionnaire (Schroder et al., 2013) respectively. I also assessed the relationship between these self-reported measures and behavioural measures of goal-directed actions and habits. I predicted that relative to controls, cocaine use disorder patients would show higher scores on the subscales of the Creature of Habit Scale, but have lower scores on the Habitual Self Control Questionnaire.

In the same vein, Chapter 7 tested the habit theory of addiction with a novel behavioural paradigm, specifically its generalisation to a large online population characterised by harmful alcohol consumption. This novel behavioural paradigm places the learned instructions (goal-directed) and learned behaviours (habits) into conflict, to directly test which system prevails in influencing behaviour. I hypothesised that harmful alcohol drinkers, who are not formally diagnosed with alcohol use disorder, show enhanced habit formation, relative to a control population with only social drinking. In particular, harmful alcohol drinkers would more likely display a bias for learned (habitual) behaviours when there is a goal-habit conflict.

Finally, Chapter 8 integrates the findings from the experimental chapters and delineates the contributions of this thesis in elucidating reinforcement learning impairments frequently associated with SUD. I also discuss potential theoretical and clinical implications derived from these findings, as well as its limitations and directions for future research.

Chapter 2: Overall Methods and Analyses

This chapter outlines the general methods and analyses used in this thesis. I first discuss the overall data collection strategy and inclusion criteria applied; this is followed by a brief description of the experimental paradigms used for this thesis. Finally, I outline the data analysis strategy, which includes descriptions of the conventional and computational approaches. Detailed participant description, technical features for the experimental task and statistical analysis will be reported in each experimental chapter.

2.1 Data collection strategies

This thesis mainly reports data that were collected from participants through face-to-face assessments. These participants were recruited from the Cambridge (UK) local community either through word-of-mouth, flyer and poster advertisements or College mailing lists; participants who expressed an interest in taking part first underwent a telephone screening to collect basic demographic information, identify their drug use history and their physical and mental health status. Those who met the inclusion criteria (see section 2.2) were then invited to physically attend the lab to complete a series of questionnaires and tasks in the presence of a researcher.

Face-to-face assessment is the default mode for cognitive research, but this was not possible when the COVID-19 pandemic brought about lab closures and restricted human face-to-face contact. Consequently, online testing became necessary as lock-down restrictions made community recruitment impossible. I turned to data collection via an online research platform, Prolific Academic (https://www.prolific.co/), as several studies noted its reliability for behavioural research (Palan & Schitter, 2018; Peer et al., 2017). In general, eligible participants were identified through a custom filter tool and a pre-screening questionnaire, which probed for basic demographic information, drug use history and mental and physical health history. Those who fulfilled the inclusion criteria (see section 2.2) were then sent a link to complete a series of computerised tasks and questionnaires. All participants, recruited via the Cambridge community and online, were paid upon completion of the study procedures. Both face-to-face and online studies were approved by the local Cambridge Psychology Research Ethics Committee.

Online behavioural testing is increasingly popular due to the growing ubiquity and accessibility of the internet, but there are notable differences between online and face-to-face behavioural assessments. The primary advantage of online behavioural testing is the ability to reach diverse populations and collect large samples within a relatively short span of time. However, the trade-off is that screening for psychiatric disorders and drug use were only limited to self-report, unlike face-to-face assessments which provided the opportunity to administer structured interviews over the telephone and in person. As a remedy, Prolific users were required to answer more self-reported questions before they were given access to the study link (see section 2.2 and Chapter 7 for details). Moreover, whilst the automatic administration of online testing by Prolific removes the need of researchers to physically test every participant, saving time and effort, face-to-face assessments are advantageous in that they enable the researcher to ensure the quality of data: researchers can give clear instructions on the questionnaire or task prior to assessment, and guide participants through them if needed. By contrast, for online behavioural testing, researchers must rely on participants' own initiatives to complete the task and questionnaires carefully by following the instructions, and were not simply speeding through the task. To ensure participants attend to the task diligently, I administered attention check questions (see Appendix A) to identify whether participants are responding as instructed, which are effective measures to uphold data quality (Meade & Craig, 2012; Oppenheimer et al., 2009). Furthermore, there is evidence that data collected online is comparable to that collected from face-to-face sessions, but these studies were limited to the general healthy population (Casler et al., 2013; Germine et al., 2012). Whether this is the same for a drug-using population, traditionally associated with cognitive impairments, is unclear. As both sampling methods are different by nature, it was necessary to make adjustments to the inclusion criteria for data collection approach.

2.2 General inclusion criteria

Unless specified otherwise, participants were generally included if they were aged 18 years and above, and possessed sufficient English proficiency to provide informed consent and understand the written and verbal instructions. Drug-using participants and healthy control volunteers each needed to satisfy different criteria.

The definition of addictive disorders in the Diagnostic and Statistical Manual (DSM) has transitioned from a categorical (with or without substance dependence in DSM-IV) to a dimensional diagnosis (mild, moderate or severe substance use disorder in DSM-5; American

Psychiatric Association, 2013), which reflects a conceptual change in our understanding of addiction. This shift acknowledged that individuals could have different severity levels, which likely reflects different levels of impaired control over drug use and different intervention needs. Hence, there was a need to understand the cognitive profile of problematic substance use from both mild and severe ends of the spectrum. This thesis explored both categorical and dimensional identifiers of problematic drug use. On one hand, during face-to-face assessments (Chapters 3-6), drug-user participants underwent psychiatric screening with the Structured Clinical Interview (SCID; First et al., 2002), and had to meet the Diagnostic and Statistical Manual 5th edition (DSM-5) criteria for moderate or severe stimulant use disorder. They were also actively using stimulant drugs at the time of the study – urine tests were conducted prior the study session to confirm the presence of cocaine metabolites, indicating prior use within 72 hours. On the other hand, as thorough screening was not possible during online testing (Chapter 7), I adopted a dimensional approach to studying problematic alcohol use, by identifying harmful use with the Alcohol Use Disorder Identification Test – a 10-item validated self-reported instrument that detects hazardous alcohol use within the normal population (J. B. Saunders et al., 1993). Normally, a cut-off score of 15 on the AUDIT indicates likelihood of moderate-to-severe alcohol use disorder, but I reduced the cut-off to 10, which included individuals with hazardous alcohol consumption but not necessarily with alcohol use disorder. As alcohol was the main substance of interest, I deliberately excluded individuals who reported concomitant use of stimulant drugs (including cocaine, crack-cocaine, amphetamines and methamphetamine), which could confound the behavioural performance. Minimal comorbid drug use was allowed in drug-user cohorts. Prescribed psychoactive medications such as anti-depressants or opioid maintenance therapy were allowed in drug user participants, but use of antipsychotics was exclusionary because they are largely dopaminergic by nature and could interfere with reinforcement learning.

Healthy controls in general had to fulfill the following criteria: (1) good physical and mental health; (2) minimal drug use; and (3) no personal history of substance use disorder. These were assessed differently depending on the data collection approach. During face-to-face assessments, physical health was verified through telephone interview of medical history and medication use; the MINI International Neuropsychiatric Inventory (Sheehan et al., 1998) was used to screen for mental health; drug abstinence was confirmed with negative urine screens prior assessment. These were not possible during online testing, so I required participants to complete the AUDIT, questions about current medication use, and the Depression, Anxiety and Stress scales (DASS-21; Lovibond

& Lovibond, 1995), which is sensitive to subclinical levels of affective disorders. Control participants were included if they had low levels of alcohol use (AUDIT < 6), reported no psychoactive medications and did not meet the DASS-21 cut-off scores for moderate subclinical levels of depression (subscale < 14), anxiety (subscale < 10) or stress (subscale < 19). All participants were thoroughly screened and were excluded from participation if they (1) had a prior history of traumatic brain injury or neurological illnesses; (2) had anti-psychotic or stimulant-based medication; (3) current or a history of psychotic disorders; (4) had a diagnosis of Dyslexia, Dyspraxia, Attention Deficit Hyperactivity Disorder (ADHD), language-related disorders, or Autism Spectrum Disorder. Additionally, for the face-to-face sessions, we breathalysed participants to confirm sobriety prior to assessment; intoxicated participants were excluded from the study.

As the theme for this thesis concerns the analysis of learning and motivated behaviour in substance use disorder, there are several confounding variables that merit discussion. It is plausible that learning ability is tightly linked to variances in intelligence quotient (IQ) and education levels (van den Bos et al., 2012), as well as affective states such as being depressed or anxious. In particular, clinically diagnosed depression and anxiety have been associated with distinct learning profiles (Rouhani & Niv, 2019). However, while individuals with substance use disorders are often highly depressed and anxious, their underlying cognitive profiles very likely differ to that of clinically diagnosed depression and anxiety. To determine the influence of these potential confounds on my data, I collected measures of estimated verbal IQ (using the National Adult's Reading Test; (H. E. Nelson, 1982)) or education levels, and a measure of affective state (e.g. DASS-21). I then assessed the relationship between behavioural data obtained from learning tasks and these measures; if there is a statistically significant relationship, this measure would then be statistically controlled within the analyses, but not otherwise.

Table 2.1 reports a breakdown of each sample by chapter. Participant descriptions are elaborated in detail in each respective chapter.

Table 2.1: Overview of samples by chapter.

Chapter	Sample	Gender (% male)	Age (years ± SD)	Years of stimulant use (years ± SD)	Alcohol use (AUDIT ± SD)	Prescribed medication	Comorbid diagnoses	Task administered
Chapter 3 (study 1)*	44 CUD	100	40.9 ± 9.2	13.7 ± 8.0	3.9 ± 5.9	14 methadone; 8 buprenorphine; 6 antidepressants; 4 benzodiazepines; 7 painkillers	24 opioid; 3 alcohol; 6 cannabis	Reinforcement Learning Task
	41 controls	100	40.1 ± 12.6	-	3.4 ± 1.7	none	none	
Chapter 3 (study 2)	18 StimUD (10 cocaine, 8 amphetamines)	83	34.3 ± 7.2	12.3 ± 6.7	-	none	none	
	18 controls	83	32.7 ± 6.9	-	-	none	none	
Chapter 4*	42 CUD	100	39.3 ± 8.8	12.3 ± 7.5	4.1 ± 5.7	15 methadone; 7 buprenorphine; 6 antidepressants; 3 benzodiazepines; 4 painkillers;	24 opioid; 2 alcohol; 7 cannabis	Category Learning Task
	40 controls	100	40.9 ± 12.4	-	3.5 ± 1.7	none	none	
Chapter 5	72 CUD	94	38.0 ± 8.6	15.9 ± 6.7	4.2 ± 4.8	26 methadone; 14 buprenorphine;	48 opioid; 5 alcohol; 25 cannabis	Slips of action task
	53 controls	90	41.3 ± 10.5	-	4.2 ± 2.0	none	none	
Chapter 6*	48 CUD	100	40.4 ± 9.1	13.4 ± 7.7	4.3 ± 5.8	15 methadone; 8 buprenorphine; 6 anti-depressants; 4 benzodiazepines; 7 painkillers	25 opioid; 3 alcohol; 8 cannabis	-[a]
	43 controls	100	40.0 ± 12.4	-	3.4 ± 1.6	none	none	
Chapter 7	120 alcohol	52	31.9 ± 9.6	16.0 ± 10.4[b]	16.7 ± 5.5	18 on SSRI; 2 painkillers	52 anxiety; 45 depression; 2 OCD; 9 eating disorder; 16 alcoholism; 6 problem gambling; 6 suicide attempt; 17 self-harm[c]	Goal-habit conflict task
	148 controls	32	32.9 ± 7.3	16.0 ± 8.6[b]	1.8 ± 1.5	none	10 anxiety; 4 depression; 4 eating disorder; 1 problem gambling[c]	

Note. CUD: cocaine use disorder; StimUD: stimulant use disorder; IQ: intelligence quotient; AUDIT: Alcohol Use Disorder Identification Test; SSRI: selective serotonin reuptake inhibitor; OCD: Obsessive-Compulsive Disorder; SD: standard deviation

* these chapters are from the same sample, but vary in group sizes due to different completion rates.

[a] chapter 6 assessed data on self-reported measures instead of a behavioural task.

[b] number of years since first drunk episode.

[c] these are self-disclosed mental health history through an online questionnaire, and were thus unverified.

2.3 Experimental paradigms

Task paradigms reported in this thesis measure reinforcement learning and instrumental control of behaviour. On one hand, reinforcement learning generally uses corrective feedback to inform participants of the nature of their choices. Therefore, reinforcement learning tasks involve trial-and-error learning from feedback to identify the best choice from an array of discrete options. As there are qualitative differences between learning from positive and negative feedback, the reinforcement learning task (Chapter 3) differentiated between positive and negative feedback, framed as monetary wins and losses respectively. Throughout the task, participants need to learn by trial-and-error to pick the stimulus that maximises their rewards while minimising their losses.

Optimal reinforcement learning is thought to be guided by distinct memory processes, which can be either declarative or non-declarative. Declarative memory refers to the learning of rules and patterns that can be verbalized (e.g. square is bad, circle is good). By contrast, non-declarative memory involves the incremental learning of complex actions or patterns, mostly acquired through trial-and-error. These memory systems were tested in two variants of a category learning task, also known as the weather prediction task (Chapter 4), each testing for declarative and non-declarative memory. On a trial-by-trial basis, participants learn to classify multiple stimuli combinations into one of two categories: shine or rain. These combinations were learnt either through feedback learning (non-declarative) or explicit memorization (declarative). Furthermore, it is possible to analyse the response strategy used by participants to solve this category learning task, thereby offering a richer analysis of learning functions.

On the other hand, tasks assessing instrumental regulatory control typically involve the relative expression of goal-directed and habit systems. The task used in this thesis is an outcome devaluation paradigm adapted for humans (Chapter 5). This task consists of two stages: an appetitive discrimination learning stage in which participants gradually learn stimulus-action-outcome contingencies; and an instructed outcome devaluation, where participants' learned responses were tested under extinction, but with certain outcomes devalued, and should no longer be responded to. Responding for these devalued outcomes, also known as slips of action, is an indicator of habits. The original publication found a strong habit predominance in cocaine use disorder that was not affected by initial action-outcome learning (Ersche et al., 2016). Using

more sensitive measures generated through computational modelling, I sought to perform follow-up analyses on the latent learning factors in the first stage, to see how individual differences in latent variables during appetitive learning affect the expression of habits in the second stage.

Although the outcome devaluation task is frequently used in human studies, one notable limitation of this task is that it cannot distinguish whether habit predominance is a result of an augmented habit system, or an impaired goal-directed system, or both. Each interpretation bears different clinical implications. For instance, if drug users exhibit an augmented habit system, intervention strategies should focus on training maladaptive habits by repetition. Since goal-directed and habit systems are viewed as dissociable processes (Balleine & O'Doherty, 2010; de Wit & Dickinson, 2009), one possible way to directly assess the strength of each system is to identify which system predominates behaviour under competition. This idea became the basis of a novel task: the goal-habit conflict task (Chapter 7). Here, action-outcome and stimulus-response habits are trained over time. These learned actions are tested by simultaneously providing specific instructions (goal-directed actions) against a backdrop of a conditioned stimulus (habits), thereby producing a situation of conflict. This task enables the direct assessment of goal-directed and habit strength, which tests the concept of imbalance between instrumental control systems purported by the habit theory of addiction.

2.4 Statistical analyses

Behavioural data are analysed using both conventional and computational approaches. The conventional approach here refers to null hypothesis testing methods widely used within the psychology literature. As these inferential statistics generally evaluate the probability of the null hypothesis being accepted (usually with a p=0.05 criterion), they are also known as frequentist statistics. These statistics offer a straightforward way to analyse key relationships and differences in demographic and behavioural data – usually in the form of means or summary scores – which is common within the psychology literature. As research methods advances, there is an increasing demand for addressing the complexity in behavioural data, which is limited if data is analysed in a conventional method. An emerging counterpart to address this shortcoming in behavioural analysis is to adopt computational strategies – the use of mathematical models to deconstruct behaviour into its constituent processes. This approach

offers a more fine-grained analysis of hypothetical cognitive processes that contribute to learning, which is increasingly popular. Both conventional and computational approaches will be used concurrently for data analyses in this thesis.

2.4.1 Conventional approach

Frequentists statistical tests are carried out with the Statistical Package for Social Science (SPSS v28). Before analyses, data are usually inspected to ascertain if they conform to the assumptions needed for parametric analyses (e.g. t-tests, Analysis of Variance [ANOVA]), such as the normality and the homogeneity of variance assumption; for data that were not normally distributed, log-transformations were used to reduce skewness of data; when the homogeneity of variance assumption is violated, a non-parametric alternative was used to analyse the data (e.g. Mann-Whitney U-test). Mixed ANOVA models are commonly used throughout the thesis to identify main effects and interactions within the task performance data. Where applicable, Mauchley tests were used to assess sphericity; Greenhouse-Geiser corrections to degrees of freedom were applied if the sphericity assumption was violated. Where applicable, post-hoc pairwise comparisons were evaluated, and I used the Bonferroni's method to correct for multiple comparisons. To examine the relationship between variables, Pearson's correlational analyses were often the default method, but non-parameteric Spearman's analyses were applied if the data did not meet the assumptions for parametric tests. For categorical data such as gender or frequency data, I used chi-squared tests to identify any significant associations; in cases were the expected frequency per cell was less than 5, the Fisher's Exact test was used as an alternative. Details on the computation of summary scores for task performance varied for each task, and will be described in each chapter.

2.4.2 Computational approach

Analyses on summary scores of task performance may offer some insight into the overt behaviours, but these measures are somewhat limited in their explanatory power. Many latent processes occur in the background before an action is expressed, but these variables are not always observable in overt behaviours. Thus, computational learning models are increasingly used to bridge this gap. These models adopt mathematical frameworks to model the generative process of learned actions, thereby making the latent variables quantifiable. Since it is hypothesised that aberrant behaviour is closely related with impairments in latent cognitive

processes involved in generating behaviour (Huys, Maia, et al., 2016; Maia & Frank, 2011), analyses with computational models might offer mechanistic insights these processes that are otherwise inaccessible through summarised behavioural scores alone.

Learning models are often used to approximate choice behaviours in participants. Individual differences in learning manifest within certain model parameters that capture hypothetical latent processes involved in learning. Typically, participants' trial-by-trial performances are fitted with a learning model, and the goal is to estimate the values of free parameters that most closely match the participants' actual choice behaviour. These parameters are subsequently used for further analyses (e.g. case-control comparisons) (Daw, 2011). Therefore, the pipeline for modelling behaviour consists of several important components: (1) learning model used, (2) parameter estimation methods, and (3) model selection process. Each component is described below.

2.4.2.1 Learning model used

This section describes the general learning model used in this thesis. In general, computational models of learning quantitatively describe the process where one learns to predict outcomes based on their behavioural choices (Robbins & Cardinal, 2019). The modelling of behavioural choices is usually a two-step process: value estimation and choice selection. First, value estimation is where the agent determines which stimulus/actions yield higher subjective value, which is intrinsically more valuable. The values of each action/stimulus are updated over time by the difference between expected and received outcomes, also known as prediction errors. One of the most widely used learning rule within the recent neuropsychiatric literature is the delta-rule learning algorithm, such as the Rescorla-Wagner model for classical conditioning (Rescorla & Wagner, 1972) or the Q-Learning model (Watkins & Dayan, 1992). In a typical delta rule algorithm, the value of a specific stimulus s on trial t, $V_t(s)$, is driven by the prediction error:

$$V_t(s) = V_{t-1}(s) + \alpha(R - V_{t-1}(s))$$

where α is the learning rate and R is the actual reinforcement. In other words, stimulus/actions values are updated on a trial-by-trial basis, and are dependent on the feedback received from

the last trial. This value update is done for each available stimulus / action for each trial. Despite its simplicity, the delta-rule learning algorithm thus far has been successful in generating sensitive behavioural measures for psychiatry and psychopharmacological research (Robbins & Cardinal, 2019), and thus would be the learning rule used throughout this thesis.

These expected values are then used to estimate actual choice behaviour, which, in the reinforcement learning literature, typically follows a softmax rule:

$$p(i,t) = \frac{exp(\beta V_t^i)}{\sum_{k=1}^{n} exp(\beta V_t^k)}$$

This equation gives the model's probability of choosing choice i amongst n choices on a trial t. The extent to which expected values are used to drive choices is governed by the reinforcement sensitivity parameter, β.

In its simplest form, individual differences in learning are expressed in the free parameters learning rate (α) and reinforcement sensitivity (β). The learning rate reflects the impact of reinforcement on learned values, whilst the reinforcement sensitivity (sometimes known as inverse temperature) governs the tendency to which stochastic choices are motivated by learned values. However, these algorithms are often adapted to suit the task paradigm and patient population to optimise data analyses in a hypothesis-driven manner. For instance, there is growing evidence suggesting that learning from different reinforcer types (e.g. reward or punishment) are subserved by dissociable neural circuits (Palminteri & Pessiglione, 2017; Pessiglione & Delgado, 2015). Since most of the data reported in this chapter involved feedback from opposing valences, the learning rate parameter is often fractionated based on the feedback received. For example, consider a task with four possible outcomes: reward (e.g. win 50p), non-reward (e.g. win 0p), punishment (e.g. lose 50p) and punishment omission (e.g. lose 0p). Given such task parameters, the learning rate would be fractionated based on the feedback received. Thus, the learning algorithm would be updated as follows:

$V_{t+1} = V_t + \alpha_{rew}(R - V_t)$ if feedback = "You win 50p";

$V_{t+1} = V_t + \alpha_{non-rew}(R - V_t)$ if feedback = "You win 0p";

$V_{t+1} = V_t + \alpha_{pun}(R - V_t)$ if feedback = "You lose 50p";

$$V_{t+1} = V_t + \alpha_{non-pun}(R - V_t)$$ if feedback = "You lose 0p";

where α_{rew}, $\alpha_{non\text{-}rew}$, α_{pun}, and $\alpha_{non\text{-}pun}$, refers to the learning rates from reward, non-reward, punishment, and non-punishment respectively.

Furthermore, tendency to perseverate – continuous inflexible repetition of past choices even when not reinforced – is often implicated in individuals with substance use disorder, which may negatively affect choice behaviour. To account for this factor in the learning model, a perseveration parameter τ is sometimes introduced during choice selection. This parameter is included as a softmax weight (along with β) and governs the extent to which choices are influenced by repetitive choices. Thus, the softmax rule would be modified as:

$$p(i,t) = \frac{exp\left(\beta V_t^i + \tau C_t^i\right)}{\sum_{k=1}^{n} exp\left(\beta V_t^k + \tau C_t^k\right)}$$

C denotes repeated choices, and are assigned 1 if participants repeated their choices and 0 if choices were not repeated. The tendency for perseveration to influence choice is governed by perseveration parameter, τ.

It is noteworthy that I apply model selection procedures to ensure the best fitting model is selected for further data analyses (see model selection procedures section below).

2.4.2.2 Parameter estimation

In this thesis, the free parameters from the learning models were estimated with Markov Chain Monte Carlo sampling method with the *RStan* package (version 2.17.2) (Carpenter et al., 2017). The estimation of model free parameters was implemented in a hierarchical Bayesian framework. At the top-level of the hierarchy, a group-level distribution was implemented: participants with shared group membership (e.g. healthy controls or patients) are assumed to vary within the same group-specific distribution. The next level introduces variances from each participant (i.e. individual variability) by accounting for subject-specific deviation from the group-level distributions. This two-level structure ensures that each participant-specific posteriors for free parameters, used in the fitting of the reinforcement learning algorithm, are drawn from both the group-level and individual-level distributions in a mutually-constraining

manner. Each parameter has its own priors, which are reported in Table 2.2. These priors were selected on the basis of prior literature and suitability for each parameter.

When analysing each model, I simulated 8 parallel chains, each with 2000 iterations, half of which were warmup iterations. As the chains estimate parameters in parallel, it was important that these estimations converge to produce reliable posteriors. Convergence was assessed with the Gelman-Rubin convergence statistics, R-hat. An R-hat value of 1 indicates perfect convergence, but any values less than 1.1 is acceptable (Brooks & Gelman, 1998). In this thesis, all winning models show acceptable levels of convergence as per the aforementioned criterion (R-hat < 1.1).

The primary measures of interest are the differences in the parameters between drug-user groups and healthy controls. A common approach within the literature is to directly compare individual parameter values in a case control analyses with frequentist statistics (Daw, 2011). However, since the Bayesian hierarchy method enables the estimation of group-level posteriors that partitions out individual variability, the primary outcome measures used to report computational analyses in this thesis are the posterior group differences, which were directly sampled in RStan by subtracting one group-level posterior from the other. These posterior group differences were interpreted with their 95% highest density intervals (HDI), which encompasses the likelihood of the sampled value falling within these intervals 95% of the time (akin to confidence intervals of frequentist statistics). Posterior group differences with 95% HDI that do not overlap 0 are interpreted as a credible group difference.

A Bayesian hierarchy method was favoured over the more common maximum likelihood estimation (MLE) approach for parameter estimation, which involves identifying point estimates of free parameters that produces the lowest negative log likelihood (i.e. best explains the data). Whilst the implementation of MLE is simpler and less computationally demanding than Bayesian estimation methods, the main disadvantage of the MLE approach is that it is disproportionately sensitive to initial starting values. Depending on the initial starting value, the best fitting point estimate may differ, especially in cases where there are several global minima i.e. multiple best-fitting parameter values within a given range (Daw, 2011). This

limitation is circumvented in a Bayesian hierarchy approach, where a posterior distribution of values, instead of a point estimate, is produced.

Table 2.2: Priors for each free parameter.

Parameter	Boundaries (lower, upper)	Prior for means	Prior for standard deviation	References
Learning rate	(0,1)	Beta (1.1, 1.1)	Normal(0, 0.05)	Clarke et al., 2014; Gershman, 2016; Kanen et al., 2019
Reinforcement sensitivity	(0, +∞)	Gamma (α = 4.82, β = 0.88)	Normal(0,1)	Clarke et al., 2014; Gershman, 2016; Kanen et al., 2019
Perseveration	(-∞, +∞)	Normal(0,1)	Normal(0, 0.05)	Christakou et al., 2013; Kanen et al., 2019

2.4.2.3 Model selection procedures

To ensure that the learning models adequately explain the task data, the task performance is often fitted with several variants of the learning model. The *bridgesampling* package was then used to determine the best fit model (Gronau et al., 2017). Designed for hierarchical models, Bridge sampling estimates the marginal likelihood – the probability of the observed data occurring given the model, p(data | model) – as the measure of model evidence (i.e. how well does this model produce the observed data). This is estimated by integrating over all possible parameters (1) the likelihood of the data given the fitted model parameters, p(data | parameters, model), and (2) the prior probability of the parameters given the model, p(parameters | model). By considering these quantities, bridge sampling's estimation of the marginal likelihood favours simpler models by penalising over-complex models that do not substantially contribute towards model fit. A larger marginal likelihood reflects larger model evidence, and the model with the largest marginal likelihood is taken as the winning model.

Within the cognitive modelling literature, other model comparison methods, most notably the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC), are more

widely used than *bridgesampling*. This thesis favours the latter for two reasons: (1) computationally, BIC and AIC methods, which are classed frequentist model comparison methods, are not suited for models with a hierarchical structure (i.e. individual parameters subsumed by group parameter) (Lu et al., 2017) (2) practically, although BIC and AIC methods are relatively simple, these methods can provide misleading approximations when the sample size is small, and cannot discriminate well between complex models (Hollenbach & Montgomery, 2020).

The Bayes Factor – the relative model evidence of one model over another – was also computed as a secondary index for model evidence. The Bayes factor is calculated as the ratios of marginal likelihood between two models. According to Kass and Raftery (1995), a (log10) Bayes factor value of 1 or more constitutes strong evidence for one model's superiority over another, whilst a value of 2 or more indicates decisive evidence (Kass & Raftery, 1995).

Whilst the marginal likelihood and Bayes factor measures provide relative evidence of a single model over others, it is noteworthy that these are not absolute indicators of the model's performance. To verify the winning model's abilities to model choice behaviour, posterior predictive checks were made, whereby I simulate data from the winning model to assess whether the actual choice behaviour are indeed reproducible by the model. The simulated data were then analysed using frequentist statistics to examine if key aspects of the original data can be recapitulated. Additionally, where applicable, I also included data on parameter recovery for the winning models to show the recoverability of known simulated parameter values during the fitting process (Wilson & Collins, 2019). Specifically, I simulated behaviour with known parameter values, and fitted the simulated behaviour to the winning model to recover these parameter values. I then assessed the correlation between simulated and recovered parameter values – good parameter recovery is indexed by a strong correlation between these values. Generally, parameter recovery for the winning models reported in this thesis was reasonably well. Scatterplots of simulated versus recovered parameters can be found in the appendices of the relevant chapters.

2.4.2.4 Summary

Reinforcement learning algorithms were generally used within this thesis to gain insight into the group differences between latent parameters that underpin learning from trial-and-error. These parameters were estimated in a hierarchical Bayesian framework, of which the primary output measure is the group mean difference for each estimated parameter. In most cases, several models were fitted to the behavioural data, and I selected the best fit model with *bridgesampling* package. The model with the largest marginal likelihood, indicating largest model evidence, is selected as the winning model.

Appendix A: Supplementary materials to Chapter 2

Questions for attention check

As recommended by Meade & Craig (2012), I implemented two questions to check for attention. Question 1 was placed at the mid-point of the study, whereas question 2 was embedded within a questionnaire with an identical 5-point Likert scale, administered towards the end of the study.

Question 1:

Now, to help with our understanding of you as a person, we are interested in certain factors about you. In particular, we are interested in whether you are attentive to the questions asked. Hence, please ignore the following sports participation question - do not provide any responses for this question and proceed to the next page of the study.

Based on your understanding of the text above, which of the following activities do you engage in regularly?

- Skiing
- Hockey
- Basketball
- Running
- Tennis
- Snowboarding
- Swimming
- Soccer
- Football
- Cycling
- Others

Question 2:

It is important that you pay attention during the study. Please select “Very often” if you have.

- Never
- Almost never
- Sometimes
- Fairly often
- Very often

Chapter 3: Deconstructing reinforcement learning in stimulant use disorder: dopaminergic modulation

This chapter has been published as:

Lim, T. V., Cardinal, R. N., Bullmore, E. T., Robbins, T. W., & Ersche, K. D. (2021). Impaired Learning From Negative Feedback in Stimulant Use Disorder: Dopaminergic Modulation. *International Journal of Neuropsychopharmacology*, *24*(11), 867–878. https://doi.org/10.1093/ijnp/pyab041

3.1 Introduction

Stimulant drug addiction, or stimulant use disorder (StimUD), is a major public health problem that causes significant harm to individuals, their families, and society (Degenhardt et al., 2014). The behaviour of chronic stimulant drug users often seems maladaptive and ill-judged, as they frequently behave in ways that are detrimental to their own interests, regardless of the negative consequences. One possibility is that drug-induced neuroadaptations may change how individuals learn from the consequences of their actions, an impairment that might extend beyond drug-taking (Maia & Frank, 2011).

Reinforcement learning (RL) is an influential account of adaptive instrumental behaviour that provides a normative framework of how humans use past consequences to guide future behaviour (Sutton & Barto, 1998). Optimal RL includes multiple processes such as valuation, reward prediction, and action selection (Niv, 2009), and many of these processes are suggested to be modulated by dopamine (Bayer & Glimcher, 2005; Frank et al., 2007; Steinberg et al., 2013) – a neurotransmitter affected by stimulant drugs such as cocaine and amphetamine. Chronic stimulant drug use has been associated with a downregulation in dopamine neurotransmission in fronto-striatal circuits (Volkow et al., 2004) that underpin learning and value-based decision-making (Ernst & Paulus, 2005; Glimcher, 2011; O'Doherty et al., 2017). Animal studies shown that cocaine exposure disrupts key aspects of RL, including reward prediction (A. C. Burton et al., 2018; Takahashi et al., 2019) and reinforcement value (Groman et al., 2020; Schoenbaum & Setlow, 2005). Although similar observations have also been reported in human stimulant drug users (Ersche et al., 2016; Harlé et al., 2015; Parvaz et al.,

2015), the exact profile of impairments remains elusive. While it is widely assumed that impairments in RL in StimUD patients are dopaminergic in nature, it is unclear how these disruptions are modulated by dopaminergic agents. There is some evidence for modulatory effects of dopamine manipulations on cognitive dysfunction in StimUD (Ersche et al., 2010; Ersche, Roiser, Abbott, et al., 2011; Goldstein et al., 2010). However, compared with control participants, StimUD patients show different behavioural and neural responses following dopaminergic drug challenges, suggesting that such medication alters RL differentially in StimUD patients (Ersche et al., 2010; Ersche, Roiser, Abbott, et al., 2011; Goldstein et al., 2010; Volkow et al., 2005). The precise actions of dopaminergic drugs are difficult to determine in human studies, but drug challenges may provide insight into the neurochemical underpinnings associated with RL in StimUD patients.

A conventional approach to quantify RL performance is to compute summary scores that reflect performance accuracy, and analyse them with a frequentist approach (e.g. Strickland et al., 2016). As RL impairments can also result from latent processes that are not directly measured by summary scores, such as motivational deficits, slower contingency learning, or inconsistencies in choice behaviour, complementary approaches are needed. An increasingly popular method is to use computational models to describe RL, allowing the quantification of latent RL parameters (Sutton & Barto, 1998). Individual differences in RL are then reflected in model parameters, which can be compared between groups (Daw, 2011). Although simple RL models might not perfectly capture all the RL-related cognitive processes, the model parameters can provide sensitive behavioural measures (Heinz et al., 2016; Robbins & Cardinal, 2019).

Here, I combine both conventional and computational approaches to address the following objectives: (i) to characterise the RL profile in a large community sample of StimUD patients using a behavioural task that assesses learning from reward and punishment separately; (ii) to explore the modulatory effects of a dopamine $D_{2/3}$ receptor agonist and an antagonist on RL in an independent sample of StimUD patients. Two pharmacological agents were used to selectively target the $D_{2/3}$ system, the dopamine receptor antagonist amisulpride and the dopamine receptor agonist pramipexole (Rosenzweig et al., 2002; Wright et al., 1997). For the computational analysis, I employed a well-established RL model (Watkins & Dayan, 1992)

and adopted the *learning rate*, the impact of reinforcement on choices, as the key outcome measure. I also modelled other processes that support learning, such as the extent to which behaviour is motivated by learned values (*reinforcement sensitivity*) and tendency to perseverate. I hypothesised that these latent learning parameters are impaired in StimUD patients and would be modulated differentially by dopamine agonist and antagonist agents. Since StimUD patients have abnormal dopamine transmission, I predicted that these dopaminergic agents would modulate RL performance differentially in StimUD patients compared with healthy controls. Specifically, task performance would be negatively affected by either dopaminergic receptor agent in healthy controls. By contrast, dopamine receptor blockade by amisulpride would further impair learning in StimUD patients, whilst pramipexole would ameliorate the deficits in reinforcement learning in these patients.

3.2 Methods

I studied two independent samples of stimulant-addicted individuals and matched healthy volunteers. For inclusion, participants had to be at least 18 years old, and able to read and write in English. Stimulant drug users needed to meet the DSM-IV-TR criteria for stimulant drug dependence (American Psychiatric Association, 2000), whereas control participants had to be healthy without a personal history of substance use disorders. Participants were recruited from the local community in Cambridge (UK) by advertisement and word of mouth. Both studies were approved by a Cambridge Research Ethics Committee. All participants provided written consent prior to enrolment and were screened for psychiatric disorders using the Mini-International-Neuropsychiatric-Inventory (Sheehan et al., 1998); psychopathology in drug users was further evaluated using the Structured Clinical Interview for DSM-IV (First et al., 2002). All StimUD patients were actively using stimulant drugs, which was confirmed by positive urine screens prior to testing, suggesting that they had been using the drug within the past 72 hours. All urine samples provided by control participants tested negative for all drugs; participants were also breathalysed to verify sobriety. Exclusion criteria for all participants included a lifetime history of a psychotic disorder, neurological illness, or traumatic head injury, and acute alcohol intoxication. All participants completed the National Adult Reading Test (H. E. Nelson, 1982) and the Barratt Impulsiveness Scale (Patton et al., 1995) to estimate the verbal intelligence quotient (IQ) and impulsive personality traits respectively. Participants also reported their monthly disposable income and rated their willingness to pick up £0.50 off the floor on a visual analogue scale (always—never) as a proxy for the subjective value of

monetary reward. StimUD participants additionally completed the Obsessive-Compulsive Drug Use Scale (Franken et al., 2002) as a measure of compulsive drug use.

3.2.1 Study 1

Sample: Forty-four men who met the DSM-IV-TR criteria for cocaine dependence, referred as cocaine use disorder (CUD), had been using cocaine for a mean of 13.7 years (standard deviation [SD]: ±8.0) and the majority also met criteria for dependence on another substance (55% opiates, 7% alcohol, 14% cannabis). Participants with co-morbid opiate dependence were either prescribed methadone (32%, mean daily dose: 49mg, SD: ±13.0) or buprenorphine (18%, mean daily dose: 7.5mg, SD: ±3.5). Some CUD patients were taking prescribed medication, including antidepressants (14%), benzodiazepines (9%), painkillers (16%), antibiotics (5%), and anticoagulants (5%). The 41 healthy control participants did not use prescribed mediation and reported low levels of drug and alcohol use, as reflected low total scores on the Alcohol-Use-Disorder-Identification-Test (J. B. Saunders et al., 1993) (mean score: 3.4, SD: ±1.7) and Drug-Abuse-Screening-Test (H. A. Skinner, 1982) (mean score: 0.08, SD: ±0.3). CUD patients reported a significantly lower monthly disposable income than controls (t_{83}=2.6, p=0.012; see Table 3.1).

RL Task: This task evaluated learning from financial gains and losses (Bland et al., 2016) (Figure 3.1). Participants were presented with pairs of coloured circles and asked to learn by trial-and-error to select the stimulus that maximises their overall earnings. The two conditions of reward and punishment were differentiated by feedback. Specifically, feedback was explicitly framed as wins (‘you win 50p’ and ‘you win 0p’) and losses (‘you lose 50p’ and ‘you lose 0p’) in the reward and punishment conditions respectively. Participants completed 120 learning trials, with each reinforcement condition represented by unique stimulus pairs and repeated 60 times, interspersed randomly throughout the task. Optimal choices for each stimulus pair were reinforced 70% of the time, either by winning £0.50 (reward) or avoid losing £0.50 (punishment).

3.2.2 Study 2

Sample: Thirty-six volunteers were recruited from the community: 18 fulfilled the DSM-IV-TR criteria for stimulant drug dependence (10 cocaine, 8 amphetamine), referred to as StimUD henceforth. The remaining 18 recruits were healthy with no personal drug-taking history. StimUD patients had been using stimulant drugs for an average of 12.3 years (SD: ±6.7), had no comorbid dependencies, and were not taking prescribed medication. The two groups did not differ in their disposable income (t_{33}=-0.66, p=0.514). Data from this sample have been published elsewhere (Ersche et al., 2010; Ersche, Roiser, Abbott, et al., 2011; Ersche, Roiser, Lucas, et al., 2011; Kanen et al., 2019).

RL Task: This task has a similar design with that of study 1, but has three different conditions, distinguished by distinct stimulus pairs and outcomes: reward, punishment and neutral (Murray et al., 2019). Specifically, outcomes for the reward, punishment, and neutral conditions were intentionally phrased as monetary gains (i.e. 'you win 50p'), losses (i.e. 'you lose 50p') and no financial consequences (i.e. 'no change') respectively. Unlike study 1, reward omission (i.e. 'win 0p') and punishment avoidance (i.e. 'lose 0p') were not explicitly signalled during the feedback phase; participants will not receive any explicit feedback for these outcomes (Figure 3.1). There was one stimulus pair per condition, each repeated 40 times in randomised order. Optimal choices for each stimulus pair were also reinforced 70% of the time.

Drug administration: Participants were administered a single dose of 400mg amisulpride or 0.5mg pramipexole in a double-blind, placebo-controlled, crossover design. Prior to each drug administration, participants also took a dose of domperidone (30mg), a peripheral dopamine D_2 receptor antagonist, as a pre-treatment to the potential side effect of nausea/vomiting. Initially, pramipexole was administered at a dose of 1.5mg to the first six participants (three StimUD and three control participants), which was tolerated by StimUD but not by control participants. These control participants were subsequently administered 0.5mg pramipexole on a separate session, which was well-tolerated. Thereafter, all remaining participants received 0.5mg pramipexole. In total, I included data from 18 control and 18 StimUD participants, but subsequently excluded the three StimUD participants who received higher dose of pramipexole from the analysis. Participants completed the RL task approximately 1.5 hours after dosing and blood samples were drawn at one and 2.5 hours post-dosing.

3.2.3 Statistical analyses

Conventional analyses: Demographic and performance data were analysed using SPSS v25 (IBM). I computed accuracy scores for the RL tasks, defined as the proportion of optimal choices made in 10-trial blocks. I used analysis of variance (ANOVA) models with a two-tailed alpha value of 0.05, with trial block and condition (reward versus punishment) as within-subject factors, and group (control versus StimUD) as a between-subjects factor. As the two drugs may exert differential effects, I decided *a priori* to analyse the effects of amisulpride and pramipexole separately. Sensitivity power analyses determined that the minimum effect sizes (Cohen's d) detectable with the current samples are 0.54 and 0.78 for study 1 and 2 respectively ($\alpha = 0.05$, power = 0.8).

Computational analyses: To examine latent learning parameters, I modelled trial-by-trial choice values using a delta-rule learning algorithm (Rescorla & Wagner, 1972), with the final choice selection process following a softmax rule (Sutton & Barto, 1998). Details of modelling procedures are reported in Appendix B. In its simplest form, a model consists of two parameters: learning rate (impact of feedback on choice values) and reinforcement sensitivity (how much choice values motivate actual behaviour). Since different neural systems are thought to subserve learning from different valences (Pessiglione & Delgado, 2015), I decomposed the learning rate by the feedback received on that trial. For example, if participant receives a reward ('you win 50p') or a punishment ('you lose 50p'), I modelled that trial with the learning rate from reward and punishment respectively, whereas trials with a reward omission ('you win 0p') or punishment avoidance ('you lose 0p') feedback were modelled with the learning rate from non-reward and non-punishment respectively. However, it is not possible to model the learning rate from non-reward or non-punishment in study 2, because reward and punishment omission feedback were not explicitly framed within a win/loss domain. Thus, I modelled learning from these outcomes with a general extinction rate. It is noteworthy that perseveration is frequently reported in StimUD patients (Ersche et al., 2008) and stimulant-exposed animals (Schoenbaum et al., 2004). I would not expect an RL task to be optimised for investigating perseverative responses, unlike a probabilistic reversal learning task (Cools et al., 2002; Ersche et al., 2008; Ersche, Roiser, Abbott, et al., 2011; Jentsch et al., 2002; Kanen et al., 2019; Schoenbaum et al., 2004). Nevertheless, I included parameters that model perseverative tendencies towards stimuli

and locations (i.e. left or right) because accounting for relevant biases might improve model-fit (Wilson & Collins, 2019), as demonstrated in my previous work (Lim et al., 2019). Thus, there were eight possible parameters in the models: learning rate from reward, non-reward, punishment and non-punishment, general extinction rate, reinforcement sensitivity, as well as perseveration tendencies to stimulus and location; but not all parameters were used in any given model (full details reported in Appendix B Table B2). I acknowledge that differences in task designs can change the best-fitting models (Wilson & Collins, 2019), so I fitted several model variants for each study and identified the best-fit model with bridge sampling (Gronau et al., 2017) (Appendix B, Table B2). To validate the winning model, I simulated data from the winning model to ensure key findings from the actual data were reproduced (Appendix B). Parameter recovery was also assessed for the winning models (Appendix B, Figure B2), which showed that the current model fitting procedure can recover the simulated parameters reasonably well.

I estimated the posterior distribution of the best-fit model parameters within a hierarchical Bayesian framework in RStan (Carpenter et al., 2017). In study 1, I modelled a group-level posterior distribution at the top level of the hierarchy for each free parameter. With the inclusion of drug factors in study 2, I constructed group/drug posteriors to model the drug effects on free parameters separately for each group/drug combination. I also constructed a subject-level hierarchy for each parameter to account for any individual variations. The primary outcome measure was the mean differences between the group/drug posteriors, *d*, each with its associated 95% highest density intervals (HDI). A HDI that excludes zero provides strong evidence for a group difference (non-zero-difference, p_{nz}>0.95).

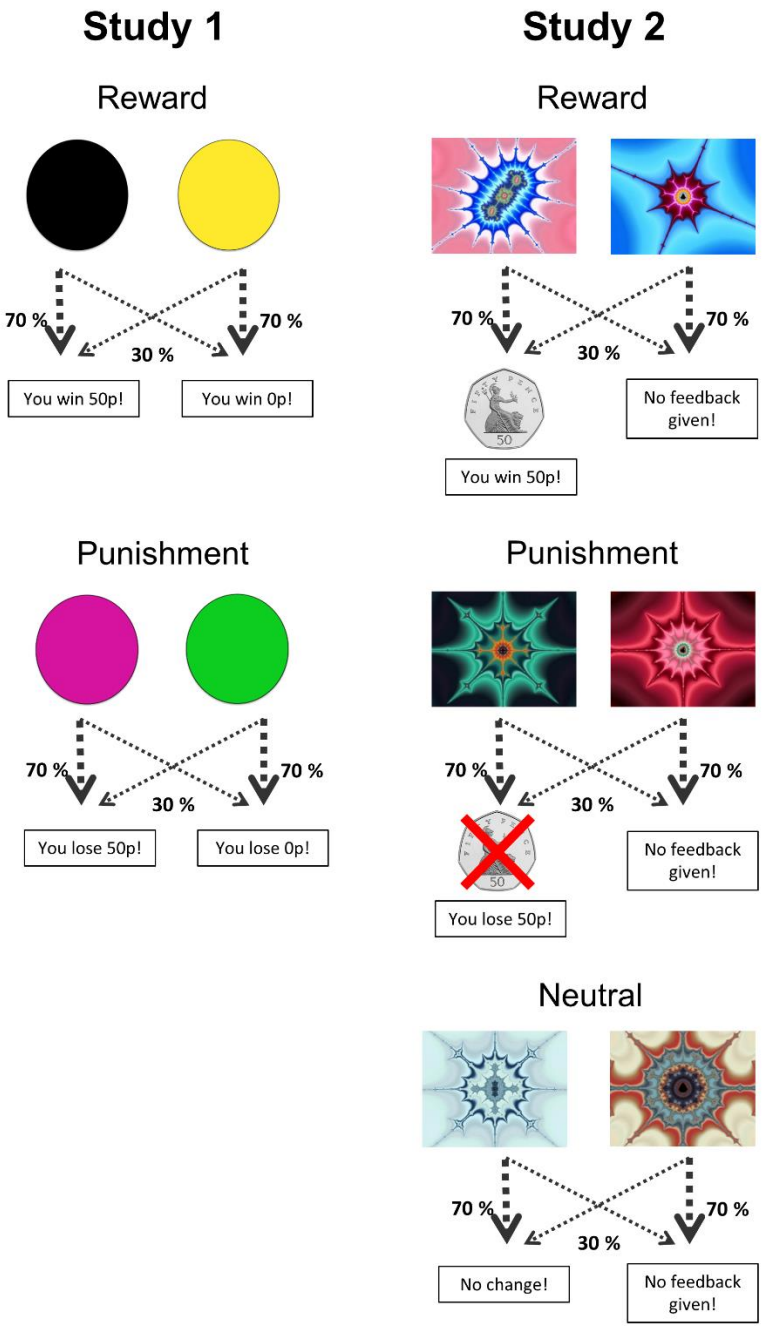


Figure 3.1: Schematics for the probabilistic reinforcement learning task of study 1 and study 2. On each trial, participants are first presented with a pair of stimuli, and required to select one stimulus. After selection, the computer will present an outcome phrased in terms of monetary gains (positive) or losses (negative); this allowed the separate assessment of learning from reward and punishment. In both studies, each condition was differentiated by unique stimulus pairs and feedback, and were interspersed across 120 trials and presented in a randomised order. Optimal choices are reinforced 70% of the time, so participants need to accrue experience over time to determine the choices that would maximise their financial gains and minimise their losses.

3.3 Results

Sample characteristics are shown in Table 3.1. In both studies, the groups were well-matched with respect to age and gender. Verbal intelligence did not differ between the groups in study 2 (t_{34}=-.235, p=0.816), but StimUD patients in sample 1 had a lower IQ scores than controls (t_{75}=8.2, p<0.001). However, IQ scores in StimUD patients were not significantly correlated with learning performance (Appendix B, Table B3), and therefore were not statistically controlled for. In both samples, the subjective value of £0.50 did not differ between the groups (Study 1: t_{83}=-1.2, p=0.232; Study 2: t_{34}=-1.7, p=0.098), suggesting that the reinforcement value of monetary rewards was similar in both groups. There were no relationships between learning performance and stimulant-related measures, including the duration or patterns of stimulant use (Appendix B, Table B3). Consistent with impulsivity being a hallmark of addiction, both patient groups scored significantly higher on the BIS-11 compared with controls (Study 1: t_{83}=-11.4, p<0.001; Study 2: t_{34}=-7.1, p<0.001).

Table 3.1: Sample demographics and task performance of the two studies.

	Study 1		Study 2	
Groups	Control	CUD	Control	StimUD
Demographics	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Sample size (n)	41	44	18	18
Age (years)	40.1 (12.6)	40.9 (9.2)	32.7 (6.9)	34.3 (7.2)
Gender (% male)	100	100	83	83
Verbal IQ (NART score)	115 (6.2)	103 (7.1)	108 (6.0)	109 (8.1)
Disposable income (£ per month)	657 (501)	387 (462)	470 (389)	621 (866)
Subjective value of 50 pence (% rating)	81.5 (23.0)	87.1 (19.6)	72.9 (31.0)	87.4 (18.8)
Trait impulsivity (BIS-11, total score)	56.1 (6.7)	79.5 (11.4)	62 (7.2)	82 (9.5)
Duration of stimulant drug use (years)	-	13.7 (8.0)	-	12.3 (6.7)
Compulsive drug use (OCDUS total score)	-	34.1 (10.1)	-	25.6 (7.9)
Task performance	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Total % correct (reward)				
Placebo	73.0 (21.6)	57.9 (22.5)	87.1 (24.5)	81.1 (18.6)
Amisulpride	-	-	87.9 (23.7)	75.8 (27.3)
Pramipexole	-	-	75.7 (30.9)	61.8 (35.2)
Total % correct (punishment)				
Placebo	63.6 (12.0)	54.3 (10.6)	73.3 (19.3)	61.7 (13.5)
Amisulpride	-	-	78.5 (17.5)	62.4 (19.8)
Pramipexole	-	-	72.9 (15.2)	64.7 (18.8)

Note. SD: standard deviation; NART: National Adult Reading Test; BIS-11: Barratt Impulsiveness Scale; OCDUS: Obsessive-Compulsive Drug Use Scale; CUD: cocaine use disorder; StimUD: stimulant use disorder

3.3.1 Study 1

Conventional analysis: Analyses of accuracy scores showed that there was a main effect of block ($F_{4.1,341}$=14.5, p<0.001) and a block-by-group interaction ($F_{4.1,341}$=3.048, p=.016), suggesting that although performance improved over time, control participants improved faster than StimUD patients. Participants learned faster from reward trials than punishment trials, reflected in a block-by-condition interaction ($F_{5,415}$=4.123, p=.001) (Figure 3.2A), but there was no group-by-block-by-condition interaction ($F_{5,415}$=0.234, p=.948). StimUD patients made more errors than controls ($F_{1,83}$=18.1, p<0.001), but no group-by-condition interaction was observed ($F_{1,83}$=1.33, p=0.252).

Computational analysis: As shown in Figure 3.3A, the best-fit learning model contained the following parameters: learning rates from reward, non-reward, punishment and non-punishment, reinforcement sensitivity and perseveration tendencies toward location and stimulus. StimUD patients showed a substantially reduced learning rate from punishment (d=-0.055, 95% HDI=-0.103 to -0.004, p_{nz}=0.973), and reinforcement sensitivity (d=-1.93, 95% HDI=-3.85 to -0.035, p_{nz}=0.953). Although the reward learning rate was reduced in StimUD patients, the difference was not credibly different between the groups (d=-0.078, 95% HDI=-0.154 to 0.007, p_{nz}=0.944). The groups did not differ on any other parameters (0 ∈ 95% HDI).

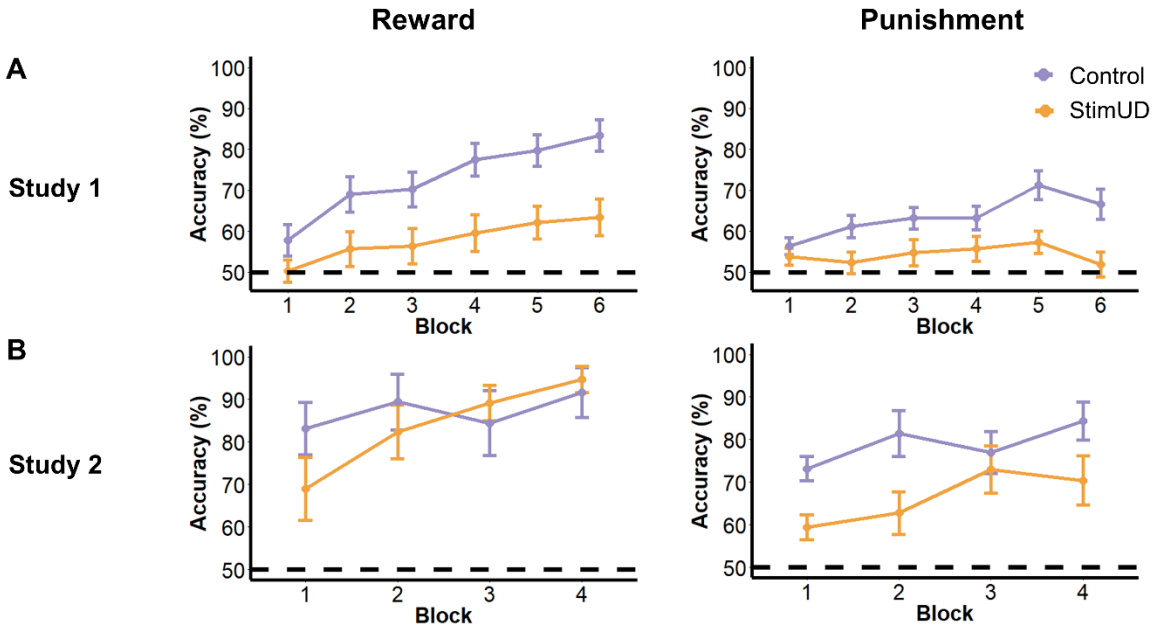


Figure 3.2: Accuracy scores for the behavioural task. These accuracy scores (defined as the proportion of optimal choices made in 10-trial blocks) are plotted separately based on condition (reward and punishment) and group (controls and StimUD). **(A)** RL performance accuracy in study 1. **(B)** RL performance accuracy for the placebo condition in study 2. [Error bars denote standard error to the mean, and the horizontal dotted line indicates accuracy at chance level (50%).]

3.3.3 Study 2

Conventional analysis: On placebo, task performance improved in all participants over time ($F_{3,102}$=6.66, p<0.001), with a significant effect of condition ($F_{1,34}$=9.83, p=0.004) again

suggesting that participants learned better from rewarding than punishing feedback (Figure 3.2B). Control participants learned faster than StimUD patients in the first two blocks, as reflected by a significant group-by-block interaction ($F_{3,102}$=3.63, p=0.016). There was neither a group effect ($F_{1,34}$=2.52, p=0.122) nor a group-by-condition interaction ($F_{1,34}$=0.610, p=0.440). No other effects reached statistical significance (ps > 0.4).

Amisulpride had no significant effect on accuracy ($F_{1,34}$=0.43, p=0.517), nor there were any group-by-drug interaction effects ($F_{1,34}$=0.619, p=0.437). There was a significant effect of block ($F_{3,102}$=18.5, p<0.001) and condition ($F_{1,34}$=15.9, p<0.001) on accuracy scores, such that all participants showed improved task performance over time, and better learning from reward than from punishment. Control participants showed improved accuracy compared with StimUD patients ($F_{1,34}$=5.41, p=0.026), but no other effects were significant (all ps > .1).

Although pramipexole also had a significant effect on accuracy ($F_{1,31}$=4.31, p=0.046), there was a significant drug-by-condition interaction ($F_{1,31}$=4.41, p=0.044). *Post-hoc* pairwise comparisons revealed a significant reduction of reward relative to punishment trial performance on pramipexole (p=0.022) but not placebo (p=0.627). Again, all participants improved performance over time ($F_{3,93}$=11.1, p<0.001), but the effects of condition ($F_{1,31}$=2.1, p=0.157), group ($F_{1,31}$=3.41, p=0.074) and group-by-drug interactions ($F_{1,34}$=0.526, p=0.474) were non-significant. Other effects were also not significant (all p>0.1).

Computational analysis: The best-fit computational model for study 2 included the following parameters: learning rates from reward and punishment, extinction rate, and reinforcement sensitivity (Figure 3.3B). On placebo, StimUD patients showed markedly reduced rates of learning from punishment (d=-0.452, 95% HDI=-0.695 to -0.199, p_{nz}>0.999) and marginally reduced learning rate from reward (d=-0.159, 95% HDI= -0.336 to 0.016, p_{nz}=0.929). The groups did not differ in terms of reinforcement sensitivity (d=-1.11, 95% HDI=-2.95 to 0.940, p_{nz}=0.797) or extinction rate (d=-0.039, 95% HDI = -0.142 to 0.070, p_{nz}=0.533).

Learning parameters were differentially affected by the dopaminergic drugs in both groups. In healthy controls, amisulpride reduced the rates of learning from reward (d=-0.142, 95% HDI = -0.263 to -0.039, p_{nz}=0.992) and punishment (d=-0.387, 95% HDI=-0.537 to -0.236, p_{nz}>0.999), and increased reinforcement sensitivity (d=1.87, 95% HDI=0.676 to 3.19, p_{nz}=0.995) (Figure 3.4A). However, amisulpride improved the rate of learning from punishment in StimUD patients (d=0.186, 95% HDI=0.020 to 0.373, p_{nz}=0.975); no other parameters were affected (0 ∈ 95% HDI) (Figure 3.4C). Similarly, pramipexole reduced punishment learning rates in controls (d=-0.270, 95% HDI= -0.440 to -0.109, p_{nz}=0.999) (Figure 3.4B) but improved the punishment learning rate (d=0.463, 95% HDI=0.199 to 0.729, p_{nz}=0.995) in StimUD patients (Figure 3.4D). Pramipexole also reduced the reinforcement sensitivity parameter in StimUD patients (d=-1.92, 95% HDI=-3.53 to -0.360, p_{nz}=0.972); other parameters were not affected (0 ∈ 95% HDI). Comparison of the drug effects between StimUD and control participants (Figure 3.5) revealed that, relative to controls, both amisulpride (d=0.573, 95% HDI = 0.341 to 0.802, p_{nz}>0.999) and pramipexole (d=0.705, 95% HDI = 0.322 to 1, p_{nz}>0.999) greatly improved the punishment learning rate parameter in StimUD patients. The drug effects on the other parameters were not credibly different between groups (0 ∈ 95% HDI).

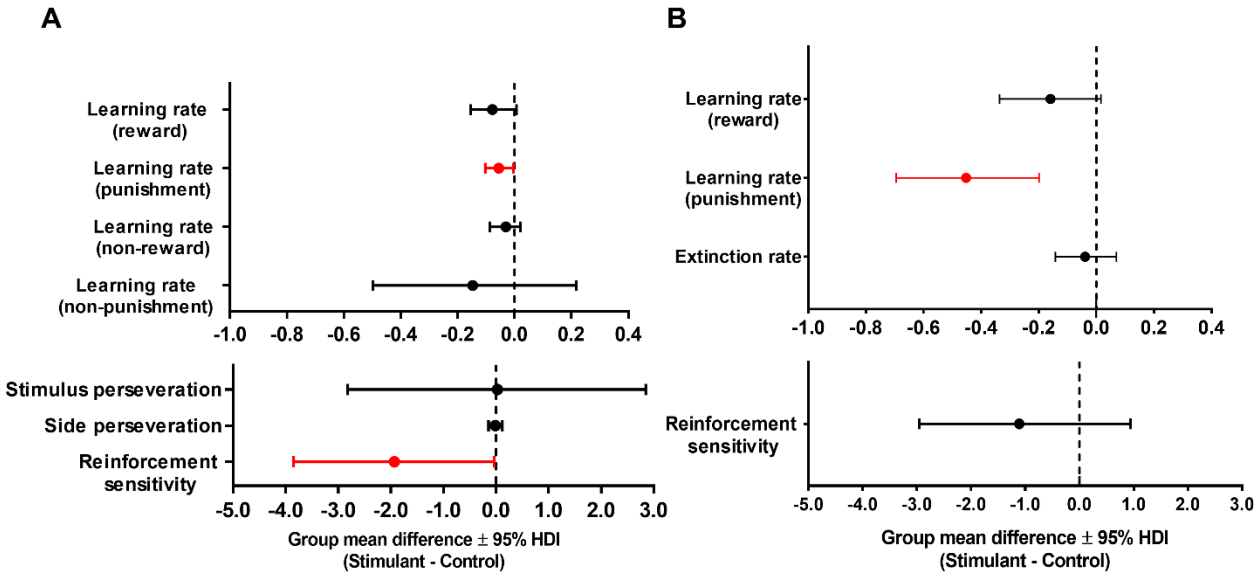


Figure 3.3: Group mean differences for the reinforcement learning parameters. (A) In study 1, the learning rate from punishment and reinforcement sensitivity were significantly reduced in the StimUD participants, while the other parameters were no different across groups. **(B)** In the placebo condition of study 2, we found markedly reduced learning rate from punishment in StimUD patients. [Error bars denote 95% highest density intervals (HDI); parameters colored in red signify a credible group difference (95% HDI excludes zero)].

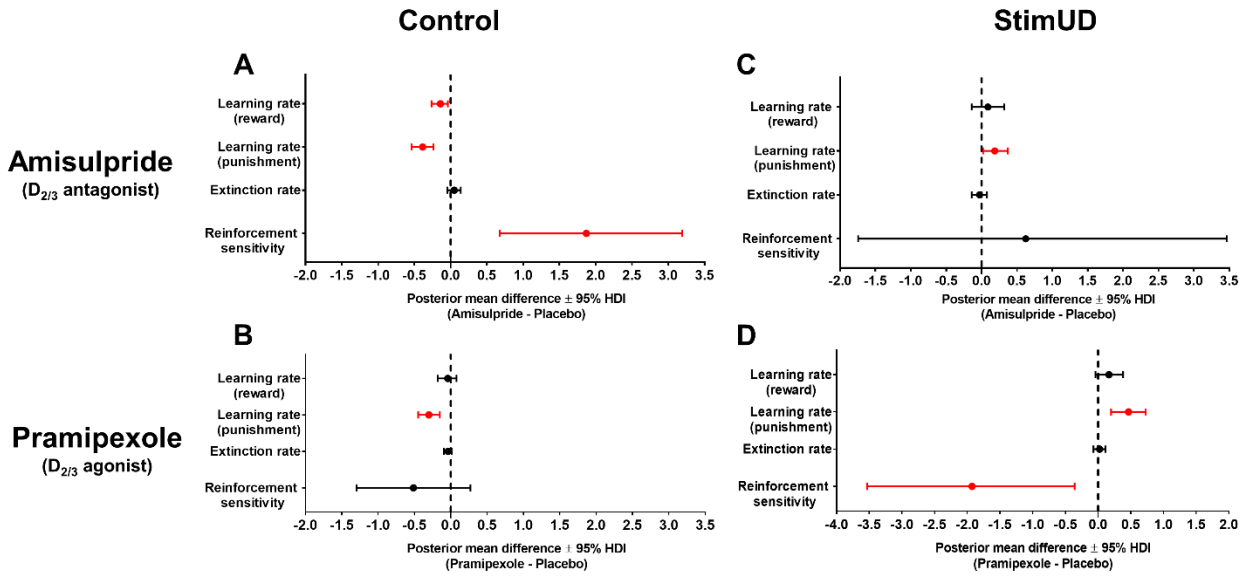


Figure 3.4: Mean differences of the reinforcement learning parameters for each drug condition. The dopaminergic agents are directly compared with placebo. **(A)** Amisulpride reduced the learning rates in healthy controls, but increased the reinforcement sensitivity parameter. **(B)** Pramipexole selectively reduced the reward learning rate parameter in control participants, but had no effect on the other parameters. **(C)** Amisulpride improved the punishment learning rate in StimUD participants. **(D)** Pramipexole significantly increased punishment learning rate and reduced reinforcement sensitivity parameters in StimUD patients [Error bars denote 95% highest density intervals (HDI); parameters colored in red indicate a credible drug effect, as their 95% HDI excludes zero.]

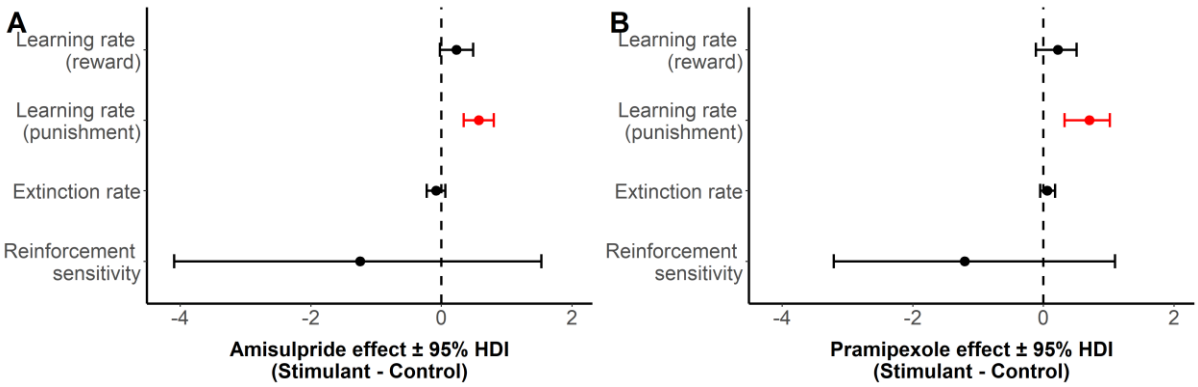


Figure 3.5: Comparison of the drug effects between the StimUD and control groups. The posteriors for drug effects were computed by sampling the group difference between the drug effects (i.e. medication minus placebo) for **(A)** amisulpride and **(B)** pramipexole. Both dopaminergic agents had a larger effect on the punishment learning rate in StimUD patients than healthy controls. [Error bars denote 95% highest density intervals (HDI); parameters coloured in red indicate a credible effect, as their 95% HDI excludes zero.]

3.4 Discussion

Behaviour in StimUD patients is thought to be driven by immediate positive outcomes, but at the expense of long-term negative consequences (Bechara et al., 2002; Verdejo-Garcia et al., 2018). I investigated RL performance in StimUD patients with a task that separately assessed learning from immediate monetary reward and punishment. As hypothesised, computational analyses revealed significant RL impairments in StimUD patients, which were driven primarily by a reduced learning rate from punishment. I also found that dopaminergic drugs differentially affected RL parameters in StimUD patients and matched controls. Whilst both dopaminergic drugs impaired the learning rates in controls, StimUD patients benefitted from them, as both drugs improved their ability to learn from punishment. Here I provide converging computational and pharmacological evidence of significant learning impairments in StimUD patients which are, at least in part, related to dopamine dysfunction.

3.4.1 Reinforcement learning profile in stimulant use disorder

RL in StimUD is characterised by significant impairments in learning from immediate punishment, which may suggest that negative outcomes have little impact on subsequent behaviour. This proposal concurs with prior research in animals, demonstrating that psychostimulant self-administration impairs the update of learned values from negative outcomes (Groman et al., 2018, 2020). Moreover, some studies in StimUD patients also reported aberrant responses towards immediate negative outcomes, whether those outcomes are electric shocks or symbolic error feedback (Ersche et al., 2016; Hester et al., 2013; Parvaz et al., 2015; Thompson et al., 2012). Negative outcomes such as monetary losses have been suggested to be important in aversive instrumental learning (Jean-Richard-Dit-Bressel et al., 2018). Consequently, the reduced impact of negative feedback during learning may hamper StimUD patients' ability to avoid negative outcomes. From a theoretical perspective, reduced learning from negative consequences may also point towards a weakness in the goal-directed system, which is sensitive to the consequences of one's actions (Balleine & Dickinson, 1998). In other words, blunted sensitivity towards negative outcomes may weaken the ability to adjust ongoing behaviour according to the situational demands and contribute to the development of compulsive behaviours in StimUD patients (R. J. Smith & Laiks, 2018). The hypothesis of a weakened goal-directed system in StimUD is supported by converging lines of evidence in

both humans (Ersche et al., 2020; Lim et al., 2019) and animals (Corbit, Chieng, et al., 2014; Zapata et al., 2010).

Although several studies report reduced responses to punishment in StimUD patients (Ersche et al., 2016; Hester et al., 2013; Thompson et al., 2012), inconsistent findings have also been observed. For example, a computational analysis by Kanen and colleagues reported i*ncreased* learning rate for punishment in StimUD patients in a serial probabilistic reversal learning task (Kanen et al., 2019). While this task also involves RL, it is important to consider the task context when interpreting these findings. In a probabilistic serial reversal learning task, participants are instructed to expect learned contingencies to change from time to time, and thus need to balance between ignoring and responding to punishment i.e. staying with or switching their choices, respectively. An increased learning rate from punishment in this context could thus also reflect an impaired ability to use negative feedback to guide behaviour amidst a volatile environment, leading to more errors in StimUD patients. Since there were no contingency reversals in my tasks, such divergence in the behavioural profile could be due to intrinsic differences in task design. Indeed, when I fitted the winning model from Kanen et al to the present data, we obtained results consistent with our model – StimUD patients still show a reduced learning rate from punishment (Appendix B).

Compared with learning from punishment, learning from reward was less impaired in StimUD patients, indicating that monetary reward remains a salient reinforcer amongst stimulant drug users. This may suggest that behaviour in StimUD patients is more amenable to positive than to negative feedback and could explain why treatments based on positive reinforcement such as contingency management (Petry, 2000; Petry et al., 2017) are effective in StimUD. Accumulating evidence further suggests that contingency management with monetary incentives is as effective (Festinger et al., 2014), or even more effective (Stoops et al., 2010; Vandrey et al., 2007), in promoting cocaine abstinence and treatment retention than non-monetary incentives (Stitzer et al., 2010). These studies jointly imply that the prospective knowledge of more salient rewards, such as monetary gains, improves contingency learning. Indeed, studies that adopted non-salient feedback (e.g. points or artificial stimuli) in RL tasks reported impairments in learning from reward in StimUD patients (Lim et al., 2019; Strickland et al., 2016), possibly reflecting the lack of a motivating reinforcer. This stands in stark contrast

to learning from negative consequences, which is significantly impaired regardless of its magnitude (Thompson et al., 2012). However, whether different modes of punishment differentially affect behaviour in StimUD patients remains an open question.

3.4.2 Dopaminergic modulation of RL in healthy participants

Although the involvement of dopamine in RL is undisputed, the exact mechanistic role of D_2 receptors in learning remains controversial, as reflected in the conflicting findings reported in the literature. For example, some studies showed that pharmacological modulation of D_2 receptors affects only reward but not punishment (Eisenegger et al., 2014; Pessiglione et al., 2006; Pizzagalli et al., 2008), suggesting that D_2 receptor signalling selectively affects reward learning. However, other evidence from humans (Cox et al., 2015; Frank & Hutchison, 2009) and preclinical studies (Alsiö et al., 2019; Hikida et al., 2010; Kravitz et al., 2012; Verharen et al., 2019) suggest that D_2 receptor signalling plays a specific role in avoiding negative outcomes (Frank, 2005; Frank & O'Reilly, 2006). Whilst the selective impairment of punishment learning in healthy participants following the $D_{2/3}$ receptor agonist is consistent with the latter view, the observation that the $D_{2/3}$ receptor antagonist affected both reward and punishment does not support the hypothesis that the D_2 receptor has a valence- specific role in learning. Such non-selective effects of $D_{2/3}$ receptor antagonism have previously been reported (Jocham et al., 2014; McCabe et al., 2011), suggesting that these receptors are generally involved in normal feedback-based learning.

The $D_{2/3}$ receptor antagonist also increased the reinforcement sensitivity parameter in healthy participants, suggesting that amisulpride increased their motivation for higher valued choices. This proposal concurs with other pharmacological studies administering amisulpride, which found that the drug enhanced sensitivity to expected values (Burke et al., 2018) and increased medial-orbitofrontal-cortex activation during choice selection (Jocham et al., 2011; Kahnt et al., 2015), a region commonly associated with value representation (O'Doherty, 2004).

When interpreting the drug effects, it is important to consider that dopaminergic $D_{2/3}$ drugs may exert presynaptic actions. At low doses, $D_{2/3}$ agents preferentially bind to pre-synaptic autoreceptors (Schoemaker et al., 1997), which inhibit dopamine transmission (Ford, 2014).

Thus, the D_2 presynaptic autoreceptor blockade by a dopamine antagonist may actually enhance dopamine transmission, whereas stimulation of D_2 autoreceptors by a dopamine agonist may result in a net reduction of dopaminergic transmission. It is therefore tempting to speculate whether the pramipexole-induced impairments in the learning rate and the amisulpride-induced enhancements in reinforcement sensitivity, as seen in the healthy participants, reflect such pre-synaptic actions.

3.4.3 Impaired RL associated with altered dopamine system in stimulant use disorder

The dopaminergic agents had the opposite effect in StimUD patients compared with healthy controls, which suggests an altered dopaminergic system in StimUD. There is considerable evidence from positron-emission-tomography studies that point towards downregulation of striatal D_2 receptors and dopaminergic neurotransmission in StimUD patients (Martinez et al., 2004, 2007; Volkow et al., 1993, 1997). Repeated stimulant drug exposure has also been proposed to upregulate the inhibitory activity of D_2 presynaptic autoreceptors, which in turn may suppress dopamine signalling below normal levels (Grace, 1995). However, it is not possible to determine precisely the nature of the dopamine system, and the mode of action of the dopamine agents in StimUD patients, which depends on dopamine levels at baseline (Cools et al., 2001, 2009). I thus interpreted the effects of dopaminergic agents in light of StimUD patients' possibly reduced dopamine activity and potential pre-synaptic effects of these agents.

If D_2 receptors are assumed to be important in learning from negative feedback (Frank & O'Reilly, 2006; Nakanishi et al., 2014), the downregulation of D_2 receptors in StimUD would explain their reduced learning from negative outcomes, which is mirrored in healthy individuals with low D_2 receptor levels (Jocham et al., 2009; Klein et al., 2007). It is therefore conceivable that amisulpride improved punishment learning in StimUD by blocking presynaptic D_2 autoreceptors, and thus increasing dopamine signalling. Pramipexole also improved punishment learning in StimUD, possibly by enhancing dopamine signalling through post-synaptic mechanisms. It remains, however, unclear why two opposing drugs work in the same direction. It is noteworthy that the reinforcement sensitivity parameter, which measures how much choices are motivated by learned values, was reduced on pramipexole. This may suggest that altering the dopamine balance reduced StimUD patients' tendency to engage in the RL task as the choice values became less motivating. This concurrent reduction in

motivation might also explain why StimUD patients did not show improvements in overall performance, despite an improved punishment learning. The effects of dopaminergic agents seem to confirm altered dopaminergic activity in StimUD patients, which have been associated with learning difficulties, but the precise pharmacological actions are likely to depend on task context, drug dosage and baseline dopamine transmission.

3.4.4 Strengths, weaknesses and outlook

The current data provide compelling evidence for impaired learning from punishment in two independent samples of StimUD patients, one of which had comorbid dependencies whilst the other had none. Concurrent use of other drugs such as opiate, alcohol or cannabis is therefore unlikely to have affected the observed performance profiles. Limitations include the uncertainty of the nature of the drug effects i.e. whether they reflect pre-synaptic or post-synaptic effects, which is difficult to determine in human research. Therefore, any inferences on the drug effects should be cautiously interpreted. Further neuroimaging evidence (e.g. positron-emission-tomography) is warranted to clarify the action of the dopaminergic drugs, as responses to dopaminergic drug may vary according to baseline dopamine synthesis capacity (Cools et al., 2009) and dopamine receptor density (Cohen et al., 2007; Eisenegger et al., 2014). Although I focused exclusively on dopamine, it is important to acknowledge that other neurotransmitter systems such as serotonin (Seymour et al., 2012) and glutamate (Groman et al., 2020) are also implicated in RL. There is also evidence that amisulpride has an affinity for serotonin receptors (Abbas et al., 2009), which may also modulate sensitivity to aversive events (Cools et al., 2011; Daw et al., 2002). Future studies using a longitudinal design are needed to investigate these factors. Additionally, the smaller sample size of study 2 (n=36) may be under-powered to detect certain significant effects. Furthermore, the group difference in verbal IQ in study 1 (but not study 2) is noteworthy, as it might confound learning performance, since higher IQ would likely lead to better learning performance. Upon reflection, the difference in IQ-matching between study 1 and study 2 may be related to differences in the volunteers. Whilst both studies recruited from Cambridgeshire, most control volunteers for study 1 were members of the University with undergraduate and postgraduate degrees i.e. a population with relatively higher IQ. By contrast, study 2 had access to volunteer panels for research (e.g. GSK volunteer panel) and was well-funded. Thus, we were able to recruit control participants from the general public beyond the University environment. However, in the context of this chapter, it is likely that the confounds from verbal IQ in study 1 would not significantly affect the interpretation

of the data, for two reasons: (1) there were no statistically significant associations between verbal IQ and task performance measures in the StimUD group, which suggests that IQ is unlikely to alter performance here; (2) the main finding in study 1 was replicated in study 2 (with a matched IQ sample), which was reassuring. Nonetheless, my findings present novel evidence for selective learning impairments in StimUD, and highlight the utility of computational modelling in deconstructing complex cognitive processes, with promising prospects for psychiatry and psychopharmacology research (Huys et al., 2021).

Appendix B: Supplementary materials to Chapter 3

Supplementary methods

Behavioural task: additional description

The probabilistic reinforcement learning tasks were designed to investigate learning from monetary feedback. In both versions, I assessed learning from reward and punishment separately. On each trial, participants were presented with pairs of stimuli and were required to learn by trial and error to select the stimulus that minimizes their financial losses or maximises their winnings. I differentiated between reward and punishment conditions by using distinct stimulus pairs and different feedback type. In the reward condition, the favourable choices received a 'you win 50p' outcome 70% of the time and a 'you win 0p' outcome 30% of the time. By contrast, unfavourable choices were punished (i.e. 'you lose 50p') 70% of the time, and not punished ('you lose 0p') 30% of the time.

As these data were part of two separate larger studies, there were several differences between the task designs:

1 I administered an additional neutral condition to the second sample as a baseline for performance. Like the reward and punishment trials, the neutral trials consisted of a unique stimulus pair, and also maintained a 70% contingency, but choosing either stimuli from this condition had no financial consequences (refer to Figure 3.1).
2 There were two unique stimulus pairs for each reinforcement condition in study 1 (four in total), but only one unique stimulus pair for each condition in study 2 (three in total).
3 In study 2, no explicit feedback was provided for reward and punishment omission outcomes i.e. only a blank screen.
4 In study 2, participants were informed that their task earnings would be translated to actual monetary bonuses. By contrast, participants did not receive earnings for their performance in study 1, but they were nevertheless instructed to do their best.

Statistical analyses

Computational modelling of behaviour

I modelled the expected values of each choice, trial by trial, using a delta rule (Rescorla & Wagner, 1972):

$$V_{t+1} = V_t + \alpha(R - V_t)$$

V_t denotes the expected value of the chosen stimulus on trial t. The update of this value is determined by the product of α, the learning rate, and the prediction error, the discrepancy between expected value and actual reinforcement received on trial t, R. Mathematically, R is assigned 1 for reward, -1 for punishment, and 0 otherwise. Since there is evidence for different neural systems subserving learning from reward and punishment (Pessiglione & Delgado, 2015), I fractionated α based on the feedback received. In study 1, there are four possible outcomes: reward (i.e. you win 50p), reward omission (i.e. you win 0p), punishment (i.e. you lose 50p) and punishment omission (i.e. you lose 0p), hence the learning from each outcome was modelled with a different α as follows:

$V_{t+1} = V_t + \alpha_{rew}(R - V_t)$ if feedback = "You win 50p";

$V_{t+1} = V_t + \alpha_{non-rew}(R - V_t)$ if feedback = "You win 0p";

$V_{t+1} = V_t + \alpha_{pun}(R - V_t)$ if feedback = "You lose 50p";

$V_{t+1} = V_t + \alpha_{non-pun}(R - V_t)$ if feedback = "You lose 0p";

where α_{rew}, $\alpha_{non\text{-}rew}$, α_{pun}, and $\alpha_{non\text{-}pun}$, refers to the learning rates from reward, non-reward, punishment, and non-punishment respectively.

In study 2, reward and punishment omission outcomes were not explicitly framed within a win/loss domain; instead participants were not provided with explicit feedback (Figure 3.1). Therefore, it would not be apparent from this feedback whether participants experienced a loss/gain from their selection. Thus, I modelled learning from these outcomes with a general extinction rate, α_{ext}:

$V_{t+1} = V_t + \alpha_{rew}(R - V_t)$ if feedback = "You win 50p";

$$V_{t+1} = V_t + \alpha_{pun}(R - V_t)$$ if feedback = "You lose 50p";

$$V_{t+1} = V_t + \alpha_{ext}(R - V_t)$$ if no feedback given;

Increases in α would indicate an increased rate of stimulus value update from the corresponding feedback.

I used the expected values to model actual choice behaviour, via a softmax rule:

$$p(i,t) = \frac{exp(\beta V_t^i)}{\sum_{k=1}^{n} exp(\beta V_t^k)}$$

This equation gives the model's probability of choosing choice i amongst n choices on a trial t. The extent to which expected values are used to drive choices is governed by the reinforcement sensitivity parameter, β. Cocaine use disorder has been associated with the tendency to perseverate responses irrespective of reinforcement (Ersche et al., 2008; Ersche, Roiser, Abbott, et al., 2011), and this may affect learning. Hence, in models where I accounted for perseveration, I modified the softmax equation to include the term C_{stim} and C_{loc} to indicate perseveration towards the previously chosen stimulus and towards the previously chosen location (e.g. left, right) respectively, regardless of choice value:

$$p(i,t) = \frac{exp(\beta V_t^i + \tau_{stim} {C_{stim}}_t^i + \tau_{loc} {C_{loc}}_t^i)}{\sum_{k=1}^{n} exp(\beta V_t^k + \tau_{stim} {C_{stim}}_t^k + \tau_{loc} {C_{loc}}_t^k)}$$

C_{stim} and C_{loc} are assigned 1 if participants repeated their choices for the same stimulus (e.g. chose stimulus A on previous trial, and choose stimulus A again on current trial) and location (e.g. responded 'left' last trial, and respond with 'left' again this trial) respectively, and 0 if choices were not repeated. The tendency for perseveration to influence choice is governed by perseveration parameters for stimulus, τ_{stim}, and for location, τ_{stim}. There were in total eight possible free parameters in the learning models (not all used in any given model): learning rate from reward, non-reward, punishment and non-punishment, general extinction rate, reinforcement sensitivity, and well as perseveration tendencies to stimulus and side. I used bridge sampling procedures to identify the best-fitting model from several variants of the learning models (see section below and Appendix B, Table B2).

Parameter estimation

I estimated the posterior distribution of free parameters by analysing the best-fit model within a hierarchical Bayesian framework. For each parameter, I modelled a group-level distribution at the top of the hierarchy. For study 2, I used each group/drug combination. The posterior distributions of group-level (or group/drug) parameters were the main measures of interest. Prior distributions were assigned to all parameters (see Appendix B, Table B1). I also modelled inter-subject variability and accounted for the within-subject aspects of the design: for each parameter, subject-specific deviations from the group mean were drawn from a normal distribution with mean 0 and a parameter-specific standard deviation (itself estimated). Subject-specific parameters were then used in the reinforcement learning model to predict choice (what does the model predict is the probability of the subject choosing stimulus A?), and the model was fitted by comparing these to actual choices made (did the subject choose stimulus A?). I implemented the procedure in RStan (version 2.17.2), which uses a Markov chain Monte Carlo approach. I simulated eight parallel chains, each with 2000 iterations (including warmup), and directly sampled posterior distributions of the group/drug mean differences, d, as the primary outcome measure. Measures of dispersion for the group differences are indicated by the 95% highest density intervals (HDI); a 95% HDI that excludes zero provides strong evidence for a group difference (non-zero-difference, p_{nz}>0.95). No multiple comparison corrections were applied, because Bayesian hierarchical analyses tend to produce more conservative comparisons by shifting point estimates towards each other ('partial pooling'), making intervals more likely to include zero (Gelman et al., 2012).

Model selection

I used bridge sampling, implemented via the "*bridgesampling*" package in R, to determine the best-fit model. Given that the correct model is one of the models being considered, the posterior probability of a model, *P*(model | data), can be measured directly by taking (1) the prior probability of the model itself, *P*(model), and (2) the marginal likelihood of, or evidence for, the model, *P*(data | model), which can be estimated via bridge sampling. The marginal likelihood integrates, over all possible parameter values, the product of (2a) the likelihood of the data given the fitted model (how well the model fits the data), *P*(data | parameters, model), and (2b) the probability of the parameters given the model, *P*(parameters | model), thus incorporating Occam's razor by penalizing over-complex models. We assumed all models were

equiprobable *a priori*. I also report the Bayes factor, defined as the ratios of marginal likelihood of a pair of models, as a secondary indicator for model evidence. Model comparison results are reported in Appendix B Table B2. Generally, a Bayes factor of more than one constitutes sufficient evidence that one model is better than the next (Kass & Raftery, 1995).

Winning model validation

It is important that the winning model is able to recover key aspects of the behavioural findings (Wilson & Collins, 2019). To that end, I performed a posterior predictive check by simulating data for 50 'virtual' subjects per group for each study, using the posterior mean values for each group-level parameter. I did not incorporate subject variance, as I was interested in group effects instead of individual variability. The simulated data were analysed using analysis of variance (ANOVA) in the same way as reported in the main manuscript text, with block and condition (reward versus punishment) as within-subject factors, and group (StimUD versus controls) as a between-subjects factor.

I assessed parameter recovery for the winning model to ensure that the model fitting procedure was able to recover the simulated parameter values. To that end, I simulated 1000 datasets with known parameter values (randomly generated from priors in Table B1), and fitted these data with the same model fitting process to recover these parameter values. Parameter recovery was assessed by inspecting the correlations between simulated and recovered parameter values. Strong correlations between simulated and recovered parameters indicate good recoverability (Wilson & Collins, 2019).

As an extra validation step, I also fitted my data to the winning model of Kanen et al., who recently analysed behavioural performance of stimulant use disorder (StimUD) and obsessive-compulsive disorder (OCD) patients on a serial probabilistic reversal learning task (Kanen et al., 2019). The main purpose of this analysis is to determine whether the divergence in the current results and Kanen et al.'s, specifically in the punishment learning rate, is due to intrinsic differences in the behavioural tasks or modelling procedures. The winning model in Kanen et al. had five parameters: reward learning rate, punishment learning rate, reinforcement sensitivity, stimulus stickiness and location stickiness (Kanen et al., 2019). The main difference between the winning models and that of Kanen et al.'s is the inclusion of extinction rate

parameters into the models i.e. I included parameters that modelled separately the effects of reward omission and punishment avoidance events. This method of modelling was not possible in Kanen et al's study as their task only contained two types of feedback (reward and non-reward). Thus, to mimic Kanen et al's winning model as closely as possible, I removed the extinction rate parameters and modelled all reward trials (reward delivery + omission) with the reward learning rate, and all punishment trials (punishment + avoidance) with the punishment learning rate. The marginal likelihood for this model was also estimated with bridge sampling to compare it against my winning model. If this analysis produce results consistent with my winning model, it would suggest that any intrinsic differences between current findings and that of Kanen et al's are likely attributed to difference between behavioural tasks (e.g. the inclusion of reversals); if not, then the divergence with the current results might be a product of different modelling procedures.

Supplementary results

Model selection and winning model validation

The winning model in study 1 included the following parameters: learning rates from reward, punishment, non-reward, and non-punishment; perseveration towards location and stimulus; and reinforcement sensitivity. Based on the criteria of Kass and Raftery (Kass & Raftery, 1995), there was overwhelming evidence that the winning model is superior to the next best model (Table B2). The winning model for study 2 consisted of four parameters: learning rates from reward and punishment, extinction rate, and reinforcement sensitivity. Again the Bayes factor suggests that the winning model was far superior to the next-ranked model.

I was able to validate the winning models for each study. For study 1, analyses on simulated data from the winning model were able to recover the main effects of group ($F_{1,98}$=38.5, p<.001), block ($F_{5,490}$=60.3, p<0.001), condition ($F_{1,98}$=91.1, p<0.001) as well as the lack of group-by-condition interaction ($F_{1,98}$=1.93, p=0.168) as reported in the main text. For study 2, analyses on accuracy scores of simulated data from placebo recovered the main effect of condition ($F_{1,98}$=64.4, p<0.001), block ($F_{3,294}$=22.7, p<.001), and group-by-block interaction ($F_{3,294}$=4.22, p=.006), and the lack of group effect ($F_{1,98}$=0.666, p=.416) and group-by-condition interaction ($F_{1,98}$=1.05, p=.308). Simulation of amisulpride data was also able to reproduce all the main effects: I was able to recover the main effects of group ($F_{1,98}$=9.06, p=.003), condition ($F_{1,98}$=61.1, p<.001), block ($F_{3,294}$=93.9, p<.001) as well as the lack of drug effect ($F_{1,98}$=1.10, p=.297). The simulated data additionally yielded a group-by-drug interaction ($F_{1,98}$=15.3, p<.001) such that amisulpride significantly reduced accuracy in StimUD patients (t_{98}=3.81, p=.001), but not controls (t_{98}=2.2, p=.123); this difference could be due to the removal of participant variability in the simulated data, which might have increased the effect size. Simulations of pramipexole data recaptured the main effect of block ($F_{3,294}$=88, p<.001) and drug-by-condition interaction ($F_{1,98}$=5.82, p=.018). The simulated pramipexole data also found a main effect of condition ($F_{1,98}$=84, p<.001), which again may be due to reduced variance in simulated data; all other effects were not statistically significant (p > 0.1).

Figure B2 shows the scatterplots between simulated and recovered parameters for the winning models. All simulated and recovered parameters were strongly correlated, suggesting good parameter recovery.

I analysed the data with the winning model in Kanen et al. as an extra confirmatory step. It is worth noting that bridge sampling procedures indicate that this model was ranked the 6th best model in both studies (Table B2), which means that results from this model do not best describe the current data. Nevertheless, analyses with Kanen et al.'s winning model replicated the main finding for the reduced punishment learning rate in both study 1 (d = -0.064, 95% HDI = -0.118 to -0.007, p_{nz} = 0.972) and the placebo condition in study 2 (d = -0.182, 95% HDI = -0.305 to -0.055, p_{nz} = 0.995). The ability to replicate the reduced punishment learning rate with Kanen et al's winning model confirms that the divergent findings on punishment-driven learning are due to intrinsic differences between the current RL task and one with contingency reversals.

Additional correlations between demographics and task performance

Since there was a group difference in verbal IQ in study 1, as measured with the National Adult Reading Test (H. E. Nelson, 1982), I further assessed the relationship between verbal IQ and task performance in the StimUD group. I also investigated the relationship between task performance measures and stimulant use duration, as well as severity of compulsive drug use as measured by the Obsessive-Compulsive Drug Use Scale (OCDUS). Results are shown in Table B3. I did not find any significant correlations between these demographic measures and task performance in both studies, whether measured by accuracy scores or inferred computational learning parameters.

Supplementary Tables and Figures

Table B1: Prior distributions for all possible parameters.

Parameter	Range (lower bound, upper bound)	Group mean priors	Inter-subject standard deviation priors*
Reward learning rate	0, 1	Beta(1.1, 1.1)	Normal(0, 0.05)
Punishment learning rate	0, 1	Beta(1.1, 1.1)	Normal(0, 0.05)
Non-reward learning rate	0, 1	Beta(1.1, 1.1)	Normal(0, 0.05)
Non-punishment learning rate	0, 1	Beta(1.1, 1.1)	Normal(0, 0.05)
General learning rate	0, 1	Beta(1.1, 1.1)	Normal(0, 0.05)
General extinction rate	0, 1	Beta(1.1, 1.1)	Normal(0, 0.05)
Reinforcement sensitivity	0, ∞	Gamma(α = 4.82, β = 0.88)	Normal(0, 1)
Perseveration towards location (side)	-∞, +∞	Normal(0, 1)	Normal(0, 0.05)
Perseveration towards stimulus	-∞, +∞	Normal(0, 1)	Normal(0, 0.05)

Note. *all standard deviation priors are constrained to be positive.

Table B2: Variants of learning models and model comparison results.

Model parameters	Ranking	Log marginal likelihood	Log posterior p(model)	Posterior p(model)	Log_{10} Bayes Factor (relative to next-ranked model)
Study 1					
α_{rew}, $\alpha_{non\text{-}rew}$, α_{pun}, $\alpha_{non\text{-}pun}$, β, τ_{loc}, τ_{stim}	**1**	**-5586.47**	**-0.009**	**0.991**	**2.058**
α_{rew}, $\alpha_{non\text{-}rew}$, α_{pun}, $\alpha_{non\text{-}pun}$, β, τ_{loc}	2	-5591.21	-4.747	0.009	7.698
α_{rew}, $\alpha_{non\text{-}rew}$, α_{pun}, $\alpha_{non\text{-}pun}$, β, τ_{stim}	3	-5608.93	-22.5	1.74×10^{-10}	0.069
α_{rew}, $\alpha_{non\text{-}rew}$, α_{pun}, $\alpha_{non\text{-}pun}$, β	4	-5609.09	-22.6	1.49×10^{-10}	8.385
α_{rew}, α_{pun}, β, τ_{loc}	5	-5628.40	-41.9	6.12×10^{-19}	0.870
α_{rew}, α_{pun}, β, τ_{loc}, τ_{stim}	6	-5630.40	-43.9	8.27×10^{-20}	6.831
α_{rew}, α_{pun}, β, τ_{stim}	7	-5646.13	-59.7	1.22×10^{-26}	0.013
α_{rew}, α_{pun}, β	8	-5646.16	-59.7	1.19×10^{-26}	9.389
α_{rew}, α_{pun}, α_{ext}, β, τ_{loc}	9	-5667.78	-81.3	4.84×10^{-36}	1.323
α_{rew}, α_{pun}, α_{ext}, β, τ_{loc}, τ_{stim}	10	-5670.82	-84.3	2.30×10^{-37}	9.389
α_{rew}, α_{pun}, α_{ext}, β, τ_{stim}	11	-5692.44	-105.9	9.39×10^{-47}	0.005
α_{rew}, α_{pun}, α_{ext}, β	12	-5692.45	-106.0	9.29×10^{-47}	87.24
α, β, τ_{loc}, τ_{stim}	13	-5893.33	-306.9	5.37×10^{-134}	0.851
α, β, τ_{loc}	14	-5895.29	-308.8	7.57×10^{-135}	6.777
α, β	15	-5910.89	-324.4	1.26×10^{-141}	0.083
α, β, τ_{stim}	16	-5911.08	-324.6	1.04×10^{-141}	131.5
α_{rew}, $\alpha_{non\text{-}rew}$, α_{pun}, $\alpha_{non\text{-}pun}$, τ_{loc}, τ_{stim}	17	-6213.89	-627.4	3.24×10^{-273}	0.909
α_{rew}, $\alpha_{non\text{-}rew}$, α_{pun}, $\alpha_{non\text{-}pun}$	18	-6215.98	-629.5	4.00×10^{-274}	33.40
α_{rew}, α_{pun}, τ_{loc}, τ_{stim}	19	-6292.88	-706.4	1.61×10^{-307}	2.452
α_{rew}, α_{pun}, α_{ext}, τ_{loc}, τ_{stim}	20	-6298.52	-712.1	5.68×10^{-310}	5.742
α_{rew}, α_{pun}, α_{ext}	21	-6311.75	-725.3	1.03×10^{-315}	40.53
α, τ_{loc}, τ_{stim}	22	-6405.06	-818.6	$< 5 \times 10^{-324}$	288.8
None (random choice model)	23	-7070.10	-	-	-
Study 2					
α_{rew}, α_{pun}, α_{ext}, β	**1**	**-5887.29**	**-0.002**	**0.998**	**2.768**
α_{rew}, α_{pun}, α_{ext}, β, τ_{stim}	2	-5893.67	-6.374	0.002	3.201
α_{rew}, α_{pun}, α_{ext}, β, τ_{stim}, τ_{loc}	3	-5901.04	-13.75	1.07×10^{-6}	1.032
α_{rew}, α_{pun}, α_{ext}, β, τ_{loc}	4	-5903.42	-16.13	9.91×10^{-8}	33.96
α_{rew}, α_{pun}, β	5	-5981.62	-94.33	1.08×10^{-41}	4.093
α_{rew}, α_{pun}, β, τ_{stim}, τ_{loc}	6	-5991.04	-103.75	8.73×10^{-46}	1.656
α_{rew}, α_{pun}, β, τ_{loc}	7	-5994.86	-107.6	1.93×10^{-47}	5.029
α, β	8	-6006.44	-119.1	1.80×10^{-52}	6.635
α, β, τ_{loc}	9	-6021.71	-134.4	4.18×10^{-59}	450.6
α_{rew}, α_{pun}, α_{ext}, τ_{stim}	10	-7059.22	-1171.9	$< 5 \times 10^{-324}$	0.012
α_{rew}, α_{pun}, α_{ext}	11	-7059.25	-1171.0	$< 5 \times 10^{-324}$	6.414
α_{rew}, α_{pun}, α_{ext}, τ_{stim}, τ_{loc}	12	-7074.02	-1186.7	$< 5 \times 10^{-324}$	4.869
α_{rew}, α_{pun}, α_{ext}, τ_{loc}	13	-7085.23	-1197.9	$< 5 \times 10^{-324}$	608.7
None (random choice model)	14	-8486.89	-	-	-

Note. Unless otherwise stated, log refers to the natural logarithm. α_{rew}: learning rate from reward; $\alpha_{non\text{-}rew}$: learning rate from non-reward; α_{pun}: learning rate from punishment; $\alpha_{non\text{-}pun}$: learning rate from non-reward; α: general learning rate; α_{ext}: extinction rate; β: reinforcement sensitivity; τ_{loc}: perseveration by location ("side"); τ_{stim}: perseveration by stimulus

Table B3: Correlations between demographics and task performance in stimulant use disorder patients.

Task performance measure	Verbal IQ (NART score)		Duration of stimulant use (years)		Compulsive drug use severity (OCDUS score)	
Study 1						
Conventional measures	*r*	*p*	*r*	*p*	*r*	*p*
Accuracy score (reward)	-.130	.939	-.166	.281	-.067	.664
Accuracy score (punishment)	.076	.647	-.157	.310	-.211	.169
Computational parameters	*r*	*p*	*r*	*p*	*r*	*p*
Learning rate from reward	.106	.521	-.209	.173	-.141	.361
Learning rate from punishment	-.039	.812	-.194	.206	-.284	.061
Learning rate from non-reward	-.091	.580	-.005	.975	-.111	.474
Learning rate from non-punishment	.122	.458	-.169	.272	-.185	.231
Reinforcement sensitivity	.146	.377	-.140	.366	-.010	.950
Perseveration by location	-.257	.114	.263	.084	.057	.714
Perseveration by stimulus	-.204	.214	-.002	.989	-.031	.839
Study 2 (placebo)						
Conventional measures	*r*	*p*	*r*	*p*	*r*	*p*
Accuracy score (reward)	.003	.991	-.086	.734	-.253	.311
Accuracy score (punishment)	-.179	.476	-.240	.337	.298	.230
Computational parameters	*r*	*p*	*r*	*p*	*r*	*p*
Learning rate from reward	.066	.795	-.457	.057	-.255	.308
Learning rate from punishment	.262	.294	.013	.959	.323	.192
General extinction rate	.196	.435	.054	.832	.124	.625
Reinforcement sensitivity	.192	.446	.053	.834	-.128	.614

Note. None of the pairwise correlations were statistically significant. [NART: National Adult Reading Test; OCDUS: Obsessive-Compulsive Drug Use Scale]

Figure B1: **Group posterior distributions for each parameter of the winning model.** (A) Group posteriors for study 1. (B) Group posteriors for the placebo condition in study 2. [Note: StimUD: stimulant use disorder; α_{rew}: learning rate from reward; $\alpha_{non\text{-}rew}$: learning rate from non-reward; α_{pun}: learning rate from punishment; $\alpha_{non\text{-}pun}$: learning rate from non-reward; α_{ext}: general extinction rate; β: reinforcement sensitivity; τ_{loc}: perseveration by location ("side"); τ_{stim}: perseveration by stimulus]

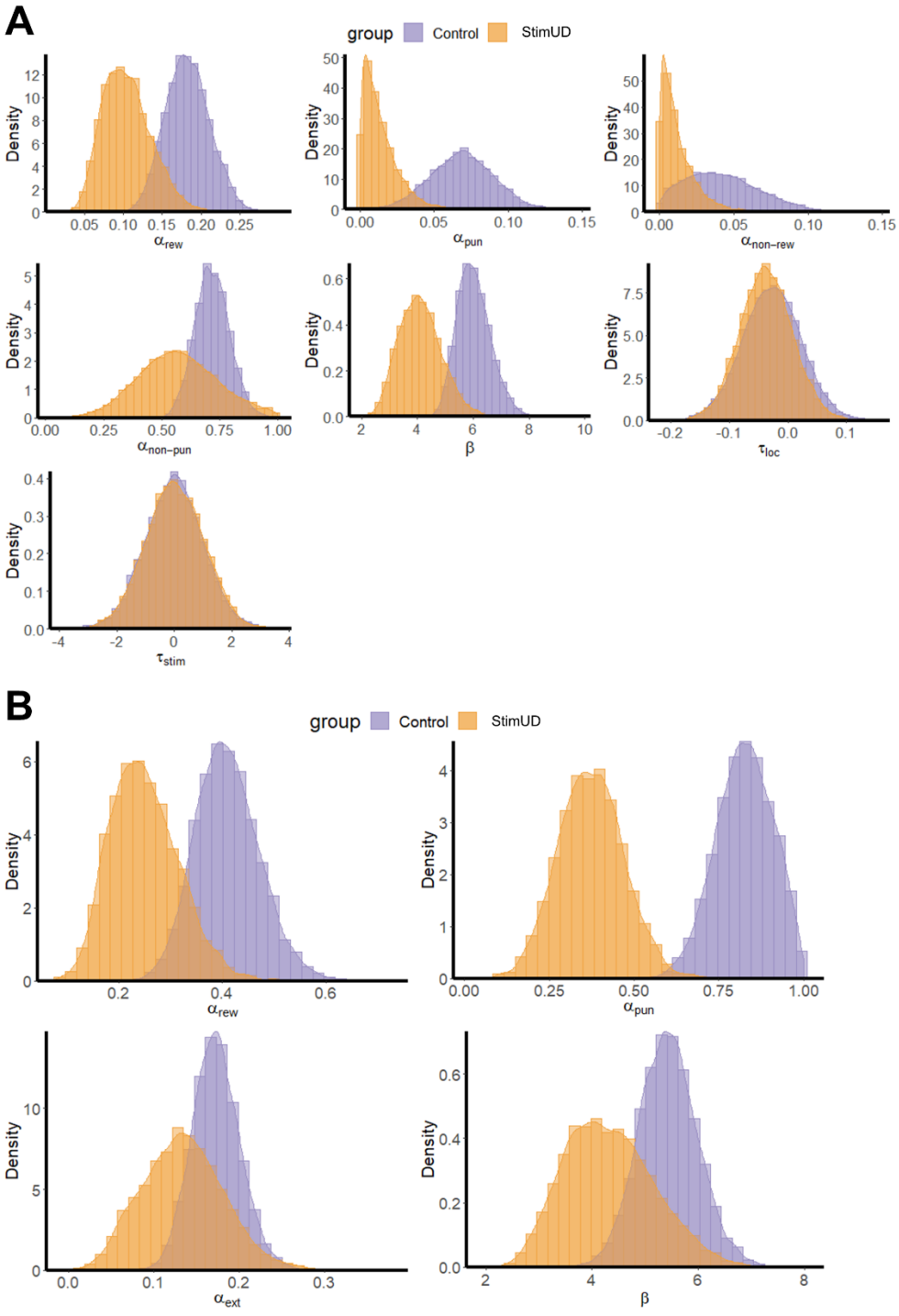
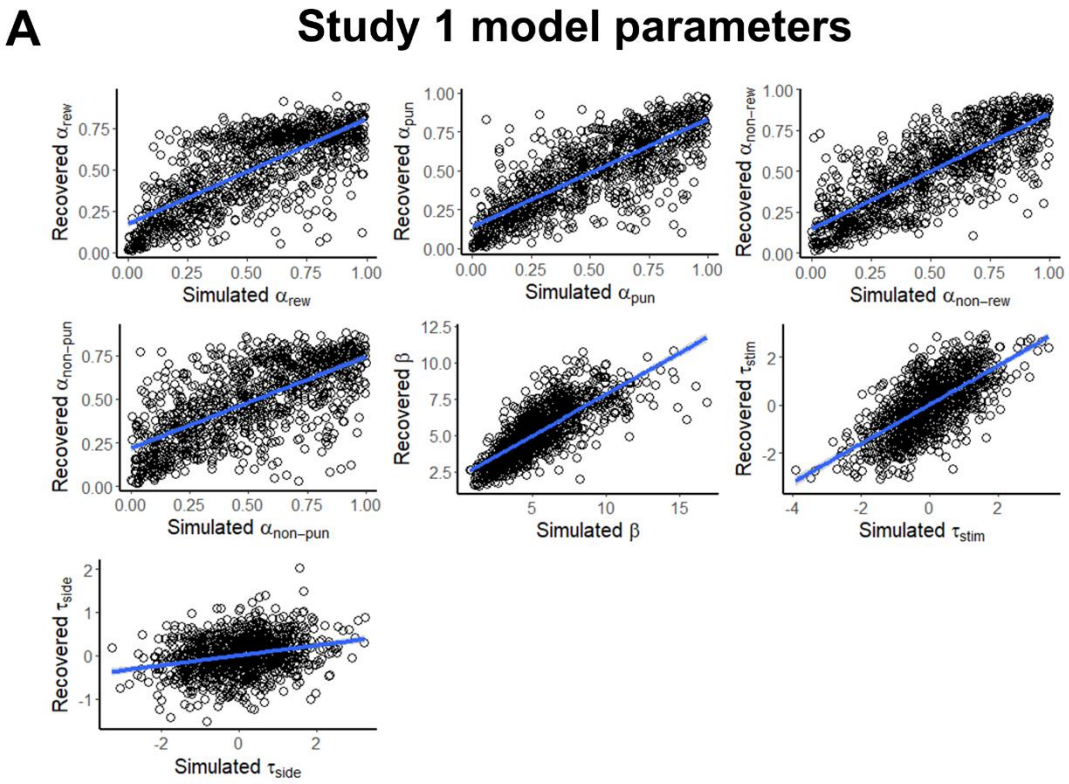
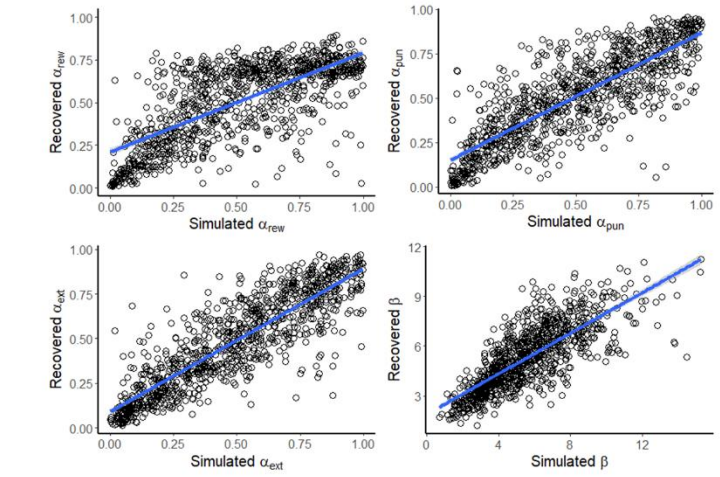


Figure B2: **Parameter recovery for the winning models in Chapter 3.** The scatterplots between simulated and recovered parameter values for each model parameter of (A) Study 1 and (B) Study 2. (C) The simulated-recovered pairwise correlations for each parameter.





C

Parameters	r	p
Study 1		
α_{rew}	0.75	<2.2e-16
α_{pun}	0.80	<2.2e-16
$\alpha_{non\text{-}rew}$	0.81	<2.2e-16
$\alpha_{non\text{-}pun}$	0.68	<2.2e-16
β	0.77	<2.2e-16
τ_{stim}	0.67	<2.2e-16
τ_{side}	0.27	<2.2e-16
Study 2		
α_{rew}	0.71	<2.2e-16
α_{pun}	0.82	<2.2e-16
α_{ext}	0.87	<2.2e-16
β	0.78	<2.2e-16

Chapter 4: Declarative and non-declarative memory in cocaine use disorder: behavioural analyses of probabilistic category learning

4.1 Introduction

Cocaine use disorder (CUD) is linked with alterations to learning functions that could perpetuate maladaptive drug-taking patterns (Everitt et al., 2001; Hyman, 2005; Redish et al., 2008). Converging evidence from CUD patients (e.g. Chapter 3) identified marked impairments in learning from reinforcing feedback. However, there are other processes that facilitate reinforcement learning, and are important for goal-directed behaviour. In particular, a parallel line of research suggests that declarative and non-declarative memory processes complement such learning. On the one hand, declarative memory refers to consciously accessible knowledge that is learned via memorisation of facts. This memory is thought to be dependent on the medial temporal lobe, including the hippocampus (Squire & Zola, 1996; Squire & Zola-Morgan, 1991). On the other hand, non-declarative memory encompasses knowledge that is learnt by practice and trial-and-error, and is thought to depend on the striatum (Poldrack et al., 1999). These memory processes facilitate the learning of associative knowledge, and can be tested experimentally by assessing category learning, defined as the learning of perceptual or abstract classifications that help guide decisions (Ashby & Maddox, 2005). Since CUD patients generally exhibit marked learning deficits, an open question is whether these learning impairments also reflect disruptions to memory processes that aid associative learning.

To examine this possibility, this chapter compares the performance of CUD patients and controls on a well-known probabilistic category learning task – the weather prediction task (Knowlton et al., 1996). This is a probabilistic category learning task that assesses the learning of contingencies between multiple cues and outcomes. This task comes in two versions, each with identical probabilistic structures and task objective, but they differ in terms of the route of learning. The first version, known as the feedback version, requires learning of contingencies from trial-by-trial corrective feedback, akin to standard reinforcement learning. Participants cannot merely rely on remembering the outcome of a previous encounter, but instead have to integrate information over many trials to form an understanding of what is optimal (Knowlton et al., 1996). By contrast, the second version, known as the paired-associates version, eliminates the corrective feedback, and instead requires participants to learn via observation (Poldrack et al., 2001). Participants are simultaneously presented with the cues and the outcome. They then need to memorise the relationship between cues and weather, and predict them later during a

test phase. The concurrent use of these two versions of the task have enabled the dissociation between declarative and non-declarative memory (Poldrack et al., 2001; Shohamy, Myers, Grossman, et al., 2004). Prior studies in neuropsychiatric conditions have supported this distinction. For example, Parkinson's Disease, characterised by an impaired nigro-striatal system, showed deficits in the feedback, but not the paired-associates version, suggesting a selective impairment to non-declarative memory whilst declarative memory remained intact (Shohamy, Myers, Grossman, et al., 2004). It is conceivable that CUD patients, who also show striatal deficits (Luijten et al., 2017; Yager et al., 2015), would have impaired non-declarative memory. However, the existing evidence for this is scarce and equivocal, with one study claiming deficits to non-declarative memory task performance in cocaine-addicted individuals (Vadhan et al., 2014), whilst another did not (Vadhan et al., 2008).

Since CUD patients exhibit different cognitive profiles from healthy controls, one possibility is that patients may adopt different learning strategies. In particular, although the feedback version of the weather prediction task predominantly requires non-declarative processes, it is possible to adopt other strategies to solve this task. Gluck and colleagues (2002) showed that with the feedback version, participants can either rely on single cues to guide their decision, or integrate information of multiple cues to predict the weather. These strategies are hypothesised to map onto hippocampal-based declarative (verbalised rule-based strategy) and striatal-based non-declarative (integrating cues through trial-and-error) memory respectively (Gluck et al., 2002; Shohamy, Myers, Grossman, et al., 2004). For instance, Parkinson's Disease patients were more likely to respond with single-cue simple strategies instead of multi-cue complex strategies, consistent with their cortico-striatal impairments (Shohamy, Myers, Onlaor, et al., 2004). Analyses on response strategies could potentially yield richer insights into the learning patterns associated with CUD.

The aims of this chapter are (1) to study the declarative and non-declarative memory processes in CUD patients, and (2) determine whether CUD patients use different strategies during learning compared to healthy controls. I administered two versions of the weather prediction task to CUD patients, and hypothesised that CUD patients exhibit impairments in the version that primarily relies on non-declarative memory. I further analysed the response strategies of patients during the feedback version using a modified version of Gluck et al.'s analyses. I hypothesised that poor feedback learning is linked with the use of suboptimal response

strategies during learning, and predicted that CUD patients are more likely to engage in simple but suboptimal responding strategies.

4.2 Methods

4.2.1 Sample description

Eighty-two men were recruited from the local community through flyer advertisements and word-of-mouth. Participants were included in the study if they were aged at least 18 years old and have sufficient English proficiency. Patients were required to satisfy the Diagnostic and Statistical Manual (5th edition, DSM-5) criteria for cocaine use disorder, whilst control participants had to be healthy with no prior history of substance use disorders. All participants provided informed consent upon study enrolment, and were screened with the Mini International Neuropsychiatric Inventory (MINI, Sheehan et al., 1998) and breathalysed to confirm sobriety; psychopathology in CUD patients was additionally evaluated with the Structured Clinical Interview (First et al., 2002). All participants also underwent urine screens for undeclared drug use, which confirmed active cocaine use in all CUD patients, and drug abstinence in all control participants. Acute alcohol intoxication, or a lifetime history of psychotic or neurological disorders or traumatic brain injury led to exclusion from the study. The protocol received ethical approval from the Cambridge Psychology Research Ethics Committee. The final sample included 42 male CUD patients and 40 male healthy volunteers. Patients reported using cocaine for 12.3 years on average (standard deviation [SD]: 7.5 years); all CUD patients completed the Obsessive-Compulsive Drug Use Scale as a quantitative measure of drug use severity (Franken et al., 2002). Forty healthy volunteers reported no current or past history of substance use disorder, and no heavy drug use (Table 4.1). To ascertain whether affective status affects procedural or declarative memory, participants were also administered the Depression, Anxiety and Stress Scale (DASS-21; Lovibond and Lovibond, 1995) as a measure for subclinical levels of depression, stress and anxiety.

4.2.2 Behavioural tasks

I administered two versions of the weather prediction task: the standard feedback version and the paired-associates version.

Feedback version: On each trial participants were presented a combination of one, two or three out of four available tarot cards, and were asked to predict whether "sun" or "rain" is more likely. Feedback is immediately delivered upon choice selection to inform participants of the correct answer (Figure 4.1A). The unique cards each have a constant probabilistic relationship with the weather, such that two of the four cards predicted rain and sun 87.5% of the time, whereas the other two cards predicted rain and sun 75% of the time. Thirteen possible combinations were constructed from these cards, each with varying likelihood of outcomes. We followed the probabilistic structure used within prior literature (Kemény & Lukács, 2010, 2013), which I outlined in Table 4.1. Participants have to learn by trial-and-error to accrue knowledge on the relationship between card combinations and the weather. Participants first completed 150 training trials. This was immediately followed by a 50-trial test phase, where participants were asked to predict the weather without receiving feedback on their choices. Rate of optimal choices made during learning and test phases reflect non-declarative memory performance.

Paired-associates version: This version shares an identical probabilistic structure to the feedback version. On each trial, participants were presented with a combination of cards and the weather associated with the combination (Figure 4.1B). Instead of learning by trial-and-error, participants were instructed to memorise the relationship between the card combinations and the weather. To ensure that participants were paying attention on each trial, participants needed to confirm whether the outcome displayed was sun or rain. After completing 150 memorisation trials, participants also completed 50-trial test phase, where participants predict the weather from the card combinations. Declarative memory is measured by the rate of optimal choices made during the test phase.

After the test phases for each version, I assessed participants' declarative task knowledge (Figure 4.1C). For each card, participants needed to mark on a 100mm continuous visual analogue scale (1) the likelihood of each card in predicting sun or rain (more likely to be sunny – more likely to rain) and (2) how confident they were in their estimation (not confident at all – very confident).

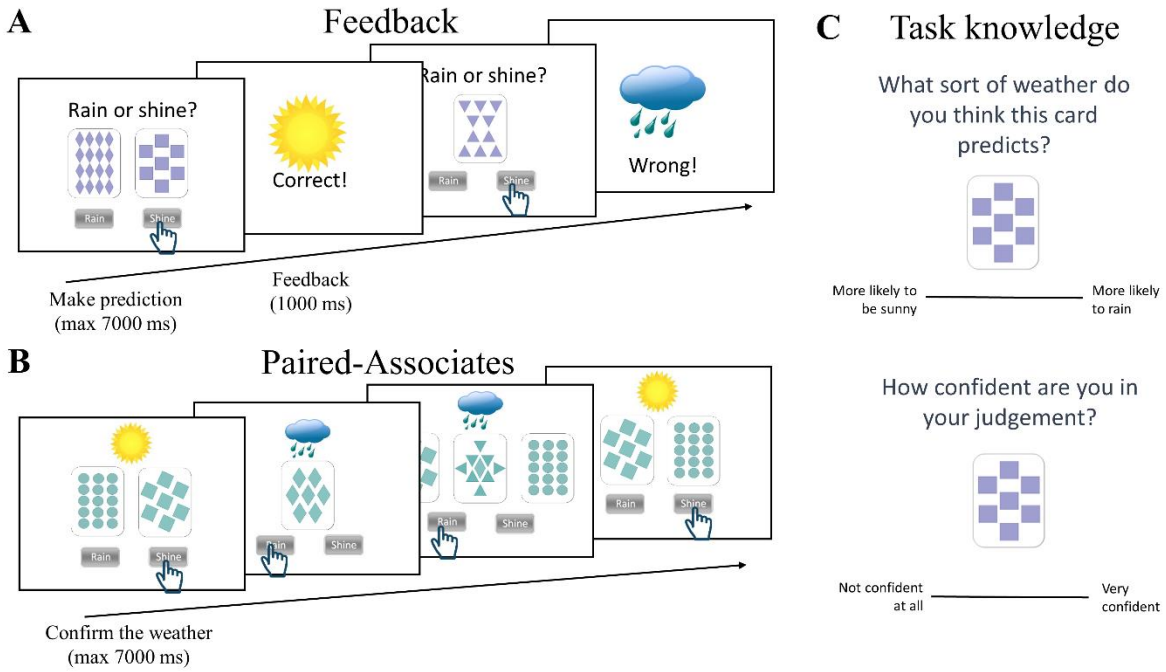


Figure 4.1: Schematics for the weather prediction tasks. (A) In the feedback version, participants need to learn by trial-and-error the optimal responses for each card combinations. Participants are shown a card combination on each trial, and must pick whether this combination predicts sun or rain; upon selecting a choice, they receive immediate feedback. (B) In the paired-associates version, participants learn by direct memorisation. On each trial, both cards and weather are presented simultaneously for participants to memorise; they need to then confirm their attention by selecting the corresponding weather. (C) After each task, participants' declarative knowledge is assessed: they were asked to rate each card on a 100 mm continuous visual analogue scale (i) how likely does each card predicts rain and (ii) how confident they are in their ratings.

Table 4.1: Probabilistic structure of the weather prediction task. Both feedback and paired-associates versions share identical probabilistic structure. For each pattern, each cue is either present (1) or absent (0). [p(sun|pattern) reflects the probability of sun being the correct response for this pattern; p(pattern) reflects how often this pattern occurs on a given trial]

Pattern	Cues present				P(pattern)	Frequency (per 150 trials)	P(sun | pattern)	Optimal response
	square	diamond	circle	triangle				
A	1	1	1	0	0.04	6	1.0	Sun
B	1	1	0	1	0.02	3	1.0	Sun
C	1	1	0	0	0.16	24	0.875	Sun
D	1	0	1	1	0.02	3	0	Rain
E	1	0	1	0	0.02	3	1.0	Sun
G	1	0	0	0	0.16	24	0.875	Sun
H	0	1	1	1	0.04	6	0	Rain
I	0	1	1	0	0.04	6	0.5	-
J	0	1	0	1	0.02	3	0	Rain
K	0	1	0	0	0.08	12	0.75	Sun
L	0	0	1	1	0.16	24	0.125	Rain
M	0	0	1	0	0.08	12	0.25	Rain
N	0	0	0	1	0.16	24	0.125	Rain

4.2.3 Statistical analysis

Demographics and performance data were analysed using analysis of variance (ANOVA). The primary performance measure for both versions is the overall rate of optimal choices during the test phase, defined as the selection of the outcome most associated with the card combinations ("Optimal Response" column in Table 4.1). For the learning phase in the feedback version, I also determined the rate of optimal choices in 50-trial blocks.

As a measure of declarative task knowledge, I computed likelihood estimation errors for both feedback and paired-associates versions. This is defined as the absolute deviation from the actual probabilities of each card (larger = less precise). For instance, if the square card has an 87.5% chance of sun, and the participant indicates that it has a 100% chance of sun, then the likelihood estimation error for that card is 12.5 (estimate minus actual likelihood). This process was repeated for each card, and a total score was computed as the average across all four cards – higher error estimates indicates lower precision. In addition to a total score, I also calculated separately the average error estimates for cards with 87.5% likelihood and cards with 75% likelihood. The same was done for confidence ratings: a total confidence rating (averaged across

four cards), as well as confidence ratings for strong and weak predictor cards were calculated; a larger value reflects more confidence in their ratings.

Relationships between declarative knowledge and task performance were explored using Spearman's correlation analyses. Rate of optimal choices during the test phases, likelihood estimation errors and self-rated confidence were log-transformed to reduce skew, but raw untransformed values were presented in figures. Any post-hoc pairwise comparisons were adjusted for multiple comparison using the Bonferroni's method. A sensitivity power analysis identified that, given the current sample, this study is sufficiently sensitive to detect a moderate effect size (Cohen's d = 0.54).

4.2.4 Strategy analysis

Different people use different strategies to solve the weather prediction task. Gluck and colleagues (Gluck et al., 2002) conceptualised three distinctive strategies: (1) singleton strategy, where participants respond definitively when only one card is present (e.g. respond "rain" when *only* the triangle card is present [pattern N]), and respond randomly when more than one card is present; (2) one-cue strategy, where responding is based on the presence or absence of a single card; and (3) multi-cue strategy, where responding takes into account the whole combination instead of single cards. Gluck and colleagues constructed for each strategy an ideal data, which were defined as the expected responses if the participant fully adhered to that response strategy. For example, if a participant reliably followed a one-cue strategy, they should have responded with 'sun' whenever a specific card (e.g. square) was present in the combination, and 'rain' if the card was absent. They then compared actual task performance with this ideal data with a least mean square method to produce a model fit score. The calculation of this score is as follows:

$$Model\ fit\ score = \frac{\sum_p (\#sun_{ideal} - \#sun_{actual})^2}{\sum_p (\#pattern)^2}$$

where p refers to the pattern shown, #sun_ideal denotes ideal response given pattern p, #sun_actual denotes actual response by participant, and #pattern presented refers to number of times pattern p was presented. This model fit score ranges from 0 to 1, with 0 indicating perfect

score i.e. participant fully adhered to that strategy. This model fit score is computed for each strategy; the strategy with the lowest model fit score is deemed the dominant strategy. Any model with a model fit score of more than 0.1 is considered as a 'non-identifiable' strategy because it means the response data does not fit well to that strategy (Gluck et al., 2002). For simplicity, these strategies have been grouped as simple (singleton and one-cue) and complex (multi-cue) strategies within prior literature (Schwabe & Wolf, 2012; Thomas & LaBar, 2008).

However, a potential shortcoming of this simple method is that it does not account for individual differences in learning. This is an important consideration factor as prior work has shown substantial differences in learning between healthy volunteers and addicted patients (Kanen et al., 2019; see also Chapter 3). Neglecting these variances could under-estimate the goodness-of-fit for each strategy in CUD patients by increasing the percentage of non-identifiable strategy (see Appendix C, Figure C1).

Hence, I modified the strategy analyses in three ways: (1) I constructed learning models that correspond to simple (single-cue based) or complex (multi-cue based) strategies, which broadly follows a reinforcement learning algorithm; (2) I included basic free parameters of learning (learning rate and reinforcement sensitivity) during the modelling process; (3) I compared model fit using bridge sampling procedures. A side-by-side comparison between my strategy analyses and that with the Gluck et al method is presented in the Appendix C (Figure C1).

1) Learning model: Consistent with prior literature, I used a simple learning model and a complex learning model to reflect response strategies (Gluck et al., 2002; Schwabe & Wolf, 2012; Thomas & LaBar, 2008). As singleton and one-cue strategies are rule-based strategies that are easily verbalised, they are treated as simple strategies (Gluck et al., 2002; Shohamy et al., 2008). By contrast, developing a multi-cue strategy requires the integration of knowledge over many trials, and is thought to be implicit (Shohamy et al., 2008), so it is treated as a complex strategy. Both simple and complex strategies are updated using delta rule:

$$V(s,a)_{t+1} = V(s,a)_t + \alpha(R - V_t)$$

$$\alpha \begin{cases} \alpha_{pos} \; when \; R = 1 \\ \alpha_{neg} \; when \; R = -1 \end{cases}$$

where *V(s,a)* reflects value of action *a* given state *s* on trial *t*, α reflects learning rate (itself fractionated according to feedback) and *R* denotes reinforcement. Actual choice selection is computed following a softmax rule:

$$p(a|s) = \frac{exp(\beta V_t^a)}{\sum_{k=1}^{n} exp(\beta V_t^k)}$$

where *p(a|s)* is the probability of making choice *a* given state *s*, and β is the inverse temperature parameter that governs sensitivity to action values.

An important distinction between simple and complex models is how card-weather relationships are learnt. In the simple model, I constructed rule-based learner models, where learning is based on the presence or absence of a single cue (Gluck et al., 2002). In other words, the choice values are updated based on only two possible state spaces: (1) when card is present, and (2) when card is absent (s = {shape absent, shape present}). The simple model was constructed for each of the four unique cards (square, diamond, circle, triangle). By contrast, in the complex model, I assumed that participants develop unique S-R contingencies towards each unique pattern. Thus, delta rule updates choice values for each of the 13 unique combinations as reported in Table 4.1 (s = {pattern 1, pattern 2, … pattern 13}).

2) Individual differences during learning: Individual differences are accounted for by modelling the learning rates, α, and the inverse temperature, β, in above mentioned equations. The fitting process for trial-by-trial data was conducted with RStan (Carpenter et al., 2017). In both models, I allowed free parameters such as the α and β to vary from each participant to account for the individual variation. The parameters α and β had priors of beta (shape 1=1.1, shape 2=1.1) and gamma (shape=4.82, rate=0.84) distributions respectively. It is noteworthy that the free parameters here only serve the purpose of accounting for individual differences in learning with each strategy, and are not of interest in this chapter. Instead, I am interested in the dominant strategy used during learning.

3) Model comparison for strategies: The primary outcome measure was the dominant strategy used during the learning and test phases of the feedback condition. The dominant strategy is the

winning model (i.e. log marginal likelihood closest to 0) as identified by bridge sampling model selection procedure described in the model selection section in Chapter 2 (Gronau et al., 2017). During model comparison, I also compared the performance of the two models against a random choice model, where selection of rain or shine is equally likely regardless of the cues presented (chance performance). Any task performance during learning for which the random choice model was the winning model suggests "guessing" behaviour, and was excluded from any subsequent analysis. I modelled strategy only for the feedback version, because the paired-associates version did not involve any feedback-based learning or overt behaviour, so the learning process cannot be modelled. To assess the reliability of the model selection process and the specificity of the models, I performed model recovery checks to assess whether a simulated response based on a specific strategy can be recovered by the model selection process (Wilson & Collins, 2019). I found that model recovery for these strategies was extremely well, suggesting that we can clearly arbitrate between these responding strategies using the current method (see Appendix C, Figure C3).

4.3 Results

4.3.1 Demographics and clinical data

Sample characteristics are reported in Table 4.2. Both groups were matched on age and gender, but the cocaine group had significantly lower verbal IQ than that of the control group. However, as verbal IQ scores were not correlated with any task performance measures in either group (all $p > 0.05$; see Table 4.3), verbal IQ was not statistically controlled for in the analyses. Consistent with prior findings, patients also demonstrated significantly higher levels of impulsivity and reduced propensity for goal pursuit. However, neither demographics nor clinical measures within CUD patients correlated with task performance (all $p > 0.1$)

Table 4.2: **Sample demographics for Chapter 4.**

Demographics	**Group**		**Group comparison**	
	Control	Cocaine	*t*	p-value
Sample size	40	42	-	-
Age (years)	40.1 (12.4)	39.3 (8.8)	0.675	.502
Gender (% male)	100	100	-	-
Verbal IQ (NART score)	115 (6.1)	103 (7.4)	7.73	< .001
Compulsive drug use (OCDUS score)	-	33.7 (9.3)	-	-
Drug use (DAST-20 score)	0.08 (0.3)	-	-	-
Alcohol use (AUDIT score)	3.5 (1.7)	4.1 (5.7)	-0.69	.492
Depression (DASS-21 subscale)	3 (3.3)	17 (11.4)	-7.72	< .001
Anxiety (DASS-21 subscale)	2 (2.4)	12 (7.0)	-9.30	< .001
Stress (DASS-21 subscale)	6 (4.9)	17 (8.7)	-7.54	< .001

Note. NART: National Adult's Reading Test; BIS-11: Barratt Impulsiveness Scale; OCDUS: Obsessive-Compulsive Drug Use Scale; HSCQ: Habitual self-control questionnaire; DAST-20: Drug and Alcohol Screening Test; AUDIT: Alcohol Use Disorder Identification Test; DASS-21: Depression, Anxiety, and Stress Scale; standard deviation reported in parentheses.

Table 4.3: Summary scores for weather prediction tasks performance measures and their associations with verbal IQ.

Task performance	**Mean (SD)**		**Pearson's r with verbal IQ (p)**	
	Control	Cocaine	Control	Cocaine
Feedback				
learning phase (% correct)	84.6 (6.9)	67.2 (11)	0.090 (0.597)	0.100 (0.557)
test phase (% correct)	93.1 (8.9)	71.3 (21)	0.107 (0.528)	0.115 (0.499)
Likelihood estimation errors (%)	12.9 (11)	24.5 (16)	0.069 (0.687)	-0.136 (0.421)
Confidence rating (%)	77.2 (9.8)	61.7 (20)	0.244 (0.145)	-0.019 (0.910)
Paired-Associates				
test phase (% correct)	91.5 (12)	70.1 (22)	0.205 (0.224)	0.285 (0.087)
Likelihood estimation errors (%)	11.3 (7.3)	29.7 (17)	-0.067 (0.694)	-0.268 (0.109)
Confidence ratings (%)	70.2 (18)	67.0 (18)	-0.047 (0.782)	0.240 (0.153)

Note. SD: standard deviation; Verbal IQ measured by the total score on the National Adult's Reading Test.

4.3.2 Task performance and knowledge

Behavioural summary scores are reported in Table 4.3. In the training phase of the feedback version (Figure 4.2A), although there was a main effect of block ($F_{2,160}$=18.3, p<0.001) that suggests overall training performance improved over time, the cocaine group performed significantly worse than the control group ($F_{1,80}$=70.1, p<0.001). There was no group-by-block interaction ($F_{2,160}$=1.1, p=0.345).

To assess participants' performance on the test phase (Figure 4.2B), I analysed the rate of optimal choices during test phase, with condition (Feedback versus Paired-Associates) as a within-subject factor and group (Control versus Cocaine) as a between-subject factor. Analyses revealed that cocaine-addicted patients performed significantly worse than healthy controls in both versions ($F_{1,80}$=78.3, p<0.001). However, the effects for condition ($F_{1,80}$=0.629, p=0.430) and group-by-condition interaction ($F_{1,80}$=0.14, p=0.709) were not statistically significant (Figure 4.2B).

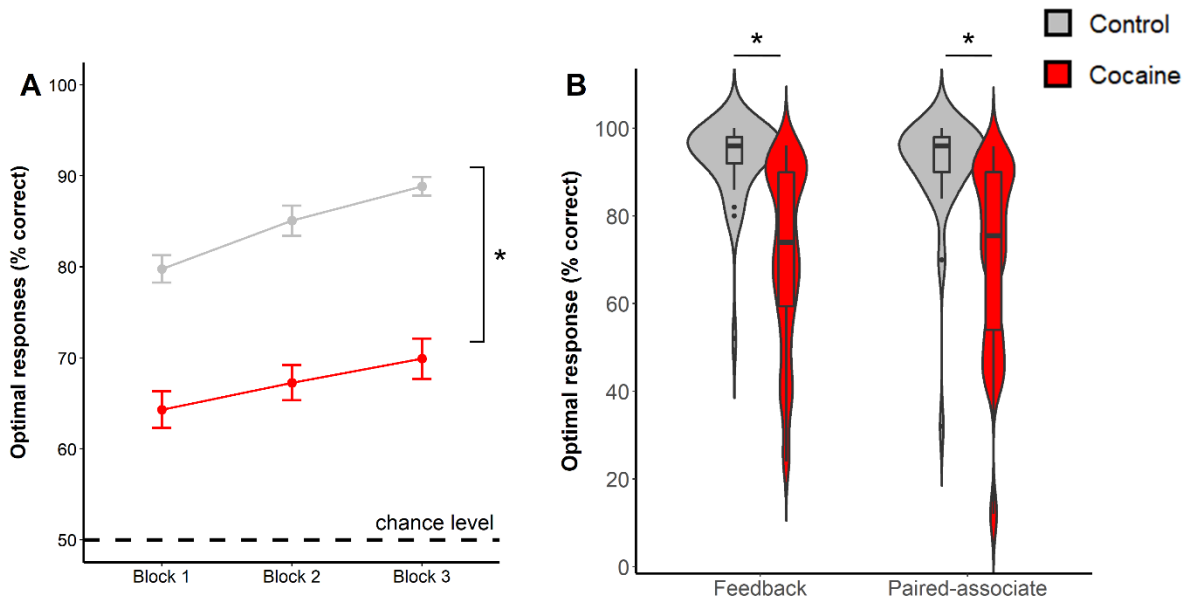


Figure 4.2: Task performance for the weather prediction tasks. (A) During the learning phase of the feedback condition, all participants steadily improved their task performance over time, but CUD patients' performance consistently fell behind that of control participants. (B) Analyses on the rate of optimal responses during the test phase revealed that CUD patients performed significantly worse than control participants, irrespective of task version. [* indicates statistical significance at p < 0.05]

I analysed explicit task knowledge as indexed by the likelihood estimation error (Figure 4.3A). A mixed ANOVA model with condition (Feedback versus Paired-Associates), card probability (87.5% versus 75%) and group (Control versus Cocaine) as factors found that patients made larger errors when estimating the likelihood than controls ($F_{1,80}$=37, p<.001), suggesting that overall declarative knowledge is weaker in patients, irrespective of the task condition. There was a main effect of card probability ($F_{1,80}$=19.5, p<.001), such that participants estimated likelihoods better for the 87.5% cards, but there was a condition-by-card probability interaction effect ($F_{1,80}$=9.27, p=.003). Post-hoc pairwise comparisons revealed that all participants made more precise estimations for the 87.5% cards than the 75% cards in the feedback condition (p<.001). However, this effect was absent in the paired-associates condition (p=.186). Other effects did not reach statistical significance (all p > 0.1). Likelihood estimation error was correlated with test phase performance on both feedback (ρ=-0.683, p<.001) and paired-associates version (ρ=-0.738, p<.001). To determine whether declarative knowledge alone can account for task performance, I re-analysed test phase performance with group as a between-subject factor, and included declarative knowledge as a covariate in an ANCOVA model. Analyses revealed no group differences in paired-associates test performance ($F_{1,79}$=1.6, p=0.208), but the group difference remains significant in the feedback phase condition ($F_{1,79}$=18.2, p<0.001) – confirming that the paired-associates condition strongly depends on declarative knowledge, but the feedback condition does not.

Analyses on the confidence ratings with a similar ANOVA model found that overall, CUD patients were less confident than controls of their knowledge ($F_{1,80}$=7.66, p=0.007) (Figure 4.3B). However, there was a condition-by-group interaction ($F_{1,80}$=7.1, p=0.009). Post-hoc analyses revealed that the groups differ in their confidence only in the feedback condition (p<0.001), but not the paired-associates condition (p=.471). There was a main effect of card probability ($F_{1,80}$=5.66, p=.020), in which all participants were more confident on their estimates for the 87.5% cards compared with the 75% cards. Additionally, I found a card probability-by-group interaction ($F_{1,80}$=4.85, p=.031), such that control participants were more confident than CUD patients on their estimates for the 87.5% predictive cards (p=.002), but were equally confident with patients for the 75% predictive cards (p=.210). There was also a condition-by-card-probability interaction ($F_{1,80}$=5.38, p =.023), which showed that the difference in confidence for the 87.5% and 75% predictive cards was significant in the feedback (p=.002), but not the paired-associates condition (p=.518). All other effects were not

statistically significant (p>0.1). A positive relationship was observed between mean knowledge confidence and task performance in both feedback (ρ=.583, p<.001) and paired-associates version (ρ=.289, p=.008) of the task. However, the group differences remained significant after statistically controlling for confidence levels in both feedback ($F_{1,79}$=16.8, p<0.001) and paired-associate versions ($F_{1,79}$=28.4, p<0.001), suggesting that confidence levels did not modulate task performance.

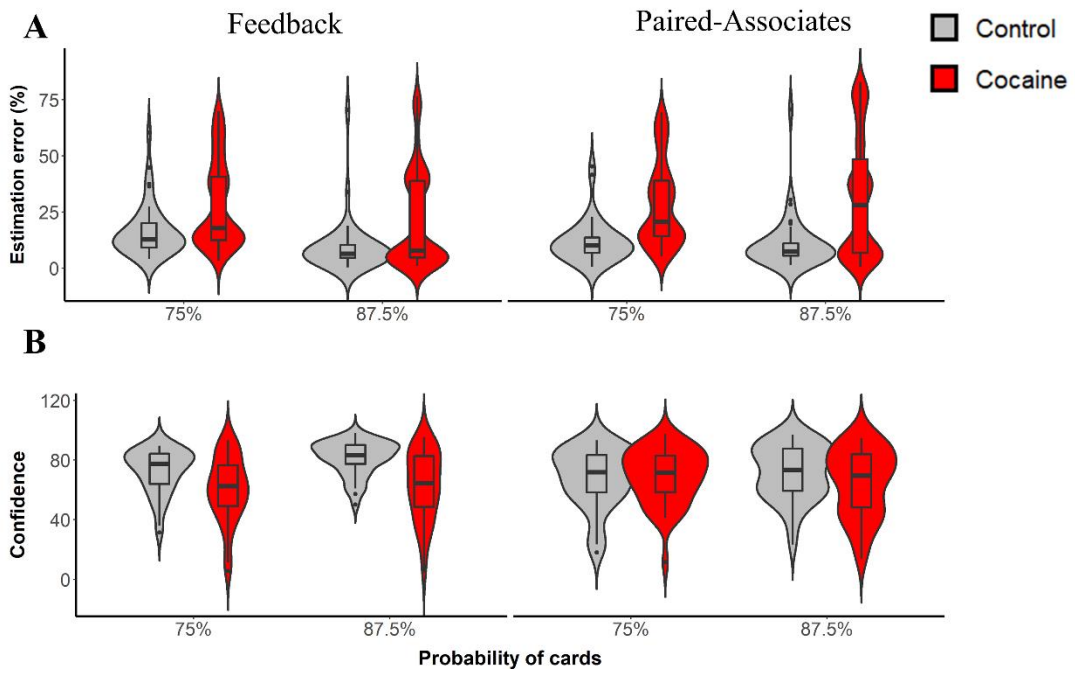


Figure 4.3: Post-task measurements of declarative knowledge. (A) Likelihood estimation error, categorised by task version, group and the card probability. (B) Confidence on ratings, categorised by task version, group and the card probability.

4.3.3 Strategy analysis

During the training phase of the feedback version, the random model was the winning model for three cocaine-addicted patients (7%), suggesting that they did not develop any strategy during the course of learning. These data were excluded from any further analyses in this section. Figure 4.4A shows the proportion of participants for each dominant strategy during learning. I identified a significant association between group and dominant strategy (Fisher's exact = 19.9, p<.001), such that 37 healthy controls (93%), as opposed to only 20 patients (51%), used a complex strategy. The remaining controls (n=3, 7%) and patients (n=19, 49%) adopted the simple strategy. It is noteworthy that analyses of block-by-block strategy during learning found

that each group consistently maintained these strategies throughout learning (see Appendix C, Figure C2). Additionally, strategy analysis on the test phase (Figure 4.4B) revealed almost identical patterns to the training phase, suggesting that most participants maintained their strategy adopted during test phase. Again, for most patients' responding, the simple strategy was the winning model (n=21, 54%), while in the majority of healthy controls', the complex strategy was the winning model (n=34, 85%) – this group difference was also statistically significant (Fisher's exact = 17.7, p<.001).

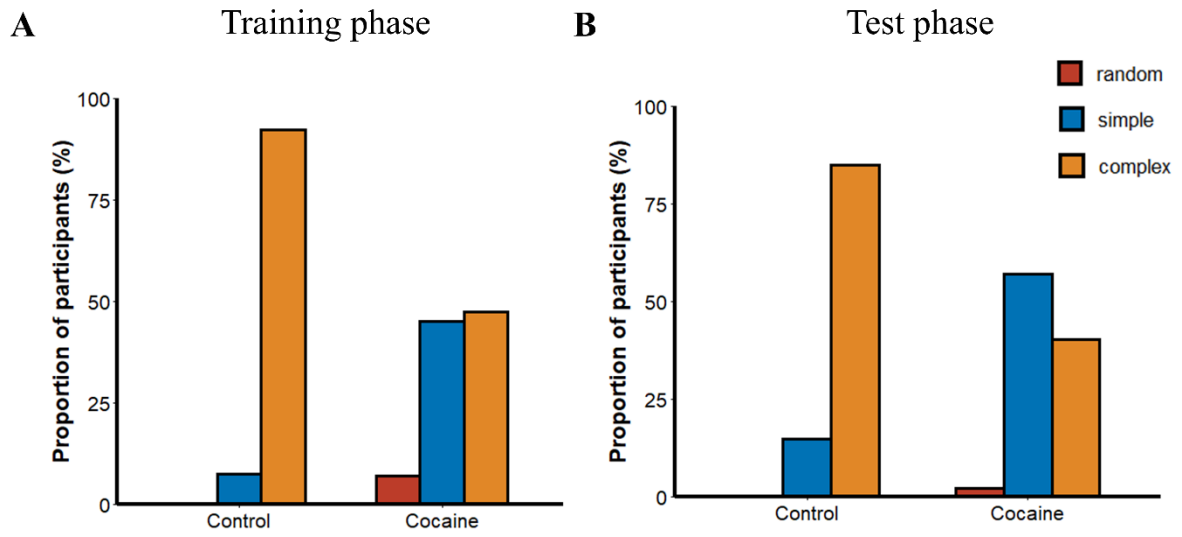


Figure 4.4: Dominant response strategies in the feedback version of the weather prediction task. (A) Strategy use during feedback learning significantly differed across groups; whilst controls mostly adopted a more complex, multi-cue strategy, almost half of CUD patients engaged in more simple memorisation strategy while learning the task. (B) Participants largely retained their strategy use during the test phase, with most controls adopting a complex strategy, whereas CUD patients were more likely to use simple strategy.

As an added exploratory analyses, I wanted to identify whether any demographic or clinical measures differed between those CUD patients who used a simple versus a complex response strategy. A multivariate ANOVA model with the between-subject factor, strategy (simple versus complex), and age, verbal IQ, compulsive drug use, and affective measures (DASS-21 subscores) as dependent variables did not find a main effect of strategy on any of these measures (all p > 0.05). This suggests that different strategy users do not differ significantly on any of the collected demographic and clinical measures.

4.4 Discussion

I sought to test declarative and non-declarative learning systems in CUD patients with two versions of the weather prediction task, each representing measures of learning from feedback or from explicit memorisation. The findings showed that regardless of the task version, CUD patients were impaired in their learning performance. Closer examination of participants' response strategy during feedback learning revealed that while most controls adopted a more optimal complex (multi-cue) strategy that is thought to reflect striatal-dependent learning, half of cocaine-addicted patients relied on the simple (single-cue) strategy, which suggests a reliance on a suboptimal strategy. Together these findings suggest that learning deficits in CUD may extend beyond feedback learning, possibly due to their use of suboptimal strategies.

4.4.1 CUD is linked with impaired declarative and non-declarative memory

In this study, the weather prediction task was manipulated such that categories were learnt either from corrective feedback or explicit memorisation. Learning from feedback is known to be impaired in cocaine addiction, so the former findings are in line with extant literature. This particular impairment was especially highlighted in Chapter 3, where stimulant-dependent individuals showed reduced learning from negative feedback. Chronic cocaine use is known to interfere with normal reward prediction error signalling (A. C. Burton et al., 2018; Saddoris et al., 2017), a reinforcement signal that "stamps-in" learned contingencies (Wise, 2004). This is also consistent with reports in human drug users who show blunted striatal prediction errors (Parvaz et al., 2015; Tanabe et al., 2013). One study has provided evidence for dopamine release during learning of the feedback version, but not the paired associates version, of the weather prediction task (Wilkinson et al., 2014), implying that dopamine is critical for feedback learning. Thus, it is logical to expect CUD, associated with dopaminergic dysfunction, to show reduced behavioural performance during feedback learning.

Reduced performance of the paired-associates version suggests that cocaine-addicted patients were also impaired in declarative memory. The declarative memory system, which is thought to be supported by the medial temporal lobe (including hippocampus), mediates acquisition of contextual information, most notably in encoding contingencies among environmental stimuli (Boorman et al., 2016; Eichenbaum et al., 1994; Garvert et al., 2017). Indeed, patients with medial temporal lobe damage are impaired in contingency learning (Bradfield et al., 2020; Palombo et al., 2019; Vilà-Balló et al., 2017). The paired-associates task is thought to simulate the learning of such contingencies (i.e. cue-weather relationship) (Poldrack et al., 2001). Since

cocaine exposure affects hippocampal-dependent declarative memory encoding (Sudai et al., 2011; Yamaguchi et al., 2004), it is conceivable that medial temporal lobe function is impaired in cocaine addiction. Further corroborating evidence can be found in prior studies which demonstrated inferior performance of stimulant-addicted patients on learning tasks that are sensitive to medial temporal lobe function (Ersche et al., 2006; van Gorp et al., 1999). However, this is speculative as I did not have evidence for medial temporal lobe impairments in this study.

Additional evidence for a weakened declarative system was also provided by patients' reduced task knowledge, as reflected by the lower precision than healthy controls while estimating the predictive values of each cue. However, whilst reduced precision in estimating the cue-weather probabilities has been interpreted as reduced task knowledge in previous studies (Lagnado et al., 2006; Price, 2005), I cannot discount the possibility of impaired probability encoding in cocaine addiction. A plethora of studies using probabilistic learning paradigms have identified impairments in cocaine-addicted subjects, but probability encoding has never been explicitly explored. It was suggested that dopamine codes for degrees of uncertainty (Fiorillo et al., 2003), and dopaminergic areas such as the anterior cingulate cortex track and update the predictive value of reward in humans (Behrens et al., 2007). Again, dopaminergic dysfunction linked with cocaine addiction also alludes to the possibility of aberrant probability encoding in cocaine-addicted patients, reflected in the estimation of cue-weather probabilities. However, this is only speculative, and future studies are needed to confirm this hypothesis.

4.4.2 CUD patients use suboptimal strategies during category learning

Half of CUD patients rely on cue-based simple declarative strategy when learning the weather prediction task. Prior data from healthy volunteers showed that while completing the weather prediction task, participants were more inclined to use simple memorisation strategies during initial stages of learning, and gradually transitioned to the complex multi-cue strategy during late learning (Gluck et al., 2002). This is also consistent with the notion that declarative knowledge was available to participants during early learning of the weather prediction task, as it is easily and rapidly acquired (Knowlton et al., 2017; Poldrack et al., 2001). Interestingly, unlike the profile reported in Gluck et al (2002), my analyses on block-by-block response strategy during learning revealed that both controls and CUD patients are consistent in their strategy use throughout the task (Appendix C, Figure C2). This lack of transition is likely due to the differences in the probabilistic relationships: the probability used in this study (strong predictor = 87.5% versus weak predictor = 75%) is much easier than the one used in (Gluck et

al., 2002) (strong predictor = 75.6% versus weak predictor = 57.5%), which may have facilitated quicker transitions to the more complex strategy during early learning. But what is clear, is that most control participants adopted the more complex multi-cue strategy during learning. By contrast, only half of the patient sample used the complex strategy, suggesting that there might be differences in how patients engage this task. It is noteworthy that I have explored whether any demographic traits might predict the use of simple strategy within cocaine-addicted patients during learning, but did not find any significant differences. Thus, the current data do not provide any clear indications why some patients adopt a simpler strategy.

I speculate that the lack of switching to the complex strategy seen in some cocaine-addicted patients may be due to their inability to integrate cue-weather information during the course of learning. Such integration is reminiscent of the striatum's function in combining well-learned motor or cognitive action sequences to allow a more efficient expression, a process known as "chunking" (Graybiel, 1998; Graybiel & Grafton, 2015). In this vein, it is plausible that the use of the complex strategy, characterised by the integration of multiple cue-weather relationships, is mediated by the striatum. Indeed, inactivation of dorsal striatum with a dopamine D_2 antagonist abolishes chunking in rats (Levesque et al., 2007), suggesting a direct role of striatal dopamine in the formation of chunks. Thus, the lack of switch from single-cue to multi-cue strategy might reflect a deficiency in chunking. This deficit in chunking is also observed in Parkinson's disease patients (Tremblay et al., 2010), mirroring the finding that these patients never switched to a more complex strategy despite extensive training (Shohamy, Myers, Onlaor, et al., 2004).

4.4.3 Limitations and conclusion

The current findings should be interpreted in light of several limitations. First, although the weather prediction task has well-established neural substrates, the lack of neuroimaging methods precludes any conclusive inferences drawn about the neural substrates that underpin cocaine-addicted patients' strategy use and task performance, and should be cautiously interpreted. Moreover, there is little evidence for the neural circuits underpinning different strategy use within the literature, so whether relying on simple strategies reflect a compensatory response for poor feedback learning in patients is unclear. Furthermore, as neither task performance nor task knowledge correlated with clinical measures, it remains an open question whether category learning deficits observed is a by-product or a predisposing risk factor of

addiction. Future longitudinal research should investigate whether these category learning deficits predate the development of maladaptive drug-seeking behaviour.

Notwithstanding these limitations, this chapter provided evidence for impaired feedback and observational learning in cocaine addiction using a well-established category learning task. These findings add to the growing body of evidence for aberrant goal-directed learning in cocaine addiction.

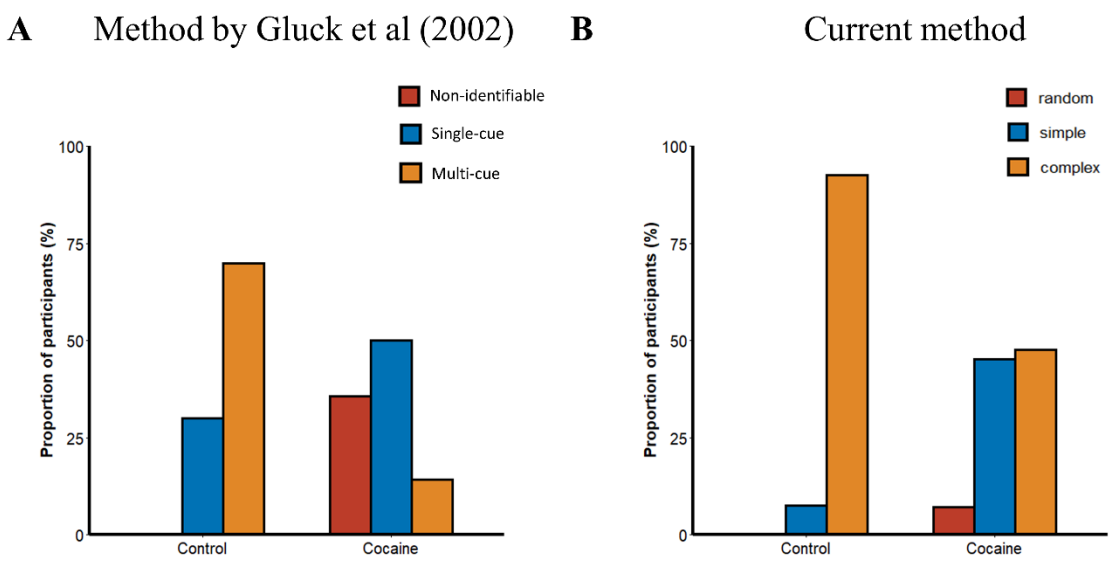


Figure C1: **Comparison of strategy analyses** between (A) original method by Gluck and colleagues, and (B) current method that accounted for individual differences for each strategy. The method by Gluck et al gave rise to a significant minority of non-identifiable strategies, which was reduced greatly when analysed after accounting for individual differences in learning.

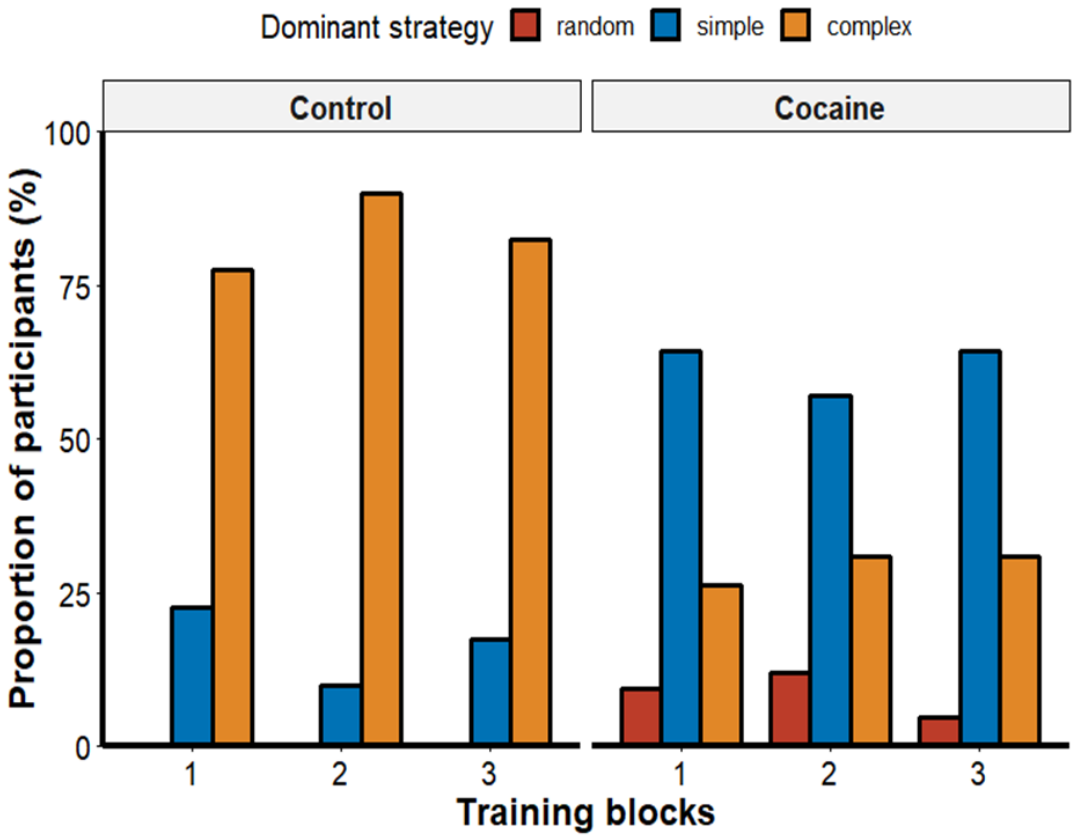


Figure C2: **Dominant strategy for each 50-trial block during feedback learning.** Control participants consistently engaged the task with a complex strategy, as opposed to cocaine use disorder patients, who mostly use a simple rule-based strategy during learning.

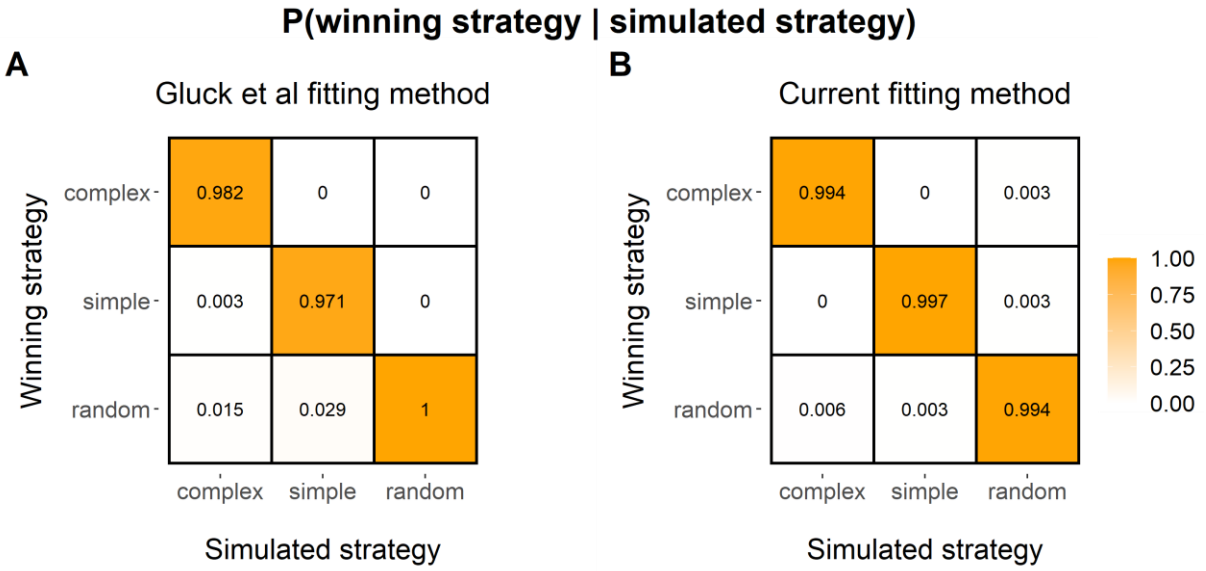


Figure C3: Confusion matrices of strategy modelling for the Weather Prediction Task as indices of model recovery. I simulated 1000 responding patterns based of the three strategies as outlined in the chapter. I then fitted these responses to each strategy and identified the winning model with two methods: (A) the original fitting and model selection methods as suggested by Gluck et al (2002) [N.B. using the terminology in Gluck et al (2002), complex, simple, and random strategies refer to multi-cue, one-cue and non-identifiable strategies respectively.]; (B) the current method as reported in the main text i.e. by accounting for individual differences in the learning parameters (randomly generated from the priors reported in section 4.2.4) during model fitting, and using *bridgesampling* to identify the winning model. The numbers in the matrices denote the probability of recovering the strategy from a simulated strategy – higher indicates better recovery. Evident from the diagonals of both matrices, both methods showed excellent recoverability. This suggests that we can arbitrate between these strategies extremely well.

Chapter 5: The relationship between reinforcement learning and habit formation in cocaine use disorder

This chapter has been published as:
Lim, T. V., Cardinal, R. N., Savulich, G., Jones, P. S., Moustafa, A. A., Robbins, T. W., & Ersche, K. D. (2019). Impairments in reinforcement learning do not explain enhanced habit formation in cocaine use disorder. *Psychopharmacology*, *236*(8), 2359–2371. https://doi.org/10.1007/s00213-019-05330-z

Pertaining to this work, I would like to declare that:

Mr P.S. Jones analysed the neuroimaging data and contributed to the following:

- 5.2.4 Neuroimaging data
- Figures 5.3A and 5.3B

Dr R.N. Cardinal performed the confirmatory modelling analyses and contributed to the following:

- 5.2.3.5 Confirmatory modelling of goal-directed action and habitual responding
- Appendix D: General two-system computational model of goal-directed and habitual responding

5.1 Introduction

Cocaine addiction is a global health problem that contributes to major economic and health burdens and is difficult to treat (Degenhardt et al., 2014). Although the initial positive reinforcing effects of cocaine are mediated by dopaminergic neurotransmission in the mesolimbic dopaminergic system, subsequent drug-seeking is guided by conditioning processes in a wider neural network (Everitt & Robbins, 2005). Instrumental learning paradigms have provided a theoretical framework of impaired behavioural control for drug addiction (Everitt & Robbins, 2005, 2016), as well as other psychiatric disorders (Heinz et al., 2016; Robbins et al., 2012). Instrumental learning is thought to be regulated by two distinct systems, namely the goal-directed and habit systems (Adams & Dickinson, 1981). The goal-directed system, which is subserved by frontostriatal regions (de Wit et al., 2009; Tanaka et al., 2008; Valentin et al., 2007), controls voluntary instrumental behaviour by evaluating the potential consequences of actions. The habit system, which is subserved by corticostriatal circuits (Brovelli et al., 2011; de Wit et al., 2012; Tricomi et al., 2009; Zwosta et al., 2018), regulates automatic impulses in response to stimulus-response associations that have been formed over repeated experiences. Both systems are needed in everyday life, and optimal

behavioural performance has been shown to require a balance between the joint regulation of these two systems (Balleine & O'Doherty, 2010). A growing body of literature suggests that drug addiction develops through drug-induced disruption in corticostriatal subsystems that underlie these learning processes (Belin & Everitt, 2008; Corbit, Chieng, et al., 2014; Gourley et al., 2013; A. Nelson & Killcross, 2006). In most cases, drug-taking is initiated in a recreational setting and used in a goal-directed manner to experience pleasure. However, prolonged drug use in the same context may become habitual. As such, the initiation of drug-taking becomes triggered by environmental cues, irrespective of whether the experience of the drug is pleasurable (Miles et al., 2003; L. Vanderschuren & Everitt, 2004). At the final stage of addiction, drug-taking habits predominate and may even continue in spite of harmful consequences (Everitt & Robbins, 2005, 2016). It has been suggested that when habits spiral out of control, drug seeking is characterized by a failure to revert control towards the goal-directed system when the situational demands require it and becomes compulsive (Ersche et al., 2012).

A classic task to assess the balance between goal-directed and habit learning is the Slip-of-Action task (de Wit et al., 2007), which is based on an outcome devaluation paradigm to model the transition between behaviours that are initiated when obtaining reward and responses to a previously learned stimulus-response association. The extent to which participants maintain their previously learned behaviour despite outcome devaluation is considered an index of habit. Chronic cocaine and alcohol users (Ersche et al., 2016; Sjoerds et al., 2013), but not chronic tobacco smokers (Luijten et al., 2019), have been shown to develop a predominance of habits on this task, but the nature of their bias remains unclear. It has been hypothesised that either difficulties with goal-directed learning facilitate the transition of control from the goal-directed toward the habit system, or an *augmented* control by the habit system results in habit predominance (Robbins & Costa, 2017; Vandaele & Janak, 2018). Whilst the bulk of prior work has focused on cocaine's influence on the transition of control from the goal-directed to the habit system, less attention has been given to its influence on goal-directed learning.

Reinforcement learning algorithms implement learning and action selection in response to motivationally relevant reinforcement (Russell & Norvig, 1995; Sutton & Barto, 1998). Basic parameters in a typical reinforcement learning model are learning rate (α) and reinforcement sensitivity (also known as choice inverse temperature, β). *Learning rates* modulate the extent to which information is learnt, with higher rates indicating that feedback is integrated more

rapidly in order to inform future choices. *Reinforcement sensitivity* regulates the influence of associative strength during action selection, with higher sensitivity reflecting a greater impact of action values on choices. Such reinforcement learning models can be fitted to the observed behaviour, yielding estimates of the model's parameters, and different models can be compared, allowing learning to be investigated in a hypothesis-driven manner (Daw, 2011). One additional parameter relevant to drug addiction is the tendency for *perseverative responding* (sometimes termed 'stickiness'). As chronic cocaine use has been associated with profound reversal learning deficits in both animals and humans exposed to cocaine (Calu et al., 2007; Ersche et al., 2008; Ersche, Roiser, Abbott, et al., 2011; Schoenbaum et al., 2004), it is possible that inflexible contingency evaluations may also contribute to their learning deficits.

In the present study, I apply an hierarchical Bayesian approach to previously published data using the Slip-of-Action task in both healthy volunteers and patients with cocaine use disorder (CUD) (Ersche et al., 2016). I hypothesise that overall poor learning performance in CUD patients can be explained by abnormalities in at least one of the following parameters: learning rate, reinforcement sensitivity, perseveration and extinction. The latter parameter, extinction, was included in the model in light of its relevance for subsequent habit learning. *Extinction* describes the ability to learn from non-rewarding events. Given that habit formation has also been described in terms of behavioural autonomy (Dickinson, 1985), it is conceivable that habits form more easily in individuals who are resistant to extinction. Previously, it has been shown that individual differences in corticostriatal structural connectivity accounted for the balance between goal-directed and habitual behaviours in an outcome devaluation task (de Wit et al., 2012). This suggests a candidate neurobiological substrate that may explain a habit bias in CUD. Therefore, as an adjunct to the behavioural analyses, I also investigated, in a subset of participants, the white matter integrity of the two networks investigated by de Wit et al (2012): the anterior caudate – medial orbitofrontal cortex and the premotor cortex – posterior putamen; these networks are thought to underpin goal-directed and habitual behaviours respectively (de Wit et al., 2012). It was hypothesised that CUD patients have abnormalities in the structural connectivity of these systems. I further predict that white matter integrity of the goal-directed system is required for successful action-outcome learning and that deficiencies would facilitate the formation of habitual responding.

5.2 Methods

5.2.1 Sample description

Fifty-five healthy control volunteers (94.3% male) and 70 patients with CUD (90.3% male) were recruited for the study. Full details of the sample can be found elsewhere (Ersche et al., 2016). All CUD patients were recruited from the local community and satisfied the DSM-IV criteria for cocaine-dependence (American Psychiatric Association, 2013). Forty-eight CUD patients also met DSM-IV criteria for opiate dependence, 25 for cannabis dependence and five for alcohol dependence. Twenty-six CUD patients were prescribed methadone (mean dose 48.7ml, SD ± 18.0) and 14 were prescribed buprenorphine (mean dose 7.2ml, SD±4.8). Although significantly more CUD patients (94%) reported smoking tobacco compared with control volunteers (11%) (Fisher's $p < 0.001$), nicotine dependence was not assessed using the DSM-IV criteria. CUD patients had been using cocaine for an average of 16 years (7.7±SD) and were at the time of the study all active users of the drug, as verified by urine screen. Two CUD patients were excluded due to incomplete data sets. Healthy control volunteers were partly recruited by advertisement and partly from the BioResource volunteer panel (www.cambridgebioresource.group.cam.ac.uk). None of the healthy volunteers had a history of drug or alcohol dependence. The following exclusion criteria applied to all participants: no history of neurological or psychotic disorders, no history of a traumatic brain injury, no acute alcohol intoxication (as verified by breath test), and insufficient English proficiency. All volunteers consented in writing and were screened for current psychiatric disorders using the Mini-International Neuropsychiatric Inventory (Sheehan et al., 1998). Psychopathology in drug users was further evaluated using the Structured Clinical Interview for DSM-IV (First et al., 2002). All participants completed the National Adult Reading Test (NART) (H. E. Nelson, 1982) to provide an estimate of verbal IQ and the Alcohol Use Disorders Identification Test (AUDIT) (J. B. Saunders et al., 1993), to evaluate the pattern of alcohol intake. The study was conducted under UK National Health Service Research Ethics Committee approvals (12/EE/0519; principal investigator: KDE).

5.2.2 Slip-of-Action Task

Details of the task are reported elsewhere (Ersche et al., 2016). In brief, in the first part of the task, participants complete an appetitive discrimination task in which they learn over 96 trials the associations between a response (left or right button press) and a rewarding outcome

(gaining points or no points). On each trial, participants were presented with one of six animal pictures and were instructed to learn by trial-and-error which button to press in order to gain points (Figure 5.1). Feedback was provided immediately. The rewards were delivered deterministically, i.e. there is only one correct response for each stimulus. Correct responses were recorded as an index of learning from positive reinforcement.

Completion of the first phase led to the second phase, in which participants were instructed to select the correct response for each animal picture as quickly as possible. However, some outcomes were devalued such that participants were told that responses for certain animal pictures were no longer valuable, and they should not be selected (i.e. participants had to withhold their response). No feedback was provided during this phase, which consisted of nine 12-trial blocks, which at the start of each block, informed participants about the devalued outcomes. Responses toward devalued animal pictures are considered 'slips of actions' and have been suggested to reflect habitual control (de Wit et al., 2007, 2009). I calculated a 'habit bias', based on responding to devalued stimuli minus responding to value stimuli. Participants who respond in a goal-directed fashion, will follow the instruction to only respond to the stimuli that carry a value. However, sometimes they may fail to do so, making a 'slip of action' such that they respond to devalued stimuli although they do not carry any more points. For these participants, their habit bias will be low or even negative. By contrast, participants who respond in a habitual manner will not make this distinction between valued and devalued outcome, as they continue responding equally often to devalued and the value stimuli, making frequent slips of action, so that their habit bias (or slip-of-action score) is likely to be high and close to zero.

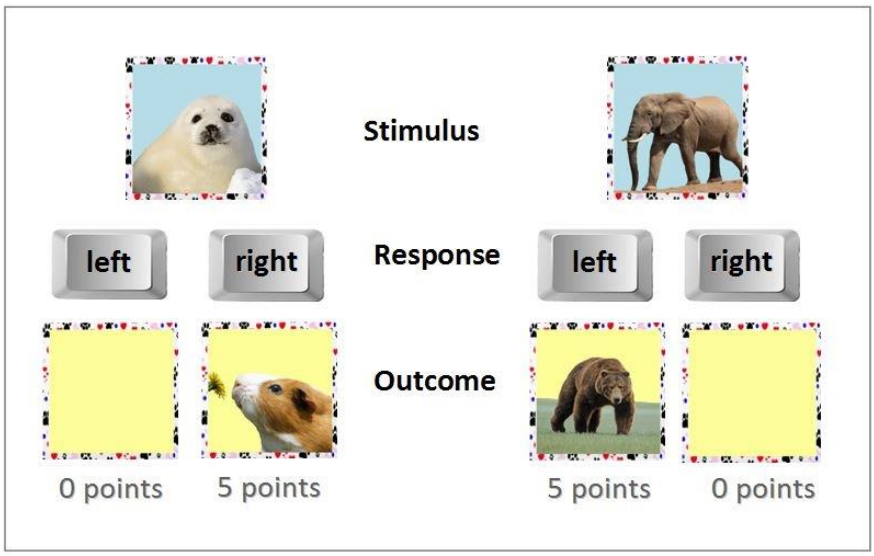


Figure 5.1: Outline of the appetitive discrimination learning task. Participants were required to learn by trial and error which response associated with an animal picture gained them points. Feedback was provided by a picture of another animal coupled with either a number of points or an empty box with no points.

5.2.3 Statistical analysis and computational modelling

5.2.3.1 Demographic and behavioural data

Data were analysed using the Statistical Package for the Social Sciences, v.22 (SPSS, Ltd.). Group differences regarding demographics and fractional anisotropy (FA) values of the goal directed, as well as the habit system pathway were analysed using independent samples t-tests. The white matter tracts between the medial orbitofrontal cortex and the anterior part of the caudate nucleus have previously been shown to underlie goal-directed control, whereas the tracts between the posterior putamen and the premotor cortex is thought to subserve habit control (de Wit et al., 2012). To determine the learning parameters that subsequently affected habitual responding, I performed a stepwise regression model, in which I included the three relevant learning parameters of the model (learning rate, reinforcement sensitivity, perseveration), group status, and white matter integrity between the medial orbitofrontal cortex and the anterior caudate nucleus (as reflected by FA values). I also calculated Pearson's correlation coefficients to evaluate putative relationships between these learning parameters, demographic variables and the duration of cocaine use. To address the question as to whether proneness to habits in CUD patients is due to deficits in goal-directed learning, I fitted an ANCOVA model and included the parameter learning rate as a covariate. All statistical tests were two-tailed and significance levels were set at 0.05. The minimum effect size detectable here is 0.51 (Cohen's d), as determined by a sensitivity power analysis.

5.2.3.2 Reinforcement learning algorithm

I fitted trial-by-trial performance on the appetitive learning phase with a delta rule to model the choice selection process. Since there are two possible responses for each stimulus (i.e. 'respond right' and 'respond left'), the associative strength for the chosen stimulus-response pairing on a given trial, V_t, was updated, using the following algorithm:

$$V_{t+1} = V_t + \alpha(R_t - V_t)$$

When a particular response is positively reinforced, the associative strength for the stimulus–response association increases. This associative strength for each stimulus–response pairing is updated on a trial-by-trial basis via prediction errors that represent discrepancies between expected outcome, V_t, and actual outcome, R_t. Larger prediction errors thus lead to greater changes in associative strength. The sensitivity to this prediction error is regulated by the free parameter, α. Higher α represents increased sensitivity to prediction errors, resulting in quicker updating of associative strengths and enhanced learning.

There is evidence for differential neural processing of reward and non-reward (Kim et al., 2006), suggesting that these two processes may be dissociable. To account for this possible distinction, I tested two classes of computational models. In one class, I fractionated α based on the context. Trials that are positively reinforced were updated by an appetitive learning rate, α^{rew}, whereas trials that were not reinforced were regulated by an extinction rate, α^{ext}. (Increases in α^{rew} would indicate increased learning from reinforcement, and increases in α^{ext} similarly from non-reinforcement.) In a second class, I used a single α value, termed learning rate, to modulate prediction errors irrespective of outcome. I also allowed for the fact that a subject may "stick with" or perseverate to the response that they selected on the previous trial. For trial t and response k, I defined C^k_t to be 1 if the subject chose response k on the previous trial (trial t – 1), and 0 otherwise. I then defined a perseveration parameter τ through which a putative tendency to perseverate influenced behaviour, alongside the reinforcement learning process.

Associative strengths and perseverative tendencies were then used to select actions. This process followed a softmax rule, according to the following equation:

$$p(i,t) = \frac{e^{\beta V_t^i + \tau C_t^i}}{\sum_{k=1}^{n} e^{\beta V_t^k + \tau C_t^k}}$$

This softmax equation gives the model's predicted probability of a given choice *i* on a given trial *t*. Associative strengths (calculated as above) drive choices, and the degree to which they influence the final choice is determined by the reinforcement sensitivity parameter β. A tendency to perseverate can also influence choice, and the degree to which this happens is determined by the perseveration parameter τ. As outlined in Table 5.1, there are four possible free parameters that were modelled: learning rate, extinction rate, reinforcement sensitivity and perseveration.

The task design involved an explicit instruction of a different task context and different performance rules in the second phase, gave no feedback, and relies for successful performance on explicit representations of instrumental value that can be instructed. These limitations prevented accurate trial-by-trial modelling of behaviour from the second phase within this model. An additional confirmatory model, representing goal-directed action and habit learning explicitly, was therefore used to check the effects of outcome devaluation (see below).

Table 5.1: Summary of the reinforcement learning models tested.

Free parameters				Model selection				
Learning rate[a]	Extinction rate, α^{ext}	Reinforcement sensitivity, β	Perseveration, τ	Log marginal likelihood	Log posterior p(model)	Posterior p(model)	Log_{10} Bayes factor (relative to next-ranked model)	Ranking
✓		✓	✓	-6718.8	-0.578	0.561	0.106	1
✓	✓	✓	✓	-6719.0	-0.823	0.439	18.03	2
✓	✓	✓		-6760.5	-42.33	0	0.407	3
✓		✓		-6761.5	-43.27	0	140.71	4
✓	✓		✓	-7085.5	-367.27	0	20.04	5
✓	✓			-7131.6	-416.40	0	492.78	6
				-8266.3	-1548.06	0	N/A	7[b]

Note. Several models with different parameter combinations were assessed via bridge sampling. I show the included posterior probabilities for each model, i.e. the probability of each model given the data (and given that they were equiprobable before the data). Models were ranked accordingly and I found that the best-fit model used three parameters: learning rate, reinforcement sensitivity and perseveration. I have also included log Bayes factors for comparisons between the ranked models. According to the criteria of Kass & Raftery (1995), there was overwhelming evidence that the top two ranked models were superior to all other models. Though the difference between the top two models was marginal, I have selected the model that was more likely, which was also the more parsimonious of the two. [Notes: Logs are natural logarithms unless stated.

[a] For some models, the learning rates were fractionated into learning from reward (α^{rew}) or non-reward (i.e. extinction rate, α^{ext}), as shown. If extinction rate is not defined in the model, then the learning rate should encompass learning from both reward and non-reward (α).

[b] To verify that these results were not spurious findings, I included a random choice model, which assumes that choices were selected at random (p = 0.5 for each of the two possible responses). These results suggest that all tested models fit the data better than the random choice model.]

5.2.3.3 Parameter estimation

Free parameters from reinforcement learning algorithms were estimated using a hierarchical Bayesian approach. This approach produces a posterior distribution for all parameters of interest. I defined prior distributions for all parameters. The learning rate parameters alpha (α, α^{rew}, α^{ext}), which have the range [0, 1], were given a prior beta (1.1, 1.1) distribution. Reinforcement sensitivity, β, was given a prior gamma(4.82, 0.88) distribution (Gershman, 2016). Perseveration, τ, was given a normal(0, 1) prior; perseverative parameters can be negative, indicating anti-perseveration (switching behaviour) (Christakou et al., 2013).

At the top level of the hierarchy, for each parameter I defined a separate distribution for each group (CUD and controls). These were the primary measures of interest. Each individual subject's parameter was drawn from a distribution about their group-level parameter, with the assumption that individual subjects' differences from their group mean had a normal distribution with mean 0 and a parameter-specific standard deviation (necessarily positive). For α and τ, this standard deviation was drawn from a prior half-normal (0, 0.17) distribution. For β, the standard deviation of inter-subject variability was drawn from a prior half-normal (0, 2) distribution. Final subject-specific parameters were bounded as follows: $\alpha \in [0,1]$; $\beta \in [0,+\infty]$; $\tau \in [-\infty, +\infty]$. These final subject-specific parameters were then used in a reinforcement learning model, whose output was the probability of selecting each of the two actions on any given trial. The model was fitted (yielding posterior distributions for each parameter) by fitting these probabilities (arbitrarily, the probability of choosing the right-hand response) to actual choices (did the subject choose the right-hand response?).

I conducted the Bayesian analysis in RStan (Carpenter et al., 2017), which uses a Markov chain Monte Carlo method to sample from posterior distributions of parameters. Primary values of interests were posterior distributions of the group difference (CUD – control) for each free parameter. Measures of dispersion of posterior distributions were denoted as 95% highest density intervals (HDI). Given the assumptions (priors, model) and data, there is a 95% probability that the true value lies within the 95% HDI. An HDI of the group difference that does not overlap with zero indicates credible group differences. Parameter recovery was assessed for the winning model, which showed satisfactory recovery (Appendix D, Figure D3).

5.2.3.4 Model selection

As shown in Table 5.1, several variants of the models were tested against each other. The best model was determined using bridge sampling (Gronau et al., 2017), which estimates model fit. The bridge sampling procedure computes the probability of the observed data given the model of interest, the marginal likelihood $P(D \mid M)$, which encompasses both the probability of the data given specific values of the model's parameters, the likelihood $P(D \mid \theta, M)$ (is there a good fit?) and the prior probability of the parameter values given the model, $P(\theta \mid M)$ (thus encapsulating a penalty for over-complex models; Occam's razor). The marginal likelihoods $P(D \mid M_i)$ can be combined with prior model probabilities $P(M_i)$ to obtain posterior model probabilities $P(M_i \mid D)$. I report posterior probabilities for the models, which indicate evidence for the model; a higher probability indicates a better model. Additionally, I also report the log Bayes factor as a second indicator of model evidence, Bayes factors being ratios of marginal likelihoods of a pair of models. I assumed models were equiprobable *a priori*.

5.2.3.5 Confirmatory modelling of goal-directed action and habitual responding

To analyse more directly the question of whether the balance between goal-directed and habitual systems was altered in the CUD group, as assessed by the outcome devaluation procedure, a full two-system model of instrumental learning was developed and simulated as an additional check. This model implemented outcome devaluation via instantaneous instruction (see Appendix D). The behavioural task (Ersche et al., 2016) was incompletely specified for this fuller instrumental model in some respects, in that it did not permit independent evaluation of the learning rate for habit and goal-directed systems, though it permitted evaluation of the relative expression of those two systems via the outcome devaluation phase. The behavioural task was also ambiguous as to whether the framing of the task was likely to have allowed further S–R habit learning (as distinct from expression) during the outcome devaluation phase, given that the response instructions were altered substantially in this phase; we therefore tested models with and without S–R learning during this test phase ("habit learning at test", HLAT, or "no habit learning at test", NHLAT; see Appendix D), with the caveat that the HLAT model had the potential to confound the effects of outcome devaluation and extinction in the measurement of learning rate.

5.2.4 Neuroimaging data

To address the critical question of whether abnormal learning performance is associated with variations in frontostriatal connectivity, neuroimaging data was obtained from almost 70% of the participants (44 controls, 44 CUD). The selection of this subgroup was based on MRI-suitability and availability for the acquisition of the scan. The subgroup was representative of the entire sample, as no significant group differences in their demographic profiles were identified.

5.2.4.1 MRI data acquisition, pre-processing and ROI generation

The brain scans were acquired at the Wolfson Brain Imaging Centre, University of Cambridge, UK. T1-weighted MRI scans were acquired at by a T3 Siemens Magenetom Tim Trio scanner (www.medical.siemens.com) using a magnetization-prepared rapid acquisition gradient-echo (MPRAGE) sequence (176 slices of 1 mm thickness, TR=2300ms, TE=2.98ms, TI=900ms, flip angle=9°, FOV=240 x256). One CUD scan was removed due to excessive movement. All images were quality controlled by radiological screening. The MPRAGE images were processed using the recon-all Freesurfer (v5.3.0, recon-all, v 1.379.2.73) pipeline to generate individually labelled brains using the standard subcortical segmentation and Destrieux atlas surface parcellations. Two regions of interest (ROIs) were created in both the left and right hemispheres: the combined caudate and nucleus accumbens, and the medial orbitofrontal cortex, as well as the premotor cortex (BA6) (thresholded version) and posterior putamen (defined as the putamen for y <= 2mm in MNI space (see de Wit, Watson et al. 2012)). A mask was created in MNI space for y>2mm. The inverse MNI transform for each individual was applied to the mask to put it in native conformed space, which was then used to split the putamen into posterior and anterior portions. The anterior caudate – medial orbitofrontal cortex mask (Figure 5.3A) and the premotor cortex – posterior putamen mask (Figure 5.3B) were based on a prior tractography analysis by de Wit and colleagues (2012), and is thought to be implicated in the goal-directed and habit systems respectively. In addition, two exclusion masks were created comprising each hemisphere and all ventricles. All ROIs were transformed into diffusion-weighted imaging data (DWI) space for the subsequent tractography analysis.

5.2.4.2 DWI data acquisition and pre-processing

Due to excessive movement, four scans had to be excluded from the analysis (1 control, 3 CUD). DWI volumes were successfully acquired from 84 participants (43 controls, 41 CUD). All DWI scans were acquired within the same scan session as the MPRAGE data set. Sequence details were as follows: TR=7800ms, TE=90ms, 63 slices of 2mm thickness, 96x96 in-plane matrix, FOV=192x192mm. DWI data were acquired with a 63 direction encoding scheme. These 63 volumes were acquired with a b-value of 1000 s/mm2 following an initial volume with a b-value of 0 s/mm2.

The DWI-images were processed using the standard FSL (FMRIB Software Library; Release 5.0.6) tractography pipeline. First, eddy correct was performed to correct head motion and distortion, and align the series to the b0 image. Next a brain mask was created by applying bet to the b0 image. Then diffusion parameters were estimated using bedpostX. BedpostX uses a Bayesian framework to estimate local probability density functions on the parameters of an automatic relevance detection multicompartment model. In this case two fibers per voxel were modelled. Following bedpostX, probabilistic tractography was applied to the diffusion parameters using probtrackx2. Probtrackx2 computed streamlines by repeatedly generating connectivity distributions from voxels in seed ROIs. The default settings of 5000 samples per voxel and 0.2 curvature threshold were used. Analyses were performed from seed ROIs to waypoint targets in each hemisphere separately with an exclusion mask defined for each analysis comprising the combined contralateral hemisphere and ventricles. The first seed-target path interrogated was caudate and nucleus accumbens to medial orbitofrontal cortex, and the second seed-target path interrogated was posterior putamen to the premotor cortex, which made a total of four analyses per participant. Each analysis generated a waytotal, which is the number of tracts surviving the inclusion and exclusion criteria. Each participant's waytotals were normalized by the individual seed ROI volumes (x5000) to produce single measures of tract strength between the seed and target.

In addition to the waytotal each tractography analysis produced a connectivity distribution path. A summary group path distribution was produce to illustrate each tract. Each individual path was thresholded above 5% or 10 hits, whichever was the higher value. These paths were then transformed into MNI-space using a non-linear warp and a mean path created. Individual seed and target regions were also transformed into MNI-space using the combined Freesurfer to

diffusion-space affine transformation and the non-linear diffusion to MNI-space warp. A summary binary region of interest was created representing the path from the combined caudate and nucleus accumbens to medial orbitofrontal cortex. The ROI comprised voxels containing thresholded paths from at least half the participants.

FA maps were created using FSL's dtifit and were then processed according to the standard Tract-based spatial statistics (TBSS) pipeline to create a 4D volume containing each participant's skeletonised FA image. Mean FA values were calculated for each participant within the group ROI from each tractography path (anterior caudate to medical OFC and putamen to premotor cortex) and imported into SPSS for post hoc analyses.

5.3 Results

5.3.1 Group characteristics

Sample demographics are presented in Table 5.2. The groups were matched in terms of age, gender, and alcohol intake but differed significantly in terms of verbal IQ. However, only in control volunteers IQ scores were correlated with learning rate (r=.29, p=0.034) and reinforcement sensitivity (r=.30, p=0.029), but not in CUD patients (both p>0.1). I also found that in CUD patients, the duration of cocaine use correlated significantly with the degree of response perseveration (r=.29, p=0.014), but prolonged cocaine use showed no relationship with either learning rate (r=-.14, p=0.254) or reinforcement sensitivity (r=-.19, p=0.118).

Table 5.2: Sample demographics for Chapter 5.

	Mean (SD)		**Statistics**	
	Control	CUD	*t*	*p*
Group size (n)	55	70	-	-
Age (years)	41.3 (10.5)	38.0 (8.6)	1.9	0.06
Gender (% male)	94.3	90.3	-	-
Alcohol use severity (AUDIT score)	4.2 (2.0)	4.3 (4.8)	-0.4	0.97
Verbal IQ (NART score)	114 (7.4)	102 (8.1)	8.8	<0.001
Duration of cocaine use (years)	-	15.9 (6.7)	-	-

Note. CUD: cocaine use disorder; AUDIT: Alcohol Use Disorder Identification Test; NART: National Adult's Reading Test; SD: Standard deviation

5.3.2 Instrumental learning performance

As shown in Table 5.1, the winning model contained three parameters: a single learning rate, reinforcement sensitivity, and perseveration ('stickiness'). Relative to healthy control volunteers, CUD patients demonstrated reduced learning rates (see Figure 5.2; posterior mean difference, d = -0.035, 95% HDI = -0.064 to -0.010, posterior probability of non-zero difference, p_{nz} = 0.999). There were no group differences for reinforcement sensitivity (d = 1.58, 95% HDI = -1.02 to 4.51, p_{nz} = 0.69) or perseverative responding (d = -0.02, 95% HDI = -0.141 to 0.089, p_{nz} = 0.367).

In light of the high prevalence of co-morbid opiate use in cocaine addiction, I also subdivided the CUD sample into CUD participants with (n=22) and without co-morbid opiate dependence (n=48), and fitted the winning model with data of these two subgroups. As shown in Table D1 (Appendix D)**,** the two subgroups did not differ on any performance parameter.

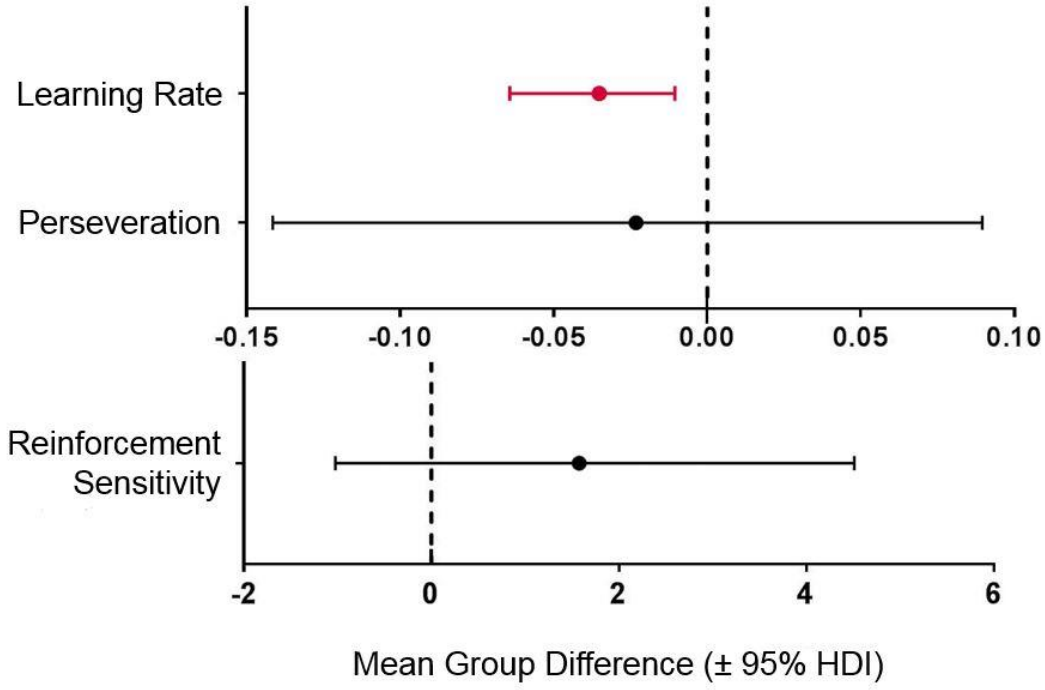


Figure 5.2: The mean group differences of the posterior distributions for each learning parameter in the model. Parameters that have group differences (indicated in red) have 95% highest density intervals that do not overlap zero. Compared with healthy control volunteers, patients with CUD show a reduced learning rate. Both mean differences in reinforcement sensitivity and perseveration did overlap with zero. (Note: the reinforcement sensitivity parameter is placed on a different axis due to scale differences).

In the additional model examining goal-directed actions and habits across both task phases, whether or not S–R learning was assumed to occur during the test (second) phase influenced the sign of the difference in learning rate observed in this two-system model (see Appendix D, Table D4), rendering interpretation of learning rates difficult. In the NHLAT model, the CUD group showed lower learning rates than controls; this is entirely consistent with the lower learning rates found via the main computational model confined to the first phase of the task (since in that model and the NHLAT model, learning rates were only measured during the initial learning phase). In the HLAT model, learning rates were higher in the CUD group; this likely reflects a confound between measuring the impact of outcome devaluation and measuring extinction in the second phase, altering the estimates of learning rates.

However, other aspects of the additional two-system models were consistent. Both the NHLAT and HLAT models showed a reduced impact of the goal-directed action system in the CUD group; no difference in the impact of the habitual system; and a somewhat greater tendency to perseverate (or lesser tendency to switch response) in the CUD group (Appendix D, Table D4**)**. These results are therefore consistent with a reduction in the relative efficacy of goal-directed action and an increase in the relative (if not absolute) efficacy of habitual learning in patients with CUD. Moreover, since the goal-directed system was consistently less effective in CUD patients, in addition to and independent of changes in learning rate, the results of both the NHLAT and HLAT models support the conclusion that excessive dominance of the habit system (due to impaired goal-directed action) in CUD patients is not explicable purely in terms of changes in learning rates.

5.3.3 Relationships between learning performance and white matter integrity

I compared the two groups with respect to white matter integrity, as reflected by fractional anisotropy (FA) values, within both the goal-directed and the habit pathways. Whilst FA values between the anterior caudate - medial OFC (goal-directed) pathway did not significant differ between CUD patients and control volunteers (t_{81}=1.57, p=0.122), I identified significant group differences in white matter integrity in the putamen - premotor cortex (habit) pathway as FA in the CUD group was significantly reduced compared with controls (t_{81}=2.19, p=0.031). I first correlated, separately for each group, the learning rates with mean FA values of the goal-directed pathway and then the slip-of-action scores with mean FA values in the habit pathway

(see Figure 5.3). Learning rates showed a positive correlation only in control volunteers (r = .406, p =.007), but not in CUD patients (r= .070, p=.668), whereas the slip-of-action score was not correlated with the FA values in either group (controls: r=-.25, CUD: r=.05; both p>0.1)

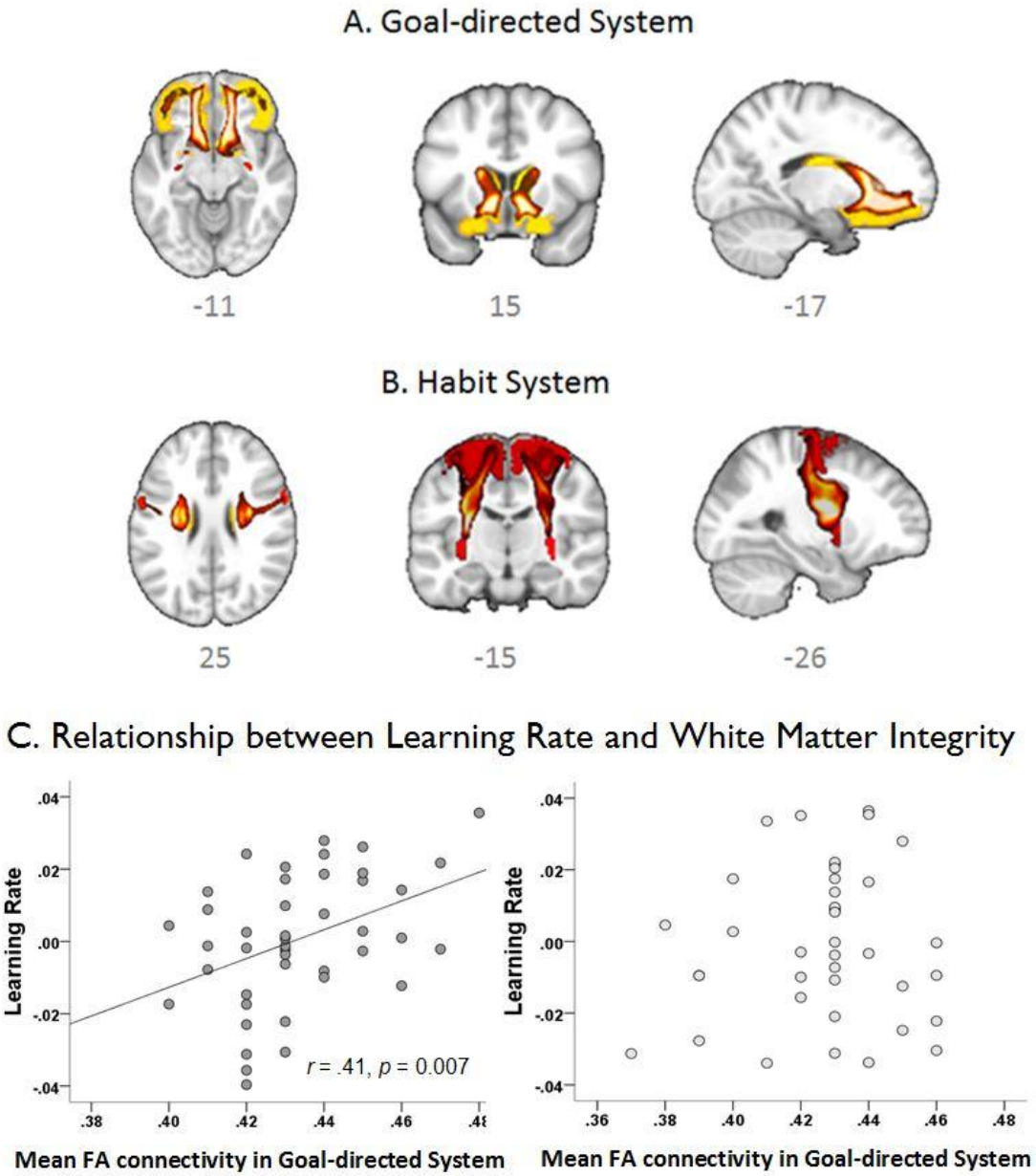


Figure 5.3: Structural connectivity of *a priori* brain networks implicated in the goal-directed and habit systems, and their relationship with the learning rate. (A) Brain regions involved in the goal-directed system has been linked with interactions between the medial prefrontal cortex, the anterior caudate nucleus and ventral parts of the striatum. (B) The habit system depends on interactions between pre-motor cortex (BA6) and the posterior putamen. (C) Scatter plot depicting the significant relationships in healthy control volunteers between learning rates and mean FA values within the neural pathway that has been suggested to underlie goal-directed learning. Scatter plot showing the lack of such a relationship in CUD patients.

To further examine the extent to which learning performance accounted for individual variation in habitual responding, I employed a stepwise regression model analysing habit bias (slip-of-action) scores. The model revealed that group status accounted for 12% of the variance in habitual responding (β_{group} = 0.362, R^2=0.12, $F_{1,121}$=18.24, p<0.001). When reinforcement sensitivity was entered in the model, about a quarter of the variance (25%) were explained by the two factors (β_{group} = 0.358, β_{reinf} = -0.355, R^2=0.25, $F_{2,120}$=20.77, p<0.001); learning rate and perseveration had no explanatory value (i.e. the addition of these parameters did not significantly improve the model). When I subsequently entered the neural correlates of the goal-directed pathway, which were available in 70% of the sample, the results did not change. In this smaller sample, group status explained 17% of the variance (β_{group} = 0.425, R^2=0.17, $F_{1,81}$=17.82, p<0.001), and together with reinforcement sensitivity, explained 30% of the variance of habitual responding (β_{group} = 0.403, β_{reinf} = -0.365, R^2=0.30, $F_{2,80}$=18.23, p<0.001), suggesting that the strong habit bias in CUD was not fully explained by the deficits in discrimination learning. This was further supported by the fact that the strong habit bias in CUD was also seen when the learning rate was included as a covariate in the analysis ($F_{1,120}$=20.2, p<0.001). Given that the groups also differed in white matter integrity in the habit pathway, I added FA values of the putamen-premotor (habit) pathway as a second covariate in the ANCOVA model, but this did not affect the significant habit bias in CUD patients ($F_{1,79}$=16.9, p<0.001).

Although the groups did not differ with respect to FA within the goal-directed pathway (t_{81}=1.57, p=0.122), I aimed to evaluate the putative relationships between the three learning parameters and FA. I calculated Pearson's correlation coefficients, which revealed relationships between the learning rate (r=.41, p=0.007) and reinforcement sensitivity (r = .34, p = 0.026) only in the control volunteers but not in CUD patients (both p > 0.5). Using Fisher's Z transformation, I found that the correlations between learning rate and FA were not significantly different between groups (Z = 1.56, two-tailed p=0.119).

5.4 Discussion

Drug addiction has been described as a disorder of learning and memory (Hyman, 2005), where behavioural choices become biased toward highly reinforcing drug-rewards which persist even

if the anticipated rewarding outcome does not materialise. Here I deconstructed the process of appetitive discrimination learning in a non-drug related context in both healthy control participants and patients with CUD using a computational modeling approach, which yielded two important findings. Firstly, I demonstrated that CUD patients exhibit significant deficits in reinforcement learning as reflected by a reduced learning rate, possibly indicating problems with making accurate reward predictions and/or updating these predictions based on feedback. Secondly, I demonstrated that the reduced learning rate in CUD patients did not, however, fully explain their proneness for stimulus-response habits during instrumental learning. Habitual response tendencies, as measured by reward devaluation, were partly explained by the diagnosis of CUD and individual variation in reinforcement sensitivity, but were not sufficiently explained by deficits in learning. These conclusions were supported by additional analyses across discrimination and devaluation phases using a two-system model representing goal-directed action and habit learning, which showed a reduced learning rate in CUD patients in the discrimination phase, and a reduced impact of the goal-directed system; changes in learning rate were not sufficient to explain the relative predominance of the habit system in CUD patients.

5.4.1 Deficits in learning from positive feedback impair appetitive discrimination learning

The findings are strikingly consistent with previous reports in both animals and humans suggesting that chronic cocaine use is associated with deficits in the processing of positive feedback (Lucantonio et al., 2015; Morie et al., 2016; Strickland et al., 2016; Takahashi et al., 2016). By changing the neuronal signaling patterns, chronic cocaine use has been suggested to alter the encoding of outcome information such as value, timing, and size of the outcome, thereby hampering predictions about the consequences of one's actions (Takahashi et al., 2019). Current findings are also consistent with work by Kanen et al. (2019), who also identified in another sample of stimulant-addicted individuals a reduced learning rate from positive feedback. It is noteworthy that those authors further showed that the learning deficits were amenable to dopaminergic modulation, thus supporting the notion of mediation via alterations in the firing patterns of dopamine neurons. The nature of the hypothesized cocaine-induced neuroadaptive changes of appetitive learning may also explain why I did not find changes in white matter integrity within the goal-directed pathway. I found a statistically significant relationship between learning from positive feedback and structural integrity in control participants, but not CUD patients. However, these correlations did not significantly differ with

one another, possibly due to lack of power, and thus need to be cautiously interpreted. There were, however, no significant structural alterations in the CUD group. It must also be emphasized that CUD patients' ability to learn from positive feedback was not entirely impaired. All participants in the study were able to learn the stimulus-reward association, but CUD patients learned them less well than healthy control participants. Their ability to learn from positive feedback also stands in stark contrast from that of learning from negative or punishing feedback, which has been repeatedly shown to be severely impaired in CUD patients (Hester et al., 2013; Tanabe et al., 2013). Such an imbalance in the ability to process reinforcing feedback has important ramifying effects on patients' decisions and behavioural choices, and therefore should be recognized as a treatment need.

5.4.2 Diagnosis of CUD and variation in reinforcement sensitivity partly explain habit bias

The mechanism that renders CUD patients prone to developing stimulus-response habits is not fully understood. The weaker white matter integrity in the habit pathway in CUD patients was, however, unrelated to behaviour, suggesting that that the increased habit bias cannot simply be attributed to structural variations. However, it has been previously suggested that a strong habit bias could reflect a compensatory response to a weakened goal directed system (Robbins & Costa, 2017; Vandaele & Janak, 2018). Here I demonstrate that reduced learning rate in CUD patients does not account sufficiently for their proneness to form stimulus-response habits. Other psychiatric disorders, such as obsessive-compulsive disorder, exhibit a habit bias on this task alongside unimpaired discrimination learning (Gillan et al., 2011). It is conceivable that the regulatory balance between goal-directed or habitual control is disrupted in CUD patents, indicating a failure to revert control to the goal-directed system following a rule change. Alternatively, but not mutually exclusively, it is also possible that habitual control is generally more predominant in cocaine addiction. Whilst there is ample evidence showing failure of CUD patients to adjust cognitive or behavioural responses to changing situational demands (Ersche et al., 2008; Ersche, Roiser, Abbott, et al., 2011; Lane et al., 1998; McKim et al., 2016; Verdejo-García & Pérez-García, 2007), far less research has addressed the predominance of the habit system.

Our data further indicates that one learning parameter in particular, reinforcement sensitivity, does seem to be involved in habit formation. This observation is not surprising given that habit

learning in this study was assessed using a reward devaluation paradigm, which deliberately manipulates the value of the expected outcome of an instrumental response to make the outcome less desirable, and the behavioural response less likely. If these manipulations, however, do not impact on performance, it may indicate that behaviour is not controlled by the anticipated consequences but by antecedent stimuli; or in other words, their behaviour has become habitual. Although reinforcement sensitivity values in this study did not differ between the groups, it is noteworthy that correct responses were reinforced by points gain, which CUD patients may not have perceived as rewarding in the first place. Future research may thus need to evaluate whether the use of more reinforcing incentives such as monetary gain or the prospects of desirable benefits would be more appropriate for a reward devaluation paradigm than points gain, possibly making devaluation more noticeable to induce a behavioural change.

5.4.3 Neural substrates of appetitive discrimination learning

These data also indicate that the diagnosis of CUD, rather than individual learning parameters, critically account for the facilitated transition from goal-directed to habitual responding. The diagnosis may thus reflect disorder-related changes within corticostriatal networks that subserve associative learning, which is likely to promote the devolution of control from the goal-directed to the habit system (A. Nelson & Killcross, 2006; Takahashi et al., 2007). Cocaine addiction has been associated with numerous changes within dopaminergic pathways such as low D2 receptor density in the striatum and reduced orbitofrontal metabolism (Volkow et al., 1993), blunted stimulant-induced dopamine release (Martinez et al., 2007), reduced white matter integrity in the inferior frontal gyrus (Ersche et al., 2012), and altered cognitive responses to dopamine agonist challenges (Ersche et al., 2010). Loss of white matter integrity specifically in the inferior frontal gyrus might also play a role in disinhibited behaviour whereas action selection is undermined by alterations in dopaminergic transmission. More research is warranted to investigate the neuromodulatory effects of specifically dopaminergic agents on associative learning. Work reported in Chapter 3 and by Kanen et al (2019) already shows some promising results, suggesting that selective learning parameters are differentially modulated by dopaminergic agonists and antagonist treatments. Functional neuroimaging may provide valuable insight into how chronic cocaine use might change the neural networks implicated in associative learning.

5.4.4 Conclusion

I show that patients with CUD have deficits in the reinforcement learning parameter of learning rate, which were neither related to structural connectivity in the 'goal-directed' pathway nor explained their strong habit bias. Moreover, I also identified significantly reduced integrity in white matter structure in brain structures implicated in habit formation, which also did not explain CUD patients' strong habit bias. These results are relevant to the hypothesis that drug addiction results in an imbalance between goal-directed and habitual control over behaviour.

Appendix D: Supplementary materials to Chapter 5

Parameters	Posterior differences [mean (95% HDI)]		
	CUD – HC	**CUD$^+$ – HC**	**CUD$^+$ – CUD**
Learning rate	-0.027 (-0.062, 0.004)	-0.039 (-0.071, -0.012)*	0.012 (-0.013, 0.041)
Reinforcement sensitivity	1.57 (-1.55, 4.88)	0.701 (-1.80, 3.69)	0.868 (-2.52, 4.57)
Perseveration	0.040 (-0.111, 0.193)	-0.058 (-0.185, 0.065)	0.098 (-0.051, 0.248)

CUD: Patients with cocaine without opioid use disorder (n=22)

HC: healthy control volunteers (n=55)

CUD$^+$: Patients with cocaine + opioid use disorder (n=48)

*probability of non-zero difference, $p_{nz} > 0.95$ ($0 \notin$ 95% HDI)

Table D1: Results for reinforcement learning analyses including patients with comorbid opioid use disorder. To ascertain whether opioid use disorder contributed towards reinforcement learning impairments, we fitted the winning reinforcement learning model, but with an extra a priori defined subgroup: patients with cocaine and opioid use disorder. Importantly, results show that there were no group differences within patients on all parameters. Although there is a group difference observed between the groups HC and CUD$^+$, this difference may largely be driven by the sample size.

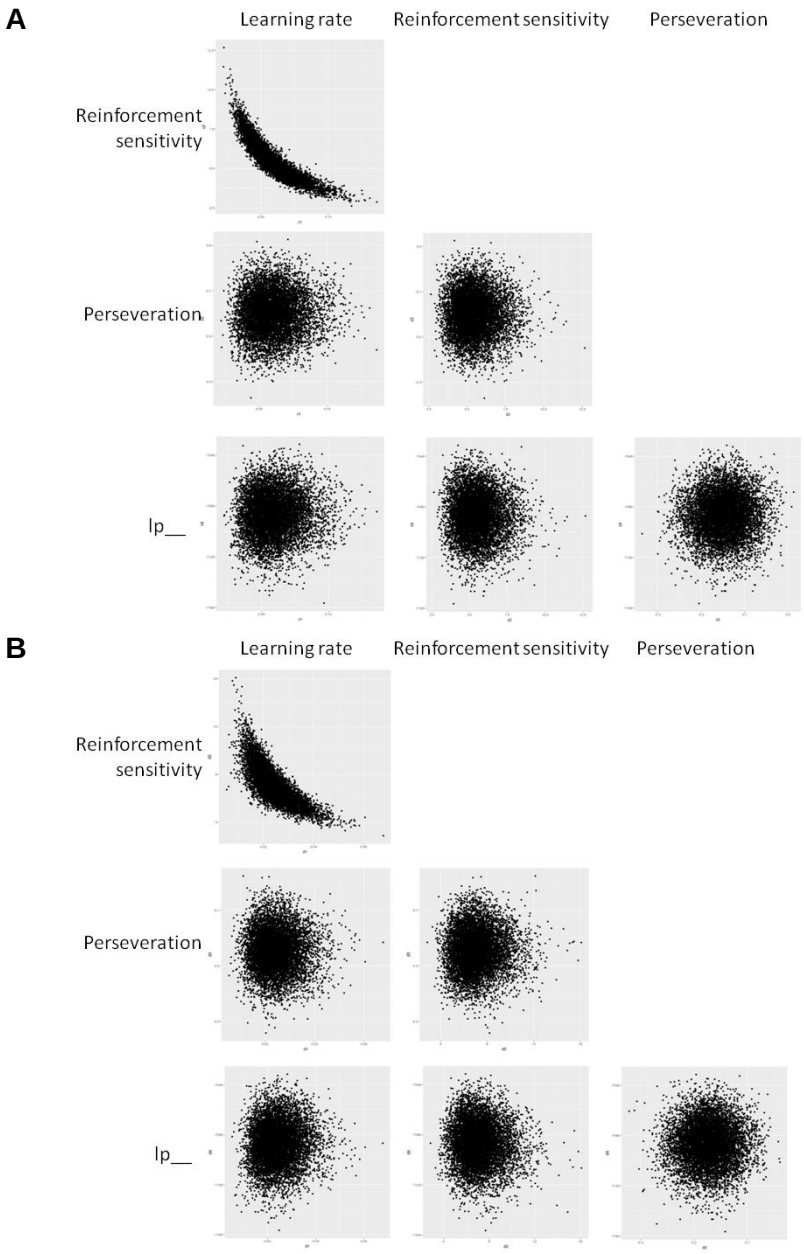


Figure D1: correlations between group-level parameter values from the winning model across iterations. Each point represents one iteration of the winning model. The "lp__" value is Stan's lp__ variable, the log posterior density up to a constant. **(A)** Control subjects. **(B)** CUD subjects.

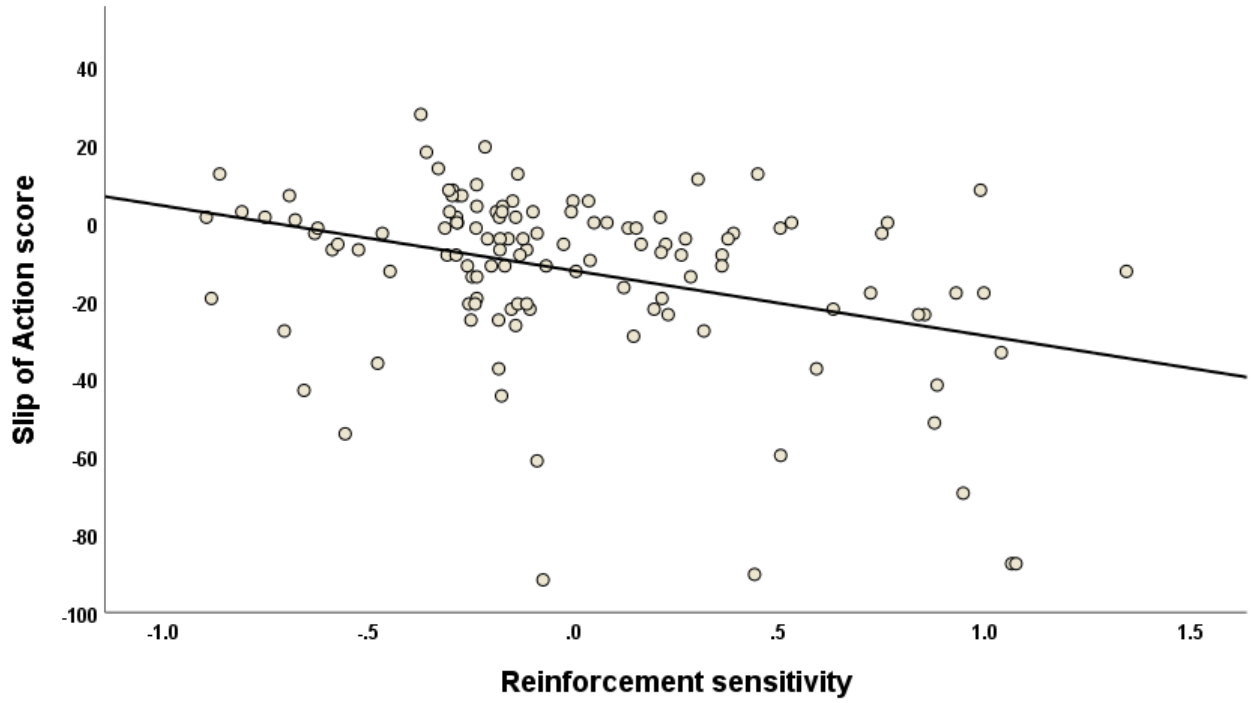


Figure D2: Scatter plot of the relationship between the reinforcement sensitivity parameter (from the winning model) and slip-of-action score (habit bias; behavioural response to outcome devaluation). As reported in the main text, reinforcement sensitivity, along with group status, jointly explained 25% of the variance in habitual responding.

Simulation of behavioural data

To determine the validity of the winning RL model, I simulated trial-by-trial data for 200 participants (100 participants per group) from the group means of the posteriors. The simulated data share an identical structure with the original task setup. Each simulated participant data has 96 trials with random assignments of stimulus–correct response mapping. Trial-by-trial responses were generated based on choice probabilities from the RL and softmax equations as reported in the main text of the manuscript. Following the analysis in Ersche et al. (2016), percentage learning accuracy was computed for each participant. Independent-samples t-test confirmed that I replicate the findings – appetitive discrimination learning performance in the CUD group was poorer to that of healthy volunteers (t_{198} = 5.08, p < .001).

Parameter recovery for the winning model

I have assessed parameter recovery for the winning model of this chapter by simulating parameter values, and repeating the model fitting procedure to recover these parameter values. The scatterplots between simulated and recovered parameters are presented in Figure D3. Overall, parameter recovery is satisfactory, as the learning rate (α, *r*=0.723, *p*<0.001), reinforcement sensitivity (β, *r*=0.672, *p*<0.001) and perseveration (τ, *r*=0.904, *p*<0.001) parameters had strong correlations between simulated and recovered values. However, the recovery for α and β seems to be better in the lower range, which corresponds to the values in our data.

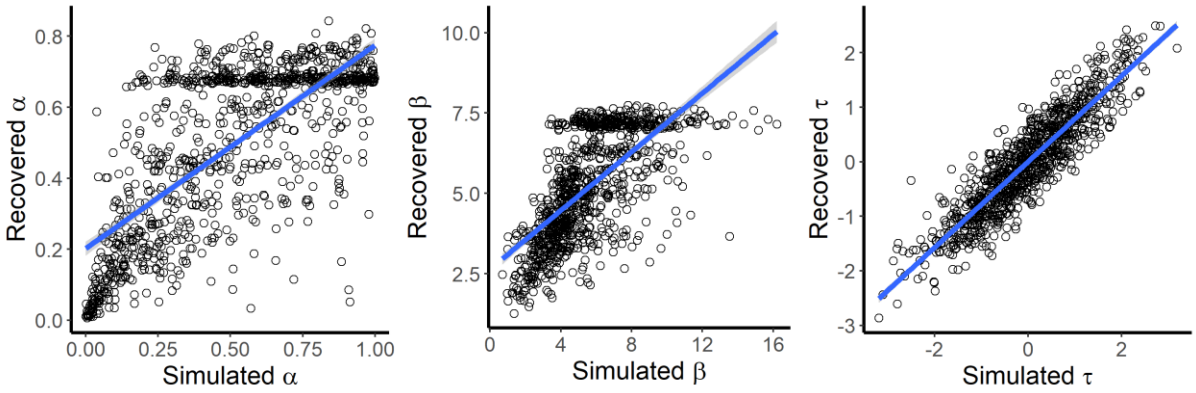


Figure D3: Parameter recovery for the winning model in Chapter 5.

General two-system computational model of goal-directed and habitual responding

We implemented a computational model representing the core features of instrumental learning, namely goal-directed action and stimulus–response (S–R) habits (but not including Pavlovian–instrumental transfer), plus response perseveration as before.

We define notation as shown in Table D2.

Table D2: notation for the two-system instrumental computational model

Term	***Description***	***Category*** *(LTM long-term memory, WM working memory)*
n_S	Number of stimuli. Integer.	Constant
n_A	Number of actions (including one for "no action", required to predict the consequences of inaction and to choose not to act, if permitted). Integer.	Constant
n_O	Number of outcomes. Integer.	Constant
α^O	Outcome (action–outcome contingency) learning rate for the goal-directed action system. Scalar.	Parameter
α^H	Learning rate for the habit system. Scalar.	Parameter
β^G	Inverse temperature parameter representing the effectiveness of the goal-directed action system at driving behaviour. Scalar.	Parameter
β^H	Inverse temperature parameter representing the effectiveness of the habit system. Scalar.	Parameter
β^P	Inverse temperature parameter representing the effectiveness of the response perseveration system. Scalar.	Parameter
t	Current trial number. Integer.	Time
$\boldsymbol{G}$	Current goal-directed action–outcome (A–O) contingencies. An $n_S \times n_A \times n_O$ matrix, mapping discriminative stimuli to A–O contingencies. Starting values are 0.	Subject LTM
$\boldsymbol{v}$	Current instrumental outcome values. Vector of size n_O. The value of outcome o is denoted $\boldsymbol{v}_o$. Starting values are 0.	Subject LTM
$\boldsymbol{H}$	Current stimulus–response (S–R, stimulus–action) habit strengths. An $n_S \times n_A$ matrix mapping stimuli to "action values" (Q values). Starting values are 0.	Subject LTM
$\boldsymbol{s}$	Stimuli: vector of length n_S representing stimulus presence (0) or absence (1) for all stimuli on trial t.	World → subject
$\boldsymbol{C}$	Action–outcome contingencies predicted on trial t by the stimuli currently present, from the tensor dot product of $\boldsymbol{s}$ and $\boldsymbol{G}$; these contingencies can exceed the conventional contingency range [–1, +1]. An $n_A \times n_O$ matrix.	Subject WM
$\boldsymbol{q}^G$	Action "values" (expected value of the action; Q values) for trial t: goal-directed component. Vector of size n_A.	Subject WM
$\boldsymbol{q}^H$	Action "values" for trial t: habit component. Vector of size n_A.	Subject WM
$\boldsymbol{q}^P$	Action "values" for trial t: perseverative component. Vector of size n_A. Defined to contain zeros for all actions, except 1 for the action chosen on the preceding trial (if there was a preceding trial).	Subject WM
$\boldsymbol{a}$	Action tendencies for trial t. Vector of size n_A.	Subject WM
$\boldsymbol{p}$	Probability of making each action on trial t. Vector of size n_A.	Subject WM
a	The selected action (as an index). Integer.	Subject → world
$\boldsymbol{o}$	Representation of which outcomes were obtained on trial t. Binary vector of length n_O containing 1 for outcomes that were obtained and 0 for those that were not.	World → subject
r	Reinforcement value of the outcome(s) obtained. Scalar.	Subject WM
d^H	Reinforcement prediction error (d for discrepancy) for the habit system on trial t. Scalar.	Subject WM
$\boldsymbol{d}^O$	Outcome prediction error for the goal-directed system on trial t. Vector of size n_A.	Subject WM

Actions were determined as follows. Discriminative stimuli (SDs) present (***s***) were combined with previous knowledge of SD-dependent contingencies (***G***) to predict the action–outcome contingencies currently operative (***C***). Combining these contingencies with the value of the outcomes (***v***) gives the declarative expected value of each goal-directed action ($\boldsymbol{q}^G$). Simultaneously, the same environmental stimuli (***s***) act via S–R associations (***H***) to drive actions habitually ($\boldsymbol{q}^H$); this is a procedural rather than a declarative representation but the quantity $\boldsymbol{q}^H$ reflects the "expected value" of actions based on past experience (in a different sense to a declarative expectation). Perseveration produces a further direct drive ($\boldsymbol{q}^P$) towards the most recently selected action.

$$C = s \cdot G$$
$$q^G = C \cdot v$$
$$q^H = s \cdot H$$
$$a = \beta^G q^G + \beta^H q^H + \beta^P q^P$$
$$p = \text{softmax}(a) = \frac{e^a}{\sum e^a}$$

For constrained choices, such as two-choice trials not permitting a "non-response" action, the softmax was calculated across valid responses only (with $\boldsymbol{p}_a = 0$ for all actions not permitted).

The goal-directed system learned as follows. Instrumental contingency learning was driven by calculating an outcome prediction error $\boldsymbol{d}^O$ for all outcomes, as the difference between obtained outcomes (***o***) and predicted outcomes (action–outcome contingencies for the chosen action: $\boldsymbol{C}_{a,*}$). A–O contingencies predicted for the chosen action by stimuli currently present ($\boldsymbol{G}_{s,a}$) were then updated using this prediction error:

$$d^O = o - C_{a,*}$$
$$\Delta G_{s,a} = \alpha^O d^O$$

A more general form might include instrumental incentive learning in which values for obtained outcomes are changed ($\Delta \boldsymbol{v_o}$) according to an outcome value error (d^V, the obtained reinforcement *r* minus the total value predicted for the obtained outcomes, $\boldsymbol{o} \cdot \boldsymbol{v}$) and a learning rate α^V:

$$d^V = r - o \cdot v$$
$$\Delta v_o = \alpha^V d^V$$

In the situation of a single outcome and $\alpha^V = 1$, this reduces to direct assignment of the reinforcement value to the obtained outcome. However, in the present task and model, the situation was even simpler: outcomes values were directly instructed, and so α^V was not considered.

Habit learning was as follows. S–R associations between stimuli present (***s***) and the action performed (*a*) were updated according to the reinforcement prediction error d^H, the difference between the reinforcement obtained (*r*) and the reinforcement predicted by the chosen action ($\boldsymbol{q}^H{}_a$):

$$d^H = r - q_a^H$$
$$\Delta H_{s,a} = \alpha^H d^H$$

Specific implementation for the slips-of-action task

We modelled data from both relevant phases of the original task (Ersche et al., 2016) (phase A, appetitive learning, and phase C, slip-of-action responding following outcome devaluation). We did not model the outcome–action contingency assessment (phase B), which did not not involve reinforcement feedback. We did not analyse the control task in phase D (responding to discriminative stimuli that themselves were or were not "devalued").

Reinforcing outcomes were given a notional and arbitrary value of +5 points. As described above, outcome devaluation was represented by direct instantaneous instruction in the model, reflecting the direct instruction in the slips-of-action task ("these animals are sick, avoid them"); these outcomes were temporarily devalued to –5 points. As feedback was not provided in this phase of the task and the behavioural task was framed to avoid learning, the goal-directed system was prevented from learning during these trials. The task was framed as a go/no-task and choosing the "originally correct" side was scored as a "response" (as per Figure 1C[right] of Ersche et al. (2016)). Choosing the other side, or not acting at all, was scored as a "non-response". A goal-directed subject will respond less when the relevant outcome is devalued; a habit-based subject will not alter its behaviour.

We constrained the general computational model further, via the restriction $\alpha^G = \alpha^H$, as the behavioural task did not permit differential assessment of the learning rates of instrumental and habitual systems (such that separate alpha values would lead to overfitting and did so in pilot modelling); different contributions of the two systems are therefore reflected primarily in β^G and β^H.

The behavioural task, involving explicit instructions to humans, is ambiguous as to whether it would lead to ongoing habit learning during the test phase (arguments might include: responding is to the same manipulanda, so ongoing habit learning might be expected; or, the instruction change sets up a different context, sharply altering the stimuli participating in S–R learning). Consequently, we tested two versions of the task: in one, habit learning was assumed to continue during the slips-of-action phase ("habit learning at test", HLAT); in another, it was assumed that no further habit learning occurred ("no habit learning at test", NHLAT).

Bayesian hierarchy and simulations

The group-level structure of the Bayesian model was as before, with a per-group mean and a common intersubject standard deviation for each parameter.

Priors were as shown in Table D3.

Table D3: priors for the two-system instrumental computational model.

Parameter	Prior for parameter	Prior for intersubject standard deviation
α	Beta(1.2, 1.2)	Half-normal(0, 0.17)
β^G, β^H	Gamma(shape = alpha = 4.82, rate = beta = 0.88)	Half-normal(0, 2)
β^P	Normal(0, 1)	Half-normal(0, 2)

We used a variational Bayes approximation to obtain posterior parameter distributions, via Stan's ADVI [automatic differentiation variational inference] algorithm.

Results

Results are shown in Table D4. The behavioural task was ill-specified as to whether further S–R learning would occur in the outcome devaluation test phase, and this had a potentially important impact on the assessment of learning rates in the two-system model: if learning was assumed to occur, the model suggested faster learning in the CUD group, and if not, it suggested slower learning. Interpretation of learning rate from this model is therefore more complex (see main text for discussion) but the results reflect slower learning in the first phase and a likely confound between the effects of outcome devaluation and those of extinction in measuring the effect of learning in the second phase. The models were consistent, however, in showing a reduced impact of the goal-directed action system (lower β^G) in the CUD group; no difference in the impact of the habitual system (no difference in β^H); and a greater tendency to perseverate (β^P) (or, strictly, a lesser tendency to avoid a recently chosen option, since β^P estimates were negative). Interparameter scatterplots are shown in Figure D3**.**

Table D4: results for the two-system instrumental computational model.

Condition	Parameter	Control group (posterior mean and 95% HDI)	CUD group (posterior mean and 95% HDI)	CUD – control (posterior mean and 95% HDI); **bold** indicates an HDI excluding zero (posterior probability >95% of a difference between groups)
No S–R learning during assessment phase ("no habit learning at test"; NHLAT)				
	α	0.100 [0.090, 0.111]	0.041 [0.037, 0.045]	**–0.059 [–0.070, –0.049]**
	β^G	0.323 [0.297, 0.351]	0.280 [0.247, 0.316]	**–0.043 [–0.083, –0.001]**
	β^H	0.004 [0.002, 0.007]	0.008 [0.004, 0.013]	0.004 [–0.001, 0.010]
	β^P	–0.215 [–0.251, –0.172]	–0.024 [–0.059, 0.011]	**0.190 [0.138, 0.240]**
S–R learning allowed during assessment phase ("habit learning at test"; HLAT)				
	α	0.035 [0.031, 0.038]	0.096 [0.085, 0.109]	**0.062 [0.050, 0.074]**
	β^G	0.653 [0.616, 0.693]	0.182 [0.166, 0.199]	**–0.471 [–0.514, –0.433]**
	β^H	0.002 [0.001, 0.004]	0.001 [0.000, 0.003]	–0.000 [–0.002, 0.002]
	β^P	–0.255 [–0.293, –0.215]	–0.096 [–0.134, –0.057]	**0.159 [0.105, 0.213]**

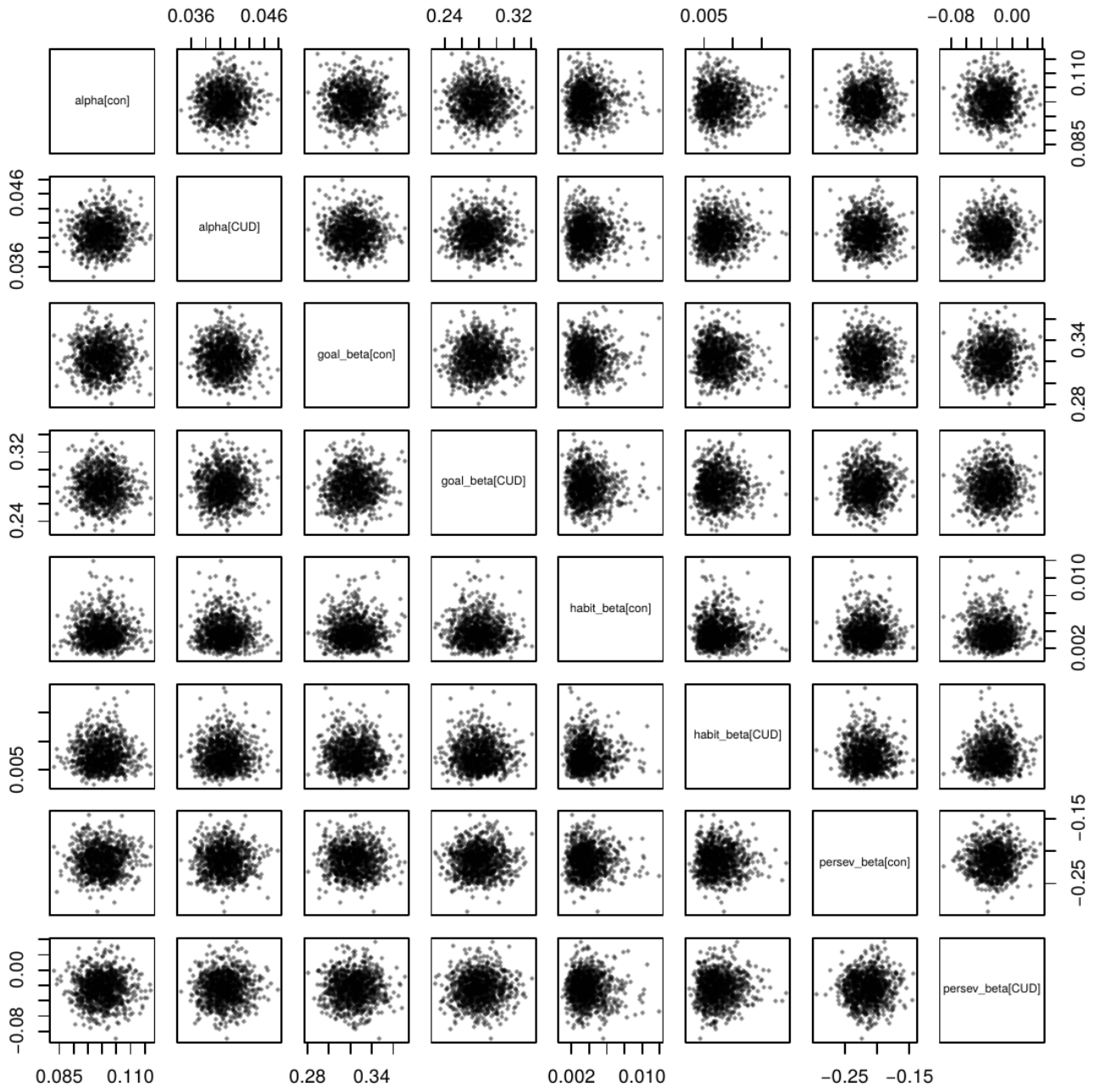


Figure D4: correlations between group-level parameter estimates from the two-system instrumental NHLAT model. Each point represents an iteration of the simulation.

Chapter 6: Assessment of goal-directed and habitual tendencies in cocaine use disorder via self-report

6.1 Introduction

Substance use disorder is linked with an imbalance in behavioural control, where habitual control is favoured over goal-directed control. Conventional methods of studying habits, such as outcome devaluation and contingency degradation, have found supporting evidence for a habit bias in cocaine use disorder, which explains, to some degree, the insensitivity to the consequences of their actions (Ersche et al., 2016, 2021). However, to what extent are patients' daily behaviours affected by habit biases, or a reduced goal-directed system? This chapter attempts to explore these constructs using self-reported questionnaires, which provide the opportunity to reflect on their day-to-day behaviours.

Although goal-directed and habit systems theoretically control instrumental action, individual differences in personality could facilitate these systems. Take goal-directed actions as an example. A more goal-oriented person would likely display higher tendencies for goal-directed actions, as they are more motivated by goals, even when said action is difficult or lacks any short-term benefits. The notion of working towards a prospective goal is consistent with the principles of goal-directed actions, whereby actions are driven by the final outcome (de Wit & Dickinson, 2009). A particularly relevant tool that captures this goal pursuit tendency is the Habitual Self-Control Questionnaire (HSCQ; Schroder et al., 2013). If goal-directed control is reduced in cocaine users, it is logical to expect that intentions or outcomes have a reduced influence in motivating actions, especially when the task at hand is difficult.

Much like how humans differ in the extent to which goals modulate their behaviour, they also vary in their proneness to develop habits. Some people identify themselves as 'creatures of habit' and find comfort in familiar environments and stable routines, whereas others hate sticking to rigid schedules and like to vary their day-to-day behaviour. It is difficult to capture habits with self-report methods due to their highly personal and unconscious nature (Robbins & Costa, 2017). However, it is possible to enquire about some aspects of habits such as routine behaviours and tendency to automatise certain actions (e.g. autopilot). These traits are captured by the Creature of Habit Scale (COHS), a self-report instrument developed to gauge habitual

engagement in everyday life (Ersche et al., 2017, 2019). The subscales of the COHS differentiates between propensity for automatic actions when entering a specific environment despite the lack of intentions (the automaticity subscale), and volitional actions done regularly for functional purposes (the routine subscale). If there is a habit predominance in behavioural control, patients with substance use disorder should demonstrate a predilection in engaging with habits in daily life, especially actions that were rewarded in the past.

The primary aim of this chapter is to determine the individual differences in habitual and goal-oriented traits with self-reported measures such as the HSCQ and the COHS. It is noted that behavioural measures might measure different constructs compared to self-reported measures (Allom et al., 2016; B. Saunders et al., 2018), so this chapter also tests whether these self-reported measures are related to experimental measures of goal-directed and habitual behaviours. Accordingly, I assess the relationship between these self-reported measures and behavioural data on goal-directed (i.e. reinforcement learning task) and habitual actions (i.e. contingency degradation task). These behavioural data were reported in Chapter 3 (study 1) and Ersche et al (2021) respectively. Based on the habit theory of addiction, I hypothesised that individuals with cocaine use disorder display increased habitual tendencies and reduced readiness for goal-directed actions. In the context of this study, I predicted that these tendencies would manifest in the form of reduced scores in the COHS subscales, and increased scores in the HSCQ scale respectively.

6.2 Methods

6.2.1 Sample description

The sample consisted of 48 male patients who met the Diagnostic and Statistical Manual (5th version; DSM-5) criteria for cocaine use disorder (CUD) and 42 male controls without a personal history of substance use disorder. Data from these participant were collected via in-person assessment as part of a larger study, some of which were reported in Chapter 3 (study 1) and Chapter 4 of this thesis, as well as in Ersche et al (2021). The cocaine group used cocaine for an average of 13.4 years (±SD: 7.7 years) and reported high levels of compulsive drug use, as indexed by the Obsessive-Compulsive Drug Use Scale (OCDUS mean total score: 34, ±SD: 10); the control group did not have any prior history of substance abuse, as indexed by the low scores on the Drug Abuse Screening Test (mean: 0.1, ±SD: 0.3).

6.2.2 Self-reported measures of goal-directed and habit tendencies

Habitual Self-Control Questionnaire (HSCQ): This 14-item scale measures the strength of persistent goal pursuit, a trait that reflects intrinsic drive to achieve planned outcomes, even under difficult circumstances. Statements describing intentions and commitments to complete difficult tasks (e.g. "I find it easy to motivate myself even if I do not enjoy a task at all.") are rated on a 5-point Likert scale (1 = disagree strongly, 5 = agree strongly). A higher total score, calculated by summing the response from all items, reflects stronger commitments to strive for goals. Questionnaire is available in Appendix E.

Creature of Habit scale (COHS): This is a 27-item self-reported questionnaire assessing proneness for habitual behaviours in daily life. Participants were required to rate each statement on a 5-point Likert scale, ranging from strongly disagree (1) to strongly agree (5). This questionnaire consists of two subscales: a routine subscale that captures preferences for regular rituals or routines (e.g. "I like to park my car or bike always in the same place."), and an automaticity subscale, which reflects proneness to eliciting automatic behaviours when in an associated environment – the central tenet of stimulus-response habits (e.g. "I often find myself finishing off a packet of biscuits just because it is lying there."). Higher scores on the automaticity and routine subscales reflect higher tendency for cue-driven behaviours and preference for routine, respectively. As reported in Ersche et al. (2016), the Mokken Homogeneity modelling and confirmatory factor analysis both converged, suggesting that the COHS exhibited good construct validity. The automaticity (ω=0.91, α=0.86) and routine subscales (ω=0.92, α=0.89) showed good reliability, as evidenced by their high McDonald's omega and Cronbach's alpha. Questionnaire is made available in Appendix E.

6.2.3 Behavioural measures for goal-directed and habitual actions

Reinforcement Learning Task: This task assesses learning from financial consequences of one's choices, a suitable measure for goal-directed learning. Detailed description of this task and its results were elaborated on in methods section in Chapter 3 (study 1). In brief, participants need to, by trial-and-error, learn to choose a stimulus that maximizes their financial gains whilst minimizing their financial losses. The primary measures used from this task are the accuracy scores for task performance and the latent RL parameters derived from computational modelling, which include parameters that account for value-driven (all learning

rates and reinforcement sensitivity) and non-value-driven processes (perseveration to stimulus and location). CUD patients in this sample are impaired on overall task performance, which were driven by reduced learning rate from negative feedback and reinforcement sensitivity parameters.

Contingency Degradation Task: Experimentally, habits are operationally defined as the absence of goal-directed actions. The contingency degradation task is one such task that tests habit strength by disrupting action-outcome relationships, a pre-requisite for goal-directed actions (Balleine & Dickinson, 1998). In this task, participants are over-trained to instrumentally respond for financial rewards, thereby developing an action-outcome contingency. After establishing an action-outcome contingency, this contingency is then degraded such that participants receive free rewards irrespective of whether they responded. In other words, instrumental responses are no longer required for financial rewards. If participants continued responding even though actions are rendered unnecessary, these actions are deemed habitual as they are not driven by the causal relationship between actions and outcomes. The primary measure from this task is the goal-to-habit ratio, calculated as the ratio of response under fully contingent condition (non-degraded) versus non-contingent condition (fully degraded). A lower value reflects elevated bias towards habitual responding. Data from this behavioural task have been analysed and reported in Ersche et al (2021), which showed that CUD patients had significantly lower goal-to-habit ratio values, suggesting that their responses are more habitual (i.e. less affected by the disruption of action-outcome contingency).

6.2.4 Statistical analysis

The primary measures are the propensity for habits and goal-directed actions, as indexed by the COHS subscales (routine and automaticity) and HSCQ total score respectively. Group differences were assessed using ANCOVA models with group as a between-subject factor and demographic variables that differed across groups entered as covariates; Pearson's correlation analyses were used to identify relationships between self-reported measures, behavioural measures, duration (years of cocaine use) and severity of drug use (OCDUS score). Given the current sample size of this study, the minimum effect size detectable is 0.60 (Cohen's d).

6.3 Results

Both groups did not differ in age (t_{88}=-0.08, p=0.936), but significantly differed in years of education (group means denoted as M) ($M_{control}$=15.7 years [±SD: 2.6], M_{CUD}=11.0 years [±SD: 1.5], $t_{65.1}$=10.5, p<0.001) and verbal IQ ($M_{control}$=116 [±SD: 6.1], M_{CUD}=103 [±SD: 7.2], t_{80}=8.3, p<0.001). An ANCOVA model with group as a between-subject factor, and years of education entered as a covariate revealed that CUD patients had significantly increased levels of COHS automaticity ($M_{control}$=30.3 [±SD: 7.3], M_{CUD}=39.2 [±SD: 7.0], $F_{1,87}$=20.9, p<0.001), and reduced goal pursuit tendency, as indicated by reduced HSCQ total score ($M_{control}$=51.8 [±SD: 5.6], M_{CUD}=43.5 [±SD: 8.5], $F_{1,87}$=6.43, p=0.013) (Figure 6.1A). CUD patients engaged in similar levels of routine behaviour to controls ($M_{control}$=54.5 [±SD: 10.3], M_{CUD}=54.3 [±SD: 9.8], $F_{1,87}$=0.22, p=0.640). As an additional check, I ran a similar ANCOVA model but with verbal IQ as the sole covariate – this did not change the main results. Within these patients, duration of cocaine use was positively correlated with automaticity levels in the cocaine group (r=0.334, p=0.020; Figure 6.1B), but were not related to either routine behaviour (r=0.230, p=0.116) or goal pursuit (r=-0.146, p=0.321). Associations between compulsive cocaine use and automaticity (r=0.250, p=0.087), routine (r=-0.133, p=0.368) and goal pursuit (r=-0.231, p=0.114) were also not significant. I assessed the relationships between self-report and behavioural measures separately in control and cocaine groups. However, there were no significant relationships found. Specifically, neither the HSCQ scores nor COHS automaticity were significantly correlated with any of the relevant behavioural measures. Table 6.1 reports the correlations coefficients for these relationships.

Table 6.1: Correlations between self-report and behavioural measures by group.

	HSCQ total score				COHS automaticity			
	Control		CUD		Control		CUD	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*
RL total accuracy	0.104	0.524	0.017	0.914	-0.021	0.897	0.198	0.197
RL parameters								
Learning rate from reward	0.107	0.510	0.133	0.390	0.003	0.987	0.129	0.405
Learning rate from punishment	0.092	0.574	0.005	0.976	-0.123	0.450	0.095	0.541
Learning rate from non-reward	0.085	0.602	0.062	0.689	0.028	0.866	0.113	0.466
Learning rate from non-punishment	-0.229	0.155	-0.270	0.076	0.295	0.064	0.006	0.972
Reinforcement sensitivity	0.246	0.127	0.105	0.498	-0.138	0.396	0.094	0.546
Perseveration (location)	-0.057	0.726	-0.096	0.537	-0.058	0.722	0.147	0.341
Perseveration (stimulus)	0.095	0.562	0.028	0.858	0.116	0.474	-0.214	0.163
Contingency Degradation Task								
Goal-to-habit ratio	-0.235	0.162	-0.175	0.234	-0.068	0.669	-0.216	0.141

Note. CUD: cocaine use disorder; RL: reinforcement learning; HSCQ: Habitual Self-Control Questionnaire; COHS: Creature of Habit Scale.

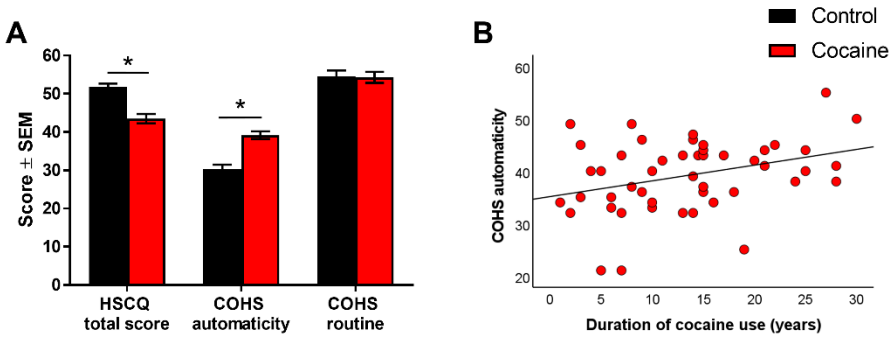


Figure 6.1: **Self-reported measures for goal-directed and habitual personalities. (A)** The cocaine group showed significantly elevated levels of automaticity, reduced levels of goal-oriented personality (HSCQ total score), but comparable levels of routine behaviour to the control group. **(B)** Increases in automaticity were positively related with duration of cocaine use. [HSCQ: Habitual Self-Control Questionnaire; COHS: Creature of Habit Scale; SEM: standard error of the mean; * denotes statistical significance at $p < 0.05$]

6.4 Discussion

Consistent with the behavioural findings, validated questionnaires of habitual and goal-directed traits identified that CUD patients had elevated automaticity, which increased as a function of cocaine use duration, and reduced tendency for goal-pursuit. These self-reported measures are also significantly correlated to the behavioural measures of goal-directed and habitual actions, demonstrating that these constructs overlap to some degree. Together these findings provide complementary evidence, in the form of self-report, that supports the habit theory of addiction.

Cocaine exposure is thought to disrupt the balance in behavioural control, either by increasing habit formation or reducing top-down goal-directed control or both (Vandaele & Janak, 2018). Studies showed that rodents exposed to cocaine over long periods had increased habits even in non-drug-related actions, such as food-seeking, in devaluation experiments (LeBlanc et al., 2013; Nordquist et al., 2007). By using an alternative self-reported approach, this study provides complementary evidence that in human CUD, environmental cues readily trigger automatic habits in daily life, suggesting an increased susceptibility for stimulus-driven actions. Indeed, prior work has shown that the automaticity subscale was jointly modulated by prior stimulant use and stressful life experience (Ersche et al., 2017), in line with the notion that exposure to stimulants and stress exacerbate habit formation in rodents (Corbit, Chieng, et al., 2014; Dias-Ferreira et al., 2009; A. Nelson & Killcross, 2006). Further, this tendency for automatic habits also increases as a function of cocaine use duration, which implies that the

habit construct is relevant to the chronicity of addiction, contrary to what has been argued (Hogarth, 2020). It is possible that cocaine use further exacerbates the tendency to engage habitually in life, but this needs further investigation. On the other hand, CUD patients also show a reduced tendency for goal-oriented behaviour, which indicates that intentions and goals are less likely to influence behaviour in human CUD – a recurring finding throughout this thesis. A recent study also reported an interesting disconnect between food choices and their perceived value in CUD patients, suggesting that at least in the appetitive domain, these patients' choices are not driven by subjective value, but by familiarity instead (Breedon et al., 2021). Originally designed to assess commitments to healthy lifestyle, the HSCQ scores have significantly predicted an increased cultivation of healthy behaviours that are related to self-control and persistence (e.g. exercise, dieting). Hence, the HSCQ score is thought to be a reasonable proxy of how strongly actions are driven by a prospective outcome (e.g. to lose weight) (Schroder et al., 2013).

The key strength of the current data is that behavioural and self-reported measures converge, thereby providing multi-layered evidence for impaired behavioural control. Although a positive relationship between years of cocaine use and self-reported automaticity suggests that increases in habitual engagement are related to cocaine use, the current study cannot definitively confirm whether increased self-reported automaticity and reduced goal pursuit precedes or follows substance use disorder development. However, the current study has some limitations. As this is a male-only sample, these results may not be generalisable to female users. Additionally, although this study is sensitive enough to detect moderate effect sizes, given its sample size, it might be under-powered to detect the associations between behavioural and self-report measures when analysed separately by group. Thus, the data should be interpreted with this in mind. Another limitation of this study is that both groups were not matched on education level (and IQ). But, covarying for these variables did not change the main results of this chapter. This suggests that, at least in this dataset, education levels (and IQ) did not have a noticeable effect on self-report measures of habitual and goal-directed tendencies. Moreover, the generalisability of this theory to other substance users remains unclear. Whilst there is an abundance of evidence for psychostimulants, surprisingly little consensus has been reached with other substances. The next chapter focuses on testing the habit theory of addiction in a community sample characterised by harmful alcohol use. Nevertheless, current findings

provide ancillary evidence for the hypothesised imbalance between goal-directed and habitual processes in CUD that could also be measured via self-report.

Appendix E: Supplementary materials to Chapter 6

The Habitual Self Control Questionnaire (Schroder et al., 2013)

HSCQ

Date: _____ ID: _____

Everyone has plans, intention and goals that they would like achieve. Although the goals may vary considerably depending on the situation, each person generally has an **attitude or way they approach things that determines how strongly they tend to stick to their intentions and carry things through**. Please tick the box that best reflects how much each statement corresponds to your own attitude.

	I disagree strongly	I disagree a bit	I neither agree nor disagree	I agree a bit	I agree strongly
1. I usually succeed in translating good intentions into action.	☐	☐	☐	☐	☐
2. I always tackle important and difficult tasks without delay.	☐	☐	☐	☐	☐
3. It would be easy for me to adopt a new habit such as doing exercise every day.	☐	☐	☐	☐	☐
4. When a goal requires controlling my behaviour over a long period of time, I tend to give up after a while.	☐	☐	☐	☐	☐
5. When I fail in a matter of some importance to me, I stick to my guns and try even harder than before.	☐	☐	☐	☐	☐
6. If something important to me turns out to be quite difficult, I just persist in my efforts.	☐	☐	☐	☐	☐
7. When I have made up my mind to complete an unpleasant task, nothing can stop me from doing so.	☐	☐	☐	☐	☐
8. Even if I am really determined to do so, I often have difficulty rejecting a tempting offer.	☐	☐	☐	☐	☐
9. When there is an opportunity to do something more enjoyable, my previous good intentions are usually lost.	☐	☐	☐	☐	☐
10. I find it easy to motivate myself even if I do not enjoy a task at all.	☐	☐	☐	☐	☐
11. Being rational and self-controlled does not seem to fit my way of life.	☐	☐	☐	☐	☐
12. If I am convinced that completing a task is really important and worth the effort, I feel that I have a lot of willpower.	☐	☐	☐	☐	☐
13. When there are more obstacles to overcome than I had expected, I tend to abandon the idea.	☐	☐	☐	☐	☐
14. I was quite successful in controlling unwanted habits in the past.	☐	☐	☐	☐	☐

Thank you!

The Creature of Habit Scale (page 1/3) (Ersche et al., 2017)

COHS

Date: _____ ID: _____

Here are some statements relating to behaviours, feelings, or preferences that some people may have. Please indicate the extent to which you agree with each statement with regard to yourself. Please answer honestly, as there are no right or wrong answers.

1. I like to park my car or bike always in the same place.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
2. I generally cook with the same spices / flavourings.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
3. When walking past a plate of sweets or biscuits, I can't resist taking one.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
4. I tend to go to bed at roughly the same time every night.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
5. I often take a snack while on the go (e.g. when driving, walking down the street, or surfing the web).				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
6. I quite happily work within my comfort zone rather than challenging myself, if I don't have to.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
7. I tend to do things in the same order every morning (e.g. get up, go to the toilet, have a coffee...).				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
8. Eating crisps or biscuits straight out of the packet is typical of me.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
9. Whenever I go into the kitchen, I typically look in the fridge.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
10. I always try to get the same seat in places such as on the bus, in the cinema, or in church.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree

The Creature of Habit Scale (page 2/3)

11. I often find myself finishing off a packet of biscuits just because it is lying there.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
12. I normally buy the same foods from the same grocery store.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
13. I rely on what is tried and tested rather than exploring something new.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
14. I generally eat the same things for breakfast every day.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
15. I tend to like routine.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
16. I usually treat myself to a snack at the end of the workday.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
17. In a restaurant, I tend to order dishes that I am familiar with.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
18. I am one of those people who get really annoyed by last minute cancellations.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
19. I often find myself eating without being aware of it.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
20. I usually sit at the same place at the dinner table.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
21. I often find myself running on 'autopilot', and then wonder why I ended up in a particular place or doing something that I did not intend to do.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
22. I always follow a certain order when preparing a meal.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree

The Creature of Habit Scale (page 3/3)

23. Television makes me particularly prone to uncontrolled eating.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
24. I tend to stick with the version of the software package that I am familiar with for as long as I can.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
25. I often find myself opening up the cabinet to take a snack.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
26. I am prone to eating more when I feel stressed.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree
27. I find comfort in regularity.				
❒ strongly disagree	❒ mildly disagree	❒ undecided	❒ mildly agree	❒ strongly agree

Thank you!

Chapter 7: Goal-directed and habitual control in problematic alcohol use

7.1 Introduction

Alcohol is one of the most commonly used substances, with an estimated 83% of the population reported recent alcohol consumption in the UK alone (R. Burton et al., 2017). Whilst occasional drinking has some positive aspects, most notably in facilitating social interactions and relaxation, excessive drinking is consistently linked with poor health outcomes and social as well as economic burden to society at large (World Health Organization, 2018). Recent UK estimates suggest that over 10 million people drink alcohol at harmful levels (Public Health England, 2017), which puts them at a higher risk in developing alcohol use disorder, a condition characterised by a loss of control over their drinking (American Psychiatric Association, 2013). However, not everyone who drinks large amounts of alcohol develop alcohol use disorder. In that respect, behavioural markers that predict the transition from harmful drinking habits into alcohol use disorder would be useful for early interventions. One such candidate may be the regulation of instrumental actions. Problematic alcohol use is hypothesised to disrupt goal-directed and habitual control processes, such that behaviours are biased towards habitual control after prolonged alcohol use (Barker et al., 2015; Corbit & Janak, 2016). However, the question of whether problems in regulatory control exist already in a community population that drinks hazardously, but not formally diagnosed with alcohol use disorder, remains elusive. This chapter aims to test this hypothesis in a UK population characterised by harmful drinking.

Animal studies provided widespread support for disrupted regulatory control after extended exposure to alcohol. Rats that were chronically exposed to ethanol demonstrate reduced sensitivity to outcome devaluation, which is indicative of habitual control over seeking responses (Corbit et al., 2012; Corbit, Nie, et al., 2014; Hopf et al., 2010; Lesscher et al., 2010; Lopez et al., 2014; Mangieri et al., 2012). Habitual seeking responses only developed following extensive alcohol exposure (Corbit et al., 2012), but these habits occur irrespective of training with sucrose or alcohol rewards, which implies an overall increase in habitual control that extends beyond alcohol (Corbit, Nie, et al., 2014; Lopez et al., 2014). Whether this overall shift towards habitual control is caused by an impaired goal-directed system or by an overall augmented habit system is still a matter of ongoing debate. There is some evidence to suggest that alcohol impairs brain regions implicated in goal-directed actions. A recent study showed

that ethanol administration attenuates orbitofrontal input to the striatum, a region that putatively mediates value and contingency learning (Renteria et al., 2018).

Although evidence for habitual predominance is compelling in the animal literature, findings from human studies have been equivocal. Extant research in humans has largely focused on patients with alcohol use disorder (AUD), most of which probed instrumental control with one of two paradigms. The first is the outcome devaluation paradigm, which generally tracks sensitivity to outcome values by devaluing certain outcomes. Two studies used this paradigm, each reported conflicting findings. Sjoerds and colleagues (2013) used the slips of action task (de Wit et al., 2007; de Wit & Dickinson, 2009), a version of a devaluation test which uses instructed devaluation. They found that patients with AUD had diminished sensitivity towards devalued outcomes, as indexed by persistent responding towards devalued outcomes. By contrast, van Timmeren et al. (2020) used taste aversion as their devaluation strategy but found that sensitivity to devaluation was intact in AUD. Such equivocal findings were also present when studying instrumental control with the second paradigm: the two-step task (Daw et al., 2011). The task is thought to simulate goal-directed and habitual processes in computational terms, enabling the dissociation of the two. Goal-directed control is formalised as a model-based process, which accurately tracks values and the transitions that lead to rewards, akin to sensitivity towards values and contingency; habitual control is formalised as a model-free process, which only learns the most rewarding response based on past reinforcement, and is more rigid, thus less sensitive to immediate value updates (Dayan & Daw, 2008; Dolan & Dayan, 2013). Some studies using this task found, in AUD patients and binge drinking cohorts, a reduced model-based control but intact model-free control, suggesting that an impaired goal-directed system but not an increased habit system in these individuals (Doñamayor et al., 2018; Sebold et al., 2014). However, other studies with larger samples sizes and more heterogeneous AUD patients found neither an impaired model-based, nor augmented model-free control over behaviour (Sebold et al., 2017; Voon et al., 2015). Moreover, a handful of studies did not find any evidence for a relationship between model-based or model-free control and the severity/chronicity of alcohol use (Doñamayor et al., 2018; Nebe et al., 2018; Patzelt et al., 2019), though there are exceptions (see Gillan et al. (2016)).

One of the challenges in studying instrumental control in humans is to simulate reliably goal-directed actions and habits in behavioural tasks. Whilst outcome devaluation and the two-step tasks are popular paradigms to study instrumental learning, these tasks have noteworthy limitations in approximating goal-directed and habitual processes. First, outcome devaluation procedures operationalise habits indirectly by the absence of goal-directed actions. As such, the main test for habits (i.e. devaluation phase) is designed to measure how likely the conditioned stimulus elicits habits when the associated outcome has lost its original value. This design can only test for the relative balance between, but not the contributions from, the two systems (Watson & de Wit, 2018). Consequently, an augmented habit system or an impaired goal-directed system should both lead to the same behaviour profile. Second, the two-step task defines goal-directed and habitual actions computationally with model-based and model-free processes respectively (Dolan & Dayan, 2013). This allows the dissociation of goal-directed and habitual processes, but the construct validity of these computational processes have recently been questioned in at least three ways: First, several studies found that increased habitual tendencies, as modelled by outcome devaluation tests, were not related to model-free processes in the two-step task, suggesting that these constructs do not overlap with one another, as previously assumed (Friedel et al., 2014; Gillan et al., 2015; Sjoerds et al., 2016). Second, the development of model-free control does not rely on extensive repetition, which is identified as a key prerequisite for the development of habits (Dickinson, 1985). Third, there are also conceptual differences between habits and model-free learning. Specifically, habits, once established, are thought to be elicited by environmental cues and have no direct relationship with outcome values. By contrast, model-free learning, even after long periods of learning, still depends on the expected outcome value, albeit less flexible than model-based processes (Miller et al., 2019). Thus, considering the shortcomings of existing behavioural tasks, there is a need for a novel behavioural paradigm that fractionates the nature of goal-directed and habitual processes in humans more clearly.

To address these issues, I co-developed a novel behavioural task – the goal-habit conflict task – that can model the expression of both goal-directed actions and habits directly. Goal-directed and habit systems usually work in conjunction to regulate behaviour, but in some cases, especially when both systems are placed in conflict, these systems can compete for expression (Balleine & O'Doherty, 2010; Bradfield & Balleine, 2013; Zwosta et al., 2018). Based on this aspect, our novel task challenges behavioural systems by creating a conflict between habits and

goal-directed responses. These conflicts would enable us to observe which system prevails. For instance, if the habit system is stronger, I would expect habitual responses made during this conflict (and vice versa). In the goal-habit-conflict task, participants first learn instrumental responses to a set of instructions (goal-directed actions), then acquires stimulus-response (S-R) habits through overtraining using monetary rewards and punishments. Once these responses have been learned, they are placed in conflict with one another, i.e. participants are told to follow instructional cues (goal-directed) in the presence of conditioned stimuli, which have previously elicited a habitual response. Goal-directed responses are defined as responses associated with the instructional cues; habits are defined as the responses consistent with learned S-R habits.

The aim now is to test whether behavioural control has shifted towards the habit system in participants who drink alcohol at harmful levels. Based on animal studies, it is hypothesised that chronic alcohol use facilitates the shift of behavioural control towards the habit system. If this hypothesis is true, I would predict that heavy drinkers exhibit more habitual responses during a conflict situation.

7.2 Methods

7.2.1 Sample description

I recruited participants with and without harmful alcohol use via Prolific Academic (https://www.prolific.co/) – an online crowdsourcing platform designed for academic research. For an online study, the possibilities to screen for psychiatric disorders and drug use were limited. Prolific Academic provides a screening tool to identify the targeted population, which I set as follows: (1) age range between 18 and 48 years, (2) both genders, (3) native English speaker, (4) UK residence, (5) no diagnosis of Dyslexia, Dyspraxia or Attention Deficit Hyperactivity Disorder (ADHD), including literacy difficulties, (6) no language-related disorders (e.g. aphasia), (7) no diagnosis of Autism Spectrum Disorder, and (8) no diagnosis of mild cognitive impairments or dementia. As the Prolific tool does not include alcohol use disorder, I decided to ask all eligible prolific users to complete the Alcohol Use Disorder Identification Test (AUDIT, (J. B. Saunders et al., 1993)), the Depression, Anxiety and Stress Scale (DASS-21, (Lovibond & Lovibond, 1995)) and to report their current medication. This allowed me to identify healthy individuals and chronic alcohol users. Chronic alcohol users

were participants who exhibit harmful levels drinking in the AUDIT (score > 10). By contrast, controls were only recruited if they drink socially; social use was defined as an AUDIT score of 6 or less. Participants were excluded if they had any lifetime exposure to stimulant drugs (e.g. cocaine, crack-cocaine, amphetamines and methamphetamine), or indicated that they had a history of addiction. Control participants were excluded if they reported any use of psychoactive medication (e.g. antidepressants, antipsychotics), or met the cut-off for subclinical levels of depression, anxiety or stress, as indexed by the DASS-21. Chronic alcohol users who use antidepressant were not excluded, but those who reported antipsychotic medication were excluded. The final sample consisted of 120 harmful alcohol users with a mean AUDIT score of 16.7 (±SD: 5.5) and 148 controls with an AUDIT score of 1.8 (±SD: 1.5). As I have argued in Chapter 6, individual differences in personality related to goal-directed and habitual actions may modulate behavioural performance. Thus, all participants also completed self-reported questionnaires of goal-directed and habitual tendencies, namely the Habitual Self Control Questionnaire (HSCQ, (Schroder et al., 2013)) and the Creature of Habit Scale (COHS, (Ersche et al., 2017)) respectively.

7.2.2 Goal-habit conflict task (The Fishing Expedition Task)

This is a novel behavioural task that measures preferences for goal-directed and habitual actions by placing them in conflict. This task adopted a cover story of a beginner fisher learning to sort various catches into different categories, and consisted of three stages:

Stage 1: Establishing goal-directed behaviour (Figure 7.1A). Goal-directed responses are deliberate responses made with a goal in mind (Balleine & Dickinson, 1998). In this stage, participants were trained to respond according to instructions. Participants were instructed to sort their catches (fishes or crabs) based on specific coloured cues (e.g. “orange accept” or “blue reject”; the colours were counterbalanced across all participants). On each trial, participants were first informed whether they had to accept or reject a catch, followed by one of two types of catches (fish or crab). Depending on the instruction and the catch, participants were required to respond with one of four keyboard buttons: “X” or “M” to accept fishes or crabs respectively, or “C” and “N” to reject fishes and crabs respectively (Figure 7.1A). Positive feedback was provided if participants sorted their catch correctly, negative feedback if they made an error. For each wrong response, that trial was repeated until participants made a correct response – this was to ensure that participants knew the instructed response. This

stage consisted of 120 trials and had 12 stimuli (6 fishes and 6 crabs), which were presented in random order.

Stage 2: Habit formation (Figure 7.1B). Habits are actions that, though initially learned with a reinforcement schedule, develops autonomy after extensive training over a prolonged period (Dickinson, 1985). Here, participants develop stimulus-response (S-R) habits via overtraining. Participants were told to sell their catches at a fish market, and should learn to maximise their earnings by gaining subsidies and avoiding taxes. Instrumental responses were over-trained either with rewarding (subsidy) or punishing (tax) feedback, thus facilitating the development of appetitive and avoidance habits respectively. On each trial, participants were shown a picture of their catch, labelled with the term "tax" or "subsidy". For subsidised catches, correct responses were reinforced by winning money (+50p), whereas wrong responses were not (+0p). By contrast, for taxed catches, correct responses led to the avoidance of money losses (-0p), whereas wrong responses led to money loss (-50p). Feedback in this stage was purely deterministic i.e. only one correct response for each fish or crab, and participants had to learn by trial-and-error which were the correct button presses (X, C, N or M). Participants were encouraged to respond quickly (< 2000 ms), as quick responses were met with larger gains (+50p instead of +10p) or smaller losses (-10p instead of -50p). This manipulation is meant to reduce tendency for conscious deliberation, which should promote habit formation (Watson & de Wit, 2018). This stage consisted of 240 trials and had 12 stimuli (6 tax, 6 subsidy). Of the 12 stimuli, 8 stimuli (4 fishes and 4 crabs) were re-used from stage 1, while 4 stimuli (2 fishes and crabs) were novel.

Stage 3: Goal-habit conflict (Figure 7.1C). Here, goal-directed responses from stage 1 and learned S-R habits from stage 2 were placed in competition with one another to assess which system dominates over overt behaviour. Participants were told to continue sorting their catches, as they did in stage 1, but would no longer receive trial-by-trial feedback on their responses. The trial design is similar to stage 1: participants first receive coloured cue instructions ("accept" or "reject"), followed by a picture of their catch (fish or crab), and they would need to respond with the appropriate button. However, stimuli from phase 2 were re-used here, such that the learned S-R habit from phase 2 could either be congruent or incongruent with the instructed response. In the congruent condition, the instructed response is consistent with the learned habit from phase 2. By contrast, the goal-directed and habitual responses differ in the incongruent trials (see Figure 7.1C for schematics). Incongruence between the two responses enables us to test which system prevails during situations of conflict. Additionally, there is also another trial

condition, termed free choice trials. In these trials, participants were shown a picture of their catch, which has a conditioned S-R habit from phase 2, but no instructional cues. In these trials, participants could respond with whichever key they preferred. The rationale for this condition is to elicit stimulus-triggered habits in the absence of any interfering goal-directed instructions. In all trials during this stage, participants were given only 2000ms to respond – time pressure is thought to facilitate the expression of habits (Watson & de Wit, 2018). Trials with missed responses were repeated at the end of each block. There were in total six trial conditions with 20 trials each: trials where the instructed response is congruent with previously rewarded response (approach congruent) or avoidance response (avoid congruent); trials where the instructed response is incongruent with previously rewarded (approach incongruent) or avoidance response (avoid incongruent); and free choice trials which showed stimuli linked with either appetitive (free choice approach) or avoidance S-R habits (free choice avoid). This phase comprised of 12 stimuli (6 fished and 6 crabs): eight stimuli were used in both phase 1 and 2 for the congruent and incongruent trials; the remaining 4 stimuli (2 fishes and 2 crabs) for the free choice trials were only used in stage 2. Examples of trials are shown in Figure 7.1C.

Post-task questionnaire: At the end of task, participants were assessed of their S-R and A-O knowledge by indicating via button presses the correct responses pertaining to each stimulus and outcome.

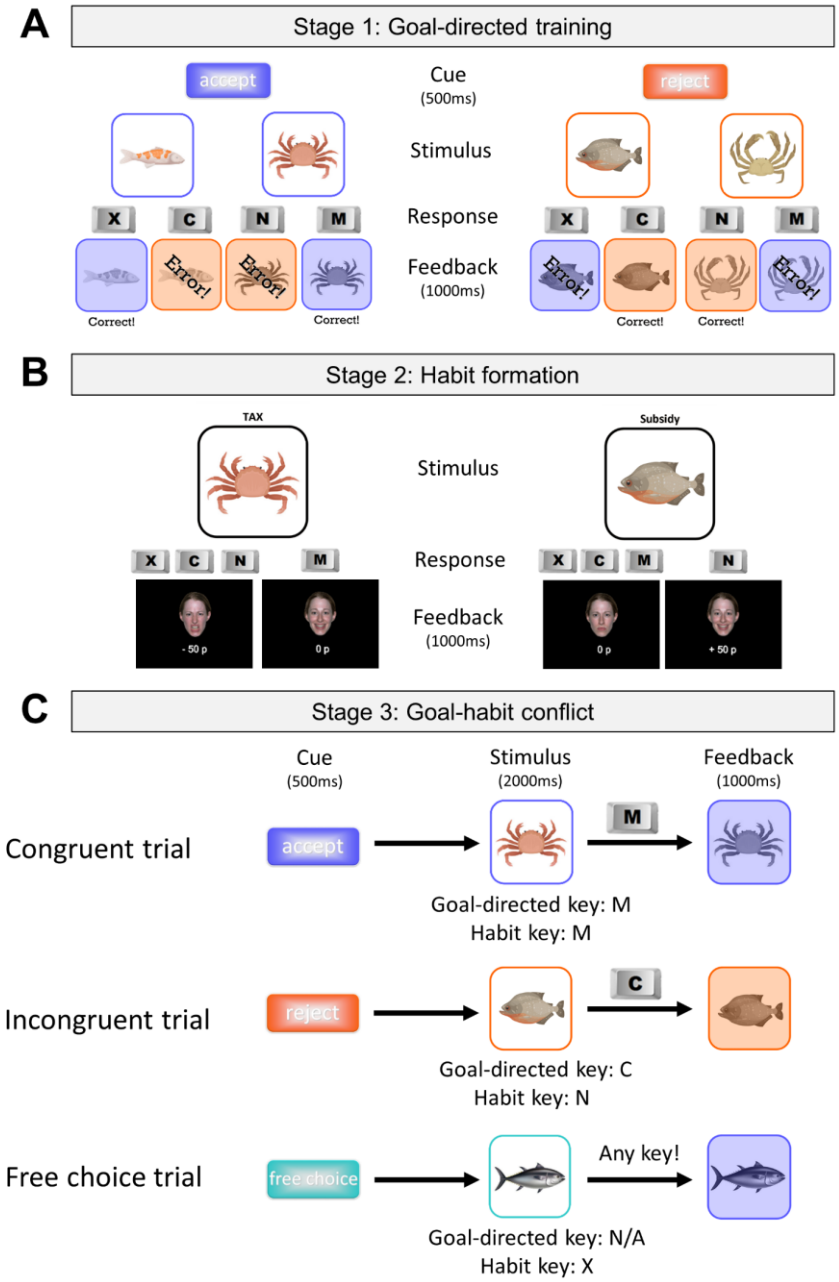


Figure 7.1: Schematics for the Goal-Habit Conflict Task. (A) In the first phase, participants learn to sort the fishes and crabs based on instructional cues. Correct responses are deterministic, and feedback is provided for each trial. **(B)** Participants are told that they will be selling their catches at the market, and need to learn the responses that would maximise their subsidies (reward) and minimise their taxes (loss). Thus, they need to learn by trial-and-error the correct response associated with each stimulus, which are deterministic. Quick responses are encouraged as this could increase their subsidies or reduce their taxes. **(C)** In the third phase, participants return to sorting fishes like in phase 1, but this time, they do not receive feedback on their responses. Here, instructed responses from cues (accept or reject) could either be consistent with learnt S-R habits (congruent trials) or differ from learnt S-R habits (incongruent trials), which allows the assessment of which system is stronger when there is a conflict. In addition, there is a free choice condition, where participants were presented with stimuli associated with an S-R habit, but without any instructional cues. Thus, participants can freely respond with any key.

7.2.3 Statistical analysis

I assessed the self-reported questionnaires and demographic measures for group differences with frequentist statistics. I also assessed the relationship between alcohol use severity (AUDIT score) and task performance.

I assessed goal-directed performance by computing the accuracy rate during the first phase, defined as the percentage of correct responses. A-O knowledge was also assessed in participants by computing the accuracy score for the post-task questionnaires. These measures were entered into an ANOVA with group as a between-subject factor.

To assess reinforcement learning performance in phase 2, block-by-block accuracy rates were computed in 24-trial bins. As learning from positive and negative feedback are thought to be dissociable processes, accuracy rates of both tax and subsidies conditions were evaluated separately. It is important for participants to acquire the correct contingencies for a fair interpretation of their performance during a goal-habit conflict. Thus, participants who performed at or below chance level during the last block (accuracy $\leq 25\%$) were excluded from the analyses. I also computed an accuracy score for explicit S-R knowledge. The accuracy scores for reinforcement learning and explicit S-R knowledge were entered into a mixed ANOVA with group as a between-subject factor, and block (blocks 1-5) and valence (tax versus subsidies) as within-subject factors. A reinforcement learning algorithm was also fitted to trial-by-trial learning performance to ascertain the latent parameters that underpin task performance. Parameters of interest in this phase include value-driven processes such as the learning rates from reward, non-reward, punishment, and non-punishment, reinforcement sensitivity; and non-value-driven process such as the tendency to perseverate – to account for repeated responses that are not driven by learned values. Parameter recovery for a similar model with these parameters was reasonably well, as demonstrated in Chapter 3 Appendix B (Figure B2). Parameter estimation and model fitting procedures followed those reported in computational methods section in Chapter 2.

Three primary measures were devised to reflect task performance in stage 3, namely switching score, number of goal-directed to habitual actions, and habit proneness. Each measure will be explained in turn below:

Switching scores: When situational demands change, it is imperative for the goal-directed system to take control over behaviour from the automatic habit system. The switching score reflects the ability to switch from habitual to goal-directed responding, which is computed as the difference in errors between trials with a goal-habit conflict (incongruent trials) and those without (congruent trials). I also computed a switching score for response times for correctly responded trials. If an individual is able to effectively switch between goal-directed and habitual responses, their switching errors and response times would be close to zero, as there would be little to no switch costs involved i.e. no difference between conditions. By contrast, a habitual person would find overriding a habitual response more effortful – this would result in a higher (more positive) switching error rates or response time when responding under conflict relative to baseline.

Number of habits/goal-directed responses: This measure tests the strength of the goal-directed and habit systems under conflict by counting the number of goal-directed and habitual actions made during the incongruent condition. A response is considered goal-directed when the participants' response is consistent with the instructions (i.e. correctly accepting or rejecting the stimulus); if participants instead respond with a learned response during phase 2, then that action is considered habitual, as it is not driven by the instructions, but instead by the learned S-R contingency. This measure was calculated separately for stimuli associated with approach and avoidance behaviours.

Habit proneness: The key feature of a S-R habit is that these well-learnt actions are automatically elicited by an associated cue. The habit proneness score measures this aspect by calculating the number of habitual responses made under the free choice conditions – when there is no need to deliberately select an instructed response. A habitual person would more readily elicit well-learnt responses when prompted with its associated cue, and thus would have a higher score. Again, this measure was calculated separately for approach and avoidance stimuli.

All measures were analysed with mixed ANOVA models, with group as the between-subject factor, and stimulus type (approach versus avoid) as the within-subject factor. For each habit measure, I also assessed its relationship with goal-directed learning (stage 1), reinforcement

learning parameters (stage 2) and individual differences in self-reported automaticity (COHS subscale). Post-hoc sensitivity power analysis indicated that this study is sensitive to a relatively small effect size (Cohen's d = 0.34).

7.3 Results

7.3.1 Sample characteristics and questionnaire data

Thirty-seven control participants (25%) and 27 alcohol users (23%) performed at or lower than chance level during the final learning block of stage 2, and were thus excluded from any subsequent analyses. The final sample consisted of 111 controls and 93 alcohol users. There was no statistically significant difference between the number of controls or alcohol users excluded (χ^2=0.2, p=0.622). The excluded group also did not differ with the resultant sample in terms of education level (χ^2=3.7, p=0.446) or AUDIT scores (t_{266}=-0.814, p=0.416), but the excluded group was older (t_{266}=2.2, p=0.026) and had more women (χ^2=4.5, p=0.034) than the final sample.

Sample demographics of the final sample are reported in Table 7.1. The groups were comparable on age, education levels and employment status, but there were more females in the control group than the alcohol group. Consequently, gender was included as a covariate for subsequent frequentist analyses. Fourteen individuals in the alcohol group (15%) considered treatment for their alcohol use, but these individuals did not differ with other alcohol users in terms of behavioural performance (all $p > 0.1$), and thus were not excluded from the analyses. Although the alcohol group showed higher levels of depression, anxiety and stress than controls, as measured by the DASS-21, these were not statistically controlled for, because (1) increased levels of depression, anxiety and stress often co-occur with increased alcohol use (Swendsen et al., 1998); (2) these measures are not significantly correlated with any task performance measures (all $p > 0.05$), and hence did not affect learning in this sample. However, whilst none of the control participants were on any psychoactive medication, 15 individuals from the alcohol group (16%) reported current use of antidepressants (5 sertraline, 5 citalopram, 1 fluoxetine, 2 mirtazapine, 1 imipramine, 1 duloxetine). These individuals also did not differ with other alcohol users in their behavioural performance and demographics (all $p > 0.1$), and were included in the main analysis. As reported in Table 7.1, the alcohol group showed

marginally increased automaticity, reduced goal directedness and comparable routine behaviour to controls.

Table 7.1: Sample demographics for Chapter 7.

	Mean (SD)		Statistics	
	Control	Alcohol	t / χ^2	p
Group size (n)	111	93	-	-
Age (years)	32.3 (7.0)	31.1 (9.5)	0.971	0.333
Gender (% male)	36	55	7.24	0.007
Education level (% completed)			7.75	0.101
Completed secondary school	14	10		
Completed Sixth form	15	16		
Started university	11	25		
Bachelor's degree	38	33		
Postgraduate degree	22	16		
Employment status (%)			0.18	0.669
Not in paid work	11	14		
Paid work / studying	89	86		
Considered treatment (n)	0	14	-	-
Alcohol use severity (AUDIT)	1.8 (1.4)	16.9 (5.4)	-28.4	< 0.001
Depression (DASS-21 subscore)	3.0 (3.3)	14.0 (11.5)	-8.97	< 0.001
Anxiety (DASS-21 subscore)	1.5 (2.1)	7.4 (7.1)	-7.83	< 0.001
Stress (DASS-21 subscore)	4.5 (4.7)	12.3 (9.2)	-7.98	< 0.001
Automaticity (COHS subscore)	31.6 (9.0)	34.0 (9.1)	-1.84	0.068
Routine (COHS subscore)	58.2 (9.6)	57.2 (9.1)	0.753	0.452
Goal-directedness (HSCQ)	49.9 (8.9)	42.3 (8.4)	6.23	< 0.001

Note. COHS: Creature of Habit Scale; HSCQ: Habitual Self-Control Questionnaire; AUDIT: Alcohol Use Disorder Identification Test; DASS-21: Depression, Anxiety and Stress Scale; SD: Standard deviation

7.3.2 Goal-directed learning (stage 1)

All participants demonstrated high levels of accuracy during action-outcome learning and adequate action-outcome knowledge during the post-task questionnaire (Figure 7.2). Importantly, task performance ($F_{1,201}$=2.6, p=0.107) and explicit knowledge ($F_{1,201}$=0.568, p=0.452) did not significantly differ between groups, indicating that action-outcome learning per se is not impaired in harmful alcohol users. Alcohol use severity was not correlated with either action-outcome learning (r=-0.100, p=0.155) or action-outcome knowledge (r=-0.029, p=0.685).

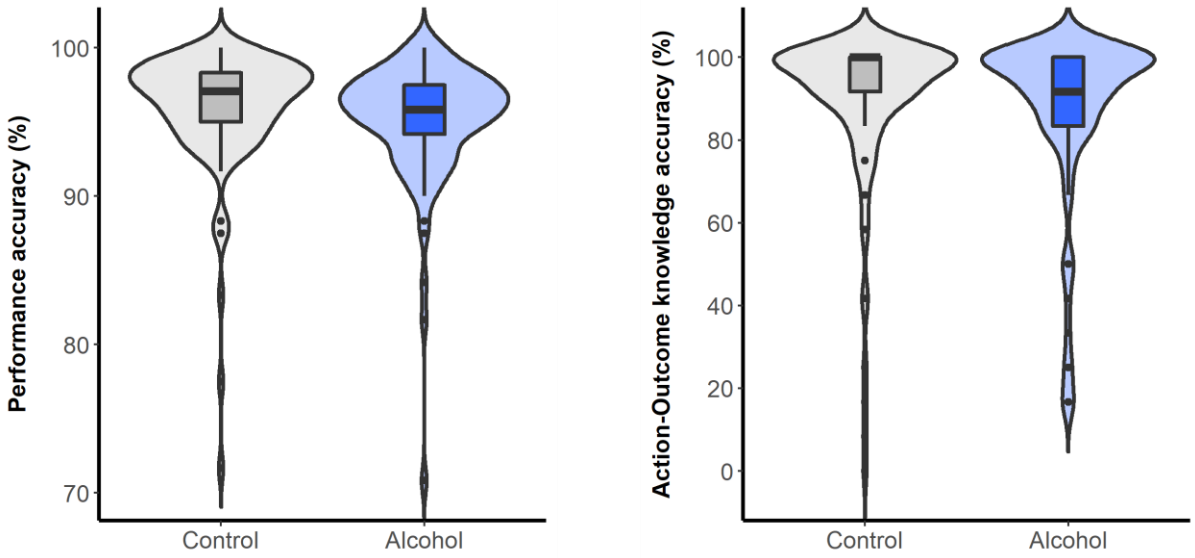


Figure 7.2: Goal-directed learning performance (stage 1). High levels of both performance accuracy and action-outcome knowledge indicate all participants learnt the responses associated with the instructional cues well. However, the groups did not significantly differ in their learning, indicating that action-outcome instructions were well learnt in both groups.

7.3.3 Habit formation (stage 2)

As shown in Figures 7.3A and 7.3B, all participants improved their performance over time ($F_{2.1,429.1}$=22.6, p < 0.001), and learned better in response to reward (subsidy) relative to punishing (tax) feedback ($F_{1,201}$=10.7, p=0.001). Both groups learned at similar speeds, as evidenced by a lack of group-by-block interaction ($F_{2.1,429.1}$=0.267, p=0.780). There were no effects of group ($F_{1,201}$=0.002, p=0.968) or group-by-valence interaction ($F_{1,201}$=1.21, p=0.307). All remaining interactions were also not statistically significant (all p > 0.1). Tests on explicit S-R knowledge also did not reveal any effects of group ($F_{1,201}$=0.64, p=0.800), valence

($F_{1,201}$=1.02, p=0.313) or interaction ($F_{1,201}$=0.188, p=0.665). Both task performance (r=-0.002, p=0.972) and explicit S-R knowledge (r=0.012, p=0.870) also did not correlate with AUDIT scores.

Computational modelling of reinforcement learning did not reveal any group differences on any learning parameters (Figure 7.3C), which means that the alcohol group was not impaired on any latent parameters of reinforcement learning. A multiple regression with all learning parameters as predictors and accuracy scores as the outcome variable revealed that the parameters accounted for 78% of the variance, suggesting that the model explains the data well. The AUDIT scores were not significantly related with any reinforcement learning parameters in either the control or the alcohol group (all $p > 0.05$).

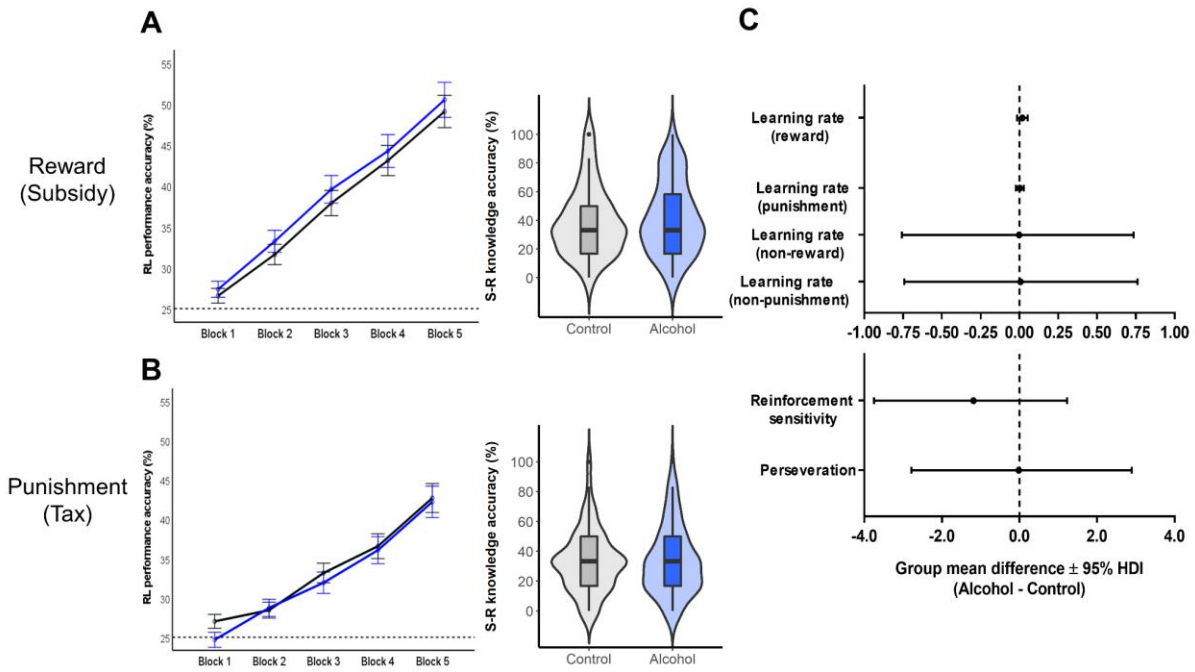


Figure 7.3: Reinforcement learning task performance and explicit S-R knowledge (stage 2). Both the alcohol and the control group did not differ in task performance and explicit S-R knowledge in both learning from rewarding (A) or punishing feedback (B). (C) Modelling of trial-by-trial reinforcement learning also did not reveal and significant impairments in the free parameters of individuals with harmful alcohol use. [Error bars in (A) and (B) denote standard error to the mean, whereas horizontal error bars in (C) denote 95% highest density intervals.]

7.3.4 Test for habit predominance (stage 3)

One-sample t-tests identified that the overall switching error rates did not significantly differ from zero for both control and alcohol group (both $p > 0.5$), suggesting that all participants generally were able to switch to goal-directed responding as instructed. A mixed ANCOVA with group (control versus alcohol), stimulus (approach versus avoidance) as factors and gender as a covariate revealed neither a main effect of group ($F_{1,201}$=0.038, p=0.845), valence ($F_{1,201}$=0.005, p=0.942) or group-by-stimulus interaction ($F_{1,201}$=0.03, p=0.861) on switching error rates (Figure 7.4A). The overall switching response time also did not significantly differ from zero (both groups $p > 0.5$), nor there were significant effects of group ($F_{1,201}$=0.54, p=0.464), valence ($F_{1,201}$=0.49, p=0.486), or group-by-stimulus interaction ($F_{1,201}$=0.70, p=0.404) on switching response times (Figure 7.4B). Switching error rates were positively correlated with goal-directed learning performance in stage 1 (r=0.184, p=0.008), but not with reinforcement learning performance in stage 2 (all $p > 0.1$); whereas switching response time was not correlated to any stage 1 or stage 2 measures (all $p > 0.1$). However, neither measures were related to self-reported automaticity, nor AUDIT scores (all $p > 0.05$).

In terms of habitual / goal-directed responses (Figures 7.4C,D), a mixed ANCOVA with the factors type (habit versus goal-directed behaviour), group (control versus alcohol) and stimulus (approach versus avoidance) showed that whilst participants' responses were overwhelmingly goal-directed ($F_{1,201}$=482, p<0.001), there were no statistically significant effects of group ($F_{1,201}$=1.19, p=0.278), valence ($F_{1,201}$=0.256, p=0.613), or group-by-stimulus interaction ($F_{1,201}$=1.24, p=0.267); all other effects were also not statistically significant (all $p > 0.2$). Within the overall sample, those who performed better during goal-directed learning in stage 1 were more likely to respond in a goal-directed manner (ρ=0.411, p<0.001) and less likely to elicit habitual responses (ρ=-0.233, p=0.001). Those with higher reward learning rates also had higher goal-directed responses (ρ=0.161, p=0.021); there were no other relationships with reinforcement learning parameters (all $p > 0.05$). Individuals with higher self-reported automaticity had lower number of goal-directed responses (ρ=-0.153, p=0.029), but were not related to the number of habits (ρ=0.076, p=0.204). The relationships between AUDIT scores and number of goal-directed actions (ρ=0.031, p=0.660) and habits (ρ=-0.047, p=0.501) were, however, not significant.

Similarly, the number of habitual responses under free choice conditions also did not differ between stimuli ($F_{1,201}$=1.97, p=0.162) or groups ($F_{1,201}$=0.244, p=0.622; Figure 7.4E). There was also no evidence for an interaction effect ($F_{1,201}$=0.098, p=0.754). Those who performed better during goal-directed learning were less likely to make habitual responses (ρ=-0.152, p=0.030); interestingly, the reinforcement sensitivity parameter from stage 2 was positively related to this habit measure (ρ=0.179, p=0.010). The total number of habitual responses was positively correlated with self-reported automaticity but only in the alcohol group (ρ=0.258, p=0.013; Figure 7.4F), not in the control group (ρ=-0.154, p=0.108); comparison between these two correlations with Fisher's Z-transformation identified a statistically significant difference between the alcohol and control groups (Z = 2.94, p=0.003). Additionally, the total number of habitual responses was not significantly associated with the AUDIT scores in both control (ρ=0.014, p=0.886) and alcohol groups (ρ=-0.014, p=0.896).

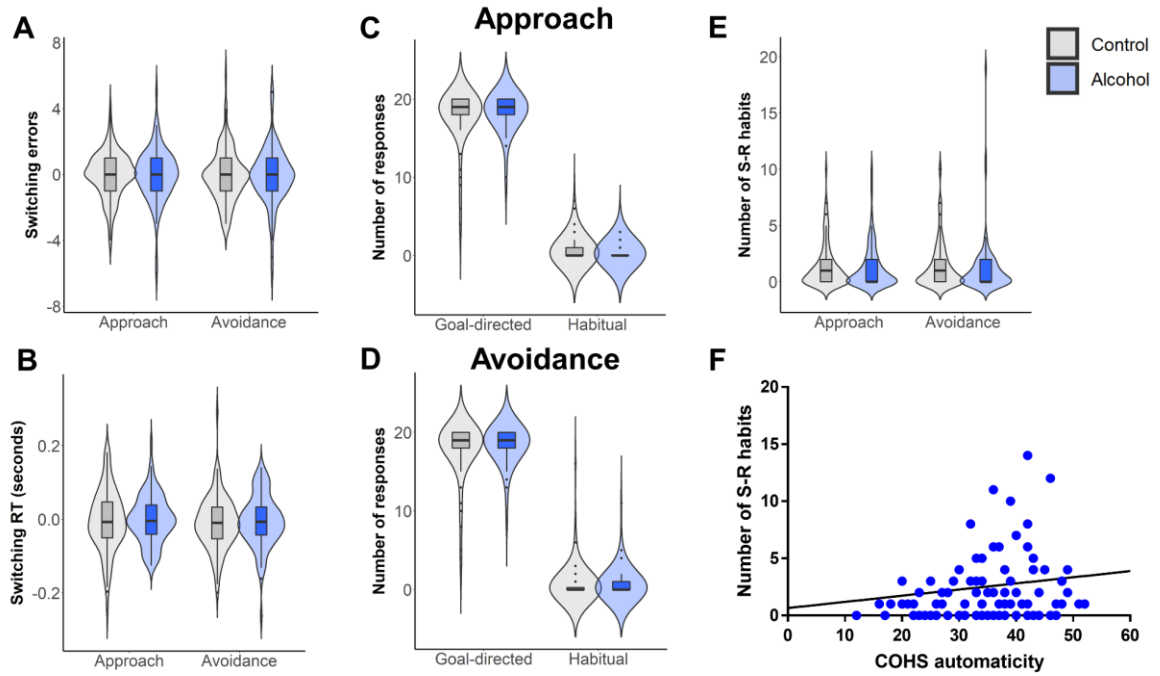


Figure 7.4: Task performance during the goal-habit conflict (stage 3). (A, B) Switching scores showed that errors and response time (RT) not significantly differ between groups (control versus alcohol) or condition (approach versus avoidance). (C, D) Both control and alcohol participants responded to the incongruent trials in a goal-directed manner, irrespective of approach or avoidance stimuli. (E) The groups also did not differ in terms of freely elicited S-R habits during the free choice trials. (F) The number of stimulus-response (S-R) habits elicited during free choice conditions is positively correlated with self-reported automaticity in harmful alcohol users. [all error bars denote one standard error to the mean]

7.4 Discussion

Changes to the regulatory systems of goal-directed and habitual actions are thought to underpin altered behaviour following chronic alcohol use (Corbit & Janak, 2016). I tested this hypothesis with a task that could directly measure goal-directed and habitual actions in an online sample characterised by harmful alcohol use, but not formally diagnosed with alcohol use disorder. Contrary to my predictions, this study did not find behavioural evidence for impaired regulatory control in harmful alcohol users. Three different task measures of habit predominance did not find any habit biases in the alcohol group, nor were they related to alcohol use severity. This suggests that when prompted with conflicting responses, these users were able to respond in a goal-directed manner, and were not more likely to elicit habitual responses than controls. Whilst there is some evidence that individuals with better goal-directed learning were more likely to respond in a goal-directed manner during the habit tests (irrespective of group membership), neither the learning of goal-directed instructions nor S-R habits through reinforcement learning significantly differ between alcohol and control groups, nor were they related to alcohol use severity. Interestingly, the tendency to elicit habitual responses (free choice trials) was significantly related with self-reported automaticity only in alcohol users, alluding to the notion that increased habit biases might be related to habitual traits.

7.4.1 Goal-directed and habitual actions not measurably affected in problematic alcohol users

The goal-directed system plays a key role in modifying behaviours such that they align with situational demands. This is especially important when actions are no longer beneficial, or even cause harm. This process is thought to be disrupted following long term alcohol use, which could pave the way for developing uncontrolled use in due course (Gillan, Robbins, et al., 2016). The current task operationalizes goal-directed actions as the ability to initiate an instructed response, despite the interference from habits, and found that alcohol users who drink at hazardous levels retain the ability to regulate behaviours in a goal-directed manner. It is noteworthy that the goal-directed measures were not related to AUDIT scores, suggesting that alcohol use is unrelated to the ability to adapt to situational demands. This is consistent with prior studies operationalising goal-directed control in terms of model-based behaviour, which also did not find any alterations to this process as a function of alcohol use severity (Doñamayor et al., 2018; Nebe et al., 2018; Patzelt et al., 2019). Even when there is no need to

obey any goal-directed instructions (free choice trials), alcohol users were also not more likely than controls to elicit habitual behaviours. In other words, there is no evidence that the habit system dominates behaviour after excessive alcohol consumption within the current sample. Notably, this null finding could not be explained by differences in the acquisition of S-R habits, as both task performance and S-R knowledge (i.e. phase 2) were comparable between groups. It is possible that dysregulation between goal-directed actions and habits only occurs at the severe AUD, where alcohol use dominates their lifestyle. Indeed, the vast majority of the alcohol users in the current data were under paid employment or studying, and not actively considering treatment, suggesting that their alcohol use does not interfere substantially with their lives, unlike severe AUD. However, current data and the mixed findings in the literature might suggest that this disrupted balance in instrumental regulatory control in alcohol users might not be as straightforward as previously conceived. Alcohol exposure alone may not be sufficient to increase habitual control, and other modulatory factors may be involved.

One candidate process that could subserve enhanced habit formation in alcohol users is the individual differences in habitual traits. This is supported by the positive correlation between self-reported automaticity and number of stimulus-elicited habits within the alcohol group, but not controls. Automaticity, as measured by the Creature of Habit Scale, reflects the tendency to elicit automatic behaviours within an associated environment (Ersche et al., 2017). Previously, automaticity was shown to be jointly modulated by stimulant exposure and negative childhood experience, but not by stimulant exposure alone (Ersche et al., 2017). In this case, it is likely that habitual traits exacerbated habitual behaviours only in individuals with harmful alcohol users. Nevertheless, such interactions are only speculative, and further investigations are required to clarify the mechanisms involved.

7.4.2 Reinforcement learning intact in alcohol users

Chronic alcohol use in humans is thought to alter cognitive systems critical for learning, including reward anticipation (Wrase et al., 2007), sensitivity to negative feedback (Galandra et al., 2020; L. D. Nelson et al., 2011), cognitive flexibility (Jokisch et al., 2014; Reiter et al., 2016), and circuits involved in prediction error signaling (Deserno et al., 2015; Park et al., 2010). Disruptions to these processes would likely have some negative impact on AUD patients' ability to adapt behaviour from prior experience. This process is usually modelled with

reinforcement learning tasks in prior studies, which generally found slower learning in AUD patients (Huys, Deserno, et al., 2016; Jokisch et al., 2014; Park et al., 2010; Vanes et al., 2014). My current analyses on reinforcement learning performance, irrespective of conventional or computational approaches, did not reveal any measurable differences in learning speed or performance between individuals with hazardous levels of alcohol use and controls. While my findings stand in stark contrast with prior studies, it is noteworthy that participants in the current sample widely varied in terms of alcohol use (as evidenced by their AUDIT scores) and were (as indexed by self-report) not formally diagnosed with AUD. This is in contrast with the vast majority of the prior data that assessed aspects of reinforcement learning in patients with a confirmed AUD diagnosis. Thus, considering this difference, it is conceivable that marked reinforcement learning impairments would only emerge at the severe end of the spectrum when control over behaviour is compromised, but not as a function of alcohol chronicity. Indeed, this notion is supported by the lack of any statistically significant relationships between reinforcement learning parameters and AUDIT scores in the alcohol group. Interestingly, a recent study on at-risk youths found that whilst there were no relationships between AUDIT and reinforcement learning task performance – consistent with the current data – functional imaging of task-related activations revealed that striatal and prefrontal brain activity in youths with higher AUDIT scores did not differentiate between positive and negative feedback (Aloi et al., 2020). This suggests that possible dysfunctions in the neural regions underpinning reinforcement learning that precede AUD may not manifest behaviourally (Aloi et al., 2020). Interestingly, this finding is reminiscent of a large-scale fMRI study that identified neural, but not behavioural, markers in adolescents at risk of impulsivity disorders (Whelan et al., 2012). This relatively understudied area warrants further attention.

7.4.3 Divergence between animal and human studies of habit formation

Whilst animal studies provide convincing evidence for a general shift towards automatic habits following alcohol exposure, findings from humans with AUD have not always been consistent with this narrative. Theoretically, the current behavioural task is conceived to improve on existing habit tests (e.g. outcome devaluation) by providing a direct way to assess the extent of goal-directed and habitual control in humans. This was in contrast with the human version of the outcome devaluation task (de Wit et al., 2007, 2009), which only models habits as the *absence* of goal-directed actions. However, even with this improved setup, I did not find any evidence for increased habitual control in alcohol users, which again diverges from findings in

animals. Upon reflection, it is possible that the differences seen here may be related to the differences in training habits between animals and humans. The psychological concept of habits originated from seminal work by Dickinson and colleagues (Adams & Dickinson, 1981; Dickinson, 1985). These works defined habits as an over-trained instrumental response that develops autonomy, to the point where the trained action is reliably elicited by the environmental stimulus that previously predicted reward, and is no longer sensitive to changes in outcome value or action contingency. Whilst this seems to hold true in humans in some cases (Tricomi et al., 2009), the volume of overtraining it takes to induce habits in animals might not be practically feasible to adopt in human studies – Tricomi et al (2009) trained participants with 12 sessions over the course of three days to induce habits. Even so, with five independent sets of behavioural data from existing human outcome devaluation tasks, de Wit and colleagues (2018) failed to replicate the findings of Tricomi et al (2009), which led them to argue that habits in humans might not necessarily emerge as a function of behavioural repetition. This suggests that developing habits via overtraining in humans might not be as straightforward as in rodent studies, and existing tasks for humans are inadequate in reliably inducing habits. Moreover, habits in humans are arguably far more complex than single-lever-response habits commonly represented in rodent studies. Therefore, paradigms inspired by animal studies, such as outcome devaluation, might not be optimised to study habits in humans. Indeed, attempts to translate the outcome devaluation paradigm to humans are argued to only be partially successful in modelling habits (Watson & de Wit, 2018). Although the two-step task is a promising attempt to model the dual-processes of instrumental learning, the lack of correspondence between model-free processes and habits needs to be considered. It is likely that habits are multi-faceted in humans, as exemplified by the various constructs identified in the development of self-reported instruments of habits, such as routine and automaticity (Ersche et al., 2017), as well as compulsivity, preferences for regularity, and aversion to novelty (Ramakrishnan et al., 2021). Perhaps the complexity of the habit construct in humans is one of the reasons why inducing habits in human experiments is deemed a difficult endeavour, and the reconciliation between animal and human studies of habit formation is not perfect.

7.4.4 Limitations and conclusion

There are several noteworthy limitations that should be considered when interpreting these results. Despite the large sample sizes from online data collection, I was unable to control whether the task was completed under conducive conditions (e.g. in a quiet room with minimal

distractions). I was also unable to monitor task engagement and concentration, as one would in an in-person assessment session (e.g. participants may be multi-tasking, or distracted by social media when completing the study). As a result, these problems might compromise data quality. Furthermore, I did not have further data on the pattern or frequency of alcohol use in this sample, which could differ to those with addicted use (cf. cocaine use of drug users in previous chapters). Thus, the potential differences in usage between the current sample and those reported in prior chapters need to be considered. Furthermore, it is noteworthy that a small proportion of alcohol users (16%) were on prescribed antidepressants. Whilst this small proportion is unlikely to significantly alter the interpretations of the data, they need to be acknowledged nonetheless. Additionally, the habit learning stage, which involved the learning of many S-R contingencies, likely required a high working memory load. Consequently, some participants might only focus on some stimuli during habit learning but not others, resulting in the goal-directed performance on certain trials to be much easier (in the absence of a competing S-R habit). Indeed, working memory is thought to be important in developing initial contingencies during reinforcement learning (Collins & Frank, 2012), but this measure was not assessed in this study. Additionally, the possibility of goal-directed instructions being easier to learn than habits cannot be entirely ruled out. It might also be entirely possible that this behavioural task required more training trials in stage 2 to sufficiently induce habits, though it is notable that increasing training trials in humans have not increased habit strength in the past (de Wit et al., 2018). Future behavioural and neuroimaging studies might be needed to replicate and extend the current findings. In conclusion, using a novel task to simulate goal-habit conflicts, I did not find evidence for dysregulation between goal-directed and habitual control over instrumental actions in individuals who consume alcohol at harmful levels.

Chapter 8: General Discussion

Maladaptive behaviours in substance use disorder (SUD) can be understood in terms of disruptions to the learning processes that normally contribute towards adaptive behaviour. Specifically, it was hypothesised that (1) reinforcement learning is impaired in SUD; and (2) behaviours in SUD are associated with an imbalance between goal-directed and habitual control systems. This thesis sought to characterise these impairments in substance use disorder from multiple perspectives. In this final chapter, I start by highlighting the key findings from the previous experimental chapters. I then discuss how these findings inform our current understanding of reinforcement learning, as well as goal-directed and habitual control in addictive behaviours. As computational analyses are a burgeoning method in neuropsychiatry, I also briefly discuss its potentials and pitfalls in studying neurocognitive processes. In closing, I also outline the limitations to, and future considerations from, my thesis.

8.1 Summary of key findings

The first set of computational analyses on probabilistic reinforcement learning in Chapter 3 revealed that moderate-to-severe stimulant use disorder patients have a selective impairment in learning from negative feedback in the form of monetary losses. This impairment is further shown to be linked with dopamine D_2 receptors, as pharmacological modulation of these receptors in patients ameliorated this impairment. These findings are highly consistent with the risky behaviours reported in SUD (American Psychiatric Association, 2013), and explain in part why maladaptive behaviours are maintained. Specifically, negative outcomes that supposedly signal harm and promote avoidance (Jean-Richard-Dit-Bressel et al., 2018) have a reduced effect on patients' behaviour, which in turn makes it less likely for them to adjust behaviour accordingly. The selective effects of dopaminergic modulation on negative feedback learning in these patients confirm the neurobiological link between aberrant learning and dopamine dysfunction, which previously have only been assumed in human SUD.

Associative learning is supported by multiple memory systems, namely declarative and non-declarative memory (Poldrack & Foerde, 2008). It is possible that impairments in these memory systems affect reinforcement learning in neuropsychiatric conditions like SUD (Seger & Miller, 2010). I tested this possibility in Chapter 4 by analysing task performance and

response strategy on two variants of an established probabilistic category learning task – the weather prediction task – each variant testing the integrity of declarative and non-declarative memory. Analyses of task performance showed that cocaine use disorder (CUD) patients had reduced performance in both task variants, whether learning declaratively (through memorisation) or non-declaratively (through feedback), suggesting that both memory systems are impaired. Closer inspection of the response strategy analyses (Gluck et al., 2002) during feedback learning revealed an interesting pattern. Control participants mostly adopted a more complex integrative strategy during learning, while patients were more likely to engage with simple but suboptimal memorisation strategies during learning. These findings also highlight aberrant engagement with memory systems in SUD, which may compromise learning. In the context of this task, the striatum is critical in integrating past experiences of reinforcement, and therefore supports non-declarative memory (Poldrack et al., 1999; Shohamy, Myers, Grossman, et al., 2004). Thus, these findings are consistent with the notion that SUD patients show marked striatal dysfunctions (Yager et al., 2015). Parkinson's disease patients, characterised by striatal deficits, also exhibit the same behavioural profile as the current SUD patients, despite extensive training (Shohamy, Myers, Onlaor, et al., 2004). However, since these memory systems are dissociable (Packard & Knowlton, 2002), they allude to the possibility that learning in SUD patients may be more effective when it involves simple rule-based memorisations, rather than complex decision-making that requires averaging from prior experiences.

A contemporary theory suggests that SUD is associated with an imbalance between goal-directed and habit systems of behavioural control, favouring the latter (Everitt & Robbins, 2005, 2016). This theory is thought to explain why behaviour is not amenable to consequences in SUD. It is suggested that habit biases are a result of an impaired top-down goal-directed system, though an augmented habit system or a combination of these two hypotheses are also possible (Robbins & Costa, 2017; Vandaele & Janak, 2018). Findings from Chapters 3 and 4 have highlighted that the goal-directed system is impaired in SUD. However, upon re-analysing an appetitive instrumental learning task with computational modelling in Chapter 5, I discovered that the findings did not wholly support the notion that an impaired goal-directed learning system contributes substantially to a habit bias. Specifically, although there is evidence for impaired goal-directed learning in CUD patients, in the form of reduced impact of positive feedback on behaviour, this impairment was insufficient to account for the increased slips of action (habits) in these patients. This finding calls for the renewal of our understanding of the

mechanisms behind accelerated habit formation in SUD, suggesting that there might be other factors at play. Certainly, a reduced goal-directed system diminishes one's capacity to monitor habits (Balleine, 2019), but the current data suggest that this is unlikely the primary reason that led to increased habits in human SUD. Indeed, when completing a contingency degradation task that does not heavily depend on learning and motivation, CUD patients still exhibit increased habitual responding (i.e. reduced sensitivity to action-outcome contingencies) (Ersche et al., 2021).

I was then interested in whether habitual tendencies extend beyond laboratory paradigms, which prompted the examination of self-reported questionnaire measures of habits and goal pursuit in CUD patients in Chapter 6. These measures provided evidence that CUD patients were more likely than their healthy peers to engage in automatic habits. They were also less likely to pursue difficult goals, suggesting a generally reduced goal-directed motivation. Together, they provide converging evidence for the validity of the habit construct in SUD, which has recently been controversial (Hogarth, 2020). An increased disposition for automatic habits in CUD patients would mean that they are more prone to fall back on familiar behaviour when under stress or challenged. Consequently, this might explain why abstinence is difficult, especially under stressful conditions (Sinha, 2001), as environmental cues are more likely to elicit habitual responses in these patients. The reduced goal motivation also reflects a diminished tendency to exert goal-directed deliberate control, especially when it involves difficult actions such as changing behaviour.

While SUD has been linked to increased habit formation (as measured behaviourally and by self-report), as of yet, it is unknown whether habit predominance is also present in harmful – but not dependent – drug users. I address this knowledge gap in Chapter 7 with an online study, where I investigated instrumental regulatory control in a large, heterogeneous, online population of harmful alcohol drinkers with a novel task I co-developed. Contrary to expectations, I did not find any evidence for changes in behavioural control systems in individuals who drink hazardously. There was no evidence for impaired goal-directed learning or enhanced habit control in these individuals. The lack of statistically significant relationships between alcohol use severity (as indexed by AUDIT total scores) and behavioural measures suggest that deficits in instrumental control likely arise only when one transitions into severe

SUD. These findings suggest that harmful drug use itself may not lead to disruptions in regulatory control associated with addictive behaviours. However, it is likely that regulatory control is affected when drug use interacts with other factors such as chronicity, pattern of use, familial and environmental risk factors. These factors could potentially mediate or moderate the transition into SUD, but were not addressed in the present study.

Taken together, this thesis identified in severe SUD impairments in learning from negative feedback and integrating learned experiences to inform future choices, both of which could negatively affect adaptive behaviour. Whilst these impairments support the notion of an impaired goal-directed system in SUD, they are insufficient to fully account for the increased habit system that dominates instrumental behaviour in severe SUD. However, impairments in reinforcement and instrumental learning may be limited to severe SUD, as a population of alcohol drinkers with harmful use (not formally diagnosed with SUD) did not show any measurable deficits in either reinforcement learning or instrumental control. This alludes to the notion that the transition from volitional to habitual, and eventually compulsive drug use in humans is a complex phenomenon and likely does not depend on drug use alone.

8.2 The role of reinforcement learning in substance use disorder

Reinforcement learning is an influential model of adaptive and goal-directed behaviour. Risky drug use is prevalent in SUD and is seen as a consequence of a breakdown in reinforcement learning processes (Redish et al., 2008). In view of the extant literature and my current findings, I argue that maladaptive behaviour in SUD is associated with a selective deficit in learning from negative outcomes and difficulties in integrating memory from past experiences, which might be relevant to our understanding of inflexible behaviours seen in these patients. However, this thesis finds that reinforcement learning deficits are unlikely to arise when alcohol consumption is harmful, but does not interfere with day-to-day functioning.

8.2.1 Altered reinforcement learning processes in substance use disorder

Humans are innately driven to avoid aversive or negative consequences (B. F. Skinner, 1963). This tendency is meant to promote functional behaviours, as individuals who do not learn from these events run the risk of repeating them and risk future harm or injury. Yet in the context of

SUD, harmful use persists despite knowledge of the health, social and legal consequences that ensue (American Psychiatric Association, 2013). Considering that individuals with SUD do not modify behaviour following punishing consequences, there is an increasing body of work attempting to study and characterise the processes related to learning from punishment in SUD (Jean-Richard-Dit-Bressel et al., 2018; L. J. Vanderschuren et al., 2017), though relatively less work has been conducted in humans. Prior studies in CUD patients have shown that they do not adjust behaviour after multiple modalities of punishment, including symbolic errors (Hester et al., 2013), disgusting cues (Ersche et al., 2014), or even electric shocks (Ersche et al., 2016). However, it is noteworthy that these patients still exhibit normal autonomic signals (e.g. skin conductance responses) in response to these cues (Ersche et al., 2014, 2016). This suggests that aversive conditioning is not impaired in SUD, but rather the systems that control *instrumental* actions following negative feedback. Supporting evidence for this comes from animal studies, which identified a selective deficit in adapting instrumental behaviour from negative feedback in stimulant-exposed rats (Groman et al., 2018; Zhukovsky et al., 2019). In particular, Chapter 3 translated these findings in humans, implying that stimulant-addicted patients have a reduced tendency to use negative feedback to inform subsequent behaviour, which might sustain compulsive behaviours.

Although human and animal research seems to suggest a reduced impact of negative reinforcement on behaviours in general, there is less consensus on positive reinforcement learning. It is thought that reinforcing properties of addictive drugs over time hamper normal reward pathways in the brain (Volkow et al., 2004), which reduces the salience of non-drug reinforcers (Garavan et al., 2000; Goldstein et al., 2007). However, the behavioural evidence for an impaired positive (reward) reinforcement learning system is mixed, with some studies showing impaired reward learning (Morie et al., 2016; Strickland et al., 2016), but others showing intact performance (Stewart et al., 2014a, 2014b). The mixed findings present in the SUD literature are also reflected in this thesis, wherein Chapter 3 did not identify any measurable group difference in reward learning, but Chapter 5 identified a reduced impact of positive feedback during appetitive learning. However, one possible explanation that could reconcile these discrepant findings might be the differences in motivational salience of the reward. Multiple strands of evidence suggest that reward responses are modulated by how motivationally salient cues and rewards are to SUD patients. For example, a prior study found that although SUD patients exhibit similar behavioural and neural response towards monetary

incentive cues, SUD patients showed exacerbated behavioural and neural (striatal and prefrontal) responses towards drug incentive cues (Zhukovsky et al., 2020). Another study found that cocaine-related pictures are more favourable than other pleasant non-drug pictures in a sample of CUD patients (Moeller et al., 2009). These studies collectively suggest an altered baseline motivational factor in SUD, and that this may affect reinforced behaviour. This difference in motivation may not only be relevant for drug versus non-drug-related cues, it may extend to other modalities as well. For instance, when learning involves monetary reinforcers, a type of secondary reinforcer, SUD patients did not differ significantly with controls on their learning rate in Chapter 3, presumably because money remains a salient reinforcer to SUD patients. By contrast, the use of symbolic points, an arbitrary reinforcer specific only to the task, yielded a marked reduction of reward learning rate in Chapter 5. Other studies with arbitrary reinforcers also demonstrated this deficit in reward learning (Morie et al., 2016; Strickland et al., 2016), whereas participants incentivised with monetary payment did not show this pattern (Stewart et al., 2014a, 2014b). It seems that positive feedback can still modulate behaviour in SUD, but its effectiveness might be highly dependent on salience. Contingency management – a behavioural intervention rooted in positive reinforcement – is a real-world example that demonstrates the effectiveness of incentives in facilitating abstinence in cocaine users (Petry, 2000; Petry et al., 2017).

Another possible explanation that could account for the discrepant results in the reward learning rate is the subtle difference between the behavioural tasks used in Chapters 3 and 5. Whilst the reinforcement learning task in chapter 3 adopted a probabilistic design (the optimal choices were rewarded 70% of the time), the task in Chapter 5 had a deterministic design (correct choices were always rewarded). Therefore, the probabilistic design may have promoted more exploration during choice selection, i.e. switching between choices (Feher da Silva et al., 2017), and thus could have obscured the effects of the rewarding feedback on behaviour.

It is possible that impaired reinforcement learning in SUD is linked to suboptimal memory systems during learning. In particular, when decomposing response strategy during learning in Chapter 4, SUD patients were more likely to use a simpler, rigid rule-based strategy that depends on declarative memory, instead of flexibly integrating learned knowledge, which

depends on the non-declarative memory system. Healthy participants usually can rely on both declarative and non-declarative systems, depending on task demands (Foerde et al., 2006; Poldrack et al., 2001). However, unlike their healthy peers, SUD patients seem to rely more on simple learning strategies. Since non-declarative and declarative systems are dissociable, this response strategy could be interpreted as a compensatory response to adjust for poor non-declarative learning. This might imply that one system could compensate for another under neuropathological conditions, such as in Parkinson's Disease or Huntington's Disease (Holl et al., 2012; Shohamy, Myers, Onlaor, et al., 2004), possibly due to deficits in neural functions (discussed later).

8.2.2 Reinforcement learning and inflexible behaviour

The understanding of reinforcement learning in SUD is also relevant to perseverative behaviours, a behavioural marker of compulsivity (Figee et al., 2016) and a recurring feature in SUD (Ersche et al., 2008; Ersche, Roiser, Abbott, et al., 2011; Jentsch et al., 2002; Schoenbaum et al., 2004). Perseveration is thought to be a manifestation of impaired reinforcement learning (Lucantonio et al., 2012), yet only recently it has been understood in computational terms. Specifically, in an uncertain environment, there is a trade-off between maximising reward values and ignoring spurious outcomes that could be noise (Gershman, 2020). As such, perseveration could be expressed as aberrant persistence towards prior reward values irrespective of changes in current value (i.e. stickiness), or an inability to update or override learned responses in light of new information (Gershman, 2020); both of which implicate the reinforcement learning systems. A prior analysis on probabilistic reversal learning task performance in populations of OCD and SUD patients supported both views: those with compulsive disorders not only aberrantly updated choice values in response to symbolic positive and negative feedback, but they also showed increased stickiness (Kanen et al., 2019). Another recent study has also identified impaired contingency learning as the reason why behavioural change during reversal learning is difficult in methamphetamine-addicted users (A. H. Robinson et al., 2021). Although the behavioural tasks reported in my thesis are not optimised to measure perseverative behaviours, unlike probabilistic reversal learning paradigms, the current findings generally align well with the notion of impaired adjustments to behaviour, even when negative feedback instructs us otherwise.

8.2.3 Reinforcement learning impairments not present in early stages of substance use disorder

Most studies that identified reinforcement learning impairments in SUD (including Chapters 3-5) were conducted in severe SUD, where drug use becomes so dysfunctional that it dominates daily life. A question I attempted to address was whether these impairments manifest themselves in individuals with early signs of pathological alcohol use, which has implications on early detection and interventions. However, analysis on a large online population of harmful alcohol drinkers, who were not formally diagnosed with SUD, did not lead to any observable reinforcement learning impairments in Chapter 7. The lack of relationship between alcohol use severity (as reflected by AUDIT scores) and reinforcement learning impairments, as well as between compulsive use (as measured by OCDUS score) and learning performances (chapters 3-5) do not support the notion of a dose-dependent effect. These findings seem to imply that while addictive drugs may interfere with reinforcement learning pathways in the brain, behavioural impairments might only be a characteristic in severe SUD. Inferring from the data presented in Chapter 7, it seems likely that harmful drinkers nevertheless retain most of their cognitive and social functioning to live functional lives, as the majority of them were studying or pursuing paid work, and were not actively considering treatment. Further, it is also possible that the COVID-19 pandemic might have contributed to temporary increases in drinking. Nevertheless, this is one of the first studies that probes reinforcement learning in an at-risk group, and future work is needed to replicate this finding.

8.2.4 Sex differences in substance use disorder and reinforcement learning

This thesis reported data on samples that were predominantly male, as most cocaine users recruited from Cambridgeshire (at the time of these studies) were overwhelmingly male. However, it is noteworthy that there are disparities between men and women with SUD (Becker et al., 2017; Greenfield et al., 2010; Lynch et al., 2002). Multiple studies have shown that women escalate from controlled to compulsive drug use more rapidly than men (Haas & Peters, 2000; Johnson et al., 2005; Piazza et al., 1989; Wagner & Anthony, 2007). Women also showed greater propensity than men to relapse following exposure to stress or triggering cues (al'Absi et al., 2015; Kennedy et al., 2013; Torres & O'Dell, 2016). These observations are suggested to be associated with sex differences in the neurobiological substrates that mediate drug reinforcement (Becker & Chartoff, 2019; Becker & Hu, 2008; Becker & Koob, 2016). Preclinical studies have shown that female rats have generally upregulated dopamine release

and uptake (Walker et al., 1999) and greater amphetamine-induced dopamine release in the nucleus accumbens (Becker, 1999), suggesting a clear sexual dimorphism in dopamine signalling. Behaviourally, animal models of psychostimulant self-administration have identified that ovariectomized female rats (i.e. removed ovaries) were more likely to exhibit increased locomotor sensitisation towards, and motivation for, psychostimulants (Hu et al., 2004; Hu & Becker, 2003; Lynch & Carroll, 1999). Moreover, the administration of estrogen further augments the reinforcing properties of psychostimulants in female rats, but not male rats (Hu & Becker, 2008; Jackson et al., 2006; Lynch et al., 2001). These studies offer compelling evidence for the organizational and activational effects of sex hormones on drug reinforcement. Considering that male and female rats have different biological responses to reward, it stands to reason that there would be sex differences in cognitive processes such as reinforcement learning and memory as well (Dalla & Shors, 2009). Human evidence for this hypothesis is scarce, but a recent study noted that women had augmented reward-prediction-error related brain activity compared to men, suggesting that there may indeed be sex differences in reinforcement learning (Joue et al., 2021). Although it remains unclear how exactly chronic use of addictive drugs differentially alters reinforcement learning pathways in men and women, these putative sex differences in reinforcement learning need to be considered when interpreting the current results. Given the male-dominant samples in my thesis, the current findings may have limited generalisability to women with SUD.

8.2.5 Section summary

In brief, reinforcement learning in SUD can be characterised by marked impairments in learning from negative feedback and deficits in integrating past experiences effectively to inform future behaviour. Deficits in these processes may lead to inflexible behaviour common in SUD, though notably these deficits may only manifest behaviourally in severe SUD, and not in the earlier stages where behaviour is largely under control. Although there are notable sex differences in SUD, most chapters only reported behavioural findings from male-dominant samples, so this should be considered when interpreting these results.

8.3 Putative neural substrates of reinforcement learning in substance use disorder

The chronic use of addictive drugs is thought to cause neuroadaptive changes within the reinforcement systems in the brain (Everitt et al., 2001; Hyman et al., 2006), which interferes

with normal reinforcement learning processes (Maia & Frank, 2011). In this section, I discuss the putative changes in the brain that are likely to underpin these deficits in SUD, which include the role of dopamine and the fronto-striatal brain pathways implicated in incentive values and integrating learned knowledge. I acknowledge that different addictive drugs may differentially affect the brain, so I also discuss the possibility of common reinforcement learning impairments across addictive disorders.

8.3.1 The role of dopamine D_2 receptors in human substance use disorder

A growing body of studies in human SUD have shown compelling evidence for a downregulation of striatal dopamine D_2 receptors in SUD (Martinez et al., 2004; Volkow et al., 1996, 1997, 2001). This downregulation putatively affects responses to reward, as individuals with alcohol, cocaine and opioid use disorder have reduced responsiveness towards non-drug reinforcers (Garavan et al., 2000; Heinz et al., 2004; Lubman et al., 2009). Whilst these studies suggest that reinforcement is blunted in SUD, they have not specifically investigated its effects on learning. The novel aspect of my findings in Chapter 3 is that it identified a relationship between aberrant dopamine D_2 receptors and impaired negative feedback learning in SUD. This is likely because D_2 receptors are implicated in aversive learning, possibly due to their role in signalling negative prediction errors (Frank & Hutchison, 2009; Frank & O'Reilly, 2006). Consequently, a downregulation in dopamine D_2 receptors may reduce the salience of negative feedback in SUD. This is mirrored in individuals with genetically determined reductions in D_2 receptor density, who also showed behavioural deficits in learning from errors (Klein et al., 2007). Other pharmacogenetic studies confirmed the link between D_2 receptor polymorphisms (e.g. Taq1A) and altered avoidance behaviours (Frank & Hutchison, 2009; Jocham et al., 2009).

Several studies have identified an allelic association between reduced D_2 receptors and increased risk for SUD (Blum et al., 1990; Noble et al., 1993). Would this imply that genetically determined deficits in learning from errors, for example from Taq1A polymorphisms, could constitute a vulnerability pathway to developing SUD? The answer to this question is not straightforward for two reasons. First, reduced D_2 receptors can be both a cause and a consequence of SUD: in non-human primates, reduced pre-morbid striatal D_2 receptor levels were negatively related to the rate of cocaine self-administration, and subsequent chronic self-

administration of cocaine reduces striatal D_2 receptors further (Moore et al., 1998; Nader et al., 2006; Nader & Czoty, 2005). Second, D_2 receptors' role in behaviour is likely multi-faceted: animal and human studies have shown that reduced D_2 receptors are also linked to increased sensitivity to drug reinforcement (Bello et al., 2011; Volkow et al., 1999) and elevated trait impulsivity (Buckholtz et al., 2010; Dalley et al., 2007) – both of which are vulnerability factors to SUD that may interact with learning too. Consequently, it is unclear to what extent reinforcement learning impairments in SUD can be explained by pre-morbid neurobiological or other (e.g. socioeconomic status, IQ) factors, and this question remains largely unexplored in the literature. To the best of my knowledge, there are no existing studies that address this question directly – a paucity also highlighted in several recent reviews (Gueguen et al., 2021; R. Smith et al., 2021). Future longitudinal studies in either humans or animals would be needed to elucidate whether impaired reinforcement learning precedes SUD development, and whether other modulatory factors such as impulsivity, which notably predicts SUD development, interact with learning.

8.3.2 Aberrant fronto-striatal systems in substance use disorder

Reinforcement learning depends on the integrity of fronto-striatal systems (Averbeck & Costa, 2017; Averbeck & O'Doherty, 2021; Haber & Behrens, 2014). Accumulating evidence from both animals and humans suggests that addictive drugs alter the structure and function of these pathways critical for incentive value and contingency learning (Calu et al., 2007; Ersche et al., 2005; Meunier et al., 2012; Moreno-López et al., 2015; Schoenbaum et al., 2004). An ancillary analysis on white matter connectivity in Chapter 5 found that whilst the structural integrity of the anterior caudate-medial OFC pathway did not differ between CUD patients and controls, the learning rate (impact of feedback on actions) was only significantly correlated in control participants, not in CUD patients. However, it should be noted that the difference between these correlational analyses was not statistically significant. One possibility could be that learned values were not effectively communicated in the brain. This interpretation would be consistent with findings from prior studies with alcohol use disorder patients, which showed that their functional connectivity within the fronto-striatal network during reward learning was reduced (Deserno et al., 2015; Park et al., 2010).

Furthermore, this thesis also provided some indirect evidence for striatal function deficits in SUD, namely from analysing response strategies in the weather prediction task, a well-established probabilistic category learning task that reliably recruits the striatum (Knowlton et al., 1996; Poldrack et al., 1999, 2001; Shohamy, Myers, Grossman, et al., 2004). Analyses of response strategy from Chapter 4 identified that half of the SUD patients were unable to effectively integrate and synthesise past experiences to guide behaviour, but instead relied on simple rule-based strategy during learning. Detecting regular occurrences and integrating learned experiences is a known function of the striatum (Seger & Spiering, 2011). In particular, my results from Chapter 4 are reminiscent of those in Parkinson's Disease patients (characterised by marked striatal impairments), who also exhibited the same behavioural profile as CUD patients in the current data (Shohamy, Myers, Grossman, et al., 2004; Shohamy, Myers, Onlaor, et al., 2004). It is unclear, though, why some SUD patients adopt a striatal-dependent strategy, given SUD patients who used a more optimal strategy did not differ in demographical or clinical characteristics from those who used a simple strategy. There could be subtle functional or structural differences in the neural pathways that discriminate between those who adopt a complex strategy and those who use a simple strategy, which could be the basis for future investigations.

8.3.3 Generalisability of impaired reinforcement learning

It is noteworthy that different classes of addictive drugs differentially affect the neurobiology in humans (Badiani et al., 2011; Lüscher & Ungless, 2006). Hence, a pertinent question is whether impairments in reinforcement learning are common across all addictive behaviours? All SUDs are equally characterised by persistent pathological drug-taking patterns despite the negative effects (American Psychiatric Association, 2013), so it is logical to presume that negative feedback learning is affected in SUD, albeit to various extents. Further, most addictive drugs modulate dopamine levels both in the short and long term (Di Chiara & Imperato, 1988; Nestler, 2005; Pierce & Kumaresan, 2006; Saal et al., 2003), which may affect reinforcement learning. I argue that it is likely that reinforcement learning is impaired in all SUDs, but a paucity in research studies precludes us from reliably addressing this question at this stage. Current evidence in SUD populations does not reveal a unifying picture. For example, although cocaine and alcohol use disorder patients are consistently impaired in learning and memory (Fernández-Serrano et al., 2011), in opioid use disorder the evidence is mixed: opioid users have been found to exhibit both exaggerated (Myers et al., 2016) and impaired punishment

learning (Myers et al., 2017) in different studies. In cannabis use disorder, only one paper showed a relationship between cannabis use disorder and impaired reward learning (Lawn et al., 2016), but none addressed whether punishment learning is affected. The issue is further complicated when we consider behavioural addictions such as gambling disorder, which lacks the neuro-adaptive influences from addictive drugs (L. Clark et al., 2019). The aetiology of brain-induced changes in gambling disorder is not clear, but recent studies do support the notion of impaired reinforcement learning in gambling disorder patients (Perandrés-Gómez et al., 2021; Wiehler et al., 2021). Ultimately, it is premature to draw any conclusions about a common reinforcement learning profile in addictive disorders, as more work is needed.

8.3.4 Section summary

Reinforcement learning impairments in SUD can be, at least in part, attributed to dopaminergic dysfunctions such as D_2 receptor downregulation and abnormal structure and function in fronto-striatal brain pathways that mediate reinforcement learning. However, whether these behavioural and neural impairments are common across addictive behaviours warrants further investigation.

8.4 Goal-directed and habitual control in substance use disorder

Chronic exposure to addictive drugs is suggested to alter the balance between goal-directed and habitual control over learned actions, favouring the latter (Everitt & Robbins, 2005, 2016). This hypothesis has been used to explain why drug-taking in SUD becomes autonomous and not amenable to the consequences associated with it. However, despite a host of evidence supporting the notion of habit predominance in animal and human SUD (Corbit et al., 2012; Ersche et al., 2016, 2021; Nordquist et al., 2007; Zapata et al., 2010), some studies suggest that habits are less essential in the development of addictive-like behaviours (Hogarth et al., 2019; Singer et al., 2018; Vandaele et al., 2019). In this section, I discuss the ongoing debate about the relevance of habits in understanding pathological behaviour in SUD, and suggest some potential pathways that lead to habit predominance in SUD.

8.4.1 Goal-habit controversy in substance use disorder

The role of habits in SUD has generated considerable debate. Recently, Hogarth (2020) suggested that an enhanced habit bias does not explain addictive behaviours, and that in the extant human data, habit predominance may be a by-product of poor contingency learning. In lieu of maladaptive habits, Hogarth posited that addictive behaviours are driven by an excessive goal-directed preference for drugs because drug choices are highly valued and sought after. As choices to initiate and maintain drug use are driven by value (Hogarth & Field, 2020), maladaptive drug use cannot be construed as a habit. I argue that whilst addictive behaviours can reflect an exaggerated motivation towards drugs, there is convincing evidence for increased habitual control in SUD, which cannot be fully accounted for by poor contingency learning, as Hogarth argued (2020). So, discarding the relevance of habits in SUD now might be premature.

On the one hand, there is indeed evidence to suggest that drug-seeking can be goal-directed. For example, rats that compulsively self-administer cocaine can flexibly navigate through a complex puzzle in order to gain access to cocaine (Singer et al., 2018). The effort and flexibility demonstrated in puzzle-solving requires careful planning and deliberation, and thus cannot be a habit-like behaviour. This is also consistent with human drug users beyond the lab. For instance, certain news outlets have reported that drug users use creative means to score and distribute while evading the authorities (Munro, 2019; Schrager, 2015). During data collection, some SUD patients even recounted that they willingly "fasted" (i.e. withheld drugs to themselves) and endured withdrawal symptoms in order to experience a more intense euphoria when using. Again, these planned behaviours cannot be habitual either. It is possible that these behaviours might actually reflect goal narrowing, where the drugs have "hijacked" the goal-directed system to intensify efforts to obtain and use them, while displacing all other non-drug related aspects (Volkow et al., 2019). As such, drugs are highly sought after, but other reinforcer types are less motivating. This goal-narrowing phenomenon has been shown in a recent study, where CUD patients valued drugs more highly than food, and were willing to incur more costs to procure them (Breedon et al., 2021).

On the other hand, there is compelling evidence that human SUD is associated with increased habit formation, and that the observed habit bias in SUD cannot simply be accounted for by poor learning. In particular, the experimental findings of Chapter 5 have demonstrated that

impaired appetitive learning is insufficient to account for habit formation in CUD. Additionally, in a contingency degradation paradigm that involves minimal reinforcement learning, CUD patients still demonstrated a habit bias (Ersche et al., 2021). Beyond the lab, SUD patients in real life also display habit-like behaviours when it comes to drug use. For example, I noticed during data collection that patients often describe drug-taking as a very 'mindless' and automatic behaviour, and they show no explicit desire to use drugs. Indeed, when asked, most drug users even reported spontaneously that they "don't know" why they use drugs (Breedon et al., 2021), suggesting an autonomous quality to that action. I also showed in Chapter 6 that this bias towards automatic habits was also found when measured by self-report. This self-reported automaticity also increased the longer they were using cocaine. These results indicate that a habitual propensity does pervade non-drug-using behaviour in human SUD as a function of cocaine use. However, it is notable that the evidence for habit predominance is consistent in cocaine users, but less so for other substances such as alcohol (Sjoerds et al., 2013; van Timmeren et al., 2020). This discrepancy may be due to potential differences in habit formation between human cocaine and alcohol users, but the equivocal state of research precludes any conclusive inferences. Regardless, future studies are needed to replicate these effects.

Taken together, it seems that the habit theory of addiction does not perfectly explain all addictive behaviours, but it should not be discarded prematurely. Some stages of SUD (e.g. drug-seeking) may be goal-directed, whilst others (e.g. drug-taking) may be driven by habitual processes. As Epstein (2020) argues, there may not be a one-size-fits-all theory to address the incredibly complex and multi-faceted nature of SUD, and theories should therefore move away from the "winner-take-all" approach (Epstein, 2020).

8.4.2 Possible pathways to a habit predominance in substance use disorder

I alluded earlier that enhanced habit formation is assumed to be a consequence of poor goal-directed learning (Hogarth, 2020; Vandaele & Janak, 2018). However, data from this thesis suggests otherwise. Specifically, findings in Chapter 5 suggest that, despite impaired goal-directed learning during initial stages, these deficits do not account for the enhanced habit formation seen in CUD. This is also supported by findings in Chapter 7, which found a lack of relationship between learned habits and goal-directed control. These data imply that there are

other factors at play that could contribute to the predominance of the habit system. Here, I speculate on two candidate factors: individual differences in automaticity and effort.

The current data seem to suggest that individual differences in automaticity might underpin habitual behaviours in SUD. As I have demonstrated in Chapter 6 with self-reported measures, CUD patients were more susceptible to elicit automatic habits when in an associated environment. Moreover, I showed a similar relationship in Chapter 7: the self-reported propensity for automatic habits is positively correlated with task-related habitual responses in users who drink at harmful levels. These findings provide supporting evidence for a relationship between enhanced propensity for stimulus-driven habits and lab measured habits. However, although an elevated habit system may contribute to the development of addictive behaviours, thus far the mechanisms that support this remain unclear, and to the best of my knowledge, no studies have so far attested to this relationship. In particular, I offer two caveats with respect to my current findings: (1) whether self-reported automaticity and habitual behaviours reflect the same construct needs further validation; and (2) it is unclear whether self-reported automaticity precedes or follows the development of SUD.

Alternatively, habit biases in SUD could be related to a reduction in effort or engagement with difficult tasks. A distinctive feature of habits is that they are easily executed, as opposed to goal-directed actions which require more cognitive resources (Balleine, 2019; Balleine & O'Doherty, 2010). The tendency to disengage with difficult activities would logically result in the falling back on well-learned behaviours that are more easily executed. This is most apparent when under stress, which is known to bias actions towards the habit system, plausibly by reducing top-down prefrontal engagement (Schwabe & Wolf, 2009). I showed with the Habitual Self Control Questionnaire (HSCQ) (Chapter 6) that CUD patients were less willing or uninterested to engage in difficult activities that require will and motivation, which is also dependent on the goal-directed system. Further supporting evidence can be seen in Chapter 4 where, during category learning, approximately half of CUD patients adopted the easier strategy instead of the difficult but more optimal one. The interpretation of effort coincides with the computational framework on model-based versus model-free behaviour. According to this perspective, the complexity of the actions is a crucial determinant of behavioural control (Dolan & Dayan, 2013). Model-based behaviours are effortful, require taxing computations,

and are meant to represent goal-directed action; by contrast, model-free behaviours only rely on prior rewards, and easily executable, and are meant to reflect habits (Daw, 2014; Dayan & Daw, 2008). In tasks that simulate model-based and model-free behavioural strategies (e.g. the two-step task, (Daw et al., 2011; Gläscher et al., 2010)), SUD patients are thought to engage more in model-free behaviours, perhaps because they are simpler and easily executed, and require lower cognitive demand (Voon et al., 2017). However, as of now, it is not clear whether this tendency to "take the easy road" reflects changes in disposition or motivational circuits in SUD.

8.4.3 Section summary

The habit theory of addiction is useful for explaining certain aspects of SUD, but contrary to assumptions, deficits in goal-directed control cannot fully account for the increased tendency for habits. Extenuating factors such as individual differences in habitual propensities and willingness to expend effort may contribute towards habit biases, but future studies are needed to test these hypotheses.

8.5 Computational modelling of behaviour in neuropsychiatry

Computational modelling of behaviour capitalises on the advances in data analytics to decompose cognitive processes into its constituent latent processes (Daw, 2011), which was previously not possible. The use of precise mathematical frameworks enables us to define "algorithmic hypotheses" about how behaviour is specifically generated (Wilson & Collins, 2019). This thesis in particular used reinforcement learning algorithms to identify individual differences within the latent parameters of those with or without SUD, which arguably has provided more insight than that of summary scores alone (Robbins & Cardinal, 2019). Perhaps one of the major advantages of computational modelling is its ability to test mechanistic hypotheses that previously could only be conducted in preclinical studies (Robbins & Cardinal, 2019). Ample examples are provided in this thesis: the use of a delta rule learning algorithm isolated a specific deficit in the impact of feedback on overt behaviour that translated prior animal work (Chapters 3 and 5). I also used model comparison methods to identify models that best explained choice behaviour (Chapter 4). There are also many examples beyond this thesis, some of which significantly advanced our knowledge of human cognition. For instance, the advent of reinforcement learning algorithms have allowed the identification of prediction errors

in the human brain (D'Ardenne et al., 2008; Pagnoni et al., 2002) that were previously found in non-human primates (Schultz et al., 1997). Another apt example is the use of modelling to tease apart the dissociable contributions of striatal versus prefrontal dopamine in multiple aspects of reinforcement learning (Doll et al., 2016; Frank et al., 2007). Indeed, the burgeoning application of computational methods in psychiatry has led to the emergence of computational psychiatry: the use of computational models to make mechanistic inferences on behaviour in psychiatric disorders (Corlett & Fletcher, 2014; Huys, Maia, et al., 2016).

However, as much as computational modelling has the potential to significantly advance the field of psychiatry and psychopharmacology, there is a need to acknowledge certain caveats within the field. First, the interpretation of similar model parameters may be different across studies, and needs to be contextualised within the behavioural task. For example, a higher learning rate parameter – i.e. a larger and more rapid update from feedback – allows the participants to reach a learning asymptote quickly, which could be beneficial in a stable environment (Behrens et al., 2007), such as the probabilistic reinforcement learning task. However, for a more volatile situation (e.g. where learned contingencies are constantly switching), such as those in serial probabilistic reversal learning paradigms, optimal performance requires a trade-off between ignoring rare events and flexibly updating actions when contingencies change. In this respect, a higher learning rate may not be beneficial, as a stable representation may not be reached if one acts on all spurious events. This is aptly demonstrated in Chapter 3 (Appendix B) – fitting the winning model of Kanen et al's (2019) reversal learning task to the current data revealed contrasting results, presumably due to the absence of a reversal component, which reduced volatility in my task. Second, for models to be informative, it is imperative for mathematical models to be grounded in psychological and neurobiological theory, which is not necessarily the case in the field of SUD. In particular, some theories that attempt to explain aberrant drug use in SUD (e.g. actor-critic models (Takahashi et al., 2008) and hierarchical control models (Keramati et al., 2017; Schwöbel et al., 2021)) are inspired by theoretical abstractions posited in computer science, which have not yet been shown to be biologically plausible in humans (Mollick & Kober, 2020). Given the complexity and heterogeneity of addictive disorders, it is likely that there will not be a unifying "one-size-fits-all" computational framework that addresses SUD holistically. Nevertheless, computational models, especially those grounded in psychological theory e.g. the Rescorla-

Wagner model (Rescorla & Wagner, 1972), have proved useful in psychiatry research, and are likely here to stay for the long haul.

8.6 General implications

Findings from this thesis argue against the notion of SUD as a consequence of moral failure, but a complex brain disorder with significant neurobiological and psychological underpinnings. Yet, many of the current drug policies continue to be motivated by the extremely damaging moral failure narrative, which focuses on enacting punitive measures against drug users (Volkow, 2021). Impaired learning from negative feedback in SUD (Chapter 3) highlights the ineffectiveness of punitive measures against addicted users. While negative consequences may act as a preventative deterrent under normal circumstances, they do little in the context of SUD. On the contrary, sentencing addicted individuals with certain punishments (e.g. imprisonment), or socially ostracising them so they will "learn their lesson" might do more harm than good. A clear example of this is the failure of the "war on drugs" campaign by the United States Government, of which punitive measures have not only failed to reduce the prevalence of drug addiction, but also resulted in an increase of preventable deaths by drug overdose. Imprisonment has been shown to actually increase the risk of overdose in former inmates (Binswanger et al., 2007). Moreover, a more recent analysis suggests that there is no relationship between punishment and drug misuse (The Pew Charitable Trusts, 2018). Instead of punitive measures, it may be more productive to adopt positive reinforcement-based treatment approaches with suitable incentives that promote abstinence and self-control. One such example is contingency management, an intervention based on positive reinforcement, which showed some efficacy in promoting abstinence and behavioural change in addicted individuals (Petry et al., 2017).

In light of the habit predominance in SUD, it has been suggested that behavioural interventions can focus on the training of new habits to replace maladaptive ones. Since SUD patients have an elevated tendency for automatic habits (Chapter 6) and a reduced goal-directed system (Chapters 3, 4, 5), interventions that promote positive habits may be more effective than strategies that that require the goal-directed system (e.g. punishment or inhibition). Further, simple but repetitive strategies may be easier to apply in SUD, considering that a subset of these patients favour a simpler rule-based strategy during learning.

8.7 Limitations

This thesis mainly focused on the behavioural aspects of reinforcement learning in human SUD. However, there are several methodological and conceptual limitations that need to be considered when interpreting the findings. First, the lockdown that followed the COVID-19 pandemic has forced me to conduct an online study on alcohol and habits (Chapter 7). Although the convenience and speed of online recruitment is advantageous, it should be noted that the data quality might differ to other studies where in-person screening and data collection were more stringent.

Second, the work presented here lacks functional neuroimaging methods to elucidate the neural processes of reinforcement learning. This is especially important, given that the hypothesis for reinforcement learning impairments in SUD is derived from the putative neuroadpative changes in learning circuits in the brain. At times, I have speculated about the neural substrates involved in these impairments, but without functional neuroimaging, these speculations need to be cautiously interpreted.

Third, this thesis reported primarily on data in cocaine use disorder. Although cocaine use disorder is representative of SUD, it may be qualitatively different to other substances such as alcohol and opioids. Although the prevailing assumption is that addictive behaviours in SUD share certain neurobiological (e.g. altered dopaminergic pathways) and psychological characteristics (e.g. risky use) (Fernández-Serrano et al., 2011; Nestler, 2005), the current findings need to be interpreted in view of the potential differences between different classes of substances.

Fourth, in most chapters, the IQ levels were not matched between the control and drug-user groups. This posed a problem for the interpretation of the reinforcement learning data, given that higher IQ plausibly leads to better learning performance (van den Bos et al., 2012). However, in the context of this thesis, statistically controlling for IQ during analyses would not be appropriate because lower verbal IQ seems to be an inherent feature of the drug-user groups.

Most SUD patients in my samples originated from socio-economically disadvantaged cohorts, and some dropped out of school early due to drug abuse problems – a common observation amongst youths with problematic substance use (Annis & Watson, 1975; Townsend et al., 2007). Whilst statistical treatments to account for IQ differences may not be strategic here, the replication of the main findings in an IQ-matched sample (i.e. Chapter 3) and the lack of association between IQ and learning performance in several chapters is reassuring. Nevertheless, these IQ differences need to be borne in mind when interpreting the data.

Fifth, my work here on reinforcement learning only investigates symbolic or monetary reinforcement, which, although common within the literature, is a limited view of reinforcement. The effects of primary reinforcers (e.g. food) and social reinforcers (e.g. peer approval) are sparsely studied in SUD. This is worth considering because drug-taking in SUD can sometimes involve a social component, which could also differ depending on the classes of substances (e.g. cocaine at parties, heroin at home) (Badiani et al., 2011). Whether social reinforcement, for example, modulates behaviour more successfully than conventional reinforcement, remains an open question.

Sixth, although this thesis investigates reinforcement learning carefully, the role of working memory was not assessed here. Working memory is thought to play a role in reinforcement learning, particularly in early learning (Collins & Frank, 2012). One study has found that SUD patients have impairments in working memory, presumably due to prefrontal deficits (Goldstein et al., 2004). Other work also showed that individuals with poorer working memory capacity performed less well in complex reinforcement learning when under stress, which in itself is known to impair prefrontal working memory (Otto et al., 2013). Whilst I am primarily interested in reinforcement learning as a whole, the influence of working memory on learning should be acknowledged.

Lastly, it should be acknowledged that aberrant learning does not explain every aspect of maladaptive behaviours in SUD. Although work in this thesis has contributed to our understanding from multiple perspectives, it only addresses the tip of the iceberg, as addictive disorders are multidimensional. For example, the mechanisms behind dysregulated affective

states in SUD (Albein-Urios et al., 2014; Fox et al., 2007) may not be sufficiently understood with reinforcement learning alone.

8.8 Future outlook

In view of the conceptual and methodological limitations, as well as unanswered questions from this thesis, I propose the following to be considered as the premise for future work:

(1) Future studies of reinforcement learning in SUD should expand beyond categorical diagnoses and adopt a more continuous approach. Whilst categorical diagnostic tools have been immensely useful in facilitating classification and treatment of psychiatric disorders, they are limited by the lack of predictive value in treatment prognosis, possibly due to the vast heterogeneity within a singular diagnosis (Insel, 2014). Further, these symptoms-based diagnoses may not capture the underlying mechanisms that could be dysfunctional (Insel, 2014). One promising approach is to adopt a trans-diagnostic approach in studying reinforcement learning, such as those proposed by the Research Domain Criterion (RDoC) (Insel et al., 2010).

(2) The effects of different classes of drugs on reinforcement learning (i.e. cannabis, opioids) should not be ignored, and there is a need to study them extensively as evidence is scarce.

(3) As alluded to earlier, the characterisation of reinforcement learning would not be complete without understanding the effects of different modalities of reinforcers. Future reinforcement learning studies in SUD should thus consider comparing the effects of social (e.g. peer approval or exclusion) or primary reinforcers (e.g. palatable foods or bitter solutions).

(4) The influence of contexts in modulating learned behaviours, as exemplified by Pavlovian-to-Instrumental Transfer (PIT), is unexplored in human SUD. This is especially relevant in the study of relapse, where environmental cues associated with drugs are thought to attain motivational salience, which triggers drug use (Hogarth et al., 2013). Given the role of Pavlovian-conditioned drug cues in invigorating drug-taking in animal models (Corbit & Janak, 2007; LeBlanc et al., 2012), it is of particular importance to translate these findings in human SUD. There is already preliminary work is this direction in alcohol use disorder (Chen et al., 2021; Garbusow et al., 2014), which is promising, but a more concerted and substantive effort is needed.

(5) The accelerated habit formation in SUD is thought to underpin the transition to compulsions in SUD (Everitt & Robbins, 2005). Yet how or whether this transition occurs in human SUD is unknown. One approach that can be used to test this hypothesis is to study, with neuroimaging, the process of habit formation in SUD. I acknowledge that studying habit formation in humans is notoriously difficult, but promising efforts have been made, at least in healthy humans, to understand the dynamics of goal-directed and habit systems over the course of learning (Zwosta et al., 2018).

(6) Not everyone who uses drugs becomes addicted. Given that there are genetic vulnerabilities that could exacerbate SUD development (e.g. weakened learning from errors (Klein et al., 2007)), it is possible that there are reinforcement learning impairments that precede the development of SUD. To the best of my knowledge, this has not been explored in the literature.

8.9 Concluding remarks

This thesis attempted to address a gap in the literature, namely the characterisation of maladaptive behaviour in SUD with a reinforcement learning framework. The synthesis of my current findings led me to conclude that learning and memory processes that facilitate goal-directed behaviour are impaired in severe SUD, which provides a putative explanation for why maladaptive behaviours persist in SUD. However, these deficits are not only insufficient to account for habit predominance in human SUD, they also do not manifest behaviourally at early stages of harmful drug use. Future work should be focused on elucidating the neurobiological mechanisms that underpin behaviour, which could lead to promising findings that are clinically informative.

Bibliography

Abbas, A. I., Hedlund, P. B., Huang, X.-P., Tran, T. B., Meltzer, H. Y., & Roth, B. L. (2009). Amisulpride is a potent 5-HT7 antagonist: Relevance for antidepressant actions in vivo. *Psychopharmacology*, *205*(1), 119–128. https://doi.org/10.1007/s00213-009-1521-8

Adams, C. D., & Dickinson, A. (1981). Instrumental Responding following Reinforcer Devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *33*(2b), 109–121. https://doi.org/10.1080/14640748108400816

Ahn, W.-Y., Vasilev, G., Lee, S.-H., Busemeyer, J. R., Kruschke, J. K., Bechara, A., & Vassileva, J. (2014). Decision-making in stimulant and opiate addicts in protracted abstinence: Evidence from computational modeling with pure users. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00849

al'Absi, M., Nakajima, M., Allen, S., Lemieux, A., & Hatsukami, D. (2015). Sex Differences in Hormonal Responses to Stress and Smoking Relapse: A Prospective Examination. *Nicotine & Tobacco Research*, *17*(4), 382–389. https://doi.org/10.1093/ntr/ntu340

Albein-Urios, N., Verdejo-Román, J., Asensio, S., Soriano-Mas, C., Martínez-González, J. M., & Verdejo-García, A. (2014). Re-appraisal of negative emotions in cocaine dependence: Dysfunctional corticolimbic activation and connectivity. *Addiction Biology*, *19*(3), 415–426. https://doi.org/10.1111/j.1369-1600.2012.00497.x

Allom, V., Panetta, G., Mullan, B., & Hagger, M. S. (2016). Self-report and behavioural approaches to the measurement of self-control: Are we assessing the same construct? *Personality and Individual Differences*, *90*, 137–142. https://doi.org/10.1016/j.paid.2015.10.051

Aloi, J., Blair, K. S., Crum, K. I., Bashford-Largo, J., Zhang, R., Lukoff, J., Carollo, E., White, S. F., Hwang, S., Filbey, F. M., Dobbertin, M., & Blair, R. J. R. (2020). Alcohol Use Disorder, But Not Cannabis Use Disorder, Symptomatology in Adolescents Is Associated With Reduced Differential Responsiveness to Reward Versus Punishment Feedback During Instrumental Learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(6), 610–618. https://doi.org/10.1016/j.bpsc.2020.02.003

Alsiö, J., Phillips, B. U., Sala-Bayo, J., Nilsson, S. R. O., Calafat-Pla, T. C., Rizwand, A., Plumbridge, J. M., López-Cruz, L., Dalley, J. W., Cardinal, R. N., Mar, A. C., & Robbins, T. W. (2019). Dopamine D2-like receptor stimulation blocks negative feedback in visual and spatial reversal learning in the rat: Behavioural and computational evidence. *Psychopharmacology*, *236*(8), 2307–2323. https://doi.org/10.1007/s00213-019-05296-y

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders—Text Revision* (4th ed.). American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Association.

Annis, H., & Watson, C. (1975). Drug Use and School Dropout: A Longitudinal Study. *Canadian Journal of Counselling and Psychotherapy*, *9*(3–4), Article 3–4. https://cjc-rcc.ucalgary.ca/article/view/60006

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Annual Review of Psychology*, *56*(1), 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

Averbeck, B. B., & Costa, V. D. (2017). Motivational neural circuits underlying reinforcement learning. *Nature Neuroscience*, *20*(4), 505–512. https://doi.org/10.1038/nn.4506

Averbeck, B. B., & O'Doherty, J. P. (2021). Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*, 1–16. https://doi.org/10.1038/s41386-021-01108-0

Badiani, A., Belin, D., Epstein, D., Calu, D., & Shaham, Y. (2011). Opiate versus psychostimulant addiction: The differences do matter. *Nature Reviews Neuroscience*, *12*(11), 685–700. https://doi.org/10.1038/nrn3104

Baler, R. D., & Volkow, N. D. (2006). Drug addiction: The neurobiology of disrupted self-control. *Trends in Molecular Medicine*, *12*(12), 559–566. https://doi.org/10.1016/j.molmed.2006.10.005

Balleine, B. W. (2019). The Meaning of Behavior: Discriminating Reflex and Volition in the Brain. *Neuron*, *104*(1), 47–62. https://doi.org/10.1016/j.neuron.2019.09.024

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4), 407–419. https://doi.org/10.1016/S0028-3908(98)00033-1

Balleine, B. W., & O'Doherty, J. P. (2010). Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. *Neuropsychopharmacology*, *35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Barker, J. M., Corbit, L. H., Robinson, D. L., Gremel, C. M., Gonzales, R. A., & Chandler, L. J. (2015). Corticostriatal circuitry and habitual ethanol seeking. *Alcohol*, *49*(8), 817–824. https://doi.org/10.1016/j.alcohol.2015.03.003

Baumeister, R. F. (2003). Ego Depletion and Self-Regulation Failure: A Resource Model of Self-Control. *Alcoholism: Clinical and Experimental Research*, *27*(2), 281–284. https://doi.org/10.1097/01.ALC.0000060879.61384.A4

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

Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: A neurocognitive perspective. *Nature Neuroscience*, *8*(11), 1458.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1), 7–15. https://doi.org/10.1016/0010-0277(94)90018-3

Bechara, A., & Damasio, H. (2002). Decision-making and addiction (part I): Impaired activation of somatic states in substance dependent individuals when pondering decisions with negative future consequences. *Neuropsychologia*, *40*(10), 1675–1689. https://doi.org/10.1016/S0028-3932(02)00015-5

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding Advantageously Before Knowing the Advantageous Strategy. *Science*, *275*(5304), 1293–1295. https://doi.org/10.1126/science.275.5304.1293

Bechara, A., Dolan, S., Denburg, N., Hindes, A., Anderson, S. W., & Nathan, P. E. (2001). Decision-making deficits, linked to a dysfunctional ventromedial prefrontal cortex, revealed in alcohol and stimulant abusers. *Neuropsychologia*, *39*(4), 376–389. https://doi.org/10.1016/S0028-3932(00)00136-6

Bechara, A., Dolan, S., & Hindes, A. (2002). Decision-making and addiction (part II): Myopia for the future or hypersensitivity to reward? *Neuropsychologia*, *40*(10), 1690–1705. https://doi.org/10.1016/S0028-3932(02)00016-7

Becker, J. B. (1999). Gender Differences in Dopaminergic Function in Striatum and Nucleus Accumbens. *Pharmacology Biochemistry and Behavior*, *64*(4), 803–812. https://doi.org/10.1016/S0091-3057(99)00168-9

Becker, J. B., & Chartoff, E. (2019). Sex differences in neural mechanisms mediating reward and addiction. *Neuropsychopharmacology*, *44*(1), 166–183. https://doi.org/10.1038/s41386-018-0125-6

Becker, J. B., & Hu, M. (2008). Sex differences in drug abuse. *Frontiers in Neuroendocrinology*, *29*(1), 36–47. https://doi.org/10.1016/j.yfrne.2007.07.003

Becker, J. B., & Koob, G. F. (2016). Sex Differences in Animal Models: Focus on Addiction. *Pharmacological Reviews*, *68*(2), 242–263. https://doi.org/10.1124/pr.115.011163

Becker, J. B., McClellan, M. L., & Reed, B. G. (2017). Sex differences, gender and addiction. *Journal of Neuroscience Research*, *95*(1–2), 136–147. https://doi.org/10.1002/jnr.23963

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Belin, D., & Everitt, B. J. (2008). Cocaine Seeking Habits Depend upon Dopamine-Dependent Serial Connectivity Linking the Ventral with the Dorsal Striatum. *Neuron*, *57*(3), 432–441. https://doi.org/10.1016/j.neuron.2007.12.019

Bello, E. P., Mateo, Y., Gelman, D. M., Noaín, D., Shin, J. H., Low, M. J., Alvarez, V. A., Lovinger, D. M., & Rubinstein, M. (2011). Cocaine supersensitivity and enhanced motivation for reward in mice lacking dopamine D_2 autoreceptors. *Nature Neuroscience*, *14*(8), 1033–1038. https://doi.org/10.1038/nn.2862

Berridge, K. C., & Robinson, T. E. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *American Psychologist*, *71*(8), 670–679. https://doi.org/10.1037/amp0000059

Binswanger, I. A., Stern, M. F., Deyo, R. A., Heagerty, P. J., Cheadle, A., Elmore, J. G., & Koepsell, T. D. (2007). Release from Prison—A High Risk of Death for Former Inmates. *New England Journal of Medicine*, *356*(2), 157–165. https://doi.org/10.1056/NEJMsa064115

Bland, A. R., Roiser, J. P., Mehta, M. A., Schei, T., Boland, H., Campbell-Meiklejohn, D. K., Emsley, R. A., Munafo, M. R., Penton-Voak, I. S., Seara-Cardoso, A., Viding, E., Voon, V., Sahakian, B. J., Robbins, T. W., & Elliott, R. (2016). EMOTICOM: A Neuropsychological Test Battery to Evaluate Emotion, Motivation, Impulsivity, and Social Cognition. *Frontiers in Behavioral Neuroscience*, *10*. https://doi.org/10.3389/fnbeh.2016.00025

Blum, K., Noble, E. P., Sheridan, P. J., Montgomery, A., Ritchie, T., Jagadeeswaran, P., Nogami, H., Briggs, A. H., & Cohn, J. B. (1990). Allelic Association of Human Dopamine D2 Receptor Gene in Alcoholism. *JAMA*, *263*(15), 2055–2060. https://doi.org/10.1001/jama.1990.03440150063027

Bogdanov, M., Timmermann, J. E., Gläscher, J., Hummel, F. C., & Schwabe, L. (2018). Causal role of the inferolateral prefrontal cortex in balancing goal-directed and habitual control of behavior. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-27678-6

Bolla, K. I., Eldreth, D. A., London, E. D., Kiehl, K. A., Mouratidis, M., Contoreggi, C., Matochik, J. A., Kurian, V., Cadet, J. L., Kimes, A. S., Funderburk, F. R., & Ernst, M. (2003). Orbitofrontal cortex dysfunction in abstinent cocaine abusers performing a decision-making task. *NeuroImage*, *19*(3), 1085–1094. https://doi.org/10.1016/S1053-8119(03)00113-7

Boorman, E. D., Rajendran, V. G., O'Reilly, J. X., & Behrens, T. E. (2016). Two Anatomically and Computationally Distinct Learning Signals Predict Changes to Stimulus-Outcome Associations in Hippocampus. *Neuron*, *89*(6), 1343–1354. https://doi.org/10.1016/j.neuron.2016.02.014

Boureau, Y.-L., & Dayan, P. (2011). Opponency Revisited: Competition and Cooperation Between Dopamine and Serotonin. *Neuropsychopharmacology*, *36*(1), 74–97. https://doi.org/10.1038/npp.2010.151

Bradfield, L. A., & Balleine, B. W. (2013). Hierarchical and binary associations compete for behavioral control during instrumental biconditional discrimination. *Journal of Experimental Psychology. Animal Behavior Processes*, *39*(1), 2–13. https://doi.org/10.1037/a0030941

Bradfield, L. A., Leung, B. K., Boldt, S., Liang, S., & Balleine, B. W. (2020). Goal-directed actions transiently depend on dorsal hippocampus. *Nature Neuroscience*, *23*(10), 1194–1197. https://doi.org/10.1038/s41593-020-0693-8

Breedon, J. R., Ziauddeen, H., Stochl, J., & Ersche, K. D. (2021). Feeding the addiction: Narrowing of goals to habits. *European Neuropsychopharmacology*, *42*, 110–114. https://doi.org/10.1016/j.euroneuro.2020.11.002

Brevers, D., Bechara, A., Cleeremans, A., Kornreich, C., Verbanck, P., & Noël, X. (2014). Impaired Decision-Making Under Risk in Individuals with Alcohol Dependence. *Alcoholism: Clinical and Experimental Research*, *38*(7), 1924–1931. https://doi.org/10.1111/acer.12447

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Brovelli, A., Nazarian, B., Meunier, M., & Boussaoud, D. (2011). Differential roles of caudate nucleus and putamen during instrumental learning. *NeuroImage*, *57*(4), 1580–1590. https://doi.org/10.1016/j.neuroimage.2011.05.059

Brown, J. W., & Braver, T. S. (2005). Learned Predictions of Error Likelihood in the Anterior Cingulate Cortex. *Science*, *307*(5712), 1118–1121. https://doi.org/10.1126/science.1105783

Buckholtz, J. W., Treadway, M. T., Cowan, R. L., Woodward, N. D., Li, R., Ansari, M. S., Baldwin, R. M., Schwartzman, A. N., Shelby, E. S., Smith, C. E., Kessler, R. M., & Zald, D. H. (2010). Dopaminergic Network Differences in Human Impulsivity. *Science*, *329*(5991), 532–532. https://doi.org/10.1126/science.1185778

Burke, C. J., Soutschek, A., Weber, S., Raja Beharelle, A., Fehr, E., Haker, H., & Tobler, P. N. (2018). Dopamine Receptor-Specific Contributions to the Computation of Value. *Neuropsychopharmacology*, *43*(6), 1415–1424. https://doi.org/10.1038/npp.2017.302

Burton, A. C., Bissonette, G. B., Vazquez, D., Blume, E. M., Donnelly, M., Heatley, K. C., Hinduja, A., & Roesch, M. R. (2018). Previous cocaine self-administration disrupts reward expectancy encoding in ventral striatum. *Neuropsychopharmacology*, *43*(12), 2350–2360. https://doi.org/10.1038/s41386-018-0058-0

Burton, R., Henn, C., Lavoie, D., O'Connor, R., Perkins, C., Sweeney, K., Greaves, F., Ferguson, B., Beynon, C., Belloni, A., Musto, V., Marsden, J., & Sheron, N. (2017). A rapid evidence review of the effectiveness and cost-effectiveness of alcohol control policies: An English perspective. *The Lancet*, *389*(10078), 1558–1580. https://doi.org/10.1016/S0140-6736(16)32420-5

Calu, D. J., Stalnaker, T. A., Franz, T. M., Singh, T., Shaham, Y., & Schoenbaum, G. (2007). Withdrawal from cocaine self-administration produces long-lasting deficits in orbitofrontal-dependent reversal learning in rats. *Learning & Memory*, *14*(5), 325–328. https://doi.org/10.1101/lm.534807

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1), 1–29. https://doi.org/10.18637/jss.v076.i01

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009

Chen, H., Nebe, S., Mojtahedzadeh, N., Kuitunen-Paul, S., Garbusow, M., Schad, D. J., Rapp, M. A., Huys, Q. J. M., Heinz, A., & Smolka, M. N. (2021). Susceptibility to interference between Pavlovian and instrumental control is associated with early hazardous alcohol use. *Addiction Biology*, *26*(4), e12983. https://doi.org/10.1111/adb.12983

Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., & Rubia, K. (2013). Neural and Psychological Maturation of Decision-making in Adolescence and Young Adulthood. *Journal of Cognitive Neuroscience*, *25*(11), 1807–1823. https://doi.org/10.1162/jocn_a_00447

Clark, L., Boileau, I., & Zack, M. (2019). Neuroimaging of reward mechanisms in Gambling disorder: An integrative review. *Molecular Psychiatry*, *24*(5), 674–693. https://doi.org/10.1038/s41380-018-0230-2

Clarke, H. F., Cardinal, R. N., Rygula, R., Hong, Y. T., Fryer, T. D., Sawiak, S. J., Ferrari, V., Cockcroft, G., Aigbirhio, F. I., Robbins, T. W., & Roberts, A. C. (2014). Orbitofrontal Dopamine Depletion Upregulates Caudate Dopamine and Alters Behavior via Changes in Reinforcement Sensitivity. *Journal of Neuroscience*, *34*(22), 7663–7676. https://doi.org/10.1523/JNEUROSCI.0718-14.2014

Cohen, M. X., Krohn-Grimberghe, A., Elger, C. E., & Weber, B. (2007). Dopamine gene predicts the brain's response to dopaminergic drug. *European Journal of Neuroscience*, *26*(12), 3652–3660. https://doi.org/10.1111/j.1460-9568.2007.05947.x

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

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

Cools, R., Clark, L., Owen, A. M., & Robbins, T. W. (2002). Defining the Neural Mechanisms of Probabilistic Reversal Learning Using Event-Related Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *22*(11), 4563–4567. https://doi.org/10.1523/JNEUROSCI.22-11-04563.2002

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. *Journal of Neuroscience*, *29*(5), 1538–1543. https://doi.org/10.1523/JNEUROSCI.4467-08.2009

Cools, R., Nakamura, K., & Daw, N. D. (2011). Serotonin and Dopamine: Unifying Affective, Activational, and Decision Functions. *Neuropsychopharmacology*, *36*(1), 98–113. https://doi.org/10.1038/npp.2010.121

Corbit, L. H., Chieng, B. C., & Balleine, B. W. (2014). Effects of Repeated Cocaine Exposure on Habit Learning and Reversal by N-Acetylcysteine. *Neuropsychopharmacology*, *39*(8), 1893–1901. https://doi.org/10.1038/npp.2014.37

Corbit, L. H., & Janak, P. H. (2007). Ethanol-Associated Cues Produce General Pavlovian-Instrumental Transfer. *Alcoholism: Clinical and Experimental Research*, *31*(5), 766–774. https://doi.org/10.1111/j.1530-0277.2007.00359.x

Corbit, L. H., & Janak, P. H. (2016). Habitual Alcohol Seeking: Neural Bases and Possible Relations to Alcohol Use Disorders. *Alcoholism: Clinical and Experimental Research*, *40*(7), 1380–1389. https://doi.org/10.1111/acer.13094

Corbit, L. H., Nie, H., & Janak, P. H. (2012). Habitual Alcohol Seeking: Time Course and the Contribution of Subregions of the Dorsal Striatum. *Biological Psychiatry*, *72*(5), 389–395. https://doi.org/10.1016/j.biopsych.2012.02.024

Corbit, L. H., Nie, H., & Janak, P. H. (2014). Habitual responding for alcohol depends upon both AMPA and D2 receptor signaling in the dorsolateral striatum. *Frontiers in Behavioral Neuroscience*, *8*. https://doi.org/10.3389/fnbeh.2014.00301

Corlett, P. R., & Fletcher, P. C. (2014). Computational psychiatry: A Rosetta Stone linking the brain to mental illness. *The Lancet Psychiatry*, *1*(5), 399–402. https://doi.org/10.1016/S2215-0366(14)70298-6

Cox, S. M. L., Frank, M. J., Larcher, K., Fellows, L. K., Clark, C. A., Leyton, M., & Dagher, A. (2015). Striatal D1 and D2 signaling differentially predict learning from positive and negative outcomes. *Neuroimage*, *109*, 95–101. https://doi.org/10.1016/j.neuroimage.2014.12.070

Dalla, C., & Shors, T. J. (2009). Sex differences in learning processes of classical and operant conditioning. *Physiology & Behavior*, *97*(2), 229–238. https://doi.org/10.1016/j.physbeh.2009.02.035

Dalley, J. W., Fryer, T. D., Brichard, L., Robinson, E. S. J., Theobald, D. E. H., Lääne, K., Peña, Y., Murphy, E. R., Shah, Y., Probst, K., Abakumova, I., Aigbirhio, F. I., Richards, H. K., Hong, Y., Baron, J.-C., Everitt, B. J., & Robbins, T. W. (2007). Nucleus Accumbens D2/3 Receptors Predict Trait Impulsivity and Cocaine Reinforcement. *Science*, *315*(5816), 1267–1270. https://doi.org/10.1126/science.1137073

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD Responses Reflecting Dopaminergic Signals in the Human Ventral Tegmental Area. *Science*, *319*(5867), 1264–1267. https://doi.org/10.1126/science.1150605

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision Making, Affect, and Learning: Attention and Performance XXIII* (pp. 3–39). Oxford University Press.

Daw, N. D. (2014). Advanced Reinforcement Learning. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics* (2nd ed., pp. 299–320). Academic Press. https://doi.org/10.1016/B978-0-12-416008-8.00016-4

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4), 603–616. https://doi.org/10.1016/S0893-6080(02)00052-7

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

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453. https://doi.org/10.3758/CABN.8.4.429

de Wit, S., Corlett, P. R., Aitken, M. R., Dickinson, A., & Fletcher, P. C. (2009). Differential Engagement of the Ventromedial Prefrontal Cortex by Goal-Directed and Habitual Behavior toward Food Pictures in Humans. *Journal of Neuroscience*, *29*(36), 11330–11338. https://doi.org/10.1523/JNEUROSCI.1639-09.2009

de Wit, S., & Dickinson, A. (2009). Associative theories of goal-directed behaviour: A case for animal–human translational models. *Psychological Research*, *73*(4), 463–476. https://doi.org/10.1007/s00426-009-0230-6

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A. C., Robbins, T. W., Gasull-Camos, J., Evans, M., Mirza, H., & Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology: General*, *147*(7), 1043. https://doi.org/10.1037/xge0000402

de Wit, S., Niry, D., Wariyar, R., Aitken, M. R. F., & Dickinson, A. (2007). Stimulus-outcome interactions during instrumental discrimination learning by rats and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(1), 1–11. https://doi.org/10.1037/0097-7403.33.1.1

de Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., Vijver, I. van de, & Ridderinkhof, K. R. (2012). Corticostriatal Connectivity Underlies Individual Differences in the Balance between Habitual and Goal-Directed Action Control. *Journal of Neuroscience*, *32*(35), 12066–12075. https://doi.org/10.1523/JNEUROSCI.1088-12.2012

Degenhardt, L., Baxter, A. J., Lee, Y. Y., Hall, W., Sara, G. E., Johns, N., Flaxman, A., Whiteford, H. A., & Vos, T. (2014). The global epidemiology and burden of psychostimulant dependence: Findings from the Global Burden of Disease Study 2010. *Drug and Alcohol Dependence*, *137*, 36–47. https://doi.org/10.1016/j.drugalcdep.2013.12.025

Deserno, L., Beck, A., Huys, Q. J. M., Lorenz, R. C., Buchert, R., Buchholz, H.-G., Plotkin, M., Kumakara, Y., Cumming, P., Heinze, H.-J., Grace, A. A., Rapp, M. A., Schlagenhauf, F., & Heinz, A. (2015). Chronic alcohol intake abolishes the relationship between dopamine synthesis capacity and learning signals in the ventral striatum. *European Journal of Neuroscience*, *41*(4), 477–486. https://doi.org/10.1111/ejn.12802

Di Chiara, G., & Imperato, A. (1988). Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats. *Proceedings of the National Academy of Sciences*, *85*(14), 5274–5278. https://doi.org/10.1073/pnas.85.14.5274

Dias-Ferreira, E., Sousa, J. C., Melo, I., Morgado, P., Mesquita, A. R., Cerqueira, J. J., Costa, R. M., & Sousa, N. (2009). Chronic Stress Causes Frontostriatal Reorganization and Affects Decision-Making. *Science*, *325*(5940), 621–625. https://doi.org/10.1126/science.1171203

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Phil. Trans. R. Soc. Lond. B*, *308*(1135), 67–78. https://doi.org/10.1098/rstb.1985.0010

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18. https://doi.org/10.3758/BF03199951

Dickinson, A., Wood, N., & Smith, J. W. (2002). Alcohol Seeking by Rats: Action or Habit? *The Quarterly Journal of Experimental Psychology Section B*, *55*(4b), 331–348. https://doi.org/10.1080/0272499024400016

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Doll, B. B., Bath, K. G., Daw, N. D., & Frank, M. J. (2016). Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. *Journal of Neuroscience*, *36*(4), 1211–1222. https://doi.org/10.1523/JNEUROSCI.1901-15.2016

Doñamayor, N., Strelchuk, D., Baek, K., Banca, P., & Voon, V. (2018). The involuntary nature of binge drinking: Goal directedness and awareness of intention. *Addiction Biology*, *23*(1), 515–526. https://doi.org/10.1111/adb.12505

Eichenbaum, H., Otto, T., & Cohen, N. J. (1994). Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences*, *17*(3), 449–472. https://doi.org/10.1017/S0140525X00035391

Eisenegger, C., Naef, M., Linssen, A., Clark, L., Gandamaneni, P. K., Mueller, U., & Robbins, T. W. (2014). Role of Dopamine D2 Receptors in Human Reinforcement Learning. *Neuropsychopharmacology*, *39*(10), 2366–2375. https://doi.org/10.1038/npp.2014.84

Elliott, R., Dolan, R. J., & Frith, C. D. (2000). Dissociable Functions in the Medial and Lateral Orbitofrontal Cortex: Evidence from Human Neuroimaging Studies. *Cerebral Cortex*, *10*(3), 308–317. https://doi.org/10.1093/cercor/10.3.308

Elliott, R., Friston, K. J., & Dolan, R. J. (2000). Dissociable neural responses in human reward systems. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *20*(16), 6159–6165.

Epstein, D. H. (2020). Let's agree to agree: A comment on Hogarth (2020), with a plea for not-so-competing theories of addiction. *Neuropsychopharmacology*, *45*(5), 715–716. https://doi.org/10.1038/s41386-020-0618-y

Ernst, M., & Paulus, M. P. (2005). Neurobiology of Decision Making: A Selective Review from a Neurocognitive and Clinical Perspective. *Biological Psychiatry*, *58*(8), 597–604. https://doi.org/10.1016/j.biopsych.2005.06.004

Ersche, K. D., Bullmore, E. T., Craig, K. J., Shabbir, S. S., Abbott, S., Müller, U., Ooi, C., Suckling, J., Barnes, A., Sahakian, B. J., Merlo-Pich, E. V., & Robbins, T. W. (2010). Influence of Compulsivity of Drug Abuse on Dopaminergic Modulation of Attentional Bias in Stimulant Dependence. *Archives of General Psychiatry*, *67*(6), 632–644. https://doi.org/10.1001/archgenpsychiatry.2010.60

Ersche, K. D., Clark, L., London, M., Robbins, T. W., & Sahakian, B. J. (2006). Profile of Executive and Memory Function Associated with Amphetamine and Opiate Dependence. *Neuropsychopharmacology*, *31*(5), 1036–1047. https://doi.org/10.1038/sj.npp.1300889

Ersche, K. D., Fletcher, P. C., Lewis, S. J. G., Clark, L., Stocks-Gee, G., London, M., Deakin, J. B., Robbins, T. W., & Sahakian, B. J. (2005). Abnormal frontal activations related to decision-making in current and former amphetamine and opiate dependent individuals. *Psychopharmacology*, *180*(4), 612–623. https://doi.org/10.1007/s00213-005-2205-7

Ersche, K. D., Gillan, C. M., Jones, P. S., Williams, G. B., Ward, L. H. E., Luijten, M., de Wit, S., Sahakian, B. J., Bullmore, E. T., & Robbins, T. W. (2016). Carrots and sticks fail to change behavior in cocaine addiction. *Science*, *352*(6292), 1468–1471. https://doi.org/10.1126/science.aaf3700

Ersche, K. D., Hagan, C. C., Smith, D. G., Abbott, S., Jones, P. S., Apergis-Schoute, A. M., & Döffinger, R. (2014). Aberrant Disgust Responses and Immune Reactivity in Cocaine-Dependent Men. *Biological Psychiatry*, *75*(2), 140–147. https://doi.org/10.1016/j.biopsych.2013.08.004

Ersche, K. D., Jones, P. S., Williams, G. B., Turton, A. J., Robbins, T. W., & Bullmore, E. T. (2012). Abnormal Brain Structure Implicated in Stimulant Drug Addiction. *Science*, *335*, 601–604.

Ersche, K. D., Lim, T. V., Murley, A. G., Rua, C., Vaghi, M. M., White, T. L., Williams, G. B., & Robbins, T. W. (2021). Reduced Glutamate Turnover in the Putamen Is Linked With Automatic Habits in Human Cocaine Addiction. *Biological Psychiatry*, *89*(10), 970–979. https://doi.org/10.1016/j.biopsych.2020.12.009

Ersche, K. D., Lim, T.-V., Ward, L. H. E., Robbins, T. W., & Stochl, J. (2017). Creature of Habit: A self-report measure of habitual routines and automatic tendencies in everyday life. *Personality and Individual Differences*, *116*, 73–85. https://doi.org/10.1016/j.paid.2017.04.024

Ersche, K. D., Meng, C., Ziauddeen, H., Stochl, J., Williams, G. B., Bullmore, E. T., & Robbins, T. W. (2020). Brain networks underlying vulnerability and resilience to drug addiction. *Proceedings of the National Academy of Sciences*, *117*(26), 15253–15261. https://doi.org/10.1073/pnas.2002509117

Ersche, K. D., Roiser, J. P., Abbott, S., Craig, K. J., Müller, U., Suckling, J., Ooi, C., Shabbir, S. S., Clark, L., Sahakian, B. J., Fineberg, N. A., Merlo-Pich, E. V., Robbins, T. W., & Bullmore, E. T. (2011). Response Perseveration in Stimulant Dependence Is Associated with Striatal Dysfunction and Can Be Ameliorated by a D2/3 Receptor Agonist. *Biological Psychiatry*, *70*(8), 754–762. https://doi.org/10.1016/j.biopsych.2011.06.033

Ersche, K. D., Roiser, J. P., Lucas, M., Domenici, E., Robbins, T. W., & Bullmore, E. T. (2011). Peripheral biomarkers of cognitive response to dopamine receptor agonist treatment. *Psychopharmacology*, *214*(4), 779–789. https://doi.org/10.1007/s00213-010-2087-1

Ersche, K. D., Roiser, J. P., Robbins, T. W., & Sahakian, B. J. (2008). Chronic cocaine but not chronic amphetamine use is associated with perseverative responding in humans. *Psychopharmacology*, *197*(3), 421–431. https://doi.org/10.1007/s00213-007-1051-1

Ersche, K. D., Ward, L. H. E., Lim, T.-V., Lumsden, R. J., Sawiak, S. J., Robbins, T. W., & Stochl, J. (2019). Impulsivity and compulsivity are differentially associated with automaticity and routine on the Creature of Habit Scale. *Personality and Individual Differences*, *150*, 109493. https://doi.org/10.1016/j.paid.2019.07.003

Everitt, B. J., Dickinson, A., & Robbins, T. W. (2001). The neuropsychological basis of addictive behaviour. *Brain Research Reviews*, *36*(2), 129–138. https://doi.org/10.1016/S0165-0173(01)00088-1

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, *8*(11), 1481–1489. https://doi.org/10.1038/nn1579

Everitt, B. J., & Robbins, T. W. (2016). Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. *Annual Review of Psychology*, *67*(1), 23–50. https://doi.org/10.1146/annurev-psych-122414-033457

Feher da Silva, C., Victorino, C. G., Caticha, N., & Baldo, M. V. C. (2017). Exploration and recency as the main proximate causes of probability matching: A reinforcement learning analysis. *Scientific Reports*, *7*(1), 15326. https://doi.org/10.1038/s41598-017-15587-z

Fernández-Serrano, M. J., Pérez-García, M., & Verdejo-García, A. (2011). What are the specific vs. Generalized effects of drugs of abuse on neuropsychological performance? *Neuroscience & Biobehavioral Reviews*, *35*(3), 377–406. https://doi.org/10.1016/j.neubiorev.2010.04.008

Festinger, D. S., Dugosh, K. L., Kirby, K. C., & Seymour, B. L. (2014). Contingency management for cocaine treatment: Cash vs. vouchers. *Journal of Substance Abuse Treatment*, *47*(2), 168–174. https://doi.org/10.1016/j.jsat.2014.03.001

Figee, M., Pattij, T., Willuhn, I., Luigjes, J., van den Brink, W., Goudriaan, A., Potenza, M. N., Robbins, T. W., & Denys, D. (2016). Compulsivity in obsessive–compulsive disorder and addictions. *European Neuropsychopharmacology*, *26*(5), 856–868. https://doi.org/10.1016/j.euroneuro.2015.12.003

Fineberg, N. A., Chamberlain, S. R., Goudriaan, A. E., Stein, D. J., Vanderschuren, L. J. M. J., Gillan, C. M., Shekar, S., Gorwood, P. A. P. M., Voon, V., Morein-Zamir, S., Denys, D., Sahakian, B. J., Moeller, F. G., Robbins, T. W., & Potenza, M. N. (2014). New developments in human neurocognition: Clinical, genetic, and brain imaging correlates of impulsivity and compulsivity. *CNS Spectrums*, *19*(1), 69–89. https://doi.org/10.1017/S1092852913000801

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

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (2002). *Structured Clinical Interview for DSM-IV-TR Axis-I Disorders, Research Version, Patient Edition (SCID-I/P-RV)*. Biometrics Research Department, New York State Psychiatric Institute.

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences*, *103*(31), 11778–11783. https://doi.org/10.1073/pnas.0602659103

Ford, C. P. (2014). The role of D2-autoreceptors in regulating dopamine neuron activity and transmission. *Neuroscience*, *282*, 13–22. https://doi.org/10.1016/j.neuroscience.2014.01.025

Fox, H. C., Axelrod, S. R., Paliwal, P., Sleeper, J., & Sinha, R. (2007). Difficulties in emotion regulation and impulse control during cocaine abstinence. *Drug and Alcohol Dependence*, *89*(2), 298–301. https://doi.org/10.1016/j.drugalcdep.2006.12.026

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, *17*(1), 51–72. https://doi.org/10.1162/0898929052880093

Frank, M. J., & Hutchison, K. (2009). Genetic contributions to avoidance-based decisions: Striatal D2 receptor polymorphisms. *Neuroscience*, *164*(1), 131–140. https://doi.org/10.1016/j.neuroscience.2009.04.048

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

Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: Psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience*, *120*(3), 497–517. https://doi.org/10.1037/0735-7044.120.3.497

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

Franken, I. H. A., Hendriks, V. M., & van den Brink, W. (2002). Initial validation of two opiate craving questionnaires: The Obsessive Compulsive Drug Use Scale and the Desires for Drug Questionnaire. *Addictive Behaviors*, *27*(5), 675–685. https://doi.org/10.1016/S0306-4603(01)00201-5

Friedel, E., Koch, S. P., Wendt, J., Heinz, A., Deserno, L., & Schlagenhauf, F. (2014). Devaluation and sequential decisions: Linking goal-directed and model-based behavior. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00587

Galandra, C., Crespi, C., Basso, G., & Canessa, N. (2020). Impaired learning from regret and disappointment in alcohol use disorder. *Scientific Reports*, *10*(1), 12104. https://doi.org/10.1038/s41598-020-68942-y

Gallagher, M., McMahan, R. W., & Schoenbaum, G. (1999). Orbitofrontal Cortex and Representation of Incentive Value in Associative Learning. *Journal of Neuroscience*, *19*(15), 6610–6614. https://doi.org/10.1523/JNEUROSCI.19-15-06610.1999

Garavan, H., Pankiewicz, J., Bloom, A., Cho, J.-K., Sperry, L., Ross, T. J., Salmeron, B. J., Risinger, R., Kelley, D., & Stein, E. A. (2000). Cue-Induced Cocaine Craving: Neuroanatomical Specificity for Drug Users and Drug Stimuli. *American Journal of Psychiatry*, *157*(11), 1789–1798. https://doi.org/10.1176/appi.ajp.157.11.1789

Garbusow, M., Schad, D. J., Sommer, C., Jünger, E., Sebold, M., Friedel, E., Wendt, J., Kathmann, N., Schlagenhauf, F., Zimmermann, U. S., Heinz, A., Huys, Q. J. M., & Rapp, M. A. (2014). Pavlovian-to-Instrumental Transfer in Alcohol Dependence: A Pilot Study. *Neuropsychobiology*, *70*(2), 111–121. https://doi.org/10.1159/000363507

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *ELife*, *6*, e17086. https://doi.org/10.7554/eLife.17086

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. https://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, 104394. https://doi.org/10.1016/j.cognition.2020.104394

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *ELife*, *5*, e11305. https://doi.org/10.7554/eLife.11305

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 523–536. https://doi.org/10.3758/s13415-015-0347-6

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & de Wit, S. (2011). Disruption in the Balance Between Goal-Directed Behavior and Habit Learning in Obsessive-Compulsive Disorder. *American Journal of Psychiatry*, *168*(7), 718–726. https://doi.org/10.1176/appi.ajp.2011.10071062

Gillan, C. M., Robbins, T. W., Sahakian, B. J., van den Heuvel, O. A., & van Wingen, G. (2016). The role of habit in compulsivity. *European Neuropsychopharmacology*, *26*(5), 828–840. https://doi.org/10.1016/j.euroneuro.2015.12.033

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Gleich, T., Spitta, G., Butler, O., Zacharias, K., Aydin, S., Sebold, M., Garbusow, M., Rapp, M., Schubert, F., Buchert, R., Heinz, A., & Gallinat, J. (2021). Dopamine D2/3 receptor availability in alcohol use disorder and individuals at high risk: Towards a dimensional approach. *Addiction Biology*, *26*(2), e12915. https://doi.org/10.1111/adb.12915

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 15647–15654. https://doi.org/10.1073/pnas.1014269108

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology. General*, *117*(3), 227–247. https://doi.org/10.1037//0096-3445.117.3.227

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do People Solve the "Weather Prediction" Task?: Individual Variability in Strategies for Probabilistic Category Learning. *Learning & Memory*, *9*(6), 408–418. https://doi.org/10.1101/lm.45202

Goldstein, R. Z., Craig, A. D. (Bud), Bechara, A., Garavan, H., Childress, A. R., Paulus, M. P., & Volkow, N. D. (2009). The Neurocircuitry of Impaired Insight in Drug Addiction. *Trends in Cognitive Sciences*, *13*(9), 372–380. https://doi.org/10.1016/j.tics.2009.06.004

Goldstein, R. Z., Leskovjan, A. C., Hoff, A. L., Hitzemann, R., Bashan, F., Khalsa, S. S., Wang, G.-J., Fowler, J. S., & Volkow, N. D. (2004). Severity of neuropsychological impairment in cocaine and alcohol addiction: Association with metabolism in the prefrontal cortex. *Neuropsychologia*, *42*(11), 1447–1458. https://doi.org/10.1016/j.neuropsychologia.2004.04.002

Goldstein, R. Z., Tomasi, D., Alia-Klein, N., Cottone, L. A., Zhang, L., Telang, F., & Volkow, N. D. (2007). Subjective sensitivity to monetary gradients is associated with frontolimbic activation to reward in cocaine abusers. *Drug and Alcohol Dependence*, *87*(2–3), 233–240. https://doi.org/10.1016/j.drugalcdep.2006.08.022

Goldstein, R. Z., Woicik, P. A., Maloney, T., Tomasi, D., Alia-Klein, N., Shan, J., Honorio, J., Samaras, D., Wang, R., Telang, F., Wang, G.-J., & Volkow, N. D. (2010). Oral methylphenidate normalizes cingulate activity in cocaine addiction during a salient cognitive task. *Proceedings of the National Academy of Sciences*, *107*(38), 16667–16672. https://doi.org/10.1073/pnas.1011455107

Gourley, S. L., Olevska, A., Gordon, J., & Taylor, J. R. (2013). Cytoskeletal Determinants of Stimulus-Response Habits. *Journal of Neuroscience*, *33*(29), 11811–11816. https://doi.org/10.1523/JNEUROSCI.1034-13.2013

Grace, A. A. (1995). The tonic/phasic model of dopamine system regulation: Its relevance for understanding how stimulant abuse can alter basal ganglia function. *Drug and Alcohol Dependence*, *37*(2), 111–129. https://doi.org/10.1016/0376-8716(94)01066-T

Graybiel, A. M. (1998). The Basal Ganglia and Chunking of Action Repertoires. *Neurobiology of Learning and Memory*, *70*(1), 119–136. https://doi.org/10.1006/nlme.1998.3843

Graybiel, A. M., & Grafton, S. T. (2015). The Striatum: Where Skills and Habits Meet. *Cold Spring Harbor Perspectives in Biology*, *7*(8), a021691. https://doi.org/10.1101/cshperspect.a021691

Greenfield, S. F., Back, S. E., Lawson, K., & Brady, K. T. (2010). Substance Abuse in Women. *Psychiatric Clinics*, *33*(2), 339–355. https://doi.org/10.1016/j.psc.2010.01.004

Groman, S. M., Hillmer, A. T., Liu, H., Fowles, K., Holden, D., Morris, E. D., Lee, D., & Taylor, J. R. (2020). Dysregulation of Decision Making Related to Metabotropic Glutamate 5, but Not Midbrain D3, Receptor Availability Following Cocaine Self-administration in Rats. *Biological Psychiatry*, *88*(10), 777–787. https://doi.org/10.1016/j.biopsych.2020.06.020

Groman, S. M., Rich, K. M., Smith, N. J., Lee, D., & Taylor, J. R. (2018). Chronic Exposure to Methamphetamine Disrupts Reinforcement-Based Decision Making in Rats. *Neuropsychopharmacology*, *43*(4), 770–780. https://doi.org/10.1038/npp.2017.159

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Gueguen, M. C., Schweitzer, E. M., & Konova, A. B. (2021). Computational theory-driven studies of reinforcement learning and decision-making in addiction: What have we learned? *Current Opinion in Behavioral Sciences*, *38*, 40–48. https://doi.org/10.1016/j.cobeha.2020.08.007

Haas, A. L., & Peters, R. H. (2000). Development of substance abuse problems among drug-involved offenders: Evidence for the telescoping effect. *Journal of Substance Abuse*, *12*(3), 241–253. https://doi.org/10.1016/S0899-3289(00)00053-5

Haber, S. N., & Behrens, T. E. J. (2014). The Neural Network Underlying Incentive-Based Learning: Implications for Interpreting Circuit Disruptions in Psychiatric Disorders. *Neuron*, *83*(5), 1019–1039. https://doi.org/10.1016/j.neuron.2014.08.031

Hammond, L. J. (1980). The Effect of Contingency Upon the Appetitive Conditioning of Free-Operant Behavior. *Journal of the Experimental Analysis of Behavior*, *34*(3), 297–304. https://doi.org/10.1901/jeab.1980.34-297

Harlé, K. M., Zhang, S., Schiff, M., Mackey, S., Paulus, M. P., & Yu, A. J. (2015). Altered Statistical Learning and Decision-Making in Methamphetamine Dependence: Evidence from a Two-Armed Bandit Task. *Frontiers in Psychology*, *6*, 1910. https://doi.org/10.3389/fpsyg.2015.01910

Heinz, A. (2002). Dopaminergic dysfunction in alcoholism and schizophrenia – psychopathological and behavioral correlates. *European Psychiatry*, *17*(1), 9–16. https://doi.org/10.1016/S0924-9338(02)00628-4

Heinz, A., Schlagenhauf, F., Beck, A., & Wackerhagen, C. (2016). Dimensional psychiatry: Mental disorders as dysfunctions of basic learning mechanisms. *Journal of Neural Transmission*, *123*(8), 809–821. https://doi.org/10.1007/s00702-016-1561-2

Heinz, A., Siessmeier, T., Wrase, J., Hermann, D., Klein, S., Grüsser-Sinopoli, S. M., Flor, H., Braus, D. F., Buchholz, H. G., Gründer, G., Schreckenberger, M., Smolka, M. N., Rösch, F., Mann, K., & Bartenstein, P. (2004). Correlation Between Dopamine D2 Receptors in the Ventral

Striatum and Central Processing of Alcohol Cues and Craving. *American Journal of Psychiatry*, *161*(10), 1783–1789. https://doi.org/10.1176/ajp.161.10.1783

Hennigan, K., D'Ardenne, K., & McClure, S. M. (2015). Distinct Midbrain and Habenula Pathways Are Involved in Processing Aversive Events in Humans. *Journal of Neuroscience*, *35*(1), 198–208. https://doi.org/10.1523/JNEUROSCI.0927-14.2015

Hester, R., Bell, R. P., Foxe, J. J., & Garavan, H. (2013). The influence of monetary punishment on cognitive control in abstinent cocaine-users. *Drug and Alcohol Dependence*, *133*(1), 86–93. https://doi.org/10.1016/j.drugalcdep.2013.05.027

Hikida, T., Kimura, K., Wada, N., Funabiki, K., & Nakanishi, S. (2010). Distinct Roles of Synaptic Transmission in Direct and Indirect Striatal Pathways to Reward and Aversive Behavior. *Neuron*, *66*(6), 896–907. https://doi.org/10.1016/j.neuron.2010.05.011

Hogarth, L. (2020). Addiction is driven by excessive goal-directed drug choice under negative affect: Translational critique of habit and compulsion theory. *Neuropsychopharmacology*, *45*(5), 720–735. https://doi.org/10.1038/s41386-020-0600-8

Hogarth, L., Balleine, B. W., Corbit, L. H., & Killcross, S. (2013). Associative learning mechanisms underpinning the transition from recreational drug use to addiction. *Annals of the New York Academy of Sciences*, *1282*(1), 12–24. https://doi.org/10.1111/j.1749-6632.2012.06768.x

Hogarth, L., & Field, M. (2020). Relative expected value of drugs versus competing rewards underpins vulnerability to and recovery from addiction. *Behavioural Brain Research*, *394*, 112815. https://doi.org/10.1016/j.bbr.2020.112815

Hogarth, L., Lam-Cassettari, C., Pacitti, H., Currah, T., Mahlberg, J., Hartley, L., & Moustafa, A. (2019). Intact goal-directed control in treatment-seeking drug users indexed by outcome-devaluation and Pavlovian to instrumental transfer: Critique of habit theory. *European Journal of Neuroscience*, *50*(3), 2513–2525. https://doi.org/10.1111/ejn.13961

Holl, A. K., Wilkinson, L., Tabrizi, S. J., Painold, A., & Jahanshahi, M. (2012). Probabilistic classification learning with corrective feedback is selectively impaired in early Huntington's disease—Evidence for the role of the striatum in learning with feedback. *Neuropsychologia*, *50*(9), 2176–2186. https://doi.org/10.1016/j.neuropsychologia.2012.05.021

Hollenbach, F. M., & Montgomery, J. M. (2020). Bayesian Model Selection, Model Comparison, and Model Averaging*. In *The SAGE Handbook of Research Methods in Political Science and International Relations* (Vol. 1–2, pp. 937–960). SAGE Publications Ltd. https://doi.org/10.4135/9781526486387

Hopf, F. W., Chang, S.-J., Sparta, D. R., Bowers, M. S., & Bonci, A. (2010). Motivation for Alcohol Becomes Resistant to Quinine Adulteration After 3 to 4 Months of Intermittent Alcohol Self-Administration. *Alcoholism: Clinical and Experimental Research*, *34*(9), 1565–1573. https://doi.org/10.1111/j.1530-0277.2010.01241.x

Hu, M., & Becker, J. B. (2003). Effects of Sex and Estrogen on Behavioral Sensitization to Cocaine in Rats. *The Journal of Neuroscience*, *23*(2), 693–699. https://doi.org/10.1523/JNEUROSCI.23-02-00693.2003

Hu, M., & Becker, J. B. (2008). Acquisition of cocaine self-administration in ovariectomized female rats: Effect of estradiol dose or chronic estradiol administration. *Drug and Alcohol Dependence*, *94*(1), 56–62. https://doi.org/10.1016/j.drugalcdep.2007.10.005

Hu, M., Crombag, H. S., Robinson, T. E., & Becker, J. B. (2004). Biological Basis of Sex Differences in the Propensity to Self-administer Cocaine. *Neuropsychopharmacology*, *29*(1), 81–85. https://doi.org/10.1038/sj.npp.1300301

Huys, Q. J. M., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, *46*(1), 3–19. https://doi.org/10.1038/s41386-020-0746-4

Huys, Q. J. M., Deserno, L., Obermayer, K., Schlagenhauf, F., & Heinz, A. (2016). Model-Free Temporal-Difference Learning and Dopamine in Alcohol Dependence: Examining Concepts From Theory and Animals in Human Imaging. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 401–410. https://doi.org/10.1016/j.bpsc.2016.06.005

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

Hyman, S. E. (2005). Addiction: A Disease of Learning and Memory. *Am J Psychiatry*, *162*, 1414–1422.

Hyman, S. E., Malenka, R. C., & Nestler, E. J. (2006). Neural Mechanisms of Addiction: The Role of Reward-Related Learning and Memory. *Annual Review of Neuroscience*, *29*(1), 565–598. https://doi.org/10.1146/annurev.neuro.29.051605.113009

Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *American Journal of Psychiatry*, *171*(4), 395–397. https://doi.org/10.1176/appi.ajp.2014.14020138

Insel, T. R., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Izquierdo, A., Suda, R. K., & Murray, E. A. (2004). Bilateral Orbital Prefrontal Cortex Lesions in Rhesus Monkeys Disrupt Choices Guided by Both Reward Value and Reward Contingency. *Journal of Neuroscience*, *24*(34), 7540–7548. https://doi.org/10.1523/JNEUROSCI.1921-04.2004

Jackson, L. R., Robinson, T. E., & Becker, J. B. (2006). Sex Differences and Hormonal Influences on Acquisition of Cocaine Self-Administration in Rats. *Neuropsychopharmacology*, *31*(1), 129–138. https://doi.org/10.1038/sj.npp.1300778

Jean-Richard-Dit-Bressel, P., Killcross, S., & McNally, G. P. (2018). Behavioral and neurobiological mechanisms of punishment: Implications for psychiatric disorders. *Neuropsychopharmacology*, *43*(8), 1639–1650. https://doi.org/10.1038/s41386-018-0047-3

Jensen, J., Smith, A. J., Willeit, M., Crawley, A. P., Mikulis, D. J., Vitcu, I., & Kapur, S. (2007). Separate brain regions code for salience vs. Valence during reward prediction in humans. *Human Brain Mapping*, *28*(4), 294–302. https://doi.org/10.1002/hbm.20274

Jentsch, J. D., Olausson, P., De La Garza, R., & Taylor, J. R. (2002). Impairments of Reversal Learning and Response Perseveration after Repeated, Intermittent Cocaine Administrations to Monkeys. *Neuropsychopharmacology*, *26*(2), 183–190. https://doi.org/10.1016/S0893-133X(01)00355-4

Jocham, G., Klein, T. A., Neumann, J., von Cramon, D. Y., Reuter, M., & Ullsperger, M. (2009). Dopamine DRD2 Polymorphism Alters Reversal Learning and Associated Neural Activity. *Journal of Neuroscience*, *29*(12), 3695–3704. https://doi.org/10.1523/JNEUROSCI.5195-08.2009

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-Mediated Reinforcement Learning Signals in the Striatum and Ventromedial Prefrontal Cortex Underlie Value-Based Choices. *Journal of Neuroscience*, *31*(5), 1606–1613. https://doi.org/10.1523/JNEUROSCI.3904-10.2011

Jocham, G., Klein, T. A., & Ullsperger, M. (2014). Differential Modulation of Reinforcement Learning by D2 Dopamine and NMDA Glutamate Receptor Antagonism. *Journal of Neuroscience*, *34*(39), 13151–13162. https://doi.org/10.1523/JNEUROSCI.0757-14.2014

Johnson, P. B., Richter, L., Kleber, H. D., McLellan, A. T., & Carise, D. (2005). Telescoping of Drinking-Related Behaviors: Gender, Racial/Ethnic, and Age Comparisons. *Substance Use & Misuse*, *40*(8), 1139–1151. https://doi.org/10.1081/JA-200042281

Jokisch, D., Roser, P., Juckel, G., Daum, I., & Bellebaum, C. (2014). Impairments in Learning by Monetary Rewards and Alcohol-Associated Rewards in Detoxified Alcoholic Patients. *Alcoholism: Clinical and Experimental Research*, *38*(7), 1947–1954. https://doi.org/10.1111/acer.12460

Jones, B., & Mishkin, M. (1972). Limbic lesions and the problem of stimulus—Reinforcement associations. *Experimental Neurology*, *36*(2), 362–377. https://doi.org/10.1016/0014-4886(72)90030-1

Joue, G., Chakroun, K., Bayer, J., Gläscher, J., Zhang, L., Fuss, J., Hennies, N., & Sommer, T. (2021). Sex Differences and Exogenous Estrogen Influence Learning and Brain Responses to Prediction Errors. *Cerebral Cortex*, bhab334. https://doi.org/10.1093/cercor/bhab334

Kahnt, T., Weber, S. C., Haker, H., Robbins, T. W., & Tobler, P. N. (2015). Dopamine D2-Receptor Blockade Enhances Decoding of Prefrontal Signals in Humans. *Journal of Neuroscience*, *35*(9), 4104–4111. https://doi.org/10.1523/JNEUROSCI.4182-14.2015

Kalivas, P. W., & O'Brien, C. (2008). Drug Addiction as a Pathology of Staged Neuroplasticity. *Neuropsychopharmacology*, *33*(1), 166–180. https://doi.org/10.1038/sj.npp.1301564

Kanen, J. W., Ersche, K. D., Fineberg, N. A., Robbins, T. W., & Cardinal, R. N. (2019). Computational modelling reveals contrasting effects on reinforcement learning and cognitive flexibility in stimulant use disorder and obsessive-compulsive disorder: Remediating effects of dopaminergic D2/3 receptor agents. *Psychopharmacology*, *236*(8), 2337–2358. https://doi.org/10.1007/s00213-019-05325-w

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

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

Kemény, F., & Lukács, Á. (2010). Impaired procedural learning in language impairment: Results from probabilistic categorization. *Journal of Clinical and Experimental Neuropsychology*, *32*(3), 249–258. https://doi.org/10.1080/13803390902971131

Kemény, F., & Lukács, Á. (2013). Self-Insight in Probabilistic Category Learning. *The Journal of General Psychology*, *140*(1), 57–81. https://doi.org/10.1080/00221309.2012.735284

Kennedy, A. P., Epstein, D. H., Phillips, K. A., & Preston, K. L. (2013). Sex differences in cocaine/heroin users: Drug-use triggers and craving in daily life. *Drug and Alcohol Dependence*, *132*(1), 29–37. https://doi.org/10.1016/j.drugalcdep.2012.12.025

Keramati, M., Durand, A., Girardeau, P., Gutkin, B., & Ahmed, S. H. (2017). Cocaine addiction as a homeostatic reinforcement learning disorder. *Psychological Review*, *124*(2), 130–153. https://doi.org/10.1037/rev0000046

Kim, H., Shimojo, S., & O'Doherty, J. P. (2006). Is Avoiding an Aversive Outcome Rewarding? Neural Substrates of Avoidance Learning in the Human Brain. *PLOS Biology*, *4*(8), e233. https://doi.org/10.1371/journal.pbio.0040233

Klein, T. A., Neumann, J., Reuter, M., Hennig, J., von Cramon, D. Y., & Ullsperger, M. (2007). Genetically Determined Differences in Learning from Errors. *Science*, *318*(5856), 1642–1645. https://doi.org/10.1126/science.1145044

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A Neostriatal Habit Learning System in Humans. *Science*, *273*(5280), 1399–1402. https://doi.org/10.1126/science.273.5280.1399

Knowlton, B. J., & Patterson, T. K. (2016). Habit Formation and the Striatum. In R. E. Clark & S. J. Martin (Eds.), *Behavioral Neuroscience of Learning and Memory* (Vol. 37, pp. 275–295). Springer International Publishing. https://doi.org/10.1007/7854_2016_451

Knowlton, B. J., Siegel, A. L. M., & Moody, T. D. (2017). Procedural Learning in Humans☆. In J. H. Byrne (Ed.), *Learning and Memory: A Comprehensive Reference (Second Edition)* (pp. 295–312). Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21085-7

Koob, G. F., & Le Moal, M. (2008). Neurobiological mechanisms for opponent motivational processes in addiction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1507), 3113–3123. https://doi.org/10.1098/rstb.2008.0094

Koob, G. F., Stinus, L., Moal, M. L., & Bloom, F. E. (1989). Opponent process theory of motivation: Neurobiological evidence from studies of opiate dependence. *Neuroscience & Biobehavioral Reviews*, *13*(2), 135–140. https://doi.org/10.1016/S0149-7634(89)80022-3

Koob, G. F., & Volkow, N. D. (2010). Neurocircuitry of Addiction. *Neuropsychopharmacology*, *35*(1), 217–238. https://doi.org/10.1038/npp.2009.110

Kravitz, A. V., Tye, L. D., & Kreitzer, A. C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neuroscience*, *15*(6), 816–818. https://doi.org/10.1038/nn.3100

Kringelbach, M. L. (2005). The human orbitofrontal cortex: Linking reward to hedonic experience. *Nature Reviews Neuroscience*, *6*(9), 691–702. https://doi.org/10.1038/nrn1747

Kumar, D. S., Benedict, E., Wu, O., Rubin, E., Gluck, M. A., Foltin, R. W., Myers, C. E., & Vadhan, N. P. (2019). Learning functions in short-term cocaine users. *Addictive Behaviors Reports*, *9*, 100169. https://doi.org/10.1016/j.abrep.2019.100169

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*(2), 162–183. https://doi.org/10.1037/0096-3445.135.2.162

Lane, S. D., Cherek, D. R., Dougherty, D. M., & Moeller, F. G. (1998). Laboratory measurement of adaptive behavior change in humans with a history of substance dependence. *Drug and Alcohol Dependence*, *51*(3), 239–252. https://doi.org/10.1016/S0376-8716(98)00045-3

Lawn, W., Freeman, T. P., Pope, R. A., Joye, A., Harvey, L., Hindocha, C., Mokrysz, C., Moss, A., Wall, M. B., Bloomfield, M. A., Das, R. K., Morgan, C. J., Nutt, D. J., & Curran, H. V. (2016). Acute and chronic effects of cannabinoids on effort-related decision-making and reward learning: An evaluation of the cannabis 'amotivational' hypotheses. *Psychopharmacology*, *233*(19), 3537–3552. https://doi.org/10.1007/s00213-016-4383-x

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. *Proceedings of the National Academy of Sciences*, *111*(32), 11858–11863. https://doi.org/10.1073/pnas.1323586111

LeBlanc, K. H., Maidment, N. T., & Ostlund, S. B. (2013). Repeated Cocaine Exposure Facilitates the Expression of Incentive Motivation and Induces Habitual Control in Rats. *PLOS ONE*, *8*(4), e61355. https://doi.org/10.1371/journal.pone.0061355

LeBlanc, K. H., Ostlund, S. B., & Maidment, N. T. (2012). Pavlovian-to-instrumental transfer in cocaine seeking rats. *Behavioral Neuroscience*, *126*(5), 681–689. https://doi.org/10.1037/a0029534

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron*, *81*(3), 687–699. https://doi.org/10.1016/j.neuron.2013.11.028

Lesscher, H. M. B., Kerkhof, L. W. M. V., & Vanderschuren, L. J. M. J. (2010). Inflexible and Indifferent Alcohol Drinking in Male Mice. *Alcoholism: Clinical and Experimental Research*, *34*(7), 1219–1225. https://doi.org/10.1111/j.1530-0277.2010.01199.x

Levesque, M., Bedard, M. A., Courtemanche, R., Tremblay, P. L., Scherzer, P., & Blanchet, P. J. (2007). Raclopride-induced motor consolidation impairment in primates: Role of the dopamine type-2 receptor in movement chunking into integrated sequences. *Experimental Brain Research*, *182*(4), 499–508. https://doi.org/10.1007/s00221-007-1010-4

Lim, T. V., Cardinal, R. N., Savulich, G., Jones, P. S., Moustafa, A. A., Robbins, T. W., & Ersche, K. D. (2019). Impairments in reinforcement learning do not explain enhanced habit formation in cocaine use disorder. *Psychopharmacology*, *236*(8), 2359–2371. https://doi.org/10.1007/s00213-019-05330-z

Loeber, S., Duka, T., Welzel, H., Nakovics, H., Heinz, A., Flor, H., & Mann, K. (2009). Impairment of Cognitive Abilities and Decision Making after Chronic Use of Alcohol: The Impact of Multiple Detoxifications. *Alcohol and Alcoholism*, *44*(4), 372–381. https://doi.org/10.1093/alcalc/agp030

Lopez, M. F., Becker, H. C., & Chandler, L. J. (2014). Repeated episodes of chronic intermittent ethanol promote insensitivity to devaluation of the reinforcing effect of ethanol. *Alcohol*, *48*(7), 639–645. https://doi.org/10.1016/j.alcohol.2014.09.002

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-U

Lu, Z.-H., Chow, S.-M., & Loken, E. (2017). A Comparison of Bayesian and Frequentist Model Selection Methods for Factor Analysis Models. *Psychological Methods*, *22*(2), 361–381. https://doi.org/10.1037/met0000145

Lubman, D. I., Yücel, M., Kettle, J. W. L., Scaffidi, A., MacKenzie, T., Simmons, J. G., & Allen, N. B. (2009). Responsiveness to Drug Cues and Natural Rewards in Opiate Addiction: Associations With Later Heroin Use. *Archives of General Psychiatry*, *66*(2), 205–212. https://doi.org/10.1001/archgenpsychiatry.2008.522

Lucantonio, F., Kambhampati, S., Haney, R. Z., Atalayer, D., Rowland, N. E., Shaham, Y., & Schoenbaum, G. (2015). Effects of Prior Cocaine Versus Morphine or Heroin Self-Administration on Extinction Learning Driven by Overexpectation Versus Omission of Reward. *Biological Psychiatry*, *77*(10), 912–920. https://doi.org/10.1016/j.biopsych.2014.11.017

Lucantonio, F., Stalnaker, T. A., Shaham, Y., Niv, Y., & Schoenbaum, G. (2012). The impact of orbitofrontal dysfunction on cocaine addiction. *Nature Neuroscience*, *15*(3), 358–366. https://doi.org/10.1038/nn.3014

Luijten, M., Gillan, C. M., De Wit, S., Franken, I. H. A., Robbins, T. W., & Ersche, K. D. (2019). Goal-Directed and Habitual Control in Smokers. *Nicotine & Tobacco Research*. https://doi.org/10.1093/ntr/ntz001

Luijten, M., Schellekens, A. F., Kuehn, S., Machielse, M. W. J., & Sescousse, G. (2017). Disruption of Reward Processing in Addiction An Image-Based Meta-analysis of Functional Magnetic Resonance Imaging Studies. *Jama Psychiatry*, *74*(4), 387–398. https://doi.org/10.1001/jamapsychiatry.2016.3084

Lüscher, C., & Ungless, M. A. (2006). The Mechanistic Classification of Addictive Drugs. *PLOS Medicine*, *3*(11), e437. https://doi.org/10.1371/journal.pmed.0030437

Lynch, W. J., & Carroll, M. E. (1999). Sex differences in the acquisition of intravenously self-administered cocaine and heroin in rats. *Psychopharmacology*, *144*(1), 77–82. https://doi.org/10.1007/s002130050979

Lynch, W. J., Roth, M., & Carroll, M. (2002). Biological basis of sex differences in drug abuse: Preclinical and clinical studies. *Psychopharmacology*, *164*(2), 121–137. https://doi.org/10.1007/s00213-002-1183-2

Lynch, W. J., Roth, M. E., Mickelberg, J. L., & Carroll, M. E. (2001). Role of estrogen in the acquisition of intravenously self-administered cocaine in female rats. *Pharmacology Biochemistry and Behavior*, *68*(4), 641–646. https://doi.org/10.1016/S0091-3057(01)00455-5

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154–162. https://doi.org/10.1038/nn.2723

Mangieri, R. A., Cofresí, R. U., & Gonzales, R. A. (2012). Ethanol Seeking by Long Evans Rats Is Not Always a Goal-Directed Behavior. *PLOS ONE*, *7*(8), e42886. https://doi.org/10.1371/journal.pone.0042886

Martinez, D., Broft, A., Foltin, R. W., Slifstein, M., Hwang, D.-R., Huang, Y., Perez, A., Frankel, W. G., Cooper, T., Kleber, H. D., Fischman, M. W., & Laruelle, M. (2004). Cocaine Dependence and D 2 Receptor Availability in the Functional Subdivisions of the Striatum: Relationship with Cocaine-Seeking Behavior. *Neuropsychopharmacology*, *29*(6), 1190–1202. https://doi.org/10.1038/sj.npp.1300420

Martinez, D., Narendran, R., Foltin, R. W., Slifstein, M., Hwang, D.-R., Broft, A., Huang, Y., Cooper, T. B., Fischman, M. W., Kleber, H. D., & Laruelle, M. (2007). Amphetamine-Induced Dopamine Release: Markedly Blunted in Cocaine Dependence and Predictive of the Choice to Self-Administer Cocaine. *American Journal of Psychiatry*, *164*(4), 622–629. https://doi.org/10.1176/ajp.2007.164.4.622

McCabe, C., Huber, A., Harmer, C. J., & Cowen, P. J. (2011). The D2 antagonist sulpiride modulates the neural processing of both rewarding and aversive stimuli in healthy volunteers. *Psychopharmacology*, *217*(2), 271–278. https://doi.org/10.1007/s00213-011-2278-4

McClure, S. M., Daw, N. D., & Read Montague, P. (2003). A computational substrate for incentive salience. *Trends in Neurosciences*, *26*(8), 423–428. https://doi.org/10.1016/S0166-2236(03)00177-2

McKim, T. H., Bauer, D. J., & Boettiger, C. A. (2016). Addiction History Associates with the Propensity to Form Habits. *Journal of Cognitive Neuroscience*, *28*(7), 1024–1038. https://doi.org/10.1162/jocn_a_00953

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Menon, M., Jensen, J., Vitcu, I., Graff-Guerrero, A., Crawley, A., Smith, M. A., & Kapur, S. (2007). Temporal Difference Modeling of the Blood-Oxygen Level Dependent Response During Aversive Conditioning in Humans: Effects of Dopaminergic Modulation. *Biological Psychiatry*, *62*(7), 765–772. https://doi.org/10.1016/j.biopsych.2006.10.020

Meunier, D., Ersche, K. D., Craig, K. J., Fornito, A., Merlo-Pich, E., Fineberg, N. A., Shabbir, S. S., Robbins, T. W., & Bullmore, E. T. (2012). Brain functional connectivity in stimulant drug dependence and obsessive–compulsive disorder. *NeuroImage*, *59*(2), 1461–1468. https://doi.org/10.1016/j.neuroimage.2011.08.003

Miles, F. J., Everitt, B. J., & Dickinson, A. (2003). Oral cocaine seeking by rats: Action or habit? *Behavioral Neuroscience*, *117*(5), 927–938. https://doi.org/10.1037/0735-7044.117.5.927

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, *126*(2), 292–311. https://doi.org/10.1037/rev0000120

Moeller, S. J., Maloney, T., Parvaz, M. A., Dunning, J. P., Alia-Klein, N., Woicik, P. A., Hajcak, G., Telang, F., Wang, G.-J., Volkow, N. D., & Goldstein, R. Z. (2009). Enhanced Choice for Viewing Cocaine Pictures in Cocaine Addiction. *Biological Psychiatry*, *66*(2), 169–176. https://doi.org/10.1016/j.biopsych.2009.02.015

Mollick, J. A., & Kober, H. (2020). Computational models of drug use and addiction: A review. *Journal of Abnormal Psychology*, *129*(6), 544–555. https://doi.org/10.1037/abn0000503

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760–767. https://doi.org/10.1038/nature03015

Moore, R. J., Vinsant, S. L., Nader, M. A., Porrino, L. J., & Friedman, D. P. (1998). Effect of cocaine self-administration on dopamine D2 receptors in rhesus monkeys. *Synapse*, *30*(1), 88–96. https://doi.org/10.1002/(SICI)1098-2396(199809)30:1<88::AID-SYN11>3.0.CO;2-L

Moreno-López, L., Perales, J. C., Son, D. van, Albein-Urios, N., Soriano-Mas, C., Martinez-Gonzalez, J. M., Wiers, R. W., & Verdejo-García, A. (2015). Cocaine use severity and cerebellar gray matter are associated with reversal learning deficits in cocaine-dependent individuals. *Addiction Biology*, *20*(3), 546–556. https://doi.org/10.1111/adb.12143

Morie, K. P., De Sanctis, P., Garavan, H., & Foxe, J. J. (2016). Regulating task-monitoring systems in response to variable reward contingencies and outcomes in cocaine addicts. *Psychopharmacology*, *233*(6), 1105–1118. https://doi.org/10.1007/s00213-015-4191-8

Munro, V. (2019, April 22). *Buying cannabis and cocaine in London has become as "easy as ordering a pizza."* MyLondon. https://www.mylondon.news/news/health/how-buying-cannabis-cocaine-london-16143070

Murphy, F. C., Michael, A., Robbins, T. W., & Sahakian, B. J. (2003). Neuropsychological impairment in patients with major depressive disorder: The effects of feedback on task performance. *Psychological Medicine*, *33*(3), 455–467. https://doi.org/10.1017/S0033291702007018

Murray, G. K., Knolle, F., Ersche, K. D., Craig, K. J., Abbott, S., Shabbir, S. S., Fineberg, N. A., Suckling, J., Sahakian, B. J., Bullmore, E. T., & Robbins, T. W. (2019). Dopaminergic drug treatment remediates exaggerated cingulate prediction error responses in obsessive-compulsive disorder. *Psychopharmacology*, *236*(8), 2325–2336. https://doi.org/10.1007/s00213-019-05292-2

Myers, C. E., Rego, J., Haber, P., Morley, K., Beck, K. D., Hogarth, L., & Moustafa, A. A. (2017). Learning and generalization from reward and punishment in opioid addiction. *Behavioural Brain Research*, *317*, 122–131. https://doi.org/10.1016/j.bbr.2016.09.033

Myers, C. E., Sheynin, J., Balsdon, T., Luzardo, A., Beck, K. D., Hogarth, L., Haber, P., & Moustafa, A. A. (2016). Probabilistic reward- and punishment-based learning in opioid addiction: Experimental and computational data. *Behavioural Brain Research*, *296*, 240–248. https://doi.org/10.1016/j.bbr.2015.09.018

Nader, M. A., & Czoty, P. W. (2005). PET Imaging of Dopamine D2 Receptors in Monkey Models of Cocaine Abuse: Genetic Predisposition Versus Environmental Modulation. *American Journal of Psychiatry*, *162*(8), 1473–1482. https://doi.org/10.1176/appi.ajp.162.8.1473

Nader, M. A., Morgan, D., Gage, H. D., Nader, S. H., Calhoun, T. L., Buchheimer, N., Ehrenkaufer, R., & Mach, R. H. (2006). PET imaging of dopamine D2 receptors during chronic cocaine self-administration in monkeys. *Nature Neuroscience*, *9*(8), 1050–1056. https://doi.org/10.1038/nn1737

Nakanishi, S., Hikida, T., & Yawata, S. (2014). Distinct dopaminergic control of the direct and indirect pathways in reward-based and avoidance learning behaviors. *Neuroscience*, *282*, 49–59. https://doi.org/10.1016/j.neuroscience.2014.04.026

Nebe, S., Kroemer, N. B., Schad, D. J., Bernhardt, N., Sebold, M., Müller, D. K., Scholl, L., Kuitunen-Paul, S., Heinz, A., Rapp, M. A., Huys, Q. J. M., & Smolka, M. N. (2018). No association of goal-directed and habitual control with alcohol consumption in young adults. *Addiction Biology*, *23*(1), 379–393. https://doi.org/10.1111/adb.12490

Nelson, A., & Killcross, S. (2006). Amphetamine Exposure Enhances Habit Formation. *Journal of Neuroscience*, *26*(14), 3805–3812. https://doi.org/10.1523/JNEUROSCI.4305-05.2006

Nelson, H. E. (1982). *National adult reading test (NART).* Nfer-Nelson Windsor.

Nelson, L. D., Patrick, C. J., Collins, P., Lang, A. R., & Bernat, E. M. (2011). Alcohol impairs brain reactivity to explicit loss feedback. *Psychopharmacology*, *218*(2), 419. https://doi.org/10.1007/s00213-011-2323-3

Nestler, E. J. (2005). Is there a common molecular pathway for addiction? *Nature Neuroscience*, *8*(11), 1445–1449. https://doi.org/10.1038/nn1578

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. https://doi.org/10.1016/j.jmp.2008.12.005

Noble, E. P., Blum, K., Khalsa, M. E., Ritchie, T., Montgomery, A., Wood, R. C., Fitch, R. J., Ozkaragoz, T., Sheridan, P. J., Anglin, M. D., Paredes, A., Treiman, L. J., & Sparkes, R. S. (1993). Allelic association of the D2 dopamine receptor gene with cocaine dependence. *Drug and Alcohol Dependence*, *33*(3), 271–285. https://doi.org/10.1016/0376-8716(93)90113-5

Nordquist, R. E., Voorn, P., de Mooij-van Malsen, J. G., Joosten, R. N. J. M. A., Pennartz, C. M. A., & Vanderschuren, L. J. M. J. (2007). Augmented reinforcer value and accelerated habit formation after repeated amphetamine treatment. *European Neuropsychopharmacology*, *17*(8), 532–540. https://doi.org/10.1016/j.euroneuro.2006.12.005

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., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annual Review of Psychology*, *68*(1), 73–100. https://doi.org/10.1146/annurev-psych-010416-044216

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

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, *110*(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Packard, M. G., & Knowlton, B. J. (2002). Learning and Memory Functions of the Basal Ganglia. *Annual Review of Neuroscience*, *25*(1), 563–593. https://doi.org/10.1146/annurev.neuro.25.112701.142937

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

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

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., & Pessiglione, M. (2017). Opponent Brain Systems for Reward and Punishment Learning: Causal Evidence From Drug and Lesion Studies in Humans. In J. C. Dreher & L. Tremblay (Eds.), *Decision Neuroscience* (pp. 291–303). Academic Press. https://doi.org/10.1016/B978-0-12-805308-9.00023-3

Palombo, D. J., Hayes, S. M., Reid, A. G., & Verfaellie, M. (2019). Hippocampal contributions to value-based learning: Converging evidence from fMRI and amnesia. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(3), 523–536. https://doi.org/10.3758/s13415-018-00687-8

Park, S. Q., Kahnt, T., Beck, A., Cohen, M. X., Dolan, R. J., Wrase, J., & Heinz, A. (2010). Prefrontal Cortex Fails to Learn from Reward Prediction Errors in Alcohol Dependence. *Journal of Neuroscience*, *30*(22), 7749–7753. https://doi.org/10.1523/JNEUROSCI.5587-09.2010

Parvaz, M. A., Konova, A. B., Proudfit, G. H., Dunning, J. P., Malaker, P., Moeller, S. J., Maloney, T., Alia-Klein, N., & Goldstein, R. Z. (2015). Impaired Neural Response to Negative Prediction Errors in Cocaine Addiction. *Journal of Neuroscience*, *35*(5), 1872–1879. https://doi.org/10.1523/JNEUROSCI.2777-14.2015

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774. https://doi.org/10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019). Incentives Boost Model-Based Control Across a Range of Severity on Several Psychiatric Constructs. *Biological Psychiatry*, *85*(5), 425–433. https://doi.org/10.1016/j.biopsych.2018.06.018

Paulus, M. P., Tapert, S. F., & Schulteis, G. (2009). The role of interoception and alliesthesia in addiction. *Pharmacology Biochemistry and Behavior*, *94*(1), 1–7. https://doi.org/10.1016/j.pbb.2009.08.005

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Perandrés-Gómez, A., Navas, J. F., van Timmeren, T., & Perales, J. C. (2021). Decision-making (in)flexibility in gambling disorder. *Addictive Behaviors*, *112*, 106534. https://doi.org/10.1016/j.addbeh.2020.106534

Pessiglione, M., & Delgado, M. R. (2015). The good, the bad and the brain: Neural correlates of appetitive and aversive values underlying decision making. *Current Opinion in Behavioral Sciences*, *5*, 78–84. https://doi.org/10.1016/j.cobeha.2015.08.006

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Petry, N. M. (2000). A comprehensive guide to the application of contingency management procedures in clinical settings. *Drug and Alcohol Dependence*, *58*(1), 9–25. https://doi.org/10.1016/S0376-8716(99)00071-X

Petry, N. M., Alessi, S. M., Olmstead, T. A., Rash, C. J., & Zajac, K. (2017). Contingency management treatment for substance use disorders: How far has it come, and where does it need to go? *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *31*(8), 897–906. https://doi.org/10.1037/adb0000287

Piazza, N. J., Vrbka, J. L., & Yeager, R. D. (1989). Telescoping of Alcoholism in Women Alcoholics. *International Journal of the Addictions*, *24*(1), 19–28. https://doi.org/10.3109/10826088909047272

Pierce, R. C., & Kumaresan, V. (2006). The mesolimbic dopamine system: The final common pathway for the reinforcing effect of drugs of abuse? *Neuroscience & Biobehavioral Reviews*, *30*(2), 215–238. https://doi.org/10.1016/j.neubiorev.2005.04.016

Pizzagalli, D. A., Evins, A. E., Schetter, E. C., Frank, M. J., Pajtas, P. E., Santesso, D. L., & Culhane, M. (2008). Single dose of a dopamine agonist impairs reinforcement learning in humans: Behavioral evidence from a laboratory-based measure of reward responsiveness. *Psychopharmacology*, *196*(2), 221–232. https://doi.org/10.1007/s00213-007-0957-y

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546–550. https://doi.org/10.1038/35107080

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience & Biobehavioral Reviews*, *32*(2), 197–205. https://doi.org/10.1016/j.neubiorev.2007.07.007

Poldrack, R. A., Prabhakaran, V., Seger, C. A., & Gabrieli, J. D. E. (1999). Striatal activation during acquisition of a cognitive skill. *Neuropsychology*, *13*(4), 564–574. https://doi.org/10.1037/0894-4105.13.4.564

Price, A. L. (2005). Cortico-striatal contributions to category learning: Dissociating the verbal and implicit systems. *Behavioral Neuroscience*, *119*(6), 1438–1447. https://doi.org/10.1037/0735-7044.119.6.1438

Public Health England. (2017). *The Public Health Burden of Alcohol and the Effectiveness and Cost-Effectiveness of Alcohol Control Policies: An evidence review*. Public Health England Publications. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/733108/alcohol_public_health_burden_evidence_review_update_2018.pdf

Ramakrishnan, S., Robbins, T. W., & Zmigrod, L. (2021). The Habitual Tendencies Questionnaire: A tool for psychometric individual differences research. *Personality and Mental Health*, pmh.1524. https://doi.org/10.1002/pmh.1524

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction: Vulnerabilities in the decision process. *The Behavioral and Brain Sciences*, *31*(4), 415–487. https://doi.org/10.1017/S0140525X0800472X

Reiter, A. M. F., Deserno, L., Kallert, T., Heinze, H.-J., Heinz, A., & Schlagenhauf, F. (2016). Behavioral and Neural Signatures of Reduced Updating of Alternative Options in Alcohol-Dependent Patients during Flexible Decision-Making. *Journal of Neuroscience*, *36*(43), 10935–10948. https://doi.org/10.1523/JNEUROSCI.4322-15.2016

Renteria, R., Baltz, E. T., & Gremel, C. M. (2018). Chronic alcohol exposure disrupts top-down control over basal ganglia action selection to produce habits. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-017-02615-9

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

Robbins, T. W., & Cardinal, R. N. (2019). Computational psychopharmacology: A translational and pragmatic approach. *Psychopharmacology*, *236*(8), 2295–2305. https://doi.org/10.1007/s00213-019-05302-3

Robbins, T. W., & Costa, R. M. (2017). Habits. *Current Biology*, *27*(22), R1200–R1206. https://doi.org/10.1016/j.cub.2017.09.060

Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S., & Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. *Trends in Cognitive Sciences*, *16*(1), 81–91. https://doi.org/10.1016/j.tics.2011.11.009

Robinson, A. H., Perales, J. C., Volpe, I., Chong, T. T.-J., & Verdejo-Garcia, A. (2021). Are methamphetamine users compulsive? Faulty reinforcement learning, not inflexibility, underlies decision making in people with methamphetamine use disorder. *Addiction Biology*, *26*(4), e12999. https://doi.org/10.1111/adb.12999

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews*, *18*(3), 247–291. https://doi.org/10.1016/0165-0173(93)90013-P

Rolls, E. T. (2004). The functions of the orbitofrontal cortex. *Brain and Cognition*, *55*(1), 11–29. https://doi.org/10.1016/S0278-2626(03)00277-X

Rose, E. J., Salmeron, B. J., Ross, T. J., Waltz, J., Schweitzer, J. B., McClure, S. M., & Stein, E. A. (2014). Temporal Difference Error Prediction Signal Dysregulation in Cocaine Dependence. *Neuropsychopharmacology*, *39*(7), 1732–1742. https://doi.org/10.1038/npp.2014.21

Rosenzweig, P., Canal, M., Patat, A., Bergougnan, L., Zieleniuk, I., & Bianchetti, G. (2002). A review of the pharmacokinetics, tolerability and pharmacodynamics of amisulpride in healthy volunteers. *Human Psychopharmacology: Clinical and Experimental*, *17*(1), 1–13. https://doi.org/10.1002/hup.320

Rouhani, N., & Niv, Y. (2019). Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology*, *236*(8), 2425–2435. https://doi.org/10.1007/s00213-019-05322-z

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., & Wager, T. D. (2014). Representation of aversive prediction errors in the human periaqueductal gray. *Nature Neuroscience*, *17*(11), 1607–1612. https://doi.org/10.1038/nn.3832

Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Rustemeier, M., Römling, J., Czybulka, C., Reymann, G., Daum, I., & Bellebaum, C. (2012). Learning from Positive and Negative Monetary Feedback in Patients with Alcohol Dependence. *Alcoholism: Clinical and Experimental Research*, *36*(6), 1067–1074. https://doi.org/10.1111/j.1530-0277.2011.01696.x

Saal, D., Dong, Y., Bonci, A., & Malenka, R. C. (2003). Drugs of Abuse and Stress Trigger a Common Synaptic Adaptation in Dopamine Neurons. *Neuron*, *37*(4), 577–582. https://doi.org/10.1016/S0896-6273(03)00021-7

Saddoris, M. P., Sugam, J. A., & Carelli, R. M. (2017). Prior Cocaine Experience Impairs Normal Phasic Dopamine Signals of Reward Value in Accumbens Shell. *Neuropsychopharmacology*, *42*(3), 766–773. https://doi.org/10.1038/npp.2016.189

Saunders, B., Milyavskaya, M., Etz, A., Randles, D., & Inzlicht, M. (2018). Reported Self-control is not Meaningfully Associated with Inhibition-related Executive Function: A Bayesian Analysis. *Collabra: Psychology*, *4*(1). https://doi.org/10.1525/collabra.134

Saunders, J. B., Aasland, O. G., Babor, T. F., Fuente, J. R. D. L., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-II. *Addiction*, *88*(6), 791–804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x

Schoemaker, H., Claustre, Y., Fage, D., Rouquier, L., Chergui, K., Curet, O., Oblin, A., Gonon, F., Carter, C., Benavides, J., & Scatton, B. (1997). Neurochemical Characteristics of Amisulpride, an Atypical Dopamine D2/D3 Receptor Antagonist with Both Presynaptic and Limbic Selectivity. *Journal of Pharmacology and Experimental Therapeutics*, *280*(1), 83–97.

Schoenbaum, G., Saddoris, M. P., Ramus, S. J., Shaham, Y., & Setlow, B. (2004). Cocaine-experienced rats exhibit learning deficits in a task sensitive to orbitofrontal cortex lesions. *European Journal of Neuroscience*, *19*(7), 1997–2002. https://doi.org/10.1111/j.1460-9568.2004.03274.x

Schoenbaum, G., & Setlow, B. (2005). Cocaine Makes Actions Insensitive to Outcomes but not Extinction: Implications for Altered Orbitofrontal–Amygdalar Function. *Cerebral Cortex*, *15*(8), 1162–1169. https://doi.org/10.1093/cercor/bhh216

Schoenbaum, G., Takahashi, Y. K., Liu, T.-L., & McDannald, M. A. (2011). Does the orbitofrontal cortex signal value? *Annals of the New York Academy of Sciences*, *1239*(1), 87–99. https://doi.org/10.1111/j.1749-6632.2011.06210.x

Schrager, A. (2015). *Economic secrets of the dark web—The safe, easy way for anyone to be a little drug lord*. Quartz. https://qz.com/481037/dark-web/

Schroder, K. E. E., Ollis, C. L., & Davies, S. (2013). Habitual Self-Control: A Brief Measure of Persistent Goal Pursuit: Habitual self-control questionnaire. *European Journal of Personality*, *27*(1), 82–95. https://doi.org/10.1002/per.1891

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

Schwabe, L., & Wolf, O. T. (2009). Stress Prompts Habit Behavior in Humans. *Journal of Neuroscience*, *29*(22), 7191–7198. https://doi.org/10.1523/JNEUROSCI.0979-09.2009

Schwabe, L., & Wolf, O. T. (2012). Stress Modulates the Engagement of Multiple Memory Systems in Classification Learning. *Journal of Neuroscience*, *32*(32), 11042–11049. https://doi.org/10.1523/JNEUROSCI.1484-12.2012

Schwöbel, S., Marković, D., Smolka, M. N., & Kiebel, S. J. (2021). Balancing control: A Bayesian interpretation of habitual and goal-directed behavior. *Journal of Mathematical Psychology*, *100*, 102472. https://doi.org/10.1016/j.jmp.2020.102472

Sebold, M., Deserno, L., Nebe, S., Schad, D. J., Garbusow, M., Hägele, C., Keller, J., Jünger, E., Kathmann, N., Smolka, M., Rapp, M. A., Schlagenhauf, F., Heinz, A., & Huys, Q. J. M. (2014). Model-Based and Model-Free Decisions in Alcohol Dependence. *Neuropsychobiology*, *70*(2), 122–131. https://doi.org/10.1159/000362840

Sebold, M., Nebe, S., Garbusow, M., Guggenmos, M., Schad, D. J., Beck, A., Kuitunen-Paul, S., Sommer, C., Frank, R., Neu, P., Zimmermann, U. S., Rapp, M. A., Smolka, M. N., Huys, Q. J. M., Schlagenhauf, F., & Heinz, A. (2017). When Habits Are Dangerous: Alcohol Expectancies and Habitual Decision Making Predict Relapse in Alcohol Dependence. *Biological Psychiatry*, *82*(11), 847–856. https://doi.org/10.1016/j.biopsych.2017.04.019

Seger, C. A., & Miller, E. K. (2010). Category Learning in the Brain. *Annual Review of Neuroscience*, *33*(1), 203–219. https://doi.org/10.1146/annurev.neuro.051508.135546

Seger, C. A., & Spiering, B. J. (2011). A Critical Review of Habit Learning and the Basal Ganglia. *Frontiers in Systems Neuroscience*, *5*. https://doi.org/10.3389/fnsys.2011.00066

Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., & Dolan, R. (2012). Serotonin Selectively Modulates Reward Value in Human Decision-Making. *Journal of Neuroscience*, *32*(17), 5833–5842. https://doi.org/10.1523/JNEUROSCI.0053-12.2012

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential Encoding of Losses and Gains in the Human Striatum. *Journal of Neuroscience*, *27*(18), 4826–4831. https://doi.org/10.1523/JNEUROSCI.0400-07.2007

Shanks, D. R. (1995). *The Psychology of Associative Learning*. Cambridge University Press. https://doi.org/10.1017/CBO9780511623288

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview

(M.I.N.I): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59*(Suppl 20), 22–33.

Sheff, D. (2008). *Beautiful Boy: A Father's Journey through his son's Addiction*. Houghton Mifflin Company.

Shohamy, D., Myers, C. E., Grossman, S., Sage, J., Gluck, M. A., & Poldrack, R. A. (2004). Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology. *Brain*, *127*(4), 851–859. https://doi.org/10.1093/brain/awh100

Shohamy, D., Myers, C. E., Kalanithi, J., & Gluck, M. A. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Neuroscience & Biobehavioral Reviews*, *32*(2), 219–236. https://doi.org/10.1016/j.neubiorev.2007.07.008

Shohamy, D., Myers, C. E., Onlaor, S., & Gluck, M. A. (2004). Role of the Basal Ganglia in Category Learning: How Do Patients With Parkinson's Disease Learn? *Behavioral Neuroscience*, *118*(4), 676–686. https://doi.org/10.1037/0735-7044.118.4.676

Singer, B. F., Fadanelli, M., Kawa, A. B., & Robinson, T. E. (2018). Are Cocaine-Seeking "Habits" Necessary for the Development of Addiction-Like Behavior in Rats? *Journal of Neuroscience*, *38*(1), 60–73. https://doi.org/10.1523/JNEUROSCI.2458-17.2017

Sinha, R. (2001). How does stress increase risk of drug abuse and relapse? *Psychopharmacology*, *158*(4), 343–359. https://doi.org/10.1007/s002130100917

Sjoerds, Z., de Wit, S., van den Brink, W., Robbins, T. W., Beekman, A. T. F., Penninx, B. W. J. H., & Veltman, D. J. (2013). Behavioral and neuroimaging evidence for overreliance on habit learning in alcohol-dependent patients. *Translational Psychiatry*, *3*(12), e337. https://doi.org/10.1038/tp.2013.107

Sjoerds, Z., Dietrich, A., Deserno, L., de Wit, S., Villringer, A., Heinze, H.-J., Schlagenhauf, F., & Horstmann, A. (2016). Slips of Action and Sequential Decisions: A Cross-Validation Study of Tasks Assessing Habitual and Goal-Directed Action Control. *Frontiers in Behavioral Neuroscience*, *10*. https://doi.org/10.3389/fnbeh.2016.00234

Skinner, B. F. (1963). Operant behavior. *American Psychologist*, *18*(8), 503–515. https://doi.org/10.1037/h0045185

Skinner, H. A. (1982). The drug abuse screening test. *Addictive Behaviors*, *7*(4), 363–371. https://doi.org/10.1016/0306-4603(82)90005-3

Smith, R. J., & Laiks, L. S. (2018). Behavioral and neural mechanisms underlying habitual and compulsive drug seeking. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *87*, 11–21. https://doi.org/10.1016/j.pnpbp.2017.09.003

Smith, R., Taylor, S., & Bilek, E. (2021). Computational Mechanisms of Addiction: Recent Evidence and Its Relevance to Addiction Medicine. *Current Addiction Reports*. https://doi.org/10.1007/s40429-021-00399-z

Solomon, R. L., & Corbit, J. D. (1974). An opponent-process theory of motivation: I. Temporal dynamics of affect. *Psychological Review*, *81*(2), 119–145. https://doi.org/10.1037/h0036128

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences*, *93*(24), 13515–13522. https://doi.org/10.1073/pnas.93.24.13515

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, *253*(5026), 1380–1386. https://doi.org/10.1126/science.1896849

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, *16*(7), 966–973. https://doi.org/10.1038/nn.3413

Stewart, J. L., Connolly, C. G., May, A. C., Tapert, S. F., Wittmann, M., & Paulus, M. P. (2014a). Striatum and insula dysfunction during reinforcement learning differentiates abstinent and

relapsed methamphetamine-dependent individuals. *Addiction*, *109*(3), 460–471. https://doi.org/10.1111/add.12403

Stewart, J. L., Connolly, C. G., May, A. C., Tapert, S. F., Wittmann, M., & Paulus, M. P. (2014b). Cocaine dependent individuals with attenuated striatal activation during reinforcement learning are more susceptible to relapse. *Psychiatry Research: Neuroimaging*, *223*(2), 129–139. https://doi.org/10.1016/j.pscychresns.2014.04.014

Stitzer, M. L., Polk, T., Bowles, S., & Kosten, T. (2010). Drug users' adherence to a 6-month vaccination protocol: Effects of motivational incentives. *Drug and Alcohol Dependence*, *107*(1), 76–79. https://doi.org/10.1016/j.drugalcdep.2009.09.006

Stoops, W. W., Lile, J. A., & Rush, C. R. (2010). Monetary alternative reinforcers more effectively decrease intranasal cocaine choice than food alternative reinforcers. *Pharmacology Biochemistry and Behavior*, *95*(2), 187–191. https://doi.org/10.1016/j.pbb.2010.01.003

Stout, J. C., Busemeyer, J. R., Lin, A., Grant, S. J., & Bonson, K. R. (2004). Cognitive modeling analysis of decision-making processes in cocaine abusers. *Psychonomic Bulletin & Review*, *11*(4), 742–747. https://doi.org/10.3758/BF03196629

Strickland, J. C., Bolin, B. L., Lile, J. A., Rush, C. R., & Stoops, W. W. (2016). Differential sensitivity to learning from positive and negative outcomes in cocaine users. *Drug and Alcohol Dependence*, *166*, 61–68. https://doi.org/10.1016/j.drugalcdep.2016.06.022

Sudai, E., Croitoru, O., Shaldubina, A., Abraham, L., Gispan, I., Flaumenhaft, Y., Roth-Deri, I., Kinor, N., Aharoni, S., Ben-Tzion, M., & Yadid, G. (2011). High cocaine dosage decreases neurogenesis in the hippocampus and impairs working memory. *Addiction Biology*, *16*(2), 251–260. https://doi.org/10.1111/j.1369-1600.2010.00241.x

Surmeier, D. J., Ding, J., Day, M., Wang, Z., & Shen, W. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends in Neurosciences*, *30*(5), 228–235. https://doi.org/10.1016/j.tins.2007.03.008

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge.

Swendsen, J. D., Merikangas, K. R., Canino, G. J., Kessler, R. C., Rubio-Stipec, M., & Angst, J. (1998). The comorbidity of alcoholism with anxiety and depressive disorders in four geographic communities. *Comprehensive Psychiatry*, *39*(4), 176–184. https://doi.org/10.1016/S0010-440X(98)90058-X

Takahashi, Y. K., Langdon, A. J., Niv, Y., & Schoenbaum, G. (2016). Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum. *Neuron*, *91*(1), 182–193. https://doi.org/10.1016/j.neuron.2016.05.015

Takahashi, Y. K., Roesch, M., Stalnaker, T., & Schoenbaum, G. (2007). Cocaine exposure shifts the balance of associative encoding from ventral to dorsolateral striatum. *Frontiers in Integrative Neuroscience*, *1*, 11. https://doi.org/10.3389/neuro.07.011.2007

Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2008). Silencing the critics: Understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an Actor/Critic model. *Frontiers in Neuroscience*, *2*. https://doi.org/10.3389/neuro.01.014.2008

Takahashi, Y. K., Stalnaker, T. A., Marrero-Garcia, Y., Rada, R. M., & Schoenbaum, G. (2019). Expectancy-Related Changes in Dopaminergic Error Signals Are Impaired by Cocaine Self-Administration. *Neuron*, *101*(2), 294-306.e3. https://doi.org/10.1016/j.neuron.2018.11.025

Tanabe, J., Reynolds, J., Krmpotich, T., Claus, E., Thompson, L. L., Du, Y. P., & Banich, M. T. (2013). Reduced Neural Tracking of Prediction Error in Substance-Dependent Individuals. *American Journal of Psychiatry*, *170*(11), 1356–1363. https://doi.org/10.1176/appi.ajp.2013.12091257

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating Consequences: Brain Systems That Encode the Causal Effects of Actions. *Journal of Neuroscience*, *28*(26), 6750–6755. https://doi.org/10.1523/JNEUROSCI.1808-08.2008

The Pew Charitable Trusts. (2018). *More Imprisonment Does Not Reduce State Drug Problems*. https://www.pewtrusts.org/-/media/assets/2018/03/pspp_more_imprisonment_does_not_reduce_state_drug_problems.pdf

Thomas, L. A., & LaBar, K. S. (2008). Fear relevancy, strategy use, and probabilistic learning of cue-outcome associations. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *15*(10), 777–784. https://doi.org/10.1101/lm.1048808

Thompson, L. L., Claus, E. D., Mikulich-Gilbertson, S. K., Banich, M. T., Crowley, T., Krmpotich, T., Miller, D., & Tanabe, J. (2012). Negative reinforcement learning is affected in substance dependence. *Drug and Alcohol Dependence*, *123*(1), 84–90. https://doi.org/10.1016/j.drugalcdep.2011.10.017

Thorndike, E. L. (1911). *Animal intelligence; experimental studies* (pp. 1–328). The Macmillan Company. https://doi.org/10.5962/bhl.title.55072

Torres, O. V., & O'Dell, L. E. (2016). Stress is a principal factor that promotes tobacco use in females. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *65*, 260–268. https://doi.org/10.1016/j.pnpbp.2015.04.005

Townsend, L., Flisher, A. J., & King, G. (2007). A Systematic Review of the Relationship between High School Dropout and Substance Use. *Clinical Child and Family Psychology Review*, *10*(4), 295–317. https://doi.org/10.1007/s10567-007-0023-7

Tremblay, P.-L., Bedard, M.-A., Langlois, D., Blanchet, P. J., Lemay, M., & Parent, M. (2010). Movement chunking during sequence learning is a dopamine-dependant process: A study conducted in Parkinson's disease. *Experimental Brain Research*, *205*(3), 375–385. https://doi.org/10.1007/s00221-010-2372-6

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, *29*(11), 2225–2232. https://doi.org/10.1111/j.1460-9568.2009.06796.x

UNODC. (2021). *World Drug Report 2020 (set of 6 booklets).* United Nations.

Vadhan, N. P., Myers, C. E., Benedict, E., Rubin, E., Foltin, R. W., & Gluck, M. A. (2014). A decrement in probabilistic category learning in cocaine users after controlling for marijuana and alcohol use. *Experimental and Clinical Psychopharmacology*, *22*(1), 65–74. https://doi.org/10.1037/a0034506

Vadhan, N. P., Myers, C. E., Rubin, E., Shohamy, D., Foltin, R. W., & Gluck, M. A. (2008). Stimulus–response learning in long-term cocaine users: Acquired equivalence and probabilistic category learning. *Drug and Alcohol Dependence*, *93*(1–2), 155–162. https://doi.org/10.1016/j.drugalcdep.2007.09.013

Vaghi, M. M., Cardinal, R. N., Apergis-Schoute, A. M., Fineberg, N. A., Sule, A., & Robbins, T. W. (2019). Action-Outcome Knowledge Dissociates From Behavior in Obsessive-Compulsive Disorder Following Contingency Degradation. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*(2), 200–209. https://doi.org/10.1016/j.bpsc.2018.09.014

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the Neural Substrates of Goal-Directed Learning in the Human Brain. *Journal of Neuroscience*, *27*(15), 4019–4026. https://doi.org/10.1523/JNEUROSCI.0564-07.2007

van den Bos, W., Crone, E. A., & Güroğlu, B. (2012). Brain function during probabilistic learning in relation to IQ and level of education. *Developmental Cognitive Neuroscience*, *2*, S78–S89. https://doi.org/10.1016/j.dcn.2011.09.007

van Gorp, W. G., Wilkins, J. N., Hinkin, C. H., Moore, L. H., Hull, J., Horner, M. D., & Plotkin, D. (1999). Declarative and Procedural Memory Functioning in Abstinent Cocaine Abusers. *Archives of General Psychiatry*, *56*(1), 85. https://doi.org/10.1001/archpsyc.56.1.85

van Timmeren, T., Quail, S. L., Balleine, B. W., Geurts, D. E. M., Goudriaan, A. E., & van Holst, R. J. (2020). Intact corticostriatal control of goal-directed action in Alcohol Use Disorder: A Pavlovian-to-instrumental transfer and outcome-devaluation study. *Scientific Reports*, *10*(1), 4949. https://doi.org/10.1038/s41598-020-61892-5

Vandaele, Y., & Janak, P. H. (2018). Defining the place of habit in substance use disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *87*, 22–32. https://doi.org/10.1016/j.pnpbp.2017.06.029

Vandaele, Y., Vouillac-Mendoza, C., & Ahmed, S. H. (2019). Inflexible habitual decision-making during choice between cocaine and a nondrug alternative. *Translational Psychiatry*, *9*(1), 1–11. https://doi.org/10.1038/s41398-019-0445-2

Vanderschuren, L., & Everitt, B. J. (2004). Drug seeking becomes compulsive after prolonged cocaine self-administration. *Science*, *305*(5686), 1017–1019. https://doi.org/10.1126/science.1098975

Vanderschuren, L. J., Minnaard, A. M., Smeets, J. A., & Lesscher, H. M. (2017). Punishment models of addictive behavior. *Current Opinion in Behavioral Sciences*, *13*, 77–84. https://doi.org/10.1016/j.cobeha.2016.10.007

Vandrey, R., Bigelow, G. E., & Stitzer, M. L. (2007). Contingency Management in Cocaine Abusers: A Dose–Effect Comparison of Goods-Based Versus Cash-Based Incentives. *Experimental and Clinical Psychopharmacology*, *15*(4), 338–343. https://doi.org/10.1037/1064-1297.15.4.338

Vanes, L. D., Holst, R. J. van, Jansen, J. M., Brink, W. van den, Oosterlaan, J., & Goudriaan, A. E. (2014). Contingency Learning in Alcohol Dependence and Pathological Gambling: Learning and Unlearning Reward Contingencies. *Alcoholism: Clinical and Experimental Research*, *38*(6), 1602–1610. https://doi.org/10.1111/acer.12393

Verdejo-Garcia, A., Benbrook, A., Funderburk, F., David, P., Cadet, J.-L., & Bolla, K. I. (2007). The differential relationship between cocaine use and marijuana use on decision-making performance over repeat testing with the Iowa Gambling Task. *Drug and Alcohol Dependence*, *90*(1), 2–11. https://doi.org/10.1016/j.drugalcdep.2007.02.004

Verdejo-Garcia, A., Chong, T. T.-J., Stout, J. C., Yücel, M., & London, E. D. (2018). Stages of dysfunctional decision-making in addiction. *Pharmacology Biochemistry and Behavior*, *164*, 99–105. https://doi.org/10.1016/j.pbb.2017.02.003

Verdejo-García, A., & Pérez-García, M. (2007). Profile of executive deficits in cocaine and heroin polysubstance users: Common and differential effects on separate executive components. *Psychopharmacology*, *190*(4), 517–530. https://doi.org/10.1007/s00213-006-0632-8

Verharen, J. P. H., Adan, R. A. H., & Vanderschuren, L. J. M. J. (2019). Differential contributions of striatal dopamine D1 and D2 receptors to component processes of value-based decision making. *Neuropsychopharmacology*, *44*(13), 2195–2204. https://doi.org/10.1038/s41386-019-0454-0

Vilà-Balló, A., Mas-Herrero, E., Ripollés, P., Simó, M., Miró, J., Cucurell, D., López-Barroso, D., Juncadella, M., Marco-Pallarés, J., Falip, M., & Rodríguez-Fornells, A. (2017). Unraveling the Role of the Hippocampus in Reversal Learning. *Journal of Neuroscience*, *37*(28), 6686–6697. https://doi.org/10.1523/JNEUROSCI.3212-16.2017

Volkow, N. D. (2021). Addiction should be treated, not penalized. *Neuropsychopharmacology*, *46*(12), 2048–2050. https://doi.org/10.1038/s41386-021-01087-2

Volkow, N. D., Chang, L., Wang, G.-J., Fowler, J. S., Ding, Y.-S., Sedler, M., Logan, J., Franceschi, D., Gatley, J., Hitzemann, R., Gifford, A., Wong, C., & Pappas, N. (2001). Low Level of Brain Dopamine D2 Receptors in Methamphetamine Abusers: Association With Metabolism in the Orbitofrontal Cortex. *American Journal of Psychiatry*, *158*(12), 2015–2021. https://doi.org/10.1176/appi.ajp.158.12.2015

Volkow, N. D., Fowler, J. S., Wang, G.-J., Hitzemann, R., Logan, J., Schlyer, D. J., Dewey, S. L., & Wolf, A. P. (1993). Decreased dopamine D2 receptor availability is associated with reduced frontal metabolism in cocaine abusers. *Synapse*, *14*(2), 169–177. https://doi.org/10.1002/syn.890140210

Volkow, N. D., Fowler, J. S., Wang, G.-J., & Swanson, J. M. (2004). Dopamine in drug abuse and addiction: Results from imaging studies and treatment implications. *Molecular Psychiatry*, *9*(6), 557–569. https://doi.org/10.1038/sj.mp.4001507

Volkow, N. D., Koob, G. F., & McLellan, A. T. (2016). Neurobiologic Advances from the Brain Disease Model of Addiction. *New England Journal of Medicine*, *374*(4), 363–371. https://doi.org/10.1056/NEJMra1511480

Volkow, N. D., Michaelides, M., & Baler, R. (2019). The Neuroscience of Drug Reward and Addiction. *Physiological Reviews*, *99*(4), 2115–2140. https://doi.org/10.1152/physrev.00014.2018

Volkow, N. D., Wang, G.-J., Fowler, J. S., Logan, J., Gatley, S. J., Gifford, A., Hitzemann, R., Ding, Y.-S., & Pappas, N. (1999). Prediction of Reinforcing Responses to Psychostimulants in Humans by Brain Dopamine D2 Receptor Levels. *American Journal of Psychiatry*, *156*(9), 1440–1443. https://doi.org/10.1176/ajp.156.9.1440

Volkow, N. D., Wang, G.-J., Fowler, J. S., Logan, J., Gatley, S. J., Hitzemann, R., Chen, A. D., Dewey, S. L., & Pappas, N. (1997). Decreased striatal dopaminergic responsiveness in detoxified cocaine-dependent subjects. *Nature*, *386*, 830.

Volkow, N. D., Wang, G.-J., Fowler, J. S., Logan, J., Hitzemann, R., Ding, Y.-S., Pappas, N., Shea, C., & Piscani, K. (1996). Decreases in Dopamine Receptors but not in Dopamine Transporters in Alcoholics. *Alcoholism: Clinical and Experimental Research*, *20*(9), 1594–1598. https://doi.org/10.1111/j.1530-0277.1996.tb05936.x

Volkow, N. D., Wang, G.-J., Ma, Y., Fowler, J. S., Wong, C., Ding, Y.-S., Hitzemann, R., Swanson, J. M., & Kalivas, P. (2005). Activation of Orbital and Medial Prefrontal Cortex by Methylphenidate in Cocaine-Addicted Subjects But Not in Controls: Relevance to Addiction. *Journal of Neuroscience*, *25*(15), 3932–3939. https://doi.org/10.1523/JNEUROSCI.0433-05.2005

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., Schreiber, L. R. N., Gillan, C., Fineberg, N. A., Sahakian, B. J., Robbins, T. W., Harrison, N. A., Wood, J., Daw, N. D., Dayan, P., Grant, J. E., & Bullmore, E. T. (2015). Disorders of compulsivity: A common bias towards learning habits. *Molecular Psychiatry*, *20*(3), 345–352. https://doi.org/10.1038/mp.2014.44

Voon, V., Reiter, A., Sebold, M., & Groman, S. (2017). Model-Based Control in Dimensional Psychiatry. *Biological Psychiatry*, *82*(6), 391–400. https://doi.org/10.1016/j.biopsych.2017.04.006

Wagner, F. A., & Anthony, J. C. (2007). Male–female differences in the risk of progression from first use to dependence upon cannabis, cocaine, and alcohol. *Drug and Alcohol Dependence*, *86*(2), 191–198. https://doi.org/10.1016/j.drugalcdep.2006.06.003

Walker, Q. D., Rooney, M. B., Wightman, R. M., & Kuhn, C. M. (1999). Dopamine release and uptake are greater in female than male rat striatum as measured by fast cyclic voltammetry. *Neuroscience*, *95*(4), 1061–1070. https://doi.org/10.1016/S0306-4522(99)00500-X

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

Watson, P., & de Wit, S. (2018). Current limits of experimental research into habits and future directions. *Current Opinion in Behavioral Sciences*, *20*, 33–39. https://doi.org/10.1016/j.cobeha.2017.09.012

Whelan, R., Conrod, P. J., Poline, J.-B., Lourdusamy, A., Banaschewski, T., Barker, G. J., Bellgrove, M. A., Büchel, C., Byrne, M., Cummins, T. D. R., Fauth-Bühler, M., Flor, H., Gallinat, J., Heinz, A., Ittermann, B., Mann, K., Martinot, J.-L., Lalor, E. C., Lathrop, M., … Garavan, H. (2012). Adolescent impulsivity phenotypes characterized by distinct brain networks. *Nature Neuroscience*, *15*(6), 920–925. https://doi.org/10.1038/nn.3092

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

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

Wilkinson, L., Tai, Y. F., Lin, C. S., Lagnado, D. A., Brooks, D. J., Piccini, P., & Jahanshahi, M. (2014). Probabilistic classification learning with corrective feedback is associated with in vivo striatal dopamine release in the ventral striatum, while learning without feedback is not. *Human Brain Mapping*, *35*(10), 5106–5115. https://doi.org/10.1002/hbm.22536

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, *5*(6), 483–494. https://doi.org/10.1038/nrn1406

Wise, R. A., & Robble, M. A. (2020). Dopamine and Addiction. *Annual Review of Psychology*, *71*(1), 79–106. https://doi.org/10.1146/annurev-psych-010418-103337

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological Review*, *114*(4), 843–863. https://doi.org/10.1037/0033-295X.114.4.843

Wood, W., Quinn, J. M., & Kashy, D. A. (2002). Habits in everyday life: Thought, emotion, and action. *Journal of Personality and Social Psychology*, *83*(6), 1281–1297. https://doi.org/10.1037/0022-3514.83.6.1281

World Health Organization. (2018, September 21). *Alcohol*. https://www.who.int/news-room/fact-sheets/detail/alcohol

Wrase, J., Schlagenhauf, F., Kienast, T., Wüstenberg, T., Bermpohl, F., Kahnt, T., Beck, A., Ströhle, A., Juckel, G., Knutson, B., & Heinz, A. (2007). Dysfunction of reward processing correlates with alcohol craving in detoxified alcoholics. *NeuroImage*, *35*(2), 787–794. https://doi.org/10.1016/j.neuroimage.2006.11.043

Wright, C. E., Sisson, T. L., Ichhpurani, A. K., & Peters, G. R. (1997). Steady-State Pharmacokinetic Properties of Pramipexole in Healthy Volunteers. *The Journal of Clinical Pharmacology*, *37*(6), 520–525. https://doi.org/10.1002/j.1552-4604.1997.tb04330.x

Yager, L. M., Garcia, A. F., Wunsch, A. M., & Ferguson, S. M. (2015). The ins and outs of the striatum: Role in drug addiction. *Neuroscience*, *301*, 529–541. https://doi.org/10.1016/j.neuroscience.2015.06.033

Yamaguchi, M., Suzuki, T., Seki, T., Namba, T., Juan, R., Arai, H., Hori, T., & Asada, T. (2004). Repetitive Cocaine Administration Decreases Neurogenesis in Adult Rat Hippocampus. *Annals of the New York Academy of Sciences*, *1025*(1), 351–362. https://doi.org/10.1196/annals.1316.043

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action–outcome learning in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 505–512. https://doi.org/10.1111/j.1460-9568.2005.04219.x

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2006). Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning. *Behavioural Brain Research*, *166*(2), 189–196. https://doi.org/10.1016/j.bbr.2005.07.012

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513–523. https://doi.org/10.1111/j.1460-9568.2005.04218.x

Zapata, A., Minney, V. L., & Shippenberg, T. S. (2010). Shift from Goal-Directed to Habitual Cocaine Seeking after Prolonged Experience in Rats. *Journal of Neuroscience*, *30*(46), 15457–15463. https://doi.org/10.1523/JNEUROSCI.4072-10.2010

Zhukovsky, P., Morein-Zamir, S., Meng, C., Dalley, J. W., & Ersche, K. D. (2020). Network failures: When incentives trigger impulsive responses. *Human Brain Mapping*, *41*(8), 2216–2228. https://doi.org/10.1002/hbm.24941

Zhukovsky, P., Puaud, M., Jupp, B., Sala-Bayo, J., Alsiö, J., Xia, J., Searle, L., Morris, Z., Sabir, A., Giuliano, C., Everitt, B. J., Belin, D., Robbins, T. W., & Dalley, J. W. (2019). Withdrawal from escalated cocaine self-administration impairs reversal learning by disrupting the effects of negative feedback on reward exploitation: A behavioral and computational analysis. *Neuropsychopharmacology*, *44*(13), 2163–2173. https://doi.org/10.1038/s41386-019-0381-0

Zwosta, K., Ruge, H., Goschke, T., & Wolfensteller, U. (2018). Habit strength is predicted by activity dynamics in goal-directed brain systems during training. *NeuroImage*, *165*, 125–137. https://doi.org/10.1016/j.neuroimage.2017.09.062