A Computational Model of Habits in the Brain

A dissertation submitted towards the degree Doctor of Philosophy (DPhil)
at the Brain Network Dynamics Unit of Oxford University

by

Charlotte Collingwood

September, 2025

Abstract

The neural mechanism of action-selection is often modelled as a combination of goal-directed learning and habitual movements. Impairments in the function and interplay of these processes are associated with the aetiology of many neural pathologies. Until recently, a fundamental discrepancy existed in the definition of habits between experimental and computational models.

Experimentally, habits are defined as the reward-independent, stimulus-response relationships which form when an action is regularly executed in the same context, regardless of outcome. In contrast, computational models largely represent habits through model-free algorithms reliant on reward prediction errors.

This thesis expands upon proposals by Miller et al.[1] and Bogacz[2] which resolve this inconsistency through dopaminergic *action* prediction errors, rather than reward. Specifically, we extend their work to test for the existence of such signals in neural and behavioural data. To this aim, two novel computational models are developed and then compared in their ability to replicate the associated data against the gold-standard reward prediction error equivalents.

This thesis first focusses on dopaminergic signals across continuous time. We begin by outlining our 'temporal-difference action learning' algorithm, which uses biologically-plausible mechanisms to determine how dynamic changes in action intensity influence the resultant prediction errors across near-continuous time. We then demonstrate that dopaminergic data collected by Greenstreet et al.[3] from the tail of the

striatum is better represented by action prediction errors than reward.

The later chapters explore the detection of value-free habits under time-constrained conditions in human behavioural data from Hardwick et al.[4]. This required the creation of our 'two-drift race diffusion model' and corresponding analytical solutions since, to our knowledge, no algorithms previously existed that included mid-trial drift changes in multi-alternate forced choice paradigms. We finally establish that most participants' behaviour was best described by stimulus-response relationships which evolved from action prediction errors.

Overall, our results support the existence of value-free action prediction errors and associated habitual behaviour in dopaminergic signals and human behavioural data. This thesis also provides a proof-of-concept application of our novel models and opens the avenue for future research to test their predictions more directly using data specifically collected for this purpose.

Abstract Word Count: 346.

Acknowledgements

The list of people to whom I owe thanks, whilst not endless, is long. The last four years have not always been easy, but the highs are entirely attributed to the many wonderful people who surround me.

First, I would like to express my gratitude to my supervisor, Prof. Rafal Bogacz, for his tireless enthusiasm and encouragement, as well as to everyone in the BNDU office. Thank you all for creating such an intellectually stimulating environment. I am equally grateful to the Medical Research Council for the studentship that supported my research and education.

To my OUPS community, especially Chiara, Elena, Rose, Kiera and David, between all our cafe co-working sessions where very little work occurred and our many jam sessions at the GLD, your love and support has kept me sane.

I cannot imagine my time here without Sherin, Lisa, Sophie and Blanca. Our mutual obsessions with plants, cats and jammy red roo made Oxford feel like home. I promise, Shadow is finally warming up to you all.

I would also like to thank Tom, Dan, Jo, Ben and Evie for always being ready to remind me that there is a world outside of academia that is important too. Similarly, to Scranner, whose willingness to help me escape Oxford and get to the closest patch of nature has meant more to me than I could ever say.

Most importantly, to my parents - thank you. Not only for your support during my

PhD, but also for endlessly encouraging me to keep going, even when you probably thought I should slow down and take a break. I have always been able to count on your unconditional presence, your willingness to chat about all things science (and maths, Dad), and lately, Mum's impeccable proofreading skills.

List of Abbreviations

A-O	Action-Outcome
AIC	Akaike Information Criterion
AMPA	α-amino-3-hydroxy-5-methyl-4-isooxazole-propionic acid
APE	Action Prediction Error
BG	Basal Ganglia
BIC	Bayesian Information Criterion
BMS	Bayesian Model Selection
BVP	Boundary Value Problem
cdf	Cumulative Density Function
CoT	Cloud of Tones
CSC	Complete-Serial-Compound
DDM	Drift Diffusion Model
DLS	Dorsolateral Striatum
DMS	Dorsomedial Striatum
DS	Dorsal Striatum
EAM	Evidence Accumulation Model
FI	Fixed Interval
FR	Fixed Ratio
GABA	γ-aminobutyric Acid

GPe	Globus Pallidus Externus
GPi	Globus Pallidus Internus
ICD	Impulse Control Disorder
ITI	Inter-Trial Interval
LTD	Long-Term Depression
LTP	Long-Term Potentiation
MLE	Maximum Likelihood Estimation
NAcc	Nucleus Accumbens
NLL	Negative Loglikelihood
NMDA	N-methyl-D-aspartate
PCA	Principal Component Analysis
PD	Parkinson's Disease
pdf	Probability Density Function
PE	Prediction Error
PFC	Prefrontal Cortex
RDM	Race Diffusion Model
RL	Reinforcement Learning
RL-DDM	Reinforcement-Learning Drift Diffusion Model
RL-EAM	Reinforcement-Learning Evidence Accumulation Model
RL-RDM	Reinforcement-Learning Race Diffusion Model
RPE	Reward Prediction Error
RT	Reaction Time
S-R	Stimulus-Response

SARSA	State-Action-Reward-State-Action
SNc	Substantia Nigra pars compacta
SNl	Substantia Nigra pars lateralis
SNr	Substantia Nigra pars reticulata
SOR	Schedule of Reinforcement
SPN	Spiny Projection Neuron
SSE	Sum Squared Error
STN	Subthalamic Nucleus
TD-AL	Temporal-Difference Action Learning
TD-RDM	Two-Drift Race Diffusion Model
TD-RL	Temporal-Difference Reinforcement Learning
TPE	Threat Prediction Error
TS	Tail of the Striatum
VI	Variable Interval
VR	Variable Ratio
VS	Ventral Striatum
VTA	Ventral Tegmental Area

List of Tables

List of Figures

Contents

1 Introduction

We have all experienced habits in our lives, and we often blame them when we make unintended actions. Maybe you were driving to buy some milk after work and, while you were distracted thinking about your research or mentally drafting that urgent email, you accidentally made a left-turn to go home rather than continuing on towards the shops. While this is a common scenario in the real world, it can appear strange on reflection - without any conscious control, your foot pressed the brake, your arms changed gears and turned the steering wheel, and (hopefully) you'll have even checked the mirrors. What is it that causes such *slips-in-action* in our behaviour when we aren't paying attention?
This is the question posed by researchers across many fields, including behavioural psychology, neurobiology and computational neuroscience, when they interrogate the nature of *habits*.

A key theory for how one action may be chosen over another proposes that the filtering of potential action plans is done through a two-process system[5] (Section 2.1). The *goal-directed* system takes time and resources in order to analyse the current context, the likely outcomes of all available choices, and make the choice that either maximises the positive consequences or minimises the negative. Though this goal-directed system is flexible and effective, the brain has thousands of choices to make every second and insufficient resources to calculate the goals and outcomes for all of them. This is where the habit process is useful, as it learns which choice is usually made in a given context and copies previous actions, thus allowing for rapid decisions to be made with minimal analysis. However, there can be dire consequences when habits go wrong.

1.1 Habits and Neuropathologies

Maladaptive habits are associated with a wide range of medical disorders, either forming a core part of the root aetiology or as side effects from medical treatment. For example, habits are linked to compulsive drug seeking behaviours[6], gambling addictions[7], and impulse control disorders (ICDs) associated with levodopa treatment for Parkinson's disease (PD)[8,9]. Our internal states are also strongly affected by habits - excessive rumination in depressed patients is often conceptualised as a 'mental habit' and modelled using a habit-goal framework[10]. Similar arguments are given for obsessive compulsive disorders[11] and motor compulsivity in Tourette's syndrome[12].
Gaining a deeper insight into how these harmful habits form and become compulsive is vital in helping patients manage and overcome their symptomatology.

Striatal dopamine has been particularly implicated in all of the above disorders and has long been thought to play a role in habit formation[7,9,11–14]. Aligning the evidence of dopamine's role in habit strength with the mathematical models describing their formation will likely play a key role in developing future treatments. But first, we need to understand what habits actually are.

1.2 What Are Habits?

Until recently, the exact definition, and consequent modelling, of what constitutes a habit has been inconsistent between the experimental and computational neuroscience fields.

In psychological and behavioural studies, where unconscious slips-in-action were first established, habits are defined as the reward-insensitive stimulus-response (S-R) behaviours[15]. These develop in extended instrumental conditioning experiments[16], during which animals repeatedly perform the same action in response to a stimulus until this behaviour becomes automatic. This automaticity eventually leads to inflexibility - when a choice is no longer optimal, habitual actions will take much longer to extinguish than those which are still goal-directed[16].

In contrast to S-R theory, the computational neuroscience field often represents habits with model-free reinforcement learning (RL) algorithms as an alternative to model-based goal-directed choices[17]. While the latter produces an optimal action policy by learning the state transitions which arise from different stimuli and actions[18], the former simply works to minimise the reward prediction error (RPE) by selecting choices with the maximum expected value in response to a given stimuli. These fundamentally rely on a RPE, and so, are reward-dependent.

To solve this disparity, Miller et al.[1] and Bogacz[2] have separately suggested that prediction errors (PEs) may instead be heterogeneous and that S-R associations could evolve from an *action prediction error (APE)* triggered by the occurrence of non-habitual actions. Specifically, Bogacz[2] hypothesises that reward-insensitive habits learn from a dopaminergic error signal encoded in a subset of basal ganglia (BG) neurons. However, the two models proposed by these papers worked on a trial-by-trial basis and neither developed a mechanistic model for how such APEs could be produced and evolve within a trial, nor directly related them to experimental data on the dopamine dynamics in the BG.

1.3 Research Questions

In this thesis, I extend upon Miller et al.[1] and Bogacz's[2] work and address four complementary questions:

1. Can habit learning be generalised to continuous and scalar actions?
2. Can evidence of action prediction errors be found in striatal dopamine?
3. Can the process by which habits affect choices and reaction times be mechanistically described?
4. Can the impact of habits in human behavioural data be quantified?

Each of these research questions is addressed in their own chapter, using two novel computational models developed as part of this work.

Over the course of the next chapter, I provide an overview of the current literature and

key models used in the computational neuroscience field. Specifically, I describe how this two-process action selection theory arose, alongside the discovery of dopamine as neurological correlate to the PE in RL algorithms and its role in learning and reward. Then, I outline the classic mathematical models used to calculate both the likelihood of a given action being selected and, in some cases, the timing of its execution. A particular focus is placed on the proposal of heterogeneous prediction errors, in the form of APEs, and what these models can teach us about the development of habits.

In Chapter 3, I present a biologically-plausible learning model which produces predictions of expected (future) actions, based on prior experience. These actions can be continuous and scalar; the model is able to predict *when* the action will occur, relative to predictive cues, and it produces realistic APEs which abide by standard theories of dopamine signalling.

I subsequently apply this learning model to experimental data in Chapter 4, in order to compare the model APE dynamics to real dopaminergic data provided by Greenstreet et al.[3]. In doing so, I show that these signals are consistently better described by an APE than they are by a RPE.

The third research question is addressed in Chapter 5 by my second computational model. I have extended an RL-based evidence accumulation model (EAM) to create an algorithm that applies both goal-directed and habitual values in a time-dependent manner for multiple alternative choices within a trial. I also present a near-analytical cost function, which allows us to efficiently predict the likelihood of a given choice being made at a specific reaction time, based on the model's previous experience.

In Chapter 6, I apply this EAM model to experimental human data from Hardwick et al.[4] and show that (1) model and parameter recovery is possible using my novel cost function and (2) the resulting parameters can provide insight into the inter-individual differences in how humans rely on and apply habits.

Finally, in Chapter 7, I discuss in more depth the utility of these models in the wider context of action-selection research and reflect on how potential combinations and extensions of these two models could provide further insight into the existence of value-free habits in the brain.

2
Background

Contents

This chapter provides an account of the relevant computational, behavioural and neuroscientific experiments and discoveries which define the field of *computational habit* research as it stands today.

Initially, three key topics are covered: how habits were discovered and defined, what the relevant established models of action selection are, and why dopamine presents a good neural correlate for these models. This knowledge is then used to highlight a key discrepancy in how these elements have been combined, present a recent proposal which addresses this issue and discuss the current questions this thesis specifically focusses on.

2.1 Defining Habits

In layman's terms, habits are understood as the regular actions we make without conscious intention. For example, we colloquially call automatic behaviours, such as having a morning coffee or biting our nails, habits. Equally, a smoker might describe reaching into their pocket for a new cigarette whenever they go outside as a 'habit' that they have developed[19].

The scientific definition is similar. It was established primarily using operant conditioning experiments[15] (Section 2.1.2). Specifically, habitual actions are defined as those that result from a strong stimulus-response (S-R) relationship. In this section, the meaning of this scientific definition of a habit is explored and the key behavioural and neuroscientific experiments which developed this classification are summarised.

2.1.1 Two-process action selection hypothesis

When considering how an animal selects an action, one pervasive theory can be found across psychological, behavioural, neuroscientific and computational fields, which states that two parallel processes are involved in filtering between potential actions[5].

The first, classically named the goal-directed pathway (or System 2), carefully analyses which available choice will lead to the best outcome in the current context, whether by maximising the positive or by minimising the negative consequences. This is often referred to as an action-outcome (A-O) relationship[20,21] - the animal understands the correlation and potential causation of action to outcome. This analysis provides flexibility to changing conditions but requires extra focus and attention[22,23], inherently slowing the decision process and creating a drain on limited cognitive resources[24–26]. Consequently, the goal-directed system is unable to manage the thousands of actions and choices that are made in every moment.

Instead, many of these are controlled by an alternative process, known as the habit pathway (or System 1), which learns to repeat the actions that were previously taken in a specific context. In contrast to A-O associations, these can be conceptualised as the

internal understanding of the S-R contingency - when an agent always makes the same choice in a given set of circumstances, an S-R relationship is established and calculating potential outcomes becomes unnecessary. By definition, this pathway is less flexible to changing circumstances but it is much more time- and resource-efficient[24,27].

The key evidence for the existence of habits, the factors that influence their strength and the likely candidates for their neural representation are summarised below.

2.1.2 Behavioural evidence

2.1.2.1 Operant conditioning

Habits were first defined through *operant conditioning* experiments (also called instrumental conditioning or instrumental association tests). B.F. Skinner[28,29] is often attributed as one of the founders of this experimental paradigm, which explores an animal's understanding of a stimulus-response-outcome relationship and their ability to learn the appropriate 'operation' under a given condition. Designed to explore Thorndike's Law of Effect[30], these tests place an animal (often a rodent or pigeon) into a box with access to a reinforcer and the ability to perform a trained action (such as lever pressing, button pecking or nose-pokes).

The earliest versions of these experiments allowed the animal to freely perform the trained action as often as it wished and imposed different latent rules for reinforcement. The pattern and rate of action execution were then recorded and reported. The insights provided by these studies regarding how the 'schedule of reinforcement (SOR)' affects behaviour are discussed further below (Section 2.1.2.3). Later iterations would also introduce a neutral cue (e.g., lights or a speaker) to explore the influence of discrete predictive cues on the behaviours expressed.

Although precise details of the reinforcement type, schedule and stimulus presentation vary across experiments, at their core, they explore an animal's association of a reinforcer with a neutral cue, which becomes a learnt, discriminative stimulus over training. Thus, this research studies the impact of this association on the behaviour of, or 'work' done by, the animal. Usually, habits are not measured during the learning phase of these

experiments, instead S-R relationships are classically tested using *extinction* paradigms[16].

2.1.2.2 Extinction

In the context of instrumental-association experiments, extinction refers to the reduction (or total loss) of learnt behaviours in response to the discriminative stimulus when a reinforcer is no longer received.

Extinction tests are classically applied after one of the following two manipulations. The first reduces the motivational drive or perceived value of an outcome. This 'devaluation' experiment can be done through satiety (e.g., comparing behaviour pre- and post-feeding in rats)[31,32] or by aversive training external to the operant conditioning context (e.g., creating an aversive reaction to a reinforcer by giving rodents lithium chloride immediately after receiving said reinforcer)[33,34]. The latter process was employed by Adams and Dickinson[33] in the first experiment on rodent lever-pressing that successfully distinguished A-O and S-R responses.
In contrast, rather than altering the outcome's value, the second manipulation impacts the A-O relationship directly by relaxing or completely removing the contingency between responses and outcome (e.g., a lever-press no longer produces a reward, reinforcers are given at random regardless of lever pressing, or reinforcers are only given when a lever *isn't* pressed).
Regardless of structure, these paradigms all test the strength of the animal's association between an executed action and the resulting outcome (i.e., the A-O relationship) by studying how rapidly the animal ceases to perform the trained behaviour in response to the learnt stimulus when it is no longer appropriate. The longer an animal's behaviour is resistant to extinction, the stronger the S-R contingency's control over action selection, relative to A-O.

It is worth noting that extinction is not the 'unlearning' of A-O relationships as once thought, since the original acquired behaviour can rapidly be reinstated in the correct context[35,36]. Rather, it seems that the animal instead learns that this relationship is no longer applicable in its current environment[37,38].

2.1.2.3 Building strong habits

With the classification of habits as outcome-insensitive perseverative actions, attention turned towards discovering which factors made them stronger or more likely to develop. Early extinction experiments showed that resistance develops with extensive training and repetition[39] and overtraining is now considered to be an established method to develop habits[16].

As mentioned above, B.F. Skinner's early work[40,41] explored how the SOR could influence habit formation. At their simplest, the conditions on responses required for an animal to receive a reinforcer can be classed in two ways:

1. **Ratio versus interval**: The former requires a certain number of actions to be completed before a reward is given (thus impacting the *ratio* of actions to reinforcement directly), while the latter reinforces the first action completed after a certain time *interval*, regardless of how many repeated actions were performed in the interim.
2. **Fixed versus variable**: Scientists can also adapt the consistency of the conditions for reinforcement. In a *fixed* condition, the rewarded ratio or interval remains constant throughout testing, whereas a *variable* setting will introduce uncertainty to the effort required or the delay during which acting is futile.

Initially explored in pigeon peck-rates, the combinations of these settings were found to impact response rates differently[29,41] as shown in Fig. 2.1. Fixed ratio (FR), variable

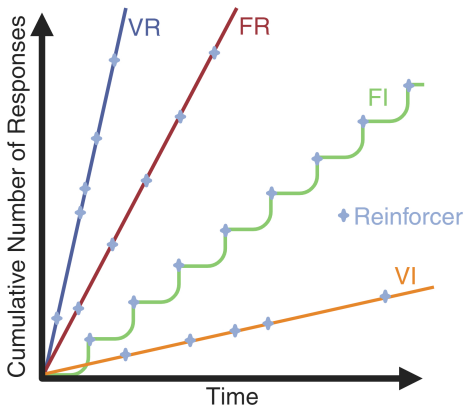


Figure 2.1: Schedules of Reinforcement. A schematic representation of classical animal behaviour in the four simple SOR, with combinations of Ratio versus Interval and Fixed versus Variable. Fixed intervals (FI) produce a scalloping effect as responding pauses following reinforcement[40]. All other SORs create linear response rates, with gradients dependent on the specific requirements (such as FI1 versus FI5).

ratio (VR) and variable interval (VI) conditions all produce linear peck rates which are maintained between reinforcers.

In contrast, fixed interval (FI) schedules lead to a 'scalloping' pattern in the animals' responses, such that pecks would cease immediately after reinforcement before slowly ramping up as the end of the interval approached, indicating that the pigeons were capable of measuring the interval duration and calculating the utility of actions relative to the time of previous cues.

Moreover, studies have shown that these learnt patterns of behaviour also survive extinction, as they will re-occur during reinstatement[38].

It has since been established that resistance to extinction is more likely to arise under interval schedules over ratio[32,42] and in variable reinforcement conditions[43]. This remains true even when the specific schedules are adapted to ensure a similar number of actions or reinforcements occur across all conditions. Though extended training is still required to produce these habits, overtraining alone cannot account for this effect of schedule.

Two key hypotheses have developed to explain these patterns of behaviour.

Dickinson[44] proposed that the key factor influencing habit formation was the 'response rate - reward rate' correlation. The rate at which rewards are given has long been assumed to directly influence response frequency[45]. Dickinson argued that VR schedules have a high correlation between reward probability and the number of actions made - the more actions an animal makes, the more likely it is to receive a reward. In contrast, there is little to no correlation for VI schedules - a single action is enough to access the reinforcer and interim actions are superfluous.

Thus, Dickinson proposed that habits result from a reduced reliance on the behaviour-reward (or A-O) correlation and that this weakened relationship is inherent to, and so rapidly learnt during, VI and FI schedules.

An alternative argument posits that it is the A-O *contiguity*, rather than rate correlation, which is relevant to habit formation. When comparing FI and VI schedules, DeRusso et al.[43] determined that, when action-reward correlation is held constant between conditions, the reduced A-O contiguity in the VI setting produces an increased

uncertainty regarding the timing of reward. They cited this as the causal factor for resistance to extinction.

DeRusso et al.'s results were replicated by Garr et al.[46], who performed a series of experiments to explore the influence of correlation, contiguity and reward density on habit expression. Their results favoured DeRusso et al.'s 'temporal uncertainty' hypothesis over Dickinson's[44] rate correlation proposal.

Interestingly, uncertainty has also been applied by many computational models when calculating the balance between goal-directed and habitual pathways[47,48]. In these proposals, each system has a measure of its own 'certainty', which determines its relative contribution to the action selection process. However, there is a discrepancy in the definition of 'uncertainty' between the experimental and computational proposals. Where DeRusso et al.[43] refers to the uncertainty in reward *timing* following an action, these computational models consider the expected *likelihood* of a reward when a choice has been made.

This difference can more clearly be seen in a fixed-trial context, where provision of a reinforcer can be probabilistic. The previously discussed studies measured *free* operant behaviour, whereby an animal learns that a new reinforcer is available (the 'trial' has restarted) when the previous reward is received. By definition, this means that every trial *must* be reinforced, and uncertainty in reward probability (or *outcome* uncertainty) is low.

This reinforcement constraint can be loosened by providing an external indicator that a trial has ended, which thus allows for the reward probability to be adapted between trials. For example, in a FI task, an external cue informs the animal that it has acted after the required delay and that a new interval has started, without any reinforcer being applied. Thrailkill et al.[49] used such a task in two VI conditions, with either a deterministic or probabilistic reward scheme. By controlling for A-O contiguity (DeRusso et al.'s[43] temporal uncertainty) between conditions, Thrailkill et al.[49] showed that habit expression was promoted under the deterministic, low outcome-uncertainty condition. This aligns with the theoretical understanding of the habitual processes acting as a low-resource

system that can be applied when the environment is predictable.

Further, uncertain contexts are stressful for animals and stress itself has a bearing on habit strength. For example, injection of anxiogenic drugs biases rats towards habitual memory systems[50] and Dias-Ferrera et al.[51] showed that when animals are chronically stressed, they are more likely to develop resistance to extinction earlier and for longer. It is well-established that stress reduces the amount of cognitive resources available, so reliance on efficient S-R relationships may be an effective coping strategy[27,52].

2.1.3 Neuroscientific evidence

One fundamental drawback in the behavioural studies of habit formation is that the evidence for goal-directed and habitual control are inextricably linked - resistance to extinction could result from either reduced cognitive control or from stronger habits. These experiments can only test their *relative* strengths. Neuroscientific studies are differently constrained, thus allowing these two processes to be disentangled to a greater degree.

Most of the initial research to establish the existence and neural location of habits was performed through lesion studies[15,53–56]. Specifically, this work determined the disassociation between action learning and action perseverance, while linking these choices to the BG and dorsal striatum (DS).

2.1.3.1 Anatomy of the basal ganglia

The impact of lesion studies is better understood when we know precisely what has been damaged, functionally and structurally. So, this section outlines the anatomy of the BG without describing how this knowledge arose[57]. The overall arrangement and relationships are schematised in Fig. 2.2. Although most of these anatomical findings apply specifically to rodents, this region's homology is strong across mammals and the rodent striatum has a medial-lateral connectivity gradient which largely mirrors the primate caudate-putamen separation[58].

The BG is composed of a series of subcortical midbrain nuclei classically associated with

reward learning and motor control.

The largest of these nuclei, the striatum, can be further subdivided into the nucleus accumbens (NAcc), or ventral striatum (VS), and the caudate and putamen (which together make up the DS). The striatum receives projections from several regions, including the majority of the neocortex[59], the thalamus and dopaminergic projections from the brainstem.

The efferent nuclei, the globus pallidus internus (GPi) and the substantia nigra pars reticulata (SNr), primarily project to the thalamus and are functionally associated with the 'release' of movement inhibition[60]. Only their wider function in action control and related connectivity are presented here.

Finally, the intrinsic nuclei include the substantia nigra pars compacta (SNc), the ventral tegmental area (VTA), the subthalamic nucleus (STN) and the globus pallidus externus (GPe). Two of these nuclei, the SNc and the VTA, are of particular interest here as they are the source of dopaminergic signalling in the BG.

The ventromedial to dorsolateral topography of the striatum is maintained throughout the BG.

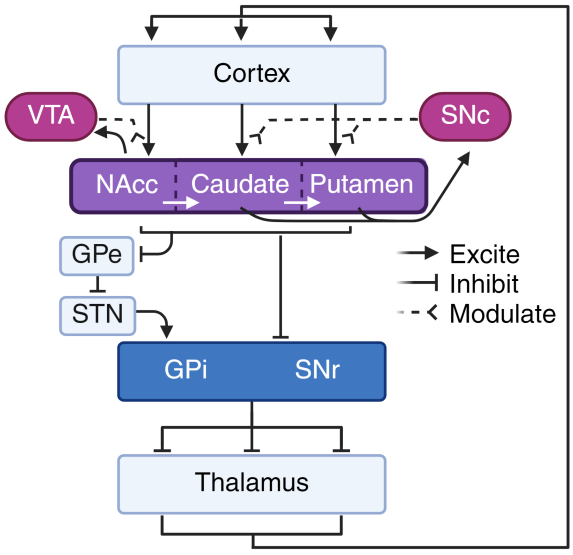


Figure 2.2: Basal Ganglia Anatomy, Simplified.
Outline of basic BG nuclei and connections. Relevant cortical regions send excitatory signals to the striatal subcomponents - the NAcc, caudate and putamen (purple). The strength of these connections is modulated by dopamine from the VTA and SNc (pink). The signals propagate through the intrinsic nuclei to reach the GPi-SNr complex (dark blue), which (dis-)inhibits thalamic projections back to the initial cortical regions. The striatum also contains internal microcircuits which project from the VS to the DS.

The details of the internal BG processing are, of course, much more complicated than a simple funnel from cortex to thalamus. They include their own parallel microcircuits of direct, indirect and hyperdirect pathways[57,61], serotonergic, cholinergic

and noradrenergic signalling[62], and interneuronal connectivity[57,63]. For this thesis, it suffices to know that, as per Fig. 2.2, dopamine release in the striatum modulates the dendritic synaptic connections between cortical projections and the GABA-ergic* spiny projection neurons (SPNs), which form roughly 90% of the striatum, and that the activity of these SPNs in the DS promote action (dis)inhibition in the thalamus, via the STN, GPe and the GPi/SNr. Section 2.2 provides more details on the role of dopamine.

The BG and cortex are connected in a series of ascending parallel corticostriatal loops. Functionally-related cortical areas project to topographically similar segments of the striatum, and the relevant thalamic regions regularly project back to the initial cortices[59,61], thus producing a closed loop. Although labelled as 'parallel', these loops are not entirely segregated and information is exchanged between them. In particular, the VS appears to feed-forward into the DS far more than the reverse[64], which led to the 'ascending spiral' hypothesis of corticostriatal connectivity[65].

Two corticostriatal circuits are relevant here[59,61]. The first, often labelled the limbic circuit, encompasses the VS, medial thalamus and medial prefrontal cortex (PFC). It receives dopaminergic input from the *mesolimbic* pathway (Section 2.2.1) and is thus influenced by the amygdala and hippocampus. This corticostriatal loop is most often associated with reward processing, following historical intracranial self-stimulation studies[66] and Schultz's seminal work on VTA dopamine neurons[67].

In contrast, the corticostriatal 'motor' circuit engages the DS and receives its projections from the associative and sensorimotor regions of the cortex. The topographical organisation of the cortical inputs is maintained at the striatal level, with the associative cortical regions projecting medially, while the dorsolateral striatum (DLS) primarily receives information from the sensorimotor cortex. The *nigrostriatal* dopamine pathway (Section 2.2.1) also forms part of this circuit. As discussed in more detail below, this motor circuit is also often functionally subdivided into the dorsomedial striatum (DMS) and lateral DLS components of the DS.

*GABA is a key inhibitory neurostransmitter.

2.1.3.2 The neural basis of habits

The rest of this section provides an overview of the key experiments which aimed to establish how habits are represented by the brain. Many brain regions play some role in action selection and thus, the expression of habits, such as the amygdala[68,69], PFC, insular cortex[56,70] and the hippocampus[71].

Here, a specific focus is given to a widely-accepted hypothesis; that (1) the DS is involved in both goal-directed and habitual learning, as well as action expression, and (2) these two processes are encoded in structurally distinct regions. Commonly, the DMS is believed to compute goal-directed relationships, while habits are proposed to exist in the DLS. This striatal paradigm has received much previous habit modelling[53,55,58,72] with an emphasis on two-process action selection algorithms and has strong neural correlates with RL models.

Given the outcome-insensitive nature of habits, it is worth noting that, when the VS is added to this gradient, a theoretical argument can be made for a transitional pattern in the BG from pure value to pure action encoding.

2.1.3.3 Animal studies

This overview begins by reporting results from three key experimental manipulations performed in animals: striatal lesions, neuronal recordings and optogenetic (de)activation.

Two entirely parallel neural processes should, in theory, be dissociable. More precisely, if two functions are anatomically separate, then it should be possible to lesion one region responsible for a given behaviour without impairing the other. This logic underlies double-dissociation lesion studies and has been applied to the two-process action selection hypothesis to great effect.

Among others, Yin et al.[54] summarise a series of experiments they performed to compare the impact of specific lesions within the DS. To begin, they trained rats under VR and VI lever-pressing schedules. The rats had either undergone an excitotoxic lesion of the

DMS or DLS, or were intact (controls). Yin et al. found that the DLS-ablated rodents expressed a reduced resistance to extinction (weaker habits) after VI training - an SOR that biases towards habit formation - for the control and DMS-lesioned animals. The opposite effect was seen following DMS ablation under VR conditions, whereby habits formed much sooner than for the other groups who remained sensitive to outcome-devaluation throughout. These results are the quintessential double-dissociation effect; the expression and formation of goal-directed action and habits are dependent on an intact DMS and DLS, respectively.

Yin et al.[54] found similar results for DMS lesions under a spatial maze task. In a 4-arm maze, an animal can either turn towards the correct direction, regardless of start-point, by understanding its relative location in a map of the world (a 'place' strategy, or A-O) or it can consistently turn in the same direction when it reaches the junction (a 'response' strategy, or S-R). Once again, disruption of the DMS resulted in animals that relied solely on 'response' strategies, suggesting an inability to apply A-O goal-directed knowledge to behaviour.

Additionally, the DLS appears to access information independently of the DMS. Yin et al.[73] showed that DMS-lesioned animals are still able to *learn* an instrumental association task, while expressing habitual behaviours. This lends further support to the discrete separation of the two processes, as the S-R associations do not appear to require prior A-O knowledge.

Overall, these striatal lesion studies provided strong evidence of the two action-selection processes being functionally separated between the DMS and DLS.

Neuronal recordings have also supported this hypothesis. For example, when an animal approaches a goal, the DMS neurons fire sequentially in a spatially-selective manner[74], implying an association of motor action to desired outcome. Further, primate recordings have found that the 'associative caudate' (DMS equivalent) neurons are activated more extensively during early learning than the 'sensorimotor' (putamen, DLS equivalent) region, which preferentially fires during established, more automatic, movements[75].

Equally, other complementary experiments have extended on neural recording studies to

directly manipulate the striatal activity via optogenetic inhibition/excitation[76,77]. Crego et al.[78] applied these methods to the DLS in a 4-arm maze task, where a left-turn was always rewarded regardless of start point. DLS excitation prior to and during a trial caused the rats to both run and make a decision faster, while inhibition lead to more deliberation before a choice was made, suggesting that the automaticity and time-efficiency of habitual actions was induced by DLS activity. When the outcome was devalued, this inhibition effect was even larger (rats ran more slowly and were more likely to quit the trial early), implying that their behaviour skewed more heavily towards goal-directed actions than in the control or excited conditions.

Taken in conjunction, these studies provide strong evidence for the existence of two discrete action-selection processes, with goal-directed control requiring an intact DMS and habitual control located within the DLS.

However, the precise mechanism through which these regions interact is still unclear. Some have suggested that they work co-operatively and that the DMS gradually recruits the DLS throughout learning[14,79], while others propose that they are in constant competition[80].

Most likely, the answer lies between the two. In a healthy brain, the balance of control early in training does weight towards the DMS, but, evolutionarily, our habits should be aligned with our goals for the majority of the decisions we make. The competition between the two likely arises from the limitations of the experimental methods used to access the S-R behaviours in isolation.

Further, it is worth restating that regions other than the striatum are involved in these action selection processes and compensatory mechanisms can be utilised. For example, Whishaw et al.[81] had previously performed a spatial task with similar results to Yin et al.[54]. However, when their paradigm was extended to increasingly complex tasks, the DMS-lesioned mice restored their goal-directed 'place' strategy. This concurs with the view that, while the DMS is not *necessary* for A-O goal-directed behaviour, overcoming its loss requires an active engagement of other regions.

As such neuronal manipulations, unsurprisingly, cannot be applied to the human brain,

scientists have turned to neuroimaging studies instead to observe the mechanisms of human habit formation.

2.1.3.4 Human habits

Animal models have been used to establish several key qualities of habitual learning and behaviour. Unfortunately, determining whether these rules hold true for humans, particularly in instrumental association studies, has not been as simple and is still at the frontier of habit research today. To this aim, neuroscientific measurements have proven promising.

De Wit et al.[82] explored five experimental paradigms in order to investigate an established failure to reproduce habits in humans using classical experimental approaches. Specifically, they showed that humans were sensitive to outcome-devaluation in all five studies, regardless of the time spent overtraining the task-associated action. Rather than interpret this as evidence supporting an absence of human S-R behaviours, they instead concluded that the paradigms previously used in animal studies were insufficient to isolate habitual behaviours in humans.

As an alternative to complete extinction tests, experimentalists are much more likely to find habits expressed as slips-of-action in humans. This definition assumes that humans are usually capable of suppressing habits cognitively, and thus, expression of habits comes through 'slips' in attention[83,84], which occur at a greater rate when the goal-directed system is taxed; for example, by contextual stress or increased cognitive requirements.

In Chapter 6, we explore one such experiment by Hardwick et al.[4], who employed time-constrictions on human behaviour to increase the likelihood of an action-slip.

Recordings under fMRI have produced further promising evidence of a DMS-DLS separation in action selection for human subjects. Guida et al.[72] performed a meta-analysis of 57 similar studies and concluded that the DLS homologue (dorsal putamen), alongside other regions, was activated by both lab-learnt and everyday expression of habits in humans. For example, Tricomi et al.[85] found that fMRI recordings

showed increased cue-related activity in the posterior putamen (equivalent to the rodent DLS) as participants learnt a VI task. Further, these subjects developed resistance to extinction with overtraining, which correlated with DLS activity. Notably, the DLS was active even when habits were not expressed. Tricomi et al.[85] interpreted this as evidence that the habitual processes continued to occur, even when the expressed behaviour was goal-directed, as one would expect from two parallel (and potentially competing) systems.

Similar studies that focussed on goal-directed control have largely found more evidence of an 'action-driven value signal' in the PFC and orbitofrontal cortex, rather than the DMS, implying a greater degree of 'top-down control'[86,87], though the caudate has likewise been implicated by some[88,89].

Human habits also appear to be influenced by stress. Schwabe and Wolf[52,90] showed that, relative to controls, cold-pressor stress prior to learning caused participants to be more resistant to extinction and have a reduced verbal understanding of the A-O contingencies. In contrast, inducing stress post-learning but prior to the extinction test had no impact on the explicit A-O knowledge, while still producing greater habitual behaviour. These studies thus imply that stress affects habit expression beyond its influence on memory formation. More recently, a similar experiment showed that stressed humans express habits after relatively little training[91].

In sum, the rodent DLS (and primate homologs) appears to play a key role in the development and expression of habits as reward-insensitive S-R associations. Next, we look to *how* habits (and goal-directed learning) could develop in these regions and explore the role that dopamine plays in this process.

2.2 Dopamine and Reinforcement Learning

One of the greatest achievements of computational neuroscience has been the discovery of striatal dopamine as a (potential) neural correlate of PEs in RL[92]. The next section presents the neurobiology of dopamine, alongside a description of its role in (and

historical associations with) both reward and movement, before relating its mechanism of action to its proposed function as a PE. This thesis contributes to the vast field of work exploring its function, where, in Chapter 4, dopamine data are directly compared with predictions from the novel RL model presented in Chapter 3.

2.2.1 Dopaminergic neuroanatomy

Dopamine is a monoamine which acts as a neurotransmitter and neuromodulator in specific brain regions. Two of the classic dopaminergic pathways, known as the *mesolimbic* and *nigrostriatal* pathways, are relevant here.

The mesolimbic pathway, whose dopamine cells originate in the VTA and project to the NAcc, alongside the PFC and amygdala[93,94], has historically been associated with reward and motivation, as part of the corticostriatal limbic circuit (Section 2.2.2). The VTA receives afferent inputs from many regions associated with emotion and motivation, including the NAcc, amygdala and lateral hypothalamus[95].

The second network of interest to action selection is the nigrostriatal pathway[93,94]. Neurons in this circuit originate in the SNc and project to the caudate and putamen. Loss of these cells is associated with kinetic pathologies, such as PD[94,96]. On account of the motor nature of these dysfunctions with the afferent signals from the somatosensory and motor cortices, together with its bidirectional communication with the DS[95], much research has focussed on the role of nigrostriatal dopamine in movement and motor control (Section 2.2.3).

The precise details of the connectivity between dopaminergic neurons and their targets strongly determines our understanding of its function as a neurotransmitter. BG dopamine neurons have vast arborisation and the projection of a single cell reaches a large number of SPNs (some estimate several thousand)[97]. The consequences of dopamine signalling is determined by the monoamine's *receptor*.

Five dopamine receptor classes are known, aptly named D1-D5, but the majority of striatal research has focussed on two G-protein coupled receptors, D1R and D2R, due to their high concentration and opposing effects within the striatum[94,98].

Activation of D1R triggers a cellular cascade with three key outcomes[99]: neuronal excitability is increased, gene expression is initiated and synaptic plasticity is promoted. In contrast, D2Rs reduce neuronal excitability and are used as presynaptic auto-receptors to regulate their own release of dopamine. D2R activation also inhibits release of glutamate at the striatum, thus reducing the excitatory input[100].

Finally, one region of the BG, the caudal tail of the striatum (TS), often excluded from simplified models, is worth noting as we will return to it in Chapter 4. The afferent dopaminergic signals to this region arise purely from the substantia nigra pars lateralis (SNl) and show a wide range of reactivity beyond reward related properties[101]. Instead, the TS has been linked to auditory association learning[102], novelty[103], and threats[104]. It is also a potential neural site in which a pure APE may be found[3], as will be discussed further in Chapter 4.

2.2.2 Reward and dopamine as a prediction error

For a long time, dopamine was considered to be 'the reward chemical'. In 1954, Olds and Milner[66] showed that activation along the mesolimbic pathway (though the precise neural circuit was only identified later) appeared to produce a subjectively hedonic experience. Combining this observation with the dopaminergic mechanisms of drug stimuli led to the 'dopamine theory of reward'[105].
In the following decades, it became increasingly apparent that, rather than encode the rewarding experience itself, mesolimbic dopamine may instead represent a RL signal.

A seminal piece of work by Schultz et al.[67] showed that primate dopamine neurons in the VTA increased their firing upon reception of an unexpected reward and that during conditioning this response would transfer to the arrival of a predictive cue. They further demonstrated that firing was mildly depressed if an expected reward was not received.
This work was striking since it established a neural signal which held all of the qualities required of a temporal PE, as proposed by Sutton and Barto[106]. Dopamine firing peaked during unexpected outcomes, transferred to predictive cues over learning and was depressed whenever an expected reward didn't arrive. Section 2.3 discusses the impact

of temporal PE algorithms on computational neuroscience further, but suffice to say for now that this research was integral in furthering the field of neural RL.

Since then, research into the temporal properties of phasic dopamine have revealed a large degree of heterogeneity[107–109], with some neurons responding to novelty[103] or aversive signals[110,111] and demonstrating adaptable sensitivities that may shift depending on context[112]. Even within the clear RPE signals, the magnitude of dopamine response is influenced by internal states, such as satiety[113] or motivation[114]. Human fMRI studies have also shown dissociable PEs within the VS[115]. Motor-related dopamine signals form their own field of study and are discussed further in Section 2.2.3.
Several explanations, which are not mutually exclusive, have been proposed for this heterogeneity. One hypothesis states that different dopamine signals report different *features* of a reward, such as sucrose or water content, that may have different motivational values depending on the animal's state[114,116]. The aversive PE evidence has led Uchida et al. to develop their 'weal and woe' theory of dopamine[104] which supposes that the TS learns when to expect threat-relevant outcomes.

Most relevant to this thesis, is the proposal that some dopamine signals may reflect an *APE*[2,3]. Indeed, nigrostriatal dopamine has been associated with motor responses for nearly as long as the reward-theory of mesolimbic dopamine has existed[117].

2.2.3 Dopamine, movement and habits

In a parallel field of dopamine studies, the neuromodulator was associated with movement initiation, motivation and vigour. This is perhaps not surprising, given the existence of a nigrostriatal dopamine pathway that so strongly links sensorimotor cortices with the DS, alongside the large body of evidence which suggests that dopamine depletion leads to hypokinetic disorders. For example, degeneration of the SNc is associated with parkinsonian symptoms in humans[118] and rats treated with 6-OHDA lesions will develop akinesias[119]. Further, the opposite also holds true as levodopa treatment can induce dyskinesias[120]. Even as they were measuring dopaminergic RPEs, Schultz et al. reported SNc activation at the initiation of arm-reaches in primates[121].

In the past 25 years, the concept of dopamine as a movement-related signal has gained traction. Much of the early research regarding the relationship between dopamine and movement looked to how altering the monoamine's concentration influenced an animal's 'willingness to work' and the vigour with which the action was carried out[122–125].

One difficulty with these studies is the interpretation of the mechanism which selects an action's 'vigour'. An increase in dopamine could induce hyperactivity[126] because of a direct influence on motor selection, or instead, an animal's motivation may be impacted by dopamine's effect on reward expectation.

Different theorists disagree on the principal cause. Berke and Hamid[127,128] have argued for a singular dopaminergic signal that incorporates motivation, effort and potential reward. Underlying this proposal is the rationalisation that, since dopamine increases excitability in SPNs, a high tonic concentration will result in a striatal network that is much more likely to initiate movement for a given input, as though the animal is more motivated to move for the same potential reward.

In contrast, others have proposed that dorsal dopamine has direct influence on (and from) movement kinematics. For example, SNc dopamine has been shown to express signals related with velocity and direction, regardless of outcome[129–131] and optogenetic stimulation of those neurons was sufficient to induce movements[129,130]. These neurons were distinct from those encoding reward and all projected to the DS[130]. Howe et al.[130] specifically demonstrated that SNc dopamine release peaked with movement initiation, was responsive to acceleration bouts during continuous locomotion and dipped when the action terminated. Further, Azcorra et al.[132] recently isolated a new genetic subtype of dopamine neuron, *anxa1*+, which showed similar locomotive response patterns, projected solely to the DLS and was completely reward insensitive. These results are in accordance with previous demonstrations of action-on and action-off dopaminergic neurons, whose activities are either up-modulated or down-modulated in response to continuous actions[133]. Elsewhere, the execution of motor plans has also been found to influence the expression of RPEs in the NAcc[134] and to suppress nigrostriatal dopamine[135] (for review, see Coddington and Dudman[136]).

Likewise, instrumental association tasks have uncovered a relationship between motor

execution, learning and dorsal dopamine. At the DMS, dopamine appears to be necessary to update A-O contingencies[137] and their axons respond specifically to contralateral cues and rewarded actions[138]. Recordings in the DLS revealed signals linked to the initiation and execution of task-specific movements[139]. Importantly, these studies show that DMS and DLS dopamine is responsive to movement *specifically* when it is a learnt A-O or S-R response, respectively, which lends further support to dopamine's role in developing these relationships and associated behaviours.

Beyond pure movements, habits and dopamine have been inextricably linked as a result of research into ICDs and addiction. Long-term use of addictive drugs is believed to produce *compulsive* habits[6,140] - maladaptive S-R relationships taken to the extreme whereby preventing action execution requires constant and active cognitive control.
The dopaminergic mechanism of action of these addictive drugs is regularly cited as a causal factor, particularly with the finding that certain D2R genotypes are linked to an increased risk of developing addiction during the early stages of drug-use[141,142]. However, addiction is a complicated disease, with many co-morbidities and confounding external and social factors. In contrast, ICDs, which have a similar pathology and apparent aetiology, can develop from a direct and known cause - treatment with a dopamine agonist[8,143]. Both disorders demonstrate maladaptive habit formation, associated with an increase in dopamine concentration.
Direct rodent studies further support a connection between dopamine and habits, as amphetamine promotes habit formation in rats[144], dopamine-specific lesions in the DLS removed the expression of learnt habits[145,146] and knocking out glutamatergic input to dopamine neurons prevented S-R learning entirely[147].

Clearly, dopamine is strongly associated with short- and long-term changes in action selection. The next section explores the mechanisms by which these effects may arise within the striatum.

2.2.4 Dopamine, synapses and Hebbian learning

As a neuromodulator, striatal dopamine's main site of action is at the synapse. These are points of (unidirectional) communication between two neurons, whereby the pre-synaptic cell can either excite (and potentially induce action potentials in) or inhibit the post-synaptic neuron. Influencing the strength of these connections, in either the short- or long-term, is a vital form of neuroplasticity and learning in the brain.

Understanding how synapses are able to alter brain function and, thus, the underlying requirements for neural learning to occur, is key when developing biologically-plausible models. Classically, synapses are believed to update using *Hebbian learning*[148–150]. First proposed by Hebb, this theory is summarised by the well-known quote: "neurons which fire together, wire together"[149]. In practice, when neurons on either side of the synapse are regularly co-activated, long-term potentiation (LTP) occurs and the connection will be strengthened in future, creating a completely local positive feedback loop.
The opposite can also hold true, such that when correlation between cell activities is low, the synapse can undergo long-term depression (LTD) and the connection between the two cells is weakened. The majority of the original research into these phenomena was largely performed in the hippocampus using rabbit eyeblink conditioning tasks[151]. These showed that regular co-activation of the perforant pathway and hippocampal granule cells resulted in the removal of NMDA-receptor† magnesium blocks in the post-synaptic cell, causing it to be more strongly impacted by pre-synaptic glutamate release. These effects appear to be maintained in the longer term through post-synaptic structural changes (e.g., increasing spine density and size) and an up-regulation in the number and sensitivity of AMPA receptors†.

Beyond pre- and post-synaptic activity, striatal plasticity is influenced by a third factor, dopamine, whose impact is so integral that corticostriatal synapses are also known as the striatal tripartite synapse[152]. One effect of dopamine is to influence the SPN excitability. SPN neurons can be in one of two states. During their 'down-state',

†NMDA and AMPA are two key receptors for glutamate, an excitatory neurotransmitter.

membrane potassium channels are open and perpetually hold the neuron far from a depolarised state. Conversely, while in the 'up-state' these ion channels close and the neuron is partially depolarised, thus requiring much less excitement to induce spiking[153]. Transition to this up-state is induced by sufficient glutamatergic signalling from the cortex to overcome the transition threshold.
Dopamine affects these synapses via D1Rs to make the post-synaptic SPN more likely to shift to an up-state and increase the probability of firing. D2Rs have the opposite effect[154].

Combining neuron co-activations with the influence of dopamine at the synapse and its apparent role as a PE produces a compelling story of striatal learning. When an RPE is produced, the striatal concentration of dopamine increases, activating D1Rs. Any post-synaptic cells receiving a glutamatergic input will be more likely to co-activate and the synapse will be strengthened. For example, SPN dendritic spines grow in the presence of dopamine during co-activation with the cortical presynaptic neuron[155]. In computational terms (described further in Section 2.3), the positive PE increases the weight of the cortical neuron's influence on SPN activity.

Overall, dopamine is a diverse neuromodulatory hormone with wide-reaching effects throughout our bodies.

2.3 Mathematical Models of Action Selection

The previous sections have presented the evidence underlying the neuropsychological definition of habits and the development of the prediction error theory of dopamine. In parallel and in conjunction with this work, computational and mathematical models have been created to replicate neural circuits, produce testable predictions and advance our understanding of how symptoms may arise when these circuits go wrong. A subset of the most influential mathematical models of action selection and the BG are outlined in the next part of this chapter.

2.3.1 Model-based or model-free?

RL models can be characterised in many ways, but one common classification separates them as either model-based or model-free learning algorithms.

The former encompasses any algorithm that holds an internal model of the environment and applies information about the current state and transition probabilities (e.g., Markov decision processes) to calculate the consequences of an action[156–158], whereas model-free systems learn an estimated value and optimised policy through trial-and-error[106,159] (for review, see Huang et al.[18]).

Daw, Niv and Dayan[17,47] argued that the dual-processes of action selection could be mapped to these RL classes, such that goal-directed actions are model-based while habitual learning is model-free. This proposal became near-ubiquitous in the field of computational habits and instrumental associations in various hierarchical formulations[25,160–162], later expanding into Pavlovian learning[163].

The justification for this is as follows. Goal-directed analysis requires an understanding of both the current context and the influence of the agent's actions on the state and future consequences. Model-based RL is commonly associated with such situations, allowing forward planning to occur before actions are undertaken. In contrast, model-free RL simply recreates the antecedent actions which have historically provided the highest average reward. The resulting inflexibility has led to a 'habitual' interpretation.

Imaging evidence aligns with this proposal, as it has been used to associate the putamen with model-free value signals[164] and link the VS with both model-based and model-free values[165]. As a result, Daw later extended this dichotomy to include both model-based and model-free representations of goal-directed learning[166].

Much deliberation has been given to determining how the brain arbitrates between model-based and model-free action selection, and to discovering correlative neural signals.

A common proposal returns to a factor believed to arbitrate between goal-directed and habitual choices - uncertainty[47,164] (Section 2.1.2.3). More specifically, this refers to

outcome-uncertainty, such that an arbiter is more likely to rely on efficient habits if the outcome is consistent and exploitation of the environment is possible with minimal forward planning.

Whether conscious or subconscious, control over this compromise is typically considered to be 'top-down' - an individual cognitively determines whether goal-directed calculation is appropriate and, if so, which actions are available. Neural imaging supports this, as activation of an 'arbiter' is often associated with (varying) regions of the PFC[164,167].

Section 2.4.1 returns to the proposal of model-free habits to interrogate this assumption further, highlight a key problem in the axioms that underlie this choice and discuss an alternative definition. However, before analysing the applicability of this field-wide assumption, a deeper comprehension of the classical model-free learning algorithms is required.

2.3.2 Temporal-difference reinforcement learning

By definition, RL encompasses a class of learning models which minimise PEs to improve future estimates of the measured outcome. Perhaps the best known model-free RL algorithm, particularly in the neuroscience field, is that of temporal-difference reinforcement learning (TD-RL).

A full review of the maths underlying this theory is provided by Jensen[158]. It can be summarised by Eq. 2.1 in which an estimate of the potential future value, V, is calculated by collating all expected rewards, R, multiplied by a discount factor, γ. This discount factor introduces the assumption that immediate rewards are perceived to have more value than delayed benefits, an effect which has been seen in numerous behavioural studies[168,169]. The optimal state-value of V is denoted by V^* and is calculated using Eq. 2.1.

$$V(t)^* = E\left[\sum_{k=1}^{\infty} \gamma_v^{k-1} R(t+k)\right] \tag{2.1}$$

where: V = expected future value,

γ_v = value discount factor,

R = reinforcement.

As RL is an entirely local form of learning, its update rule relies only on the current and antecedent state. In brief, the RPE compares the reward received at the current timestep, t, with the value it expected to receive between the current and previous timestep.

$$\delta_V(t) = R(t) + \gamma_v V(t) - V(t-1) \tag{2.2}$$

where: δ_V = reward prediction error.

However, before this RPE can be used to update the value estimate, V, a solution to the 'credit assignment problem' is needed.

2.3.2.1 Credit assignment

To be effective and produce accurate descriptions, dopamine-based RL models that work over continuous time must be able to determine which external events are salient to the outcome experienced, a requirement otherwise known as the credit assignment problem[170]. Specifically, these algorithms need to be capable of knowing which stimuli have been experienced and how long ago they occurred.

The complete-serial-compound (CSC) TD-RL model[67,171] provides one solution by representing a stimulus, i, as a series of 'component' vectors. Each timestep which occurs during the presence of a stimulus has an associated binary vector, x_{ij}, whose value is set to 1 at $t = j$ (as in Fig. 2.3D). These vector elements have their own weights, w_{ij}, and it is these *weights* that update in each timestep, using the classic RL learning rule (Eq. 2.3). The current value, $V(t)$, is then calculated as shown in Eq. 2.4 by summing the stimuli-presence values, x_{ij}, modified by their associated weight, w_{ij}.

$$w_{ij}(t+1) = w_{ij}(t) + \alpha_v \delta_v x_{ij}(t) \tag{2.3}$$

$$V(t) = \sum_{i=1}^{n} \sum_{j=1}^{t} w_{ij}(t) x_{ij}(t) \tag{2.4}$$

where: w = stimulus weight,

i = stimulus,

j = stimulus component,

x = CSC vector,

α_v = learning rate,

n = total number of stimuli.

However, Eq. 2.3 is unable to replicate a key feature of the dopaminergic signal described by Schultz[67]. Specifically, the PE slowly backpropagates across time (antecedent weights) until it arrives at the earliest predictive stimulus (Fig. 2.3B). In contrast, the DA signal

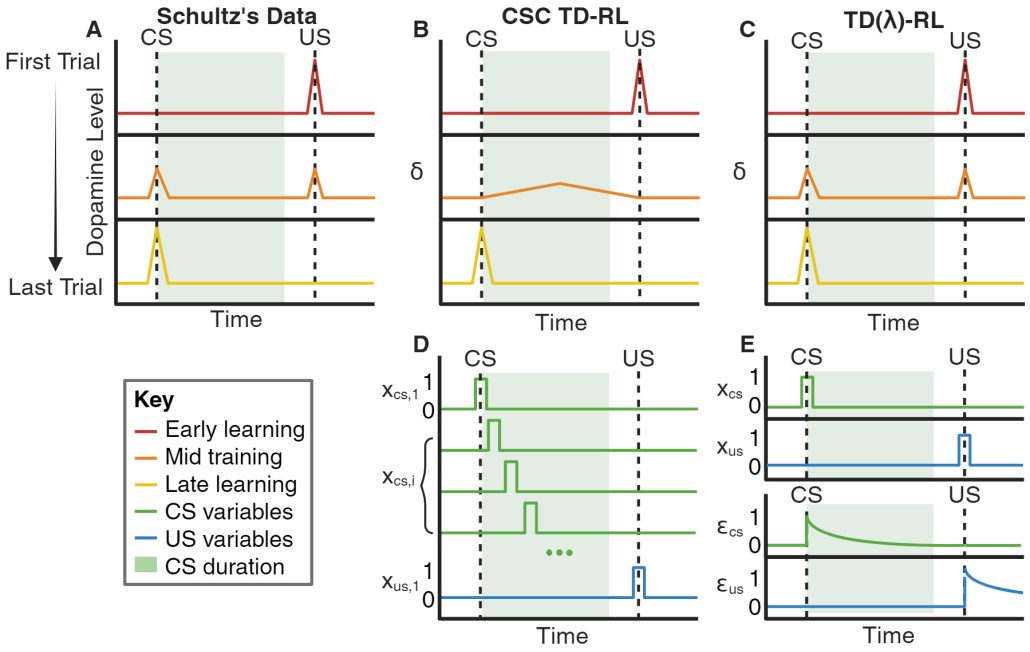


Figure 2.3: Schematic Representation of TD-RL Models.
Schematic comparison of dopamine dynamics across learning with the PE produced by the CSC TD-RL and TD(λ)-RL algorithms.
A: Dopamine dynamics from early to late in learning (top to bottom). Dopamine first responds to unexpected rewards (US). This is slowly transferred to the initiation of a predictive cue (CS). Eventually, dopamine firing only increases upon presentation of the CS.
B: CSC prediction error, δ, dynamics. Equivalent to **A** for early and late trials, but δ slowly backpropagates across the interval (orange).
C: As **B** for TD(λ)-RL, with δ transferred directly to CS during learning (orange).
D: CSC representation of stimuli. Each timestep during CS presentation (green) is represented by its own x_{ij} vector.
E: (Top) TD(λ)-RL stimulus representation, a binary vector, x_i, produced for both CS (green) and US (blue) initiation. (Bottom) The associated eligibility traces produced for CS and US initiation.

transfers directly to the CS presentation. This disparity can be solved through an *eligibility trace*[172,173].

2.3.2.2 Recording memory - λ

Sutton and Barto[172,174] adapted Eq. 2.3 in a model they named TD(λ)-RL. This algorithm replaces the individual component representations, j, with an eligibility trace, ε_i, that records the memory of previous states. The weights are then able to update proportionally to the temporal interval since the associated discriminative stimulus occurred.

In its simplest form, this is represented by an exponential decay function (Eq. 2.5) and an adaptation of the previous update rule given by Eq. 2.3 (Eq. 2.6):

$$\varepsilon_i(t+1) = \lambda\varepsilon_i(t) + x_i(t) \tag{2.5}$$

$$w_i(t+1) = w_i(t) + \alpha_v\delta_v\varepsilon_i(t) \tag{2.6}$$

where: λ = memory decay parameter,

ε = eligibility trace.

In other words, the eligibility trace controls the duration for which a weight remains *plastic* (i.e., can update) and decays at a rate of λ. Note that if λ is set to 0, Eq. 2.3 and Eq. 2.6 are equivalent for a given vector component, j (where for TD(λ)-RL, j is fixed to 1).

A larger λ value ($0 \leq \lambda < 1$) will hold the memory for longer, which results in the associated weights learning faster and being able to associate stimuli and reinforcers over greater intervals.

Then, as before, the total 'value', V, is calculated by summing the stimuli-presence values, x_i, at the current timestep, modified by their associated weight (Eq. 2.7).

$$V(t) = \sum_{i=1}^{n} w_i(t)x_i(t) \tag{2.7}$$

Fig. 2.3C and 2.3E demonstrate how the eligibility trace offers an elegant solution to the initial problem of backpropagation - now that the weights are specific to a given stimulus, i, (rather than its compound elements), credit can be assigned directly instead of travelling through the interim.
In Chapter 3, we will explore a biologically-plausible alternative to the decaying eligibility trace which better replicates many features of the dopamine RPE signals[175,176].

TD-RL, and extensions thereon, have been applied to neuroscientific models with great success[171,177–180]. The next section investigates how the BG may employ these mechanisms.

2.3.3 Modelling the basal ganglia

One of the earliest influential BG models that incorporates TD-RL is the *actor-critic* model.
Alongside his PhD thesis in which TD-RL was first developed, Sutton collaborated with Barto to propose an RL machine capable of solving a variety of tasks in an unsupervised manner[181,182]. A few years later, Houk et al.[183] explored how the BG could employ such an actor-critic model to select actions most likely to result in valuable consequences, by comparing different mathematical variables of the TD-RL RPE against the pharmacokinetics of dopamine signals (for a review of the actor-critic model's development, see Joel et al.[184]).

In the actor-critic algorithm, a critic is proposed to compare the outcomes received to those expected by the model in its current state, essentially calculating the potential value, $V(s)$, of the environment. The actor receives the state information and the critic's assessment of the environment and uses both to directly determine the policy, $\pi(a|s)$, that is most likely to maximise the utility of the action executed. Variations and adaptations of the actor-critic model have been proposed and are still employed to this day[184–188].
However, the mapping onto BG anatomy is inconsistent between successive iterations. The initial Houk et al.[183] formulation did not separate the BG into the anatomical regions described in Section 2.2.1. Rather, it defined the critic as the subset of 'striosomal SPNs'

which make up ~5% of the SPN population and project to the VTA and SNc. The actor encompassed the remaining ~95% of thalamus-projecting 'matrix SPNs'.

Later models have instead associated the critic with the 'value-pure' VS and the actor with the 'motor' DS[184,186,189]. This latter definition arises from its similarity to our understanding of the ventral-dorsal functional separation of the striatum and the mesolimbic and nigrostriatal dopamine pathways discussed in Section 2.1.3.

The actor-critic algorithm can be mathematically represented using simplified RL update rules which only care for the immediate reward. For a given policy, π, the critic and actor are respectively described by Eq. 2.8 and Eq. 2.9.

$$V^{\pi}(s,t+1) = V^{\pi}(s,t) + \alpha\Big(R(t) - V^{\pi}(s,t)\Big) \tag{2.8}$$

$$Q_a^{\pi}(s,t+1) = Q_a^{\pi}(s,t) + \alpha\Big(R(t) - V^{\pi}(s,t)\Big) \tag{2.9}$$

where: s = state,
a = action,
R = reward,
π = policy,
V = expected value in state, s,
Q_a = expected value for action, a, in state, s,
α = learning rate,
t = trial.

A second category of BG models do not separate policy and value assessment, as the actor-critic does, and instead develop the action-state relationships directly. These are the 'Q-learning' models. First proposed by Watkins[159], Q-learning algorithms learn a single function, $Q_a(s)$, which associates expected action-values to the states directly.

In Watkins' algorithm, the complete PE includes the reward in the current state, discounted by all rewards available in the next state if the optimal action is selected (Eq. 2.10):

$$Q_a(s,t) = Q_a(s,t) + \alpha\Big(R(t+1) + \gamma \max_a Q_a(s,t+1) - Q_a(s,t)\Big) \tag{2.10}$$

where: γ = discount factor,

R = reward.

Note that Eq. 2.10 can be extended to work over near continuous time by altering the interpretation of t from a trial to a timestep. This equation can also be adapted to work in a model-free manner by changing the formulation of the RPE to use simplified terms (as in Eq. 2.8 and Eq. 2.9) :

$$Q_a(s,t+1) = Q_a(s,t) + \alpha\Big(R(t) - Q_a(s,t)\Big) \tag{2.11}$$

An alternative update rule is provided by the State-Action-Reward-State-Action (SARSA) model[190], which assumes that the agent will continue to select the current action. In doing so, its learning rule is comparable to the value-based TD-RL (Eq. 2.2), now with the value dependent on the action executed:

$$\delta_{Q_a}(s,t) = R(t) + \gamma\, Q_a(s,t) - Q_a(s,t-1) \tag{2.12}$$

$$Q_a(s,t) = Q_a(s,t-1) + \alpha\delta_{Q_a}(s,t) \tag{2.13}$$

where: δ_{Q_a} = reward prediction error, given action.

Q-learning, SARSA and the actor-critic algorithms are all model-free and, in their earliest forms, did not consider the difference between habitual and goal-directed action selection.

2.3.4 Making choices

The learning models discussed above are core algorithms used to calculate the expected state/action-value. However, an RL agent also requires an 'observation model'[191] through which it can select and execute an action.

2.3.4.1 Softmax

Classically, Q-learning, SARSA and actor-critic algorithms work on a trial-by-trial basis (Eq. 2.8 - 2.13), such that each update of the Q/V variable is immediately followed by the occurrence of a new action.

Though it may appear beneficial to consistently select whichever action has the highest expected value in all trials, this method is incapable of adapting to a changing environment - the optimal decision cannot be determined without sampling the other options. This is known as the 'exploration-exploitation trade-off'[192,193]. Balancing these two approaches can be resolved through observation functions that introduce a degree of stochasticity. The softmax algorithm is one such application (Eq. 2.14).

$$P(a = c|s,t) = \frac{e^{\beta Q_c}}{\sum_{j \neq c} e^{\beta Q_j}} \tag{2.14}$$

where: β = inverse temperature parameter.

Observation functions allow for simulations of action selection to be performed, but also aid in fitting parameters to real data, as will be explored in Chapters 4 and 6. The pattern of choices made by participants in different environments, under experimental tasks or across psychopathologies, have been analysed using these models to provide multiple interpretations of their influences on action selection[53,193–201].

2.3.4.2 Introducing reaction times

The choices we make are not the only variable affected by a dual-process action selection method. Experimentalists can also measure a participant's reaction time (RT) distribution. Analysing how the RTs change across learning[202], task complexity[203–205] and environmental context[206–210] has revealed a multitude of factors influencing how individuals make choices.

One overarching feature of these studies is the well-established 'speed-accuracy

trade-off'[211,212] - the longer a participant is given to consider their options, the more likely they are to make the correct choice.

The balance of speed and accuracy is often discussed in the context of the two-process action selection theory[213,214]. In speeded-decision experiments, this effect can be parameterised by comparing the RT distributions for the correct vs. incorrect choices, as the phenomena of 'fast' errors (defined relative to 'slow' correct choices) will shift the average RT earlier.

However, as task complexity increases or if a participant is asked to prioritise their accuracy, 'slow' errors can emerge[215,216], which are believed to result from cognitive uncertainty. Recently, Damaso et al.[216] proposed that these should instead be considered as fast *response-speed* errors, versus slow *evidence-quality* errors. In unlimited time conditions, the main source of error-trials will be due to the subject being unsure of the correct response. These trials would therefore be associated with a greater degree of deliberation than when the participant strongly believes their answer to be correct and, consequently, the error RT will slow.

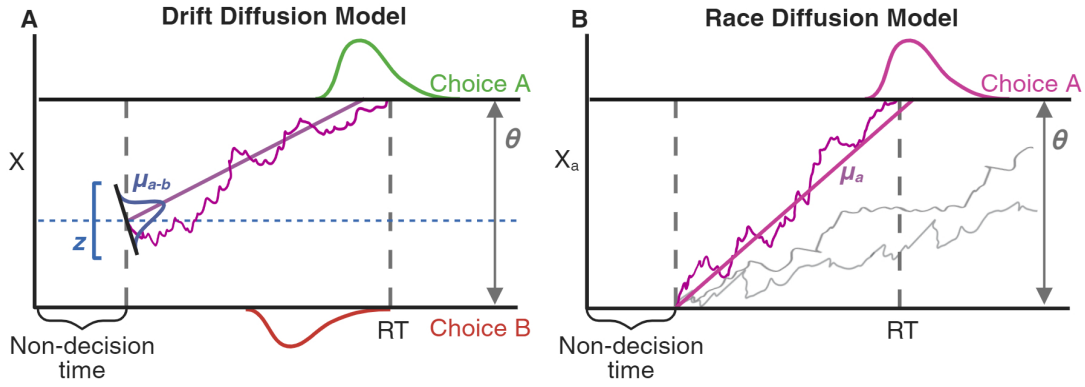


Figure 2.4: Evidence Accumulation Models.
A schematic representation of a trial for two EAMs. For both models, choice and RT are determined when the accumulator (purple) passes a threshold.
A, The Drift Diffusion Model: A single accumulator, X, (purple) drifts according to a Wiener process (Eq. 2.15). Between-trial variance comes from the start point, z, and the drift-rate, μ, whose mean value is determined by the relative evidence for choice A versus B. The distance between boundaries is a fixed value, θ.
B, The Race Diffusion Model: A single accumulator, X_a, (purple) races against many others (grey) to reach threshold, θ, first. All stochasticity is determined within the trial through the Wiener process. The drift-rate, μ_a, represents the absolute evidence for the associated choice, a, and the distance to threshold is fixed to θ.

Mechanistically, these experiments have been modelled (and these phenomena explained) using evidence accumulation models (EAMs). This section focusses on two established algorithms: the drift diffusion model (DDM) and the race diffusion model (RDM) (Fig. 2.4).

The two models address different experimental contexts but they apply the same underlying assumption; speeded-decision making results from an information accumulation process and a motor action is launched when an accumulator reaches a threshold activity level.

Some core shared components are that (1) the accumulator's drift-rate, μ, represents the degree of evidence for a choice and accumulates according to a Wiener process (Eq. 2.15), (2) the RT of any given trial is determined when the accumulator reaches a threshold value, and (3) total RT includes the accumulation process plus some 'non-decision time' which accounts for stimulus processing and motor execution.

Further, all EAM models require stochasticity within their algorithms to replicate the RT distribution, remove determinism from action selection and reproduce the existence of both fast and slow errors.

The DDM was developed first. Defined by Ratcliff in 1978[217], it works in a two-alternate forced choice system. This binary system can be described in a multitude of ways (e.g., correct/incorrect, left/right, go/no-go) but the mathematics remain the same.

A single accumulator drifts between an upper and lower threshold (often set to θ and 0, respectively), according to a Wiener-diffusion process (Eq. 2.15), and crossing a threshold determines the choice and RT of that trial. The drift-rate, μ, is defined as the *relative* evidence for the two options, which are always in direct competition. The accumulator begins each trial at a starting-point, z, that lies between the two boundaries and is able to represent an inherent bias towards either choice.

$$X(t) = X(t-1) + N(\mu, \sigma^2), \quad X(0) = z \tag{2.15}$$

where: X = accumulator activity,

t = timestep,

μ = drift-rate,

σ = noise,

z = starting point.

DDM introduces a few sources of variance. The first, *within-trial* variance, is produced through the stochastic Wiener diffusion process. It has been shown that the noise parameter, σ, at each timestep allows DDM to provide a good estimation of the RT distribution. However, the relative nature of the drift-rate renders within-trial stochasticity insufficient to replicate the existence of slow errors - each timestep is equally noisy for both choices. To solve this, *between-trial* variability in μ was introduced.
DDM algorithms can also include between-trial variability in starting-point and non-decision time, which has been shown to provide better fits in particular cases[218].

Today, DDMs are still widely influential and have been expanded to incorporate features such as collapsing thresholds[219,220] and multiple drift-rates[221–223]. However, the enforced binary system is a core limitation in the modelling of multi-alternate forced choice experiments.
The introduction of competing accumulators racing towards a single boundary resolves this issue. Whichever accumulator passes the shared threshold first 'wins' the trial and the associated choice is executed. Some early examples of these algorithms include the 'leaky, competing accumulator model'[224] and 'interactive race model'[225], which both incorporate features to replicate direct competition between accumulators.
The RDM is one such model and employs multiple random-walk accumulators. In its simplest form, these accumulators are independent and their drift-rate represents the *absolute* evidence for an outcome, rather than relative. However, independence is not a requirement and adjustments can be made to allow for competition between choices[226].
Importantly, the RDM is able to replicate RT distributions and fast/slow error effects using a *single* source of variation - the stochastic random walk. Tillman et al.[227] showed that the use of absolute evidence drift-rates renders between-trial variation in both the drift-rate and start-point obsolete.

2.3.4.3 Combining learning and action-timing

In the past decade, a new field of action-selection models have arisen - the reinforcement-learning evidence accumulation models (RL-EAMs).

Classical RL observation models (such as softmax) aim to replicate choice behaviour and accuracy, with a complete indifference to the RT distribution despite the insights this can provide on the underlying causes of inter-individual differences[226,228]. However, integrating DDM as the observation model has mathematically been shown to be functionally equivalent to softmax in its capacity to replicate choice behaviour[228,229] and produces a more mechanistic interpretation of the inverse temperature parameter, β (Eq. 2.14).

Several forms of RL-EAM exist, with different combinations of RL and EAM algorithms, but they largely follow the same format:

1. A trial begins and the EAM determines the choice and RT.
2. An action is executed and the RL model is given the external feedback.
3. The RL model updates its action- or state-value estimations (e.g., Q-value).
4. These variables are then used to determine the drift-rates on the following trial.

For this thesis, the most relevant algorithm is the reinforcement-learning race diffusion model (RL-RDM), developed by Miletić et al.[226], which defines an accumulator's drift-rate through various combinations of Q-values. Miletić's RL-RDM and other RL-EAMs will be discussed further in Chapter 6, in the context of our novel 'two-drift' RL-RDM.

2.4 Action Prediction Errors

The previous three sections of this chapter have provided a background on models of action selection within the neuroscientific and computational fields. Now, our attention turns to defining the overarching issue this thesis aims to help address - what are habits?

2.4.1 Reward-based habits - an oxymoron

We have seen that a leading computational proposal of the dual-process action selection mechanism applies a model-based/model-free dichotomy (Section 2.3.3), which is regularly implemented using either an actor-critic or Q-learning RL update rule. The justification for this approach relies principally on the relative inflexibility and reduced planning capacity of the model-free system.

However, it is in fundamental opposition to the key experimental feature which first defined habits - their reward-insensitive nature. As outlined in Section 2.1, habitual actions are synonymous with S-R relationships and placed in juxtaposition with the goal-directed A-O association. As such, they contain no memory of, and are uninfluenced by, any rewards which occur. How, then, could they learn from an RPE?

This question was posed by Miller et al.[1], who proposed that this computational understanding of action selection should be replaced by a 'value-based' versus 'value-free' dichotomy. Their argument was two-fold.

First, they expressed that the evidence for dissociable neural correlates for model-based and model-free signals is inadequate[164,230,231], especially in comparison to the DS lesion studies which cleanly separated goal-directed actions from habits[54,73,81,145]. For example, studies have reported that measures of model-free learning were uncorrelated to habit strength[232] and potentially required as much effort as model-based learning[231], which is incompatible with a resource-efficient habit. If there was a 1-1 correlation between the two definitions, as proposed by Daw et al.[47], then this dissociation should be replicated in the data.

Second, they restated an argument previously given by Dezfouli and Balleine[233]; model-free RL algorithms cannot reproduce the experimental results of resistance to A-O contingency degradation tests. The model-free system learns the action-value of a given state (as in Eq. 2.9 and Eq. 2.11). Therefore, it *should* have the capacity to learn that execution of an action now leads to a negative outcome (reward delay) and that selecting to perform 'no action' is more valuable.

It is becoming increasingly clear that the previous model-based/model-free dichotomy is insufficient to describe the separation of habits and goal-directed behaviour. This is further evidenced by Daw et al.'s[17,165] later proposal of model-free goal-directed signals and the convincing model-based habitual algorithm outlined by Dezfouli and Balleine[233]. Miller et al.[1] argue that the answer lies in replacing RPEs with *action* prediction errors (APEs).

2.4.2 Defining an action prediction error

Conceptually, APEs provide an intuitively appealing solution for an S-R learning mechanism. S-R habits learn to mimic actions which have previously been executed in the current context. So, by its nature, the PE in such a system must measure how successfully the agent expected the antecedent action, rather than the outcome. For example, in an instrumental association task, the first time an action is made in response to a cue, a large APE would occur and the habit system would update to expect that particular action slightly more when the cue is next experienced.

The precise formulation for an APE-based habit system is not fixed, but some properties are fundamental. An APE should (1) display a peak following unexpected actions, (2) shift to predictive cues during learning, and (3) be entirely reward-insensitive.

To date, two key models have been developed that employ APEs. Namely, Miller et al.'s[1] value-free habit system (henceforth referred to as the 'value-free model' for ease of communication) and Bogacz's[2] DopAct model, which directly relates APE signals and their associated habits with features of the BG's dopaminergic system. Both models are able to replicate reversal learning, devaluation and contingency degradation effects, as well as providing explanations for why VI schedules promote habit expression (Section 3.4.3). They are described in more detail below.

2.4.2.1 The value-free model

The value-free model works on a trial-by-trial basis and can be defined as a combination of three algorithms, which Miller et al.[1] label as the 'habit controller', 'goal-directed

controller' and 'arbiter' (schematised in Fig. 2.5).

In a given trial, the habit system receives information about the action executed, a, and updates its expectation of future actions, H_a, in the current state, s.

$$\delta_{H_a}(t) = a(t) - H_a(t) \tag{2.16}$$

$$H_a(t+1) = H_a(t) + \alpha_h \delta_{H_a}(t) \tag{2.17}$$

where: H_a = habit strength for action, a,
a = action,
t = trial number,
δ_{H_a} = action prediction error,
α_h = habit learning rate.

In their original paper, the goal-directed system differentiated between different types of rewards, m, allowing the model to adapt under changing homeostatic requirements. For comparability, this dimensionality is removed here, and the algorithm is reduced to a single reward type.

$$\delta_{Q_a} = R(t) - Q_a(t) \tag{2.18}$$

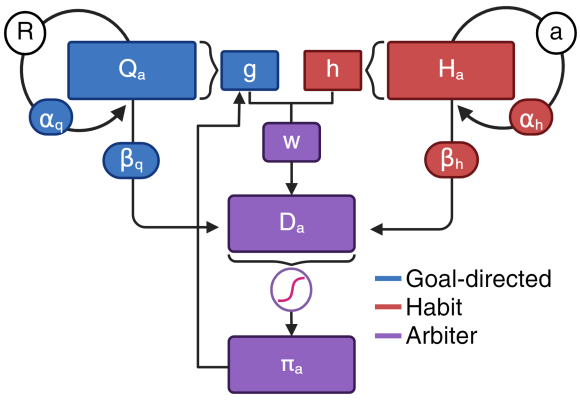


Figure 2.5: The Value-free Habit Model. A schematisation of the value-free model's dynamics within a trial. The agent receives the action (a) and reinforcement (R) that occurred on the previous trial. This information is used to update the two controllers (red and blue, for habit and goal-directed, respectively) via an APE and RPE, modified by a learning rate, α. The variables, Q_a and H_a, are then communicated to the arbiter (purple). The arbiter combines these values according to two constant scaling terms, β_q and β_h and a 'weight' variable, w, which measures the variance for both Q_a and H_a. The resulting D_a values are passed through a softmax function to produce the policies, π_a, for that trial. Finally, π_a is communicated to the goal-directed controller ahead of next trial.

$$Q_a(t+1) = Q_a(t) + \alpha_q \delta_{Q_a}(t) \tag{2.19}$$

where: Q_a = action value,
R = reinforcement magnitude,
δ_{Q_a} = reward prediction error,
α_q = learning rate for Q.

To determine the relative contributions of the two controllers, the arbiter explicitly computes A-O contingency, g, and 'habitisation', h. These are measures of the variance within Q_a and H_a values, respectively.

$$g(t) = \sqrt{\sum_a \pi_a \left(Q_a(t) - \sum_{a'} \pi_{a'} Q_{a'}(t) \right)^2} \tag{2.20}$$

$$h(t) = \sqrt{\sum_a \left(H_a(t) - \left\langle H_a(t) \right\rangle \right)^2} \tag{2.21}$$

$$w(t) = \frac{1}{1 + \exp(w_g g(t) - w_h h(t) - w_0)} \tag{2.22}$$

where: g = A-O contingency,
h = habitisation,
w = mixing weight,
$w_{g,h,0}$ = constant weight parameters for g, h, and a fixed bias,
π_a = policy for action, a.

Finally, the arbiter calculates an overall 'drive', D_a, towards a given action, which is then converted into a policy, π_a, via a softmax function.

$$D_a(t) = w(t)\left(\beta_h H_a(t)\right) + \left(1 - w(t)\right)\left(\beta_g Q_a(t)\right) \tag{2.23}$$

$$\pi_a(t) = \frac{e^{D_a(t)}}{\sum_{a'} e^{D_{a'}}(t)} \tag{2.24}$$

where: D_a = drive towards action, a,
$\beta_{g,h}$ = scaling parameters for goal-directed and habit controllers.

The key features of this algorithm are the model-based nature of the goal-directed controller, the explicitly encoded arbiter and the use of variance (i.e., uncertainty) to determine the relative weighting of the two controllers.
Miller et al.'s[1] paper later simulated free-operant tasks by extending the model to represent actions as continuous 'rates' (e.g., of lever presses), which allows it to maintain its trial-by-trial nature.

2.4.2.2 The DopAct model

Soon after, Bogacz[2] presented an alternative framework, where an APE-based habit is combined with Friston's 'active inference' algorithm[2,3,4], and directly related it to biologically-plausible computations within the BG.

The DopAct model is formed around three axioms:

1. PEs are produced constantly, both when learning from outcomes and during action planning. The model parameters update to minimise these errors.
2. Striatal dopaminergic signals are heterogeneous and the habit system learns from an APE which is encoded by distinct dopamine populations from the value-based RPE.
3. Value, goal-directed choices and habits can be mapped to the ascending spiral theory of the BG (schematised in Fig. 2.6). DopAct only explicitly encodes the latter two and assumes value to be calculated accurately according to the homeostatic needs of the agent.

Many have now proposed that the strict divisions across the striatum are manufactured and that the BG can instead be conceptualised as a gradual gradient of function, created by small parameter adaptations to a shared computational unit[1,27]. Fig. 2.6 shows how DopAct develops upon this argument when introducing habits. There is no computational difference in connectivity for the three striatal loops. Rather, the three variables arise from the disparate information contained in their inputs.

Mathematically, by applying the active inference algorithm and selecting actions through Bayesian inference[2], DopAct is able to work explicitly with probability functions and

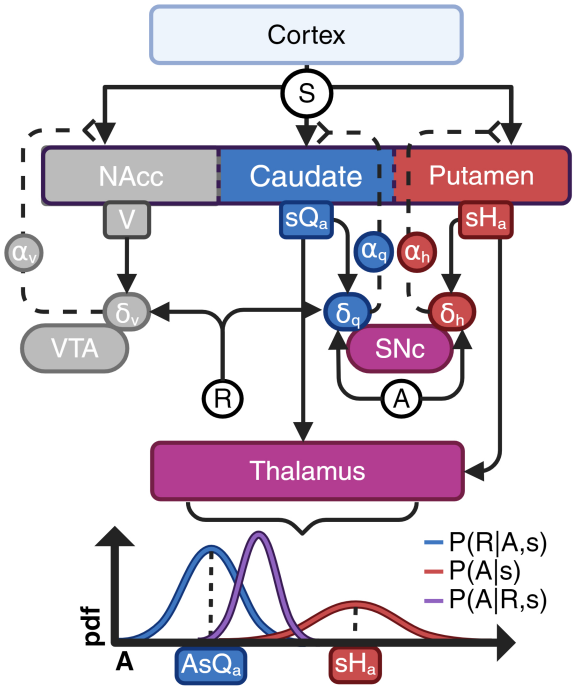


Figure 2.6: DopAct in the Basal Ganglia.
A flowchart schematisation of DopAct overlaid on BG anatomy of Fig. 2.2.
(Top) The value, V, (grey), goal-directed, Q_a, (blue), and habit, H_a, (red) systems are contained with ascending striatal spirals from the NAcc to putamen. The dopaminergic nuclei receive (1) input from the associated striatum, modulated by the cortical 'state' signal and (2) information about relevant external events (reinforcement, R, and action intensity, A). PEs are produced and dopamine signals modulate the corresponding corticostriatal synapses. The thalamus receives striatal input and applies these to select an action according to Bayesian inference.
(Bottom) A representation of the probabilistic structure for a single DopAct trial. The goal-directed $P(R|A,s)$ and habitual $P(A|s)$ distributions have mean values AsQ_a and sH_a respectively. These are convolved to produce the $P(A|R,s)$, whose maximal value determines action intensity, A.

free-energy theory.

The inference dynamics and the derivations which produce the below algorithm, well-summarised in the original paper[2], are not included here as our interest lies more specifically in the application of APEs in the learning rules.

To begin, a Bayesian rule for action selection is defined, which executes the associated action with the most probable action intensity: $P(A|R,s)$. This is then converted into an objective loss function, F, that can be minimised through parameter updates. Note that in the following discussion, this algorithm determines the optimal action *intensity* of a given action, $A_a(t)$, rather than choosing between multiple actions. DopAct can be extended to consider the latter scenario.

$$P(A|R,s) = \frac{P(R|A,s)P(A|s)}{P(R|s)} \tag{2.25}$$

$$F = \ln\Big(P(R|A,s)P(A|s)\Big) \tag{2.26}$$

where: R = reward,

A = action intensity,

s = state,

F = free energy function, to be minimised.

These probabilities are mapped to the three BG loops:

1. The value of a given state is the measure of how much reward is potentially available, $P(R|s)$, in the VS.
2. A goal-directed system calculates the predicted value which can be achieved by a given action intensity in that state, $P(R|A,s)$, in the DMS.
3. A value-free habit variable learns to predict which action intensity is likely in a state according to past experience, $P(A|s)$, in the DLS.

The choice probabilities are assumed to follow Gaussian distributions:

$$f(x;\mu,\Sigma) = X \sim N(\mu,\Sigma) \tag{2.27}$$

$$P(R|A_a,s) = f(R;sA_aQ_a,\Sigma_g), \quad P(A)a|s) = f(A_a;sH_a,\Sigma_h) \tag{2.28}$$

where: Q_a = goal-directed strength,

H_a = habit strength,

s = stimulus intensity,

$\Sigma_{g,h}$ = distribution variance (uncertainty).

In the full DopAct algorithm, the variance terms, Σ_g and Σ_h, scale the relative contributions of Q_a and H_a. This is simplified below under an assumption of $\Sigma_g = \Sigma_h = 1$, which allows F to be rewritten in straightforward prediction error terms.

$$F = -\frac{1}{2}(\delta_{Q_a}^2 + \delta_{H_a}^2) \tag{2.29}$$

$$\delta_{Q_a} = R - sA_aQ_a, \quad \delta_{H_a} = A_a - sH_a \tag{2.30}$$

where: δ_{Q_a} = reward prediction error,

δ_{H_a} = action prediction error.

Deriving the gradient of F as a function of Q_a and H_a gives the following update rules.

$$Q_a(t+1) = Q_a(t) + \alpha_q \delta_{Q_a}(t) A_a(t) s(t) \tag{2.31}$$

$$H_a(t+1) = H_a(t) + \alpha_h \delta_{H_a}(t) s(t) \tag{2.32}$$

where: α_q = goal-directed learning rate,

α_h = habit learning rate.

These update rules are near-equivalent to Eq. 2.17 and Eq. 2.19, save for an additional factor - learning is now scaled by both the action and stimulus intensity on a given trial. Under an assumption that $A_a(t)$ and $s(t)$ represent binary vectors, the value-free model's update rules are reproduced with an additional integrated gating function such that the values only update when their distribution's given variables are present.

In summary, DopAct and the value-free model share both a trial-by-trial nature and the use of variance to weight the two systems. In using Bayesian methods, the action selection and gating of learning are inherent features of DopAct's biologically-plausible dynamics. As with the value-free model, interpreting action intensity as a rate measure allows for the variable SOR effects to be understood.

The current two theoretical models for APE-based habits provide elegant solutions to simulate trial-by-trial experiments. However, further expansions on these models are required to answer several key questions.

2.4.3 Current questions

As stated in the introduction, this thesis addresses the following questions;

1. Can habit learning be generalised to continuous and scalar actions?
2. Can evidence of action prediction errors be found in striatal dopamine?
3. Can the process by which habits affect choices and RTs be mechanistically described?
4. Can the impact of habits in human behavioural data be quantified?

The first two questions are addressed in Chapters 3 and 4, and required the explicit modelling of within-trial learning, analogous to the difference between TD-RL and the trial-by-trial models described by Eq. 2.9 and Eq. 2.11. By developing a learning algorithm which closely aligns APE theory with the current mechanistic and biological understanding of dopamine signals, we are able to distinguish the effects of action 'intensity' and action 'rate' (e.g., a mouse may run very quickly to a button and press it with a lot of force, but only do it once or twice in a trial) and it becomes possible to interrogate whether APEs exist in neural data.

Further, within-trial processes also affect behaviour beyond choice. As discussed in Section 2.3.4, accounting for RT distributions reveals a wealth of information regarding the underlying processes during speeded decision making. DopAct and the value-free model use Bayesian inference and softmax as observation rules, respectively, which do not account for these details.

To answer the latter two questions, we developed an alternative observation function in the form of an RL-EAM that can account for the proposed different temporal dynamics of habit and goal-directed systems in a multi-alternate task. Incorporating this added dimensionality is particularly attractive given that imposing limits on RTs has proven to be an effective method to unmask habits in human participants (Section 2.1.3.4).

2.5 Summary

This chapter has presented an overview of the literature and research underpinning computational models of habit formation. To begin, Section 2.1 described the behavioural and neuroscientific studies which first defined habits as stimulus-response behaviours within a two-process action selection framework and linked them to the dorsal striatum. Section 2.2 expanded on the role of dopamine as a heterogeneous neural prediction error signal, involved in both reward learning and motor control. Next, Section 2.3 outlined the key mathematical models in the field, including a discussion of reinforcement learning algorithms and the benefits of different observation models. Section 2.4 then indicated how the above literature has led to the proposal of value-free, APE-based habit formation

in the striatum, before summarising the value-free[1] and DopAct[2] models.

Finally, the four research questions interrogated over the course of this thesis were restated, in the context of the current literature.

3

Temporal Difference Action Learning

Contents

3.1 Introduction

To interrogate whether striatal dopaminergic signals truly encode APEs, we must first form testable predictions regarding the expected shape and temporal dynamics of these neurotransmitters.

As described in Section 2.4.2, Miller et al.[1] and Bogacz[2] have provided clear rules on how an APE should behave: (1) spiking after an unexpected action, (2) decreasing in size as the action becomes expected, and (3) dipping below tonic levels when the expected action does not occur (see Section 2.4). However, these previous models function on a trial-by-trial basis, wherein an APE is only calculated once per trial for a static action value.

There are two key limitations to this approach.

First, this 'trial-chunked' APE cannot describe (nor provide accurate predictions for) the temporal dynamics of dopamine during the trial. Indeed, to the best of our knowledge, the algorithm we developed in the course of this project is the first mechanistic model that links the value-free APE concept to current neuroscientific understandings of the temporal dopamine dynamics in the BG. The development of such a model is crucial in determining the existence of APEs within striatal dopaminergic signals and to distinguish them from RPEs.

Second, modelling actions with a single PE and action value per 'trial' is an unrealistic portrayal of how animals actually engage with the world. Our lives do not consist of a sequence of clearly defined trials and our actions are neither instantaneous nor constant - we continuously adapt to new information and each movement is made up of a range of 'intensities', whether that be strength, speed or something else. For example, in an arm-reach task, the speed of the limb will ramp up to a peak velocity before slowing again, and the peak velocity will be influenced by the width of the target[235]. In modelling such an experiment, the impact of this variable speed on the expected APE must be considered.

In this chapter, our novel temporal-difference action learning (TD-AL) algorithm is defined, which extends the original APE proposals to evolve across continuous time and predict *when* a continuous action will be made, relative to the predictive external cues. Simulations of its behaviour are later compared to current established models of dopaminergic PEs.

TD-AL is a biologically-plausible generalisation of the mechanistic TD-RL model[106,176,236] and learns to predict the total expected (future) action in a given context, based on previous experience.

Importantly, this learning model demonstrates how a single computational algorithm can underlie learning of both reward and habits across diverse striatal regions.

3.2 The Mathematical Model

TD-AL is an extension of Sutton and Barto's[172,174] TD(λ)-RL model, described in Section 2.3.2. This next section outlines the mathematical adaptations we implemented to introduce realistic neural representations of an APE and the theoretical assumptions that motivated these choices.

3.2.1 Microstimuli - a new eligibility trace

To improve the biological-plausibility of the temporal APE model, we applied an extension that was originally designed for TD-RL to allow the timings of rewards to be learnt in a more realistic manner.

Since the development of TD(λ)-RL, several alternate formulations of eligibility traces have been introduced. One such algorithm, the *microstimulus model*, developed by Ludvig et al.[176], produces a temporal trace which better replicates the magnitude of depression in dopamine firing upon reward omission[67,176], including the absence of dopamine depression when a reward is received earlier than expected[237].

This model proposes that when a stimulus occurs in the world, a series of m microstimuli are produced, which are represented by Gaussian basis functions that track the decaying stimulus trace over time (Fig. 3.1). These microstimuli reformulate the independent elements of the CSC TD-RL binary vector, x_{ij}, (Section 2.3.2.1) as overlapping 'temporal receptive fields'.

Mathematically, a decaying memory trace, y_i, is produced for each event - similar in formulation to ε_i for TD(λ)-RL (Eq. 2.5):

$$\frac{dy_i}{dt} = -\lambda y_i, \quad y_1 = 1 \tag{3.1}$$

where: λ = memory decay parameter,

y = memory trace,

t = timestep,

i = stimulus.

This trace is convolved with a uniformly distributed sequence of Gaussian curves (Eq. 3.2) and results in the final microstimuli (Eq. 3.3).

$$f(y,\mu,\sigma) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{3.2}$$

$$x_{ij}(t) = f\left(y_i(t), \frac{j}{m}, \sigma_m\right) y_i(t) \tag{3.3}$$

where: μ = Gaussian centre,

σ = width of each curve,

i = stimulus,

j = microstimulus,

m = total number of microstimuli for any stimulus, i.

Fig. 3.1A and Fig. 3.1B demonstrate the dynamics of the microstimuli produced for a single event, i, and a cue-action-reward series, respectively. Note that this algorithm produces

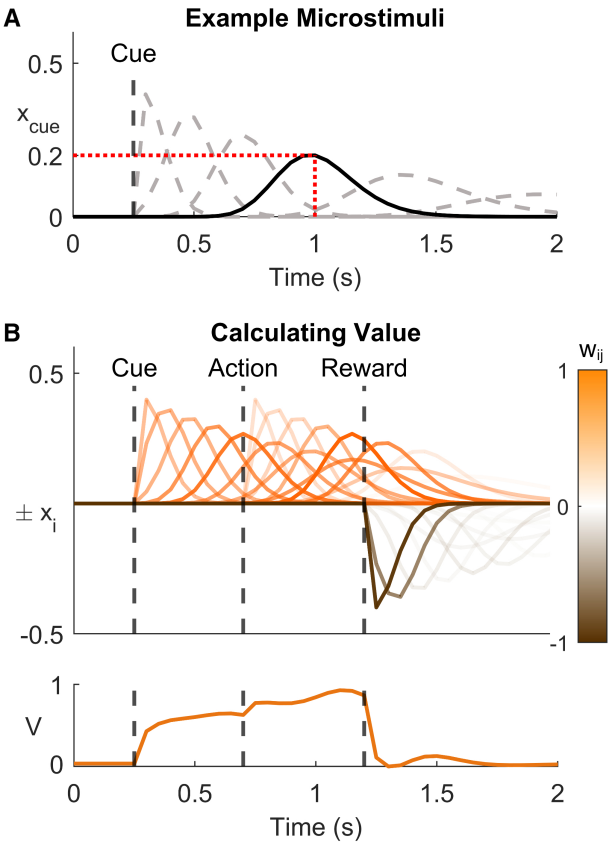


Figure 3.1: Understanding Microstimuli.
A: An example of microstimuli dynamics following a single 'cue' event. Parameters are taken from Table 3.1, with only 6 microstimuli shown here (dashed lines). The peak firing rate is uniformly distributed and decays across time. The activity of microstimulus $x_{\text{cue},4}$ (solid line) at 1 second after the event is 0.2, as demonstrated by the red line.
B: Together these two panels portrays how the value function is produced by the microstimuli.
(Top) An example simulation of a single trial, with three events: cue, action and reward. Each event produces its own microstimuli. The opacity is adjusted to their relative weighting after TD-RL is trained on 200 trials ($\alpha_v = 0.1, \gamma_v = 0.97$). Microstimuli with negative weights are plotted below the zero line for additional clarity.
(Bottom) The value function resulting from the same simulation.

curves with uniformly distributed peaks (as $\mu = j/m$) and increasing widths, which replicates the decreasing temporal accuracy as the delay increases.

As with the CSC model, the value, V, at any given time is calculated by the summed activity of all microstimuli, x_{ij}, modulated by their corresponding weights, $w_{V,ij}$.

$$V(t) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}(t) w_{V,ij}(t) \tag{3.4}$$

where: V = expected future value,
n = total number of events,
$w_{V,ij}$ = value system's weight for the j^{th} microstimulus from event, i.

As such, it is once again, the *weights* that update at each timestep, rather than a conceptual variable V, and our update rule is changed from Eq. 2.6 to Eq. 3.5. The equation to calculate δ_V is unchanged (Eq. 2.2).

$$w_{V,ij}(t+1) = w_{V,ij}(t) + \alpha_v \delta_V x_{ij}(t) \tag{3.5}$$

where: α_v = learning rate for V,
δ_V = reward prediction error.

Fig. 3.1B helps to provide an intuition for how Eq. 3.4 behaves. Each microstimulus' weight, and thus, its contribution to the final value, is represented by its relative opacity. Their weighted sum results in the 'value' curve shown. Consecutive predictive cues provide more microstimuli with steadily increasing weights and the model's expectation of future reward increases. Note that the reward microstimuli are negatively weighted, as reception of the reinforcer indicates that no future reward will arrive and they counterbalance the positive weighting of the other microstimuli to return V to 0.

In their initial formulation, Ludvig et al.[175,176] included an eligibility trace for each x_{ij}. However, we found that the impact of including ε_i was minimal given that x_{ij} is already spread across time and is sufficient to resolve the credit assignment problem (Section 3.3.1). Consequently, it was not included in this thesis' algorithms. In real terms, this

represents an assumption that synaptic plasticity (weight updates) requires pre-synaptic activity (the presence of the respective microstimulus).

The microstimulus model offers several advantages. Primarily, it aligns with our current understanding of striatal connectivity and is more mechanistic than the standard eligibility trace. For example, the microstimuli could be implemented by cortical inputs to the striatum or striatal sub-populations with different temporal dynamics following salient stimuli. Such time-dependent firing patterns have been observed in the BG's medium spiny neurons[238] and cortical oscillators have previously been proposed as the source of a striatal 'internal clock'[239,240]. Further, dopamine influences the strength of synaptic connectivity via Hebbian learning (Section 2.2.4), which can be concretely correlated to these individual weights. Finally, Ludvig[175,176] showed that microstimuli did indeed reproduce more realistic temporal dynamics than classic TD(λ)-RL, across a variety of different experimental simulations.
For the purposes of this thesis, parameter values are fixed to those used by Ludvig et al.[175] and are applied for all simulations in Chapters 3 and 4 (Table 3.1).

Parameter	Symbol	Value	Equation
Learning Rate	α_v	0-1	3.5
Discount Factor	γ_v	0-1	2.2
Number of Microstimuli	m	12	3.3
Decay Rate	λ	0.985	3.1
Microstimulus Width	σ_m	0.08	3.3

Table 3.1: TD-RL parameter values, including microstimuli.

Unless otherwise stated, all future mentions of 'TD-RL' in this thesis will specifically refer to the microstimulus algorithm.

3.2.2 Introducing action learning

The APE proposal[2] fundamentally presumes that habit formation is not unique when compared to a model-free goal-directed TD-RL system; habits are pure action predictors which learn from the dopamine signals that represent an APE. As such, the key axioms

on which the TD-AL algorithm is founded are that (1) the BG is formed of parallel computational units and (2) the information a given system represents is determined by its input.

3.2.2.1 Theoretical assumptions

As the organisational structure of the striatum utilises loops with parallel connectivity[59,61], many have posited that these loops should, logically, manipulate the contained data in an identical manner[2,186,241,242]. This formulation proposes that the information being predicted within a given region of the striatum is determined entirely by the outcome recorded by the corresponding PE signal, which can vary along a gradient from pure reward (VS) to pure action (DLS, or the TS).

Accordingly, TD-AL assumes that dopamine performs the same computational role throughout the BG. More precisely, as this project aims to produce a biologically-plausible model, we maintain that (1) phasic dopamine dynamics contain a neural PE signal which determines the direction and magnitude of synaptic plasticity changes in the striatum and (2) this PE can be produced by heterogeneous events in the environment.

3.2.2.2 Learning to predict actions

To represent an action-based value-free habit, TD-AL learns to estimate the future action intensity, H_a (Eq. 3.6), by minimising PEs. Specifically, for each potential action, a, the 'striatal' activity predicts future expected action intensity, $A_a(t)$, given the current state. As discussed in Section 3.3, A is continuous and can last (i.e., have a non-zero value) for several timesteps.

$$H_a(t)^* = E\left[\sum_{k=1}^{\infty} \gamma_h^{k-1} A_a(t+k)\right] \tag{3.6}$$

where: H_a = expected future intensity of action, a,

γ_h = action discount factor,

$A_a(t)$ = action intensity of action, a, at time, t.

Thus, the computational structure of the PE itself is unchanged, the outcome variable is

simply altered from reward to action intensity. At each timestep, the APE is produced:

$$\delta_{H_a}(t) = A_a(t) + \gamma_h H_a(t) - H_a(t-1) \tag{3.7}$$

where: δ_{H_a} = action prediction error.

One critical new feature of our model is its ability to predict *when* an action usually occurs after the associated stimulus has been perceived. As discussed in Section 2.1.2, the strength and behaviour of habits are time-dependent[15] and variable SORs are more likely to result in delayed extinction behaviour[42]. Therefore, the ability to interrogate how temporal delays influence the strength of habit versus goal-directed actions is advantageous. To achieve this, TD-AL applies the microstimulus formulation described above (Section 3.2.1, Table 3.1) to model the expected duration between cue and action. For *every* event, i, that occurs (e.g., cues, actions and rewards), m microstimuli are produced. The total habit value, H_a, at any given time is composed of the current microstimulus activity, x_{ij}, modulated by their relative weights, $w_{H_a,ij}$. So Eq. 3.4 and Eq. 3.5 become:

$$H_a(t) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}(t) w_{H_a,ij}(t) \tag{3.8}$$

$$w_{H_a,ij}(t+1) = w_{H_a,ij}(t) + \alpha_h \delta_{H_a} x_{ij}(t) \tag{3.9}$$

where: $w_{H_a,ij}$ = weight of H_a for the j^{th} microstimulus produced by event, i,
α_h = habit learning rate,
n = total number of events experienced.

Note that the microstimuli activity levels, x_{ij}, are shared across all habit and value variables as they represent the external state/context. Instead, it is the weights that contain the information learnt about the measured outcome (i.e., the specific action/reward). Again, all microstimulus parameters are fixed to the values listed in Table 3.1.

3.2.3 Making choices

In the context of habits and choice, all models require a clear definition of how an action is mathematically represented. For the purposes of this thesis, and the work presented here and in Chapter 4, a choice is defined as the action *plan* selected; for example, classifying a choice as pressing the 'left' or 'right' lever, rather than total relative position or individual changes in muscle groups. For an action in near-continuous time, this will be represented by a single continuous curve which ramps up and down as movement is initiated and ended.

This simplification is based on, and validated by, prior work into goal-directed action modelling[1,2,243,244]. It assumes that the BG are responsible for filtering between plans presented by the cortex, rather than controlling the mechanical execution of the action - a process which can occur downstream and is more often attributed to the cerebellum[183,245] (another site of much neuroscientific RL research[246–248]).

The wider implications of this assumption are further explored in Section 7.3.1 alongside a discussion of how this may be extended in future work.

Critically, TD-AL does not describe how value and habits influence the selection of actions. Rather, it demonstrates how habits can evolve by learning to copy the choices previously made by the goal-directed system.

The selection of an appropriate observation model can in many ways be orthogonal to that used for learning (Section 2.3.4), though they may influence each other. Rather, the most applicable observation model is contingent on the behaviour it aims to replicate. For example, if an experiment measures RT as its dependent variable, an EAM will be more appropriate than the softmax choice function.

The advantage of TD-AL over previous APE models is its ability to work over (near-)continuous time. As such, it is likely that future applications of this model will be interested in replicating within-trial behaviours - be that RTs, action intensities or other. Thus, TD-AL lends itself to the RL-EAM family of observation functions (Section 2.3.4). It would be particularly interesting in future research to explore how the influence of predictive cues on the temporal dynamics of H_a versus V could influence error rates and

RTs using these more mechanistic observation models.

One natural solution may be to include a time-variable RL-EAM drift-rate as a function of H_a and V. However, as will be explored in Chapter 5, it becomes exceedingly computationally expensive to apply this information in continuous time and it is largely intractable to fit such a model to real data. Chapter 7 discusses these potential algorithms as future avenues of work with TD-AL, but this thesis does not apply a continuous-time observation model. Instead, for the simulations presented in this chapter (Section 3.3), all actions are artificially represented by a fixed Gaussian curve, with a duration of 0.5 seconds and a peak value of 0.65.

3.3 Classical Experiment Simulations

This section illustrates how TD-AL behaves through the simulation of two classic experiments in which the activity of dopamine neurons projecting to the VS was recorded[67,249] (Section 2.1.3). Five models, whose key characteristics are outlined in Table 3.2, are compared to interrogate three factors:

1. How does the APE differ from the better-known RPE (Model A vs. others)?
2. How does the APE change if we consider the temporal dynamics of the action, rather than the initiation alone (Model B vs. Model C vs. Model D)?
3. What is the influence of the γ_h term on both the APE and H_a variable (Model D vs. Model E)?

Label	Algorithm	Outcome	α	γ
A	TD-RL	Reward	$\alpha_v = 0.1$	$\gamma_v = 0.97$
B	TD-AL	Action initiation	$\alpha_h = 0.1$	$\gamma_h = 0.97$
C	TD-AL	Action presence	$\alpha_h = 0.1$	$\gamma_h = 0.97$
D	TD-AL	Action intensity	$\alpha_h = 0.1$	$\gamma_h = 0.97$
E	TD-AL	Action intensity	$\alpha_h = 0.025$	$\gamma_h = 0$

Table 3.2: Key characteristics of the five models simulated in Section 3.3.

Note that α_h has been decreased for Model E to keep the size of weight updates easily comparable between models.

3.3.1 Instrumental association tasks

As discussed in Section 2.1.2, instrumental association experiments were key in developing TD-RL and S-R theory[15,29,250]. In their simplest form, an animal learns to perform an action in response to a neutral stimulus to gain a reward by repeatedly experiencing a cue-action-reward sequence (as in Fig. 3.1B) and developing an estimation of both S-R and A-O contiguities[15,34,251].

In the following simulations, all models (A-E) experienced 200 instances of the cue-action-reward sequence, with a delay of 10dt between each event (1dt = 0.05s) and a 7.5s inter-trial interval (ITI). The results of their learning are shown in Fig. 3.2.

3.3.1.1 Models A and B - Confirming RPE and APE predictions

Model A behaves exactly as predicted of a TD-RL model (see Section 2.3.2.1 and Fig. 2.3). The RPE transfers from unexpected reward to earlier predictive cues before converging on the earliest and V learns to increase with each consecutive cue until the expected reward is received. The temporary response to action initiation results from the explicit assumption that the action itself triggers its own set of microstimuli and therefore behaves as a predictive cue.

Model B uses the TD-AL algorithm to predict when an action will begin. Information about the structure of the action itself has not been provided. This is the closest to the formulation from Miller et al.[1], as the presence or absence of action is indicated by a binary vector. Fig. 3.2B shows all the features predicted for an APE in Section 2.4: (1) it is movement-locked to the unexpected action, (2) the APE shifts to the predictive cue as learning occurs, and (3) it is reward-insensitive.

3.3.1.2 Model C - Action on-off

Model C presents an intermediate between Models B and D in that it provides a measure of action duration, but receives no information about the intensity or dynamic changes of the movement.

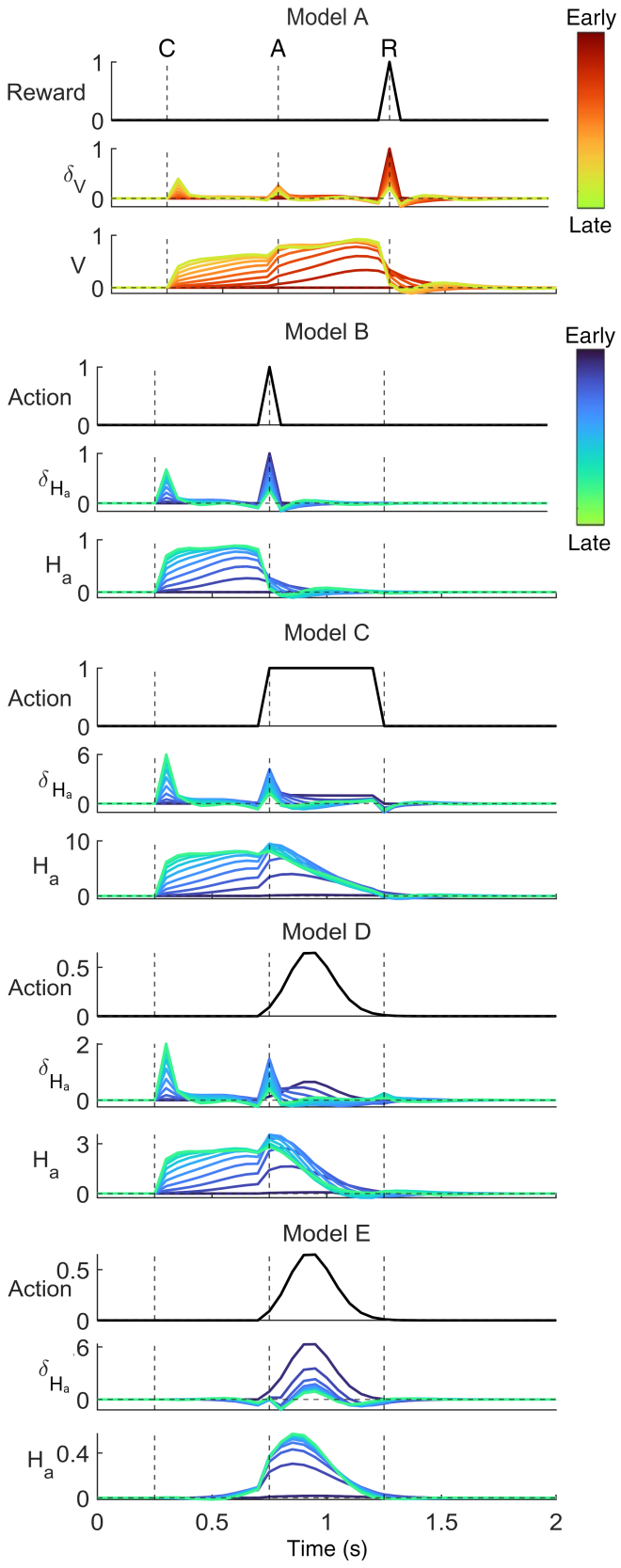


Figure 3.2: Instrumental Association Simulations.
The evolution of five models learning a simulated instrumental association task. Each model experienced 200 trials, where cue, action initiation and reward were 0.5s apart, with a 7.5s ITI. (Top) The measured outcome value. (Middle) The prediction error from early to late trials. (Bottom) The estimated outcome learnt by the model from early to late trials.
Model A: The results of a TD-RL model ($\alpha_v = 0.1, \gamma_v = 0.97$) from early trials (red) to late in learning (yellow).
Model B: The results of a TD-AL model which learns to predict action initiation ($\alpha_h = 0.1, \gamma_h = 0.97$) from early trials (blue) to late in learning (green). Both δ_{H_a} and H_a are insensitive to reward.
Model C: As **B**, but with a constant 'action presence' measure represented by a binary vector, such that $A = 1$ when the action is present.
Model D: As **C**, but learning a continuous action intensity represented by a Gaussian curve.
Model E: As **D**, but learning instantaneous action intensity rather than future action ($\alpha_h = 0.01, \gamma_h = 0$).

The resulting traces differ from B in three ways: (1) the behaviour of H_a after action initiation, (2) the pattern of δ_{H_a} following action initiation during early trials, and (3) the final magnitude of the cue-locked δ_{H_a} peak. All three effects can be explained by the difference in the sum of the action variable experienced by the model in each trial.

Model B learns to expect a single peak of $A_a(t) = 1$ at action initiation. Therefore, once the action has been initiated, H_a does not expect any future actions to begin and returns to baseline. In contrast, H_a for Model C needs to predict a temporally discounted integration of the remaining action which will occur during the trial. Until the action is completed, $A_a(t)$ has a non-zero value and H_a continues to be positive.

Similarly, during the early trials, the δ_{H_a} dynamics are locked to the duration for which $A_a(t) \neq 0$. For B, this means that once the action is initiated there is no prediction error since $A_a(t) = 0$ for all future moments, whereas for Model C, δ_{H_a} remains slightly elevated until action termination.

In late trials, the model learns to successfully predict the duration of the action and the *entire* APE transfers to the preceding cue. Thus, at convergence, the magnitude of the cue-locked peak is determined by the total sum of A_a predicted by the associated microstimulus. Model B's δ_{H_a} simply tends towards 1 at convergence, while the APE for C converges at a discounted integration of total future expected action presence.

Interestingly, in attempting to replicate the abrupt beginning and end of an 'action' using curved Gaussian basis functions, action termination results in a temporary dip in δ_{H_a}. Similar 'action-off' dips have been reported in SNc recordings, such that dopamine levels peak at action initiation and dip below baseline at action termination[130,132]. In simulating an instrumental association task, the δ_{H_a} peak transfers to the predictive cue across training. However, theoretically, the closest predictive cue for self-initiated actions would be the initiation of the action itself. As such, Model C should reproduce these dopaminergic action-on/off dynamics. Indeed, as shown in Fig. 3.2C, δ_{H_a} is temporarily responsive to action initiation.

The behaviour of TD-AL across self-initiated actions is explored further below (Section 3.3.2).

3.3.1.3 Model D - Considering action intensity

Model D introduces the continuous scalar action discussed in Section 3.2.3. Its dynamics closely resemble those of Model C, such that: (1) H_a remains positive during execution of the action and calculates a temporally discounted integration of the expected future action intensity, (2) the predictive δ_{H_a} peaks are larger in magnitude than for Model B, and (3) the pattern of δ_{H_a} following action initiation is influenced by the structure of the action itself.

However, where Model C simply estimates the duration of the action, δ_{H_a} in Model D initially copies the dynamic structure of the action intensity itself until H_a's predictions improve. Further, while it is difficult to visually differentiate the H_a dynamics during action, Model D does replicate the Gaussian structure with sigmoidal decay curve, whereas H_a for Model C approximates a linear decrease until action termination as best it can using Gaussian basis functions.

It is also worth noting that, during the final trials, Model D loses the negative prediction error at action termination, instead resulting in a minor (relatively insignificant) peak produced by a temporarily negative H_a. Unlike the consistent dip produced by Model C, this increase is an inconsistent artifact produced by replicating the continuous action with the microstimuli which causes H_a to oscillate about 0 towards the end of the action.

3.3.1.4 Model E - Interrogating the discount factor, γ

One parameter worth investigating further in our model is the discount factor, γ. In TD-RL, γ_v determines the degree to which the model expects and predicts future rewards. However, when considering actions, this parameter can arguably be considered superfluous. It is unclear whether it would be more beneficial for the striatum to predict which action should be made *right now*, or to prepare for the amount of action, and thus future effort, that is usually executed in its current context.

Given that a key theoretical proposal for the function of the basal ganglia includes a VS-DLS gradient of representations, from reward through goal-directed selection to habits[55] across a common computational unit, TD-AL makes no claim as to the likely

value of γ_h. The degree to which future outcomes are discounted could vary across striatal regions, and so, the inclusion or exclusion of γ_h are not incompatible outcomes - both versions of the learning algorithm could be utilised in the striatum to different ends. Making a prior assumption of γ_h will only be necessary if H_a is used in an observation function, as it influences our interpretation of the variable.

Comparing Models D and E (Fig. 3.2D and 3.2E) allows us to explore how a γ_h value of 0.97 or 0 will influence the evolution and final structure of H_a and δ_{H_a}. Two key features develop.

First, as shown in Fig. 3.2, H_a behaves similarly to V in TD-RL for Models C and D. Action expectation increases when the predictive cue occurs and the expected amount of future action intensity is calculated once the action has begun. Model D demonstrates this with a gradual decrease in H_a after action initiation, which integrates to the total amount of movement that has not yet occurred, rather than the sharp drop visible for Model B. In contrast, Fig. 3.2E indicates that H_a learns to predict the immediate continuous structure of the action (here a Gaussian curve), as well as it can be approximated from a Gaussian microstimulus basis function.

Second, Model E's δ_{H_a} does not show the transfer of dopamine PEs to the earliest predictive cue – despite this being a key characteristic demonstrated in TD-RL (Section 2.3.2). Instead, the APE converges to 0. However, this does not mean that experiencing the cue has no influence on action expectation, as proven by the increase in H_a just prior to initiation. The cue microstimuli have non-zero weights and the presence of these microstimuli causes an expectation of action to build. As such, the cue is still vital to the prediction of action timing, but the microstimuli are not linked to an instantaneous action change, and so, no PE occurs.

Intriguingly, this formulation therefore implies that, when an S-R relationship has been learnt, Model E predicts that no associated dopaminergic signals would be present or detectable. As such, cue-locked dopamine activity is no longer a requirement to determine that RL is occurring and so, future research may need to rely on omission trials to detect a γ_h-less APE.

3.3.2 Prior training and weight initialisation

In the above simulations, all the microstimuli weights were initiated at 0 which produces agents that begin as entirely naive to all events. However, it could alternatively be assumed that, since animals will have previously experienced the executed movement in other contexts, the action-based microstimuli weights would have already learnt and converged, especially if this action has the same temporal/intensity profile across many settings.

To explore the influence of this assumption on APE dynamics, the instrumental association simulation was repeated with another iteration of Model D, but only after the agent has experienced 200 trials of the action alone, in the absence of cues and rewards. Model D was selected for this and future simulations (Section 3.3.3) as it provides a more realistic and detailed approximation of action-learning across continuous time than B or C. Further, Model E is a special formulation of D with γ_h fixed to 0, so we elected to interrogate the more general model.

Fig. 3.3 shows the APE evolution produced by Model D over two stages.

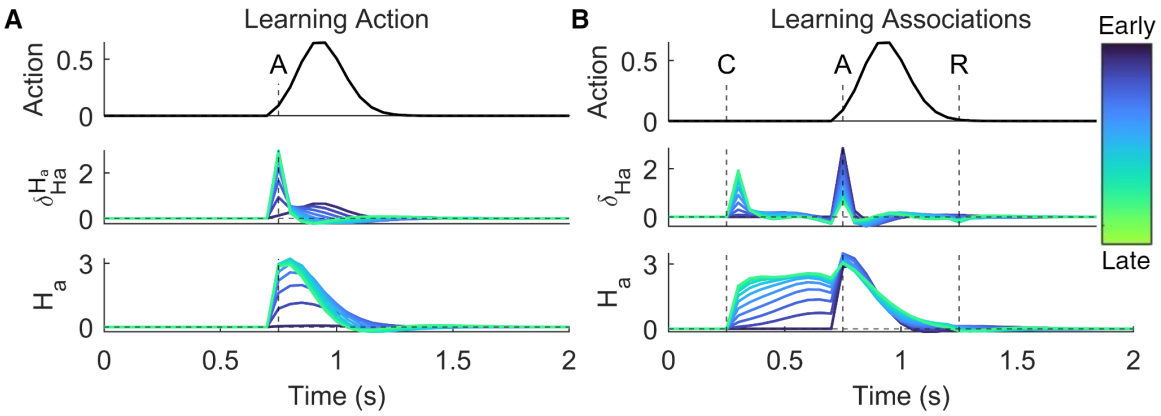


Figure 3.3: Comparing Continuous Actions.
An iteration of Model D ($\alpha_h = 0.1, \gamma_h = 0.97$) in a two-stage learning environment. (Top) The measured outcome value. (Middle) The prediction error from early to late trials. (Bottom) The estimated outcome learnt by the model from early to late trials.
A: The first 200 trials, with action as the only event that occurred. The model learns to predict the total remaining amount of action from the microstimuli produced at its initiation.
B: Simulation of an instrumental association task (as in Section 3.3.1) run using the iteration of Model D from Fig. 3.3A. H_a evolves to the same structure as Fig. 3.2C, while δ_{H_a} more closely resembles Fig. 3.2B.

1. **Prior learning:** In the first 200 trials (Fig. 3.3A), the action-associated microstimulus weights learn that the initiation itself is predictive of future movement (the δ_{H_a} peak converges to the start of the action) and H_a at every timestep represents an integration of the future action intensity remaining (modulated by γ_h).
2. **Cued Actions:** Strikingly, as a result of this early learning, δ_{H_a} is nearly indistinguishable in Fig. 3.3B from Model B in Fig. 3.2 - only the transfer of APE from one predictive cue to an earlier one is seen. The details of the movement profile are no longer surprising to the model, and therefore are irrelevant to learning.

Thus, when animals are not naive to an action, measuring dopamine during learning alone may not be sufficient to differentiate whether Model B or D underlie the data.

3.3.3 Omission responses

Next, we replicated the experimental results of reward/cue omission studies. These were presented in the initial paper by Schultz et al.[67] and have since been explored in other contexts[15,252].

One established property of TD-RL, and the dopaminergic theory of value learning, is that when an expected reward does not arrive, dopamine firing is partially suppressed[67]. Further, the absence of associated cues will decrease the predicted future reward, and so, a PE will reappear when/if the reward arrives.

The following simulations were performed on an iteration of Model D which had been previously trained on 200 instrumental association trials as in Section 3.3.1. This explores the changes in δ_{H_a} across three experimental conditions:

1. When the initial predictive cue is omitted.
2. When the expected action is not performed.
3. When the expected reward is not received.

Several predictions for how the TD-AL-based models should behave can be made for each condition and the results are shown in Fig. 3.4.

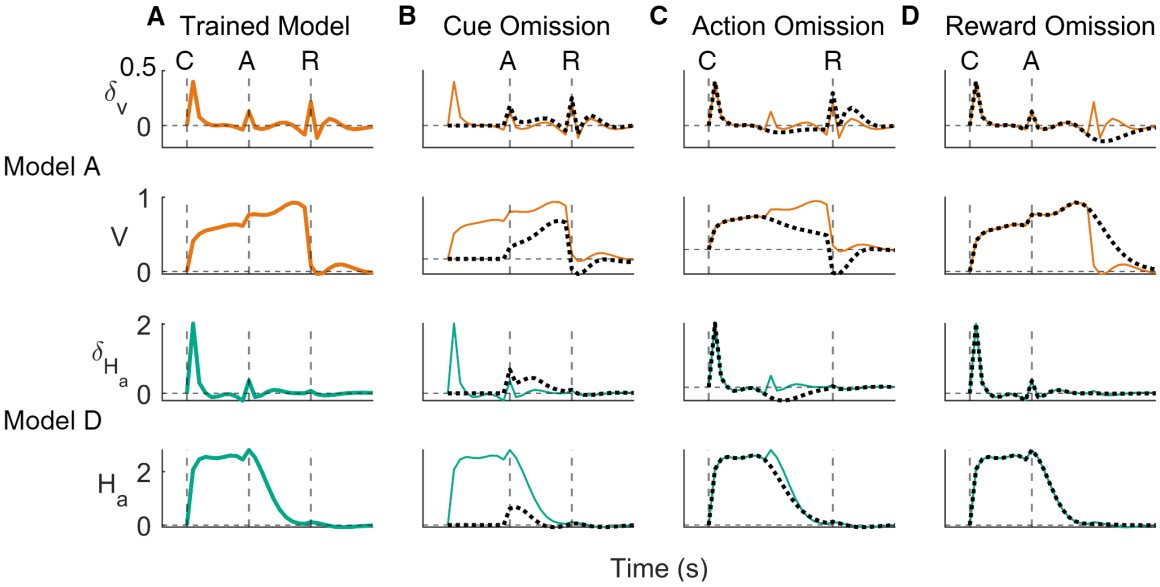


Figure 3.4: Omission Trials.
The results of simulating omission trials on the trained Model A and Model D from Fig. 3.2. From top to bottom: $\delta_V, V, \delta_{H_a}, H_a$.
A: The dynamics in the final trial from the instrumental association simulations. Both models have approached convergence.
B: A single trial with cue omitted (dotted line) overlaid on the trained dynamics from A.
C: As B, with action omitted (dotted line) rather than cue.
D: As B, with reward omitted (dotted line). H_a and δ_{H_a} are completely insensitive.

3.3.3.1 Condition 1 - Cue omission

There should be no response at the usual timing of the cue, as there was no expectation for this to occur - the initiation of a trial does not produce microstimuli, so the model has no indication that the ITI has ended. As a result, arrival of the action is surprising and should produce a larger APE than for a completely trained model. However, since Model D learns a continuous action, the difference in PE is mitigated and will be smaller than for a completely unexpected action to a naive model.

This is indeed seen in Fig. 3.4B, where the action-based microstimuli result in a small surprise for both models since the influence and weighting of the cue-based microstimuli is missing.

For Model A, the early action microstimuli are weakly weighted as, by this point in training, the cue is the greatest predictor. The minimal change in V caused by action in the trained model shows this. However, the increasing contribution of the action

microstimuli to V as the reward approaches results in only a minor difference of RPE at reward.

For Model D, there is a larger change in δ_{H_a} at action initiation, but this is still not as large as for an entirely naive model since the action itself produces microstimuli which allow H_a to predict a portion of the expected future action. By the time the reward arrives, the omission trial for Model D is once more indistinguishable from the trained model.

3.3.3.2 Condition 2 - Action omission

Action omission for an APE-based RL algorithm should behave in a comparable way to reward omission in an RPE model. As shown in Fig. 3.4C, there is no effect on δ_{H_a} or H_a until we surpass the usual time of action. At which point, the PE is temporarily depressed until the expectation of an action fades and H_a returns to baseline.

Model A treats action omission similarly to cue omission, as they both represent predictive stimuli for future reward. When the action does not arrive, the expectation built from the cue begins to decrease and the eventual reward is more surprising as a result.

3.3.3.3 Condition 3 - Reward omission

APE-based habits are reward-insensitive, and so, there should be no change from the usual behaviour of the model. Fig. 3.4D clearly confirms this, by showing that the reward-linked microstimuli weights $w_{H_a,ij}$ have not shifted from 0 - reward does not predict future or current action.

The simulation of reward omission for Model A demonstrates one of the key advantages of applying Ludvig's[176] microstimulus formulation. Rather than a sharp dip at the expected time of reward, the temporal 'smearing' of the Gaussian basis functions results in a slow decay of V and an extended mild depression of δ_V which more closely represents the results seen in real dopaminergic data[92].

3.4 Discussion

3.4.1 Summary

In this chapter, the development of the TD-AL algorithm was outlined, a biologically-plausible and mechanistic learning model which applies an APE to estimate the expected future intensity of actions based on previous experience. Next, the core assumptions made during this thesis regarding the definition of an action were described. Alternative conceptualisations will be expanded upon further in Chapter 7. Finally, simulations of two classic instrumental association experiments were used to define the key predictions that can be made regarding the model's behaviour. Specifically, the influence of action structure and naive weighting on δ_{H_a} dynamics was discussed, alongside an investigation into how γ_h inclusion alters the predictions made by H_a. These simulations also confirm that TD-AL obeys the omission patterns predicted for TD-RL.

These findings were published in the proceedings of CCN 2023[253] and presented at IBAGS XIV[254], before being later confirmed by Lee et al.[255], who included a similar algorithm when describing their feature-specific model of dopaminergic heterogeneity.

3.4.2 TD-AL predictions

Throughout this chapter, brief references have been made to previous TD-RL models which fulfil a similar purpose to TD-AL - producing testable predictions for dopamine signal dynamics.

TD-AL is a direct adaptation of Ludvig's microstimulus model[176] and so, naturally, these models share many features: (1) under the assumption that $\gamma_h > 0$, the APEs transfer directly to earlier predictive cues, (2) temporal precision worsens as the time interval between events increases, and (3) omission of the associated action produces a prolonged decrease in firing with a smaller magnitude than the positive PEs.

Further, TD-AL APEs, by definition, maintain the properties described by the DopAct model (Section 2.4.2); the dopaminergic signal should be (1) reward insensitive and (2)

movement-locked.

However, as demonstrated in Section 3.3, TD-AL also provides some unique predictions:

1. The absence of dopaminergic peaks at predictive cues is not sufficient to disprove the presence of APEs, as this can be explained through specific parameter settings ($\gamma_h = 0$).
2. An established APE-based habit will show decreased dopaminergic firing when an expected action is omitted, regardless of γ_h.
3. Action initiation acts as the final predictive cue for action completion and this is temporarily reflected in the APE signal when $\gamma_h > 0$, before earlier predictive cues dominate the signal.

3.4.3 Comparison to previous APE models

This chapter concludes with a more in-depth comparison of the assumptions made by TD-AL and the previous APE models, i.e., DopAct[2] and the value-free model[1] (Section 2.4.3).

3.4.3.1 Defining an action

As stated in this chapter's introduction, the previous APE-based habit models were designed on a trial-by-trial basis, with a single static value for action in each trial. Both Bogacz's[2] and Miller et al.'s[1] papers presented a version of their model that allowed habits to learn from a continuous scalar action *intensity* with different interpretations depending on the task being simulated. TD-AL likewise encodes a continuous action intensity, but introduces variation across time.

Understanding what is meant by action intensity may be less intuitive than recognising the 'value' of a reward. We can intuit that a bar of chocolate is often more appealing than a salad or that the concentration of sucrose provided will influence how much work a hungry rat is willing to perform[256,257]. This concept of scalar value is often implemented in behavioural experiments, where we can measure how the quantity of an expected reward will impact the amount of effort an animal will apply[258,259], the pattern of action

selection[260–262] or the changing size of corresponding dopamine spikes[252,262,263].

In the same way, an animal's environmental context may not only change which actions are made, but also *how* they are made. This property has already been implemented in previous studies where the amount of effort an animal/subject will make (e.g., the rate and force with which a mouse presses a touchscreen[264]) is measured as an indicator of the underlying motivation for the eventual outcome[265,266].

For the simulations presented in this chapter, the models experienced an identical Gaussian action in every trial. As such, the associated H_a values learnt quickly and would have high certainty in the action profile they predict once trained. Thus, the APE presented were able to converge relatively quickly. However, as we will demonstrate in Chapter 4, the ability to match action intensity to behavioural data is valuable in producing realistic predictions of dopamine dynamics between trials.

3.4.3.2 Describing SOR effects

Despite their trial-by-trial nature, both previous APE models provided theoretical accounts for the influence of VI SORs in promoting habit strength, relative to VR (Section 2.1.2).

Their explanations align with the 'response rate - reward rate correlation' theory given by Dickinson[44] which states that VR schedules produce linear relationships between press rate and reward, and therefore maintain a strong A-O contingency. The value-free model explicitly calculates this value (namely, g, in Eq. 2.20) and uses it to bias the arbiter towards the goal-directed system. In contrast, the VI SORs have a sub-linear relationship between press rate and reward, thus reducing g, so that the agent relies on habits much earlier. DopAct similarly explains this SOR effect through the accuracy of (and so, confidence in) the reward predictions produced by the goal-directed system at a given press-rate[2].

In contrast, the TD-AL model can be used to partially interpret SOR effects according to DeRusso's[43] temporal action-contiguity hypothesis, though the balance of the two systems cannot be fully explained without a specific observation function. Under a

free-operant task, the goal-directed system relies on previous actions to behave as predictive cues for later rewards. When there is a consistent temporal A-O contiguity, the prediction error is minimised, the same microstimulus weights are updated and Q_a will be large and precise. Accordingly, the greater the variation in the action-reward interval, the weaker the goal-directed system will be.
The habit system does not suffer this disadvantage, as both VI and VR schedules produce linear response rates (see Fig. 2.1) - the interval between lever-presses remains constant, and so, actions are able to serve as strong predictors for themselves. However, the scalloping of FI responses will weaken the action-to-action contiguity and H_a instead must increasingly rely on the previous rewards to behave as the primary predictor.

Overall, TD-AL provides a theoretical, biologically-plausible and mechanistic model of APE-based habits in the brain, with unique and testable predictions. In the next chapter, we explore the ability of this algorithm to describe real neural data.

4

Analysis of Neurophysiological Data

Contents

4.1 Introduction

The second goal of this PhD was to interrogate whether an APE-like signal could be found within neural data. As proposed by Bogacz[2] and discussed in Chapter 3, striatal dopamine represents the most likely neural correlate for an APE, similarly to the better established RPE.

In the previous chapter, we outlined the novel TD-AL model and posited that the continuous temporal dynamics of the associated PE could be used to test Bogacz' proposal.

Historically, research into nigrostriatal dopaminergic signals has largely been linked to motor vigour, movement kinematics and action initiation[126,127,130,131,139] (Section 2.2.3).

More recently, PE-like dynamics have been found in SNc signals and have been associated with multiple types of information, such as different reward features[107], novelty[103] and aversive outcomes[110,111]. It is likely that heterogeneity in dopamine signalling allows it to cover most, if not all, of these aspects, particularly given recent evidence that distinct genetic subtypes of dopaminergic neurons express differential sensitivities to rewards, threats and movement kinematics[132].

However, as discussed in Chapter 2, S-R-based value-free habits could not form from a reward or aversion PE as they are, by definition, outcome-insensitive. Some studies have interrogated the role of novelty in action selection[267–269] but, in isolation, novelty is not sufficient to form S-R relationships.

Further, DS recordings have shown that dopamine release (1) is specifically responsive to actions that have been learnt during an instrumental association task and (2) influences learning of both A-O and S-R contingencies[137–139,147].

Crucially, Greenstreet et al.[3] recently published dLight recordings from the TS that appear to exhibit all the properties required of a dopaminergic APE signal, i.e., they are movement-locked, decay over trials and are unchanged by the presence or omission of a reward (Section 3.4.2). Thus, their study presents a rich dataset with the potential to confirm the existence of APEs in the brain.

This chapter provides both (1) a proof-of-concept that TD-AL can be used to test real neural data for APEs and (2) evidence that supports the existence of APE signals in rodent dopaminergic recordings.

We begin with a description of the experiment and relevant results from Greenstreet et al.[3], followed by an analysis of the extent to which these data can be explained by TD-AL or TD-RL, using Bayesian model selection (BMS) methods[270].

4.2 Study by Greenstreet et al., 2025

As stated above, an APE-like signal was recently reported by Greenstreet et al.[3]. This section provides a brief overview of the pertinent experimental methods and results, which are schematised in Fig. 4.1. A full description is provided in Greenstreet et al.[3].

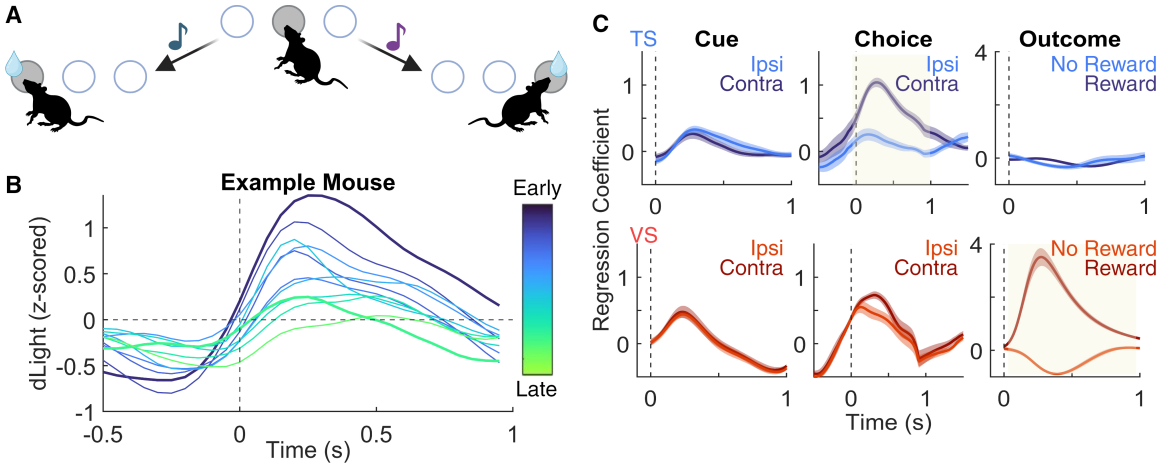


Figure 4.1: Greenstreet et al. 2025, Experimental Design and Results.
A: A schematic overview of the CoT task. Trials are self-initiated through a held nose-poke in the centre port (100-300ms), then an auditory cue indicates which port will produce a reward following a nose-poke.
B: A recreation of the example mouse dLight recordings in their original paper, aligned to movement initiation. Each line is an average of 200 consecutive trials, from early (blue) to late (green) in learning. The first and final 200 trials are indicated by thicker lines.
C: Regression coefficients across time for the TS (top) and VS (bottom) aligned to separate trial events, from their original paper.
(Left) The regression coefficients for the two cues are not significantly different. Both the TS and VS show a slight, non-specific increase 0.25s following cue.
(Centre) As with left, but aligned to action initiation and separated between ipsilateral and contralateral trials. There is a significant difference in the coefficients for the TS, indicated by the yellow highlighted region.
(Right) As centre, but aligned to expected outcome timing and separated between rewarded (correct) and unrewarded (incorrect) trials. The TS shows no response in either kernel. As expected, the VS shows a significant response to reward.

This dataset recorded the dopaminergic signals of water-deprived mice in either the VS or TS during a two-choice cloud of tones (CoT) task[271] (Fig. 4.1A).

The mice were placed in a box with ports on three walls (left, centre and right) and could self-initiate a trial by holding a nose-poke in the centre port for 100-300ms whenever a centre LED was lit. A successful initiation poke resulted in the centre LED turning off and an auditory cue sounding 0-50ms later. The physical distance between the ports and the hold requirements were designed to increase the temporal separation between cues, actions and rewards. The frequency of the auditory cue (high or low, selected between two octaves) indicated which port (left or right) would produce a reward (2μl of water) following the corresponding nose-poke. The mapping between frequency and port was counterbalanced between animals. Mice completed this task for over 10,000 trials and

successfully learned to differentiate high- and low-pitched tones. For the training task that provided the results used in Section 4.3, water-delivery was deterministic - as long as the mouse made the correct choice, it would receive a reward.

Greenstreet et al.[3] employed a fluorescence recording method to measure dopamine[272] in which an artificial dopamine receptor, dLight, is injected into the relevant neural region using a viral vector. These receptors emit a known amount of light for each molecule of dopamine that binds to them, allowing the quantity of the monoamine in the given area to be measured. Photometric dLight recordings began immediately after the initial habituation period. The researchers elected to record from two neural locations, the VS and TS. The former acted as a control and showed the clear RPE signals that have been regularly reported and are well-established[92].

Interestingly, despite the historical association of DLS with expression of S-R behaviours (Section 2.1.3), this study chose to interrogate the TS, rather than the DLS. This decision was motivated by the fact that, of all the striatal structures, the TS is located the furthest away from the reward-associated VS and receives no efferent signals from the VTA. As such, it is the least likely region to express RPE signals, which may mask the presence of APEs.

Three key results were reported, which align with Bogacz's[2] predictions for the properties of an APE (Section 3.4.2):

1. TS dopamine expressed a peak that was locked to movement initiation and decreased across trials from early to late in learning (Fig. 4.1B).
2. Regression analysis showed that this peak was significantly larger for contralateral actions in the 1s following action initiation (Fig 4.1C, top, middle). The VS response during movement was non-specific (Fig 4.1C, bottom, middle).
3. The TS signal was reward-insensitive (Fig. 4.1C, top, right).

The original paper also completed several other tests of the TS dopamine to establish a series of other results supporting its role in habits.

As part of these, they concluded that both an intact TS and efferent dopaminergic signalling were *necessary* for the mice to learn the CoT task. Specifically, TS ablations and

6-OHDA lesions caused a reduction in learning rate and performance level. The degree to which the learning rate decreased was correlated with the size of the 6-OHDA lesions and these had no influence on the kinematics of expressed behaviour. Additionally, when trained animals completed a learnt action in response to a novel cue (white noise), the movement-locked dopaminergic peak increased its magnitude once more. Overall, these results imply that TS dopamine is acting as a *learning* signal.
Further, optogenetic inhibition of D1R-expressing SPNs biased the mice towards the ipsilateral choice and vice-versa for D2R-expressing SPNs. This suggests that the TS SPNs are involved in selection and execution of learnt behaviour, not just in discrimination of auditory signals.

Thus, the dopaminergic and striatal signals are action-sensitive, play a role in habit formation and demonstrate key features of an APE. We extend on this body of work by interrogating whether the complex dopaminergic dynamics can be explained by our mechanistic APE model.

4.3 Methods

The results by Greenstreet et al.[3] present a perfect opportunity to test whether TD-AL provides a good description of apparent APE signals in dopaminergic data. Further, this data can be used to determine whether TD-AL replicates the temporal dynamics of these signals over and above a model-free RPE.
This next section first outlines the mathematical details of the models to be compared and justifies their selection. Then, it recounts the methodological analysis employed to calculate the best-fitting parameters and determine a model's goodness-of-fit. Specifically, we test the learning models' capacity to reproduce each individual mouse's dopaminergic signal over time, as presented in Fig. 4.1B.

4.3.1 Five potential models

In our study of the dopaminergic data, we aimed to interrogate two key factors:

1. Does TD-AL, rather that TD-RL, provide the best fit to the data?
2. Is there evidence for a temporal discount factor, γ, in the data?

In total, five model combinations were fit to the data. Their formulations and equations are outlined below and summarised in Table 4.1.

Label	Algorithm	γ	α	κ
Action-only	TD-AL	0	0	analytic
TD-AL$_0$	TD-AL	0	Free (0-1)	analytic
TD-AL$_\gamma$	TD-AL	Free (0-1)	Free (0-1)	analytic
TD-RL$_0$	TD-RL	0	Free (0-1)	analytic
TD-RL$_\gamma$	TD-RL	Free (0-1)	Free (0-1)	analytic

Table 4.1: An overview of the five models tested on data from Greenstreet et al.[3].

4.3.1.1 The learning models

The overarching goal of this subproject was to directly test whether the data described above could be better explained by TD-AL than the gold-standard TD-RL model. As described in Chapter 3, Ludvig et al.'s[176] microstimulus formulation was applied for both of these models (Section 3.2.1), using the fixed parameters outlined in Table 3.1. These algorithms contain two free parameters - the learning rate, α, and the discount factor, γ - which can both be fit to neural data with the fitting procedure described in Section 4.3.2. For ease of reference, the models' equations are restated below:

TD-RL

$$V(t) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}(t) w_{V,ij}(t) \tag{3.4}$$

$$\delta_V(t) = R(t) + \gamma_v V(t) - V(t-1) \tag{2.2}$$

$$w_{V,ij}(t+1) = w_{V,ij}(t) + \alpha_v \delta_V x_{ij}(t) \tag{3.5}$$

TD-AL

$$H_a(t) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}(t) w_{H_a,ij}(t) \tag{3.8}$$

$$\delta_{H_a}(t) = A_a(t) + \gamma_h H_a(t) - H_a(t-1) \tag{3.7}$$

$$w_{H_a,ij}(t+1) = w_{H_a,ij}(t) + \alpha_h \delta_{H_a} x_{ij}(t) \tag{3.9}$$

where: i = stimulus,

j = microstimulus,

m = total number of microstimuli for any stimulus, i,

n = total number of events experienced,

V = expected future value of reward, R,

x_{ij} = activity level of the individual microstimulus, j, for event, i,

$w_{V,ij}$ = V weight for microstimulus, j, produced by event, i,

H_a = expected future intensity of action, a,

$w_{H_a,ij}$ = H_a weight for microstimulus, j, produced by event, i,

δ_v = reward prediction error,

R = reward,

δ_{H_a} = action prediction error,

a = action,

A_a = action intensity of action, a,

α_v = value learning rate,

γ_v = reward discount factor,

α_h = habit learning rate,

γ_h = action discount factor.

While this data appears to contain all the elements required of an APE, i.e., a movement-locked peak that decreases over learning and is reward insensitive (Section 3.4.2), an alternative interpretation exists; the collapse in dopamine over trials could simply be a reflection of the movement kinematics, which improve with training, rather than a true representation of a learning signal.

Previous studies have detected dopaminergic responses to movement vigour and initiation[123,125,126] and performance on this CoT task speeds with experience. Thus, it is not implausible that dopamine responses to movement would appear to collapse as the time to travel between ports decreases across learning.

To account for this possible effect, a model was included in our analysis for which no learning occurs at all. This *action-only* model is equivalent to setting the learning rate, α, to 0. As a result, the microstimulus weights, w_{ij}, never update and the PE becomes a

direct measure of action intensity (Eq. 4.2).

$$H_a(t) = \sum_{i=1}^{n} \sum_{j=1}^{m} 0x_{ij} = 0 \tag{4.1}$$

$$\delta_{H_a} = A_a(t) + 0\gamma_h - 0 = A_a(t) \tag{4.2}$$

4.3.1.2 Altering γ

In the previous chapter (Section 3.3.1.4), the influence of the discount factor, γ, was explored through simulations. Classically, γ_v is set close to 1 when fitting dopaminergic data, a value that has previously provided good replications of RPE-like signals[175,273,274]. However, this parameter is not *necessary* for the model to learn. As shown in Chapter 3, when $\gamma_h = 0$, H_a becomes an estimate of the action intensity usually taken at *that point in time*, rather than a calculation of how much future action can be expected. Consequently, this formulation results in an absence of PEs in trained agents at the moment of predictive cues - a property described by Greenstreet et al.[3] whose regression model (Fig. 4.1C) shows an insensitivity to the frequency of the predictive auditory cue. Though the mice learn to use this information in order to discriminate between trials, the dopaminergic signals do not reflect a difference in reward/action expectation produced by one cue over the other. Accordingly, we elected to include two further models in our analysis, with γ fixed to 0. This was not necessary for the action-only model, in the absence of any associated learning rule.

4.3.1.3 A scaling parameter, κ

Given the unitless nature of the PEs and dopaminergic data, an additional free parameter, κ, was included in the analysis to scale the simulated PEs and minimise their difference from the true neural data. Thankfully, an analytical solution exists that calculates the optimal scaling parameter (Section 4.3.2, Eq. 4.3).

4.3.2 Fitting procedure

The dataset provided by Greenstreet et al.[3] is rich, with multiple mice completing roughly 10,000 trials while camera tracking and dLight measurements were recorded. This next section outlines the key details of the additional data preprocessing completed and fitting procedure that was applied to match model PEs to the dopaminergic signals. A schematic overview of the fitting procedure is provided by Fig. 4.2.

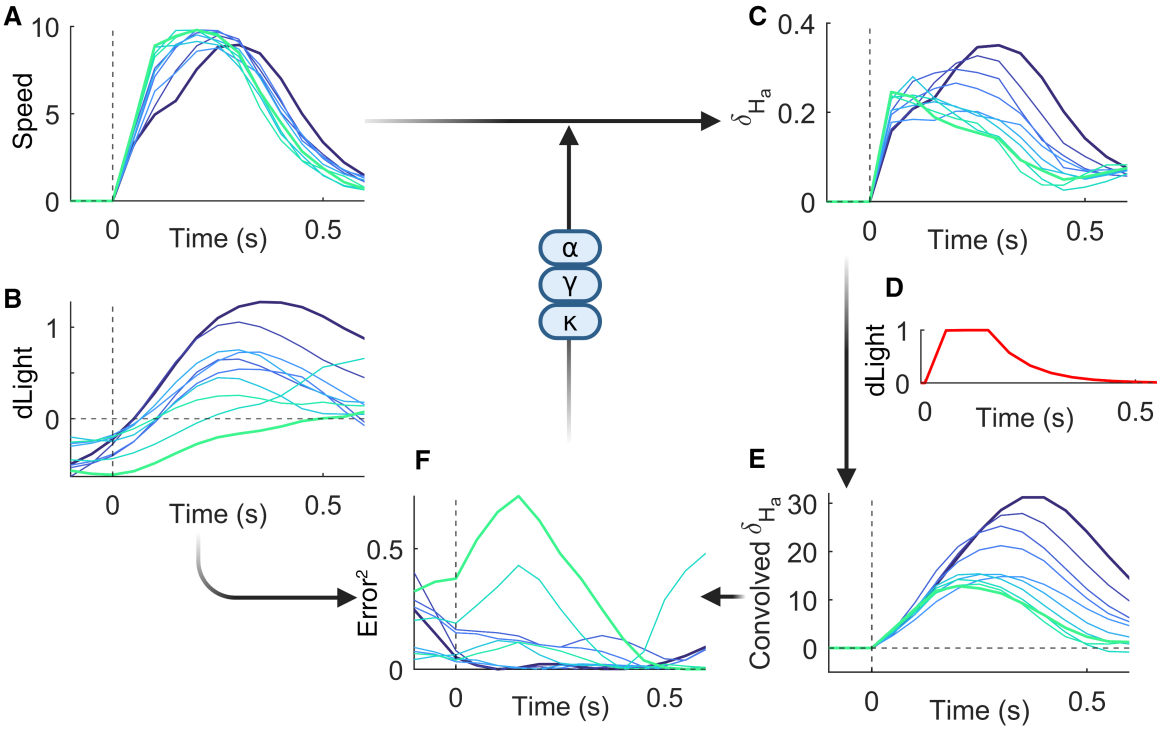


Figure 4.2: Model Fitting Procedure.
A: Camera tracking data was used to extract the speed of the mouse (averaged between nose and ears) from the time it left the centre port (0s) until it entered the side port.
B: The dLight data processed as in Fig. 4.1B.
C: For a given set of parameters, the model learns from the behavioural data and the associated PE is calculated. These trials are then averaged over 200 consecutive events.
D: The temporal dynamics of dLight fluorescence. This is convolved with the PE to produce an equivalent signal to the experimental data, as though the same methodology was used to record both the PE and dopaminergic data.
E: The resulting convolved PE.
F: The cost function is calculated as the SSE of the averaged dopaminergic signals (**B**) and convolved PEs (**E**) in the 0.5s following action initiation. The parameters update in order to minimise this cost. **C-F** repeats until the best-fitting parameters are returned.

4.3.2.1 Behavioural simulation

Each learning model was provided with a simplified series of continuous events determined by the real experiences of a given mouse. Thus, as illustrated in Fig. 4.3, trials obeyed the following sequence of steps:

1. An auditory cue sounds (either high or low).
2. The mouse exits the centre nose port (the action is initiated) and the choice is determined by the port entered (either left or right).
3. The continuous intensity of the trial's action is set to the mouse's speed, as extracted from the camera tracking data and averaged between the nose and ear movement.
4. If the mouse makes the correct action, it receives a reward.

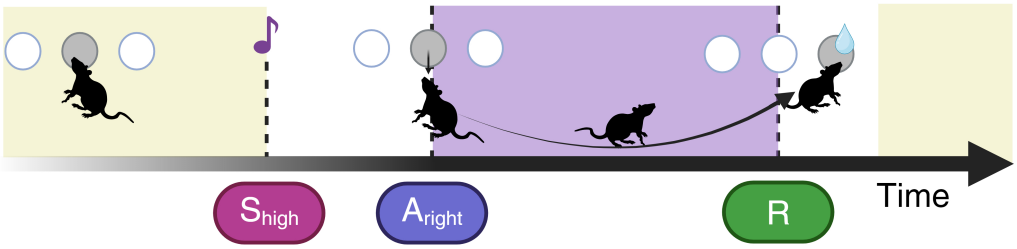


Figure 4.3: A Simulated Trial.
Following an ITI (yellow), during which all previous microstimuli are allowed to finish and the next trial begins, the 'go' tone initiates the associated cue microstimuli (either S_{high} or S_{low}). The action microstimuli (A_{left} or A_{right}) begin when the mouse leaves the centre port. The action intensity dynamics, $A_a(t)$, are determined by the speed of the mouse and continue until the mouse enters the associated port. If the correct choice is made, a reward (R) is received, the reward microstimuli are produced and the next ITI begins.

In summary, there were five potential events that produced microstimuli (two cues, two actions and a reward), resulting in 60 weights, w_{ij}, ($m = 12$ per event) which update at every timestep (dt = 0.05s) according to either Eq. 3.5 or Eq. 3.9. The delays between these events were determined by those actually experienced by the mouse on the associated trial. At the end of the trial, a fixed ITI of 30dt (1.5s) was given, such that all microstimuli from the previous trial had ended and the model could not learn to expect the subsequent auditory cue.

Speed was selected as the measure of action intensity for two reasons. First, speed is a continuous measure which remains positive regardless of the choice made, whereas

position, for example, would need to be corrected relative to a starting point. This allowed for minimal preprocessing to be required and for the two actions' H_a variables to receive equivalent information. Second, as described above (Section 4.3.1), experience and training lead to an increased speed and reduced trial duration. In using this measure directly as the learnt output, we can partially control for these training effects on the APEs and resulting dopaminergic signals.

4.3.2.2 Processing dLight data

All preprocessing described by Greenstreet et al.[3] had been completed prior to our data access. The original dopaminergic signal had a resolution of 10,000 Hz and so, following a rolling Z-score across a window of 1.6s, the frequency was reduced to a timestep of 0.05s via MATLAB's `resample` function, which uses linear interpolation to calculate new values based on averages from the true datapoints.

To reproduce Greenstreet et al.'s[3] results for each mouse (as in Fig 4.1B) the dopaminergic data was divided into individual trials comprising of a 1.5s window centred around action initiation. Using the subset of trials in which the contralateral port was cued, the dopaminergic curves were averaged across 200 consecutive trials to produce a series of movement-locked peaks. A total of six mice completed this task with TS recordings. The resulting signals are shown in Fig. 4.4.

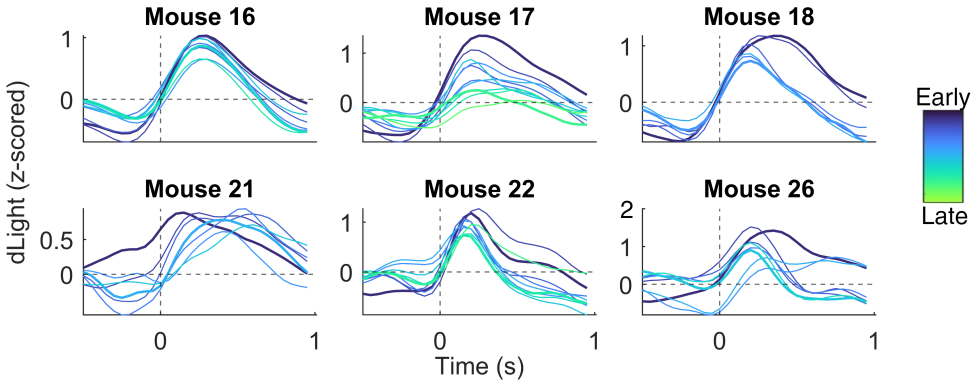


Figure 4.4: Dopamine Signals for Individual Mice.
Each panel provides the data of a different mouse, from the subset of trials where the contralateral action was cued. The dLight signal is aligned to action initiation. Each line represents an average of 200 consecutive trials in training from early (blue) to late (green), with thicker lines for the first and final 200 trials.

4.3.2.3 Fitting the PEs

Once a simulation is run with the given dataset and model parameters (α and γ), the resulting PE signal can be compared to the dopaminergic data to calculate the goodness-of-fit.

The time course of the dLight fluorescence contaminates the dopaminergic signal with a consistent artifact. Thus, to produce comparable PE signals, we convolved the δ variable with a kernel representing the temporal effect of dopamine binding on the sensor (Fig. 4.2D), using MATLAB's `conv` function. The rise and fall time of these kernels were set to $\tau_{\frac{1}{2}}$ = 9.5ms and 90ms, respectively, as reported in Patriarchi et al.[272].

The convolved PE signal is processed in an equivalent manner to the dopaminergic signal to produce trial averages over training, as shown in Fig. 4.4.

A scaling parameter, κ, is then determined such that the difference between the two figures is minimised, via Eq. 4.3.

$$\kappa = \frac{\sum_t \left(\delta_{\text{conv}}(t) D(t)\right)}{\sum_t \delta_{\text{conv}}(t)^2} \tag{4.3}$$

where: κ = scaling factor,
δ_{conv} = convolved PE signal,
D = dopaminergic signal,
t = timestep.

Finally, the cost function of the current simulation is calculated as the sum squared error (SSE) in the 0.5s (10 timesteps) following action initiation for the averaged dopaminergic data and the scaled, averaged and convolved PE data, as specified in Eq. 4.4 (and mentioned in Fig. 4.2E):

$$\text{SSE} = \sum_t \left(\left(\kappa * \delta_{\text{conv}}(t) - D(t) \right)^2 \right) \tag{4.4}$$

where: SSE = cost function for a given dataset and parameter combination.

The cost function is minimised over several iterations of the above simulations, using MATLAB's `fmincon` function, until the best-fitting parameters are established.

The recovered model for each mouse is determined via Bayesian information criterion (BIC) analysis[275], such that a lower BIC implies a better fitting model, though the Akaike information criterion (AIC)[276] is also reported.

These values represent a measure of model fit that includes a penalty for additional parameters to counteract the risk of overfitting. As our likelihood measure is approximated using an SSE cost function, the equations to calculate BIC and AIC differ slightly from the standard in order to account for the Gaussian distribution of errors[277]. Thus, the values reported in this chapter are calculated using Eq. 4.5 and Eq. 4.6:

$$BIC = t\log\left(\frac{\text{SSE}}{t}\right) + p\log(t) \tag{4.5}$$

$$AIC = t\log\left(\frac{\text{SSE}}{t}\right) + 2p \tag{4.6}$$

where: p = number of free parameters in the model,

t = number of datapoints used to calculate the SSE.

4.4 Results

4.4.1 Individual mice

To begin, we illustrate the similarity of the true data to the PEs produced from each model's best fitting parameters (Fig. 4.5, Table 4.3). This allows us to examine the success of recovery. Should none of the simulations provide qualitatively good results then the statistical BIC analysis would simply tell us which model was 'least poor', rather than providing evidence for a given underlying mechanistic model.

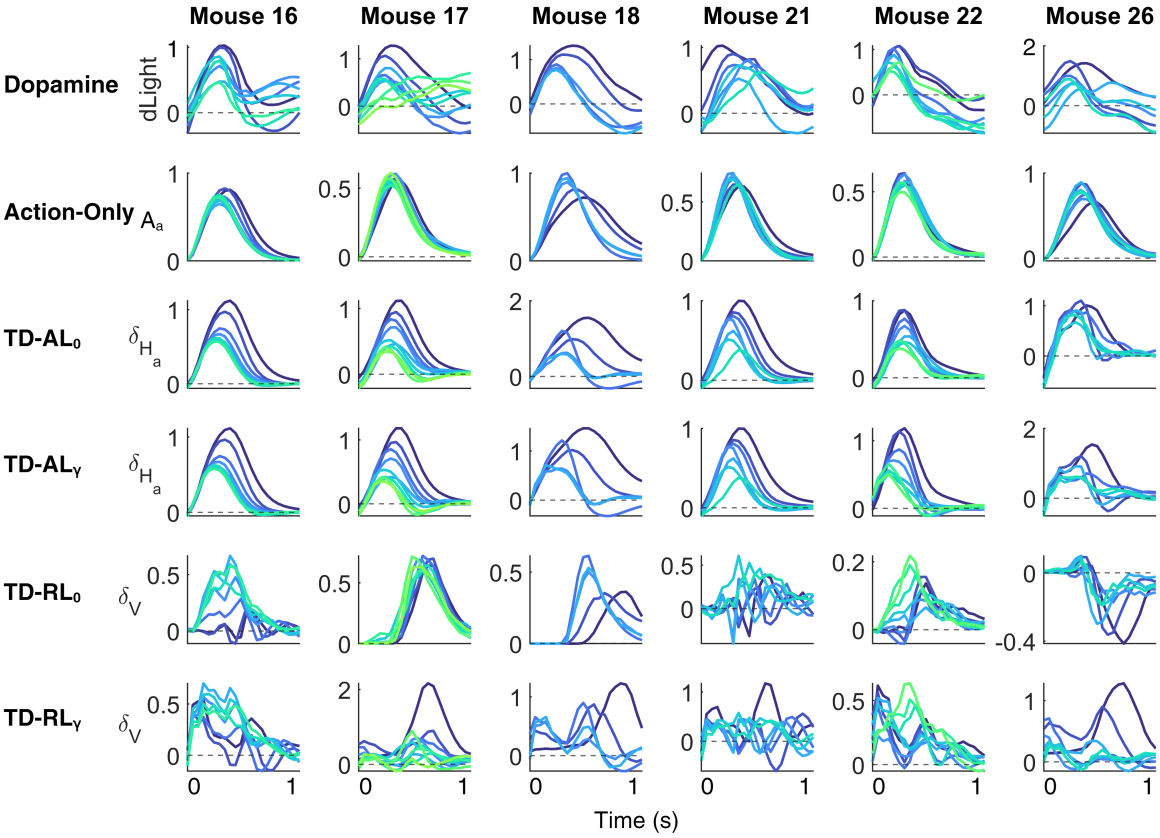


Figure 4.5: Best-fitting Simulations.
Each panel represents the corresponding signal averaged over 200 consecutive trials from early (blue) to late (green) in learning, aligned to action-initiation. Each column contains the results for a given mouse. The first row provides the true dopaminergic signal as given in Fig. 4.4. Rows 2-6 show the simulated and convolved PEs produced by the best fitting parameters (Table 4.3). From top to bottom: the action-only model, TD-AL$_0$, TD-AL$_\gamma$, TD-RL$_0$ and TD-RL$_\gamma$. Note that, as κ is optimised to average over all lines and timesteps, the scale of the y-axis varies between models.

Visually, we can see the degree to which TD-RL provides a poor fit for these data as the simulated results do not resemble the dopaminergic signal at all. A significant delay between action-initiation and reward was enforced, so the corresponding RPE peaks arrive much later than is seen in the dopaminergic data. Further, if they contain any structure at all, the peaks sometimes increase in size across learning.

In contrast, the two TD-AL models and the action-only model largely appear to qualitatively capture the timing of the peaks and the pattern of collapse. We can therefore continue and interrogate the degree to which these models quantitatively replicate the data. The best fitting model and corresponding BIC results for each mouse are provided in Table 4.2.

It is worth noting that, when the number of datapoints used to calculate BIC is large, BIC may over-penalise additional parameters and incorrectly prefer a model that is too simple, though this is partially mitigated by our use of the SSE-adapted BIC calculation (Eq. 4.5). This is a particular risk here, as we use many datapoints per trial and so the BIC values are penalised twice as strongly as AIC. Further, given that the dopaminergic data is quite complex and the maximum number of parameters for any model is three, overfitting due to increased model complexity is not likely. Thus, we also report the AIC values in Table 4.2 and will consider these results when looking at the visualisation of the best-fitting model simulations.

	Mouse	Action-only	TD-AL$_0$	TD-AL$_\gamma$	TD-RL$_0$	TD-RL$_\gamma$	Preferred Model
BIC	16	-283	**-332**	-330	-119	-138	TD-AL$_0$
	17	-220	-283	**-285**	-131	-158	TD-AL$_\gamma$
	18	-99.9	-146	**-167**	-24.3	-42.4	TD-AL$_\gamma$
	21	-169	**-194**	-190	-78.2	-97.5	TD-AL$_0$
	22	-185	-191	**-270**	-105	-138	TD-AL$_\gamma$
	26	-100	-115	**-120**	-34.8	-52.3	TD-AL$_\gamma$
AIC	16	-286	-337	**-338**	-124	-145	TD-AL$_\gamma$
	17	-223	-289	**-293**	-137	-166	TD-AL$_\gamma$
	18	-102	-150	**-172**	-28.2	-48.2	TD-AL$_\gamma$
	21	-171	**-198**	-196	-83	-104	TD-AL$_0$
	22	-188	-196	**-278**	-110	-145	TD-AL$_\gamma$
	26	-102	-120	**-127**	-39.3	-59.0	TD-AL$_\gamma$

Table 4.2: The BIC/AIC values and associated preferred model produced by the best-fitting parameters for each mouse. The lowest value for each mouse is highlighted in bold.

Overall, according to both BIC and AIC measures, *all* mice preferred a TD-AL model, thus providing strong evidence that these dopaminergic dynamics represent an action-based *learning* signal, rather than an RPE or a simple representation of movement kinematics.

The introduction of the third discount parameter, γ_h, has different impacts on the BIC and AIC of each mouse. Logically, the magnitude of the best-fitting γ_h parameter will likely influence the degree to which TD-AL$_\gamma$ is preferred relative to TD-AL$_0$. So, if a large γ_h value is required to replicate the dopaminergic data, the underlying model will potentially be

	Action-Only	**TD-AL$_0$**		**TD-AL$_\gamma$**			**TD-RL$_0$**		**TD-RL$_\gamma$**		
Mouse	κ	κ	α_h	κ	α_h	γ_h	κ	α_v	κ	α_v	γ_v
16	0.0171	0.0255	4.197×10^{-4}	0.0255	6.429×10^{-4}	0.3516	2.5243	0.6723	3.4604	0.3554	0.9381
17	0.0165	0.0369	9.264×10^{-4}	0.0396	0.0022	0.5236	1.2945	6.588×10^{-12}	5.3225	0.0288	0.9540
18	0.0209	0.0633	0.0021	0.0573	0.0036	0.5260	0.9150	6.0822×10^{-10}	4.3305	0.0307	0.9360
21	0.0146	0.0251	4.832×10^{-4}	0.0251	4.8333×10^{-4}	1.069×10^{-5}	7.8124	1	11.8394	0.3233	0.8356
22	0.0168	0.0276	5.659×10^{-4}	0.0384	0.0044	0.8366	0.7526	1	4.2435	1	1
26	0.0237	0.2421	1	0.1475	0.0534	0.6337	-1.5110	0.0312	3.0659	0.0271	0.9978

Table 4.3: The best-fitting parameters for each mouse and model.

more likely to include this term.

The primary influence of γ_h is the gradual transfer of PE to earlier time-points, which is visualised by the dopaminergic peak shifting to the left as training progresses. Qualitatively, mouse 22 shows the most notable forward shift in the TD-AL$_\gamma$ simulation, relative to TD-AL$_0$ (Fig. 4.5), and has the largest γ_h value at 0.837, which likely account for its strong preference.

At the other end of the spectrum, mouse 21 produced the only dataset to consistently favour TD-AL$_0$ with both metrics. This is perhaps unsurprising given that the best-fitting γ_h was 1.07×10^{-5}, which is magnitudes smaller than the γ_h values reported for all other mice. Such a small parameter value suggests that, even if TD-AL$_\gamma$ is the correct model used by this mouse, the difference between the outputs is statistically insignificant and so the simpler 2-parameter model will be preferred.

Accordingly, the mouse with the second lowest γ_h value (mouse 16, $\gamma_h = 0.352$) provided the only other dataset to be recovered as TD-AL$_0$ by our BIC analysis. When the parameter penalty is reduced by AIC, the preferred model returns to TD-AL$_\gamma$ in line with all other subjects.

An additional feature of interest is that mouse 26's dataset was the only one that returned a scaling factor > 0.075 for the action-based models, as $\kappa = 0.242$ and 0.148 for TD-AL$_0$ and TD-AL$_\gamma$, respectively. These models also had the largest α_h values at 1 and 0.053, respectively. All other mice had α_h values of less than 0.005 - smaller by a factor of 10. In conjunction, this suggests that the data provided for this mouse was recovered the least well, as can be confirmed with the lack of visual synchronicity between all the simulations and the true dopaminergic data. This poorer fit is further reflected in the BIC and AIC values for the three action-based models, which are less negative for mouse 26 than any of the other mice. It is difficult to determine why this dataset in particular struggled to recover, though it possibly results from the model's attempt to replicate the negative dopaminergic signal at the start of the action.

Nevertheless, these measures are still significantly better than those for the TD-RL models, providing further evidence that the recorded dopaminergic signals are more

likely to be action-based signals than reward-based, even when TD-AL struggles to recover appropriate parameters.

In sum, for all mice, we can confirm that there *is* evidence of an action-based learning signal which is not simply a record of current kinematics. The comparison of TD-AL$_\gamma$ and TD-AL$_0$ is strongly influenced by the magnitude of the best-fitting γ_h and requires a subjective consideration of whether their BIC and AIC preferences result from overfitting or capture a real difference in information contained within the signals. Unfortunately, as will be explained in Section 4.5, it has not been possible to quantitatively analyse the reliability of our model recovery, which would have been particularly useful when considering nested models such as those presented here.

Next, we extend these results to consider whether evidence for APEs and, more specifically, whether TD-AL can be uncovered at a population-wide level.

4.4.2 Group Bayesian model selection

An additional post-hoc BMS analysis was completed. Using the BIC and AIC values as a stand-in for model evidence, this method allowed us to explore the model fits at a group level (see Stephan et al.[278] and Rigoux et al.[270] for full details). It removes the common assumption that a single model is shared by all members of a population (a 'fixed effects' comparison) and, instead, calculates the likelihood of the model distribution.

BMS assumes that the ratio of models within a population takes the form of a Dirichlet distribution, such that each model, k, occurs with a probability, r, given the number of unobserved occurrences, α_k (Eq. 4.7). Additionally, this method presumes that the models compared are the only models present in the population and thus, all r sum to 1.

$$P(r|\alpha) = \mathrm{Dir}(r,\alpha) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k r_k^{\alpha_k - 1} \tag{4.7}$$

where: r = probability of model in population,

Γ = the gamma function,

α = unobserved occurrence of models in population,

k = a given model within the population set.

Stephen et al.[278] proposed that the model likelihoods can be compared according to the conditional model probability, $P(r|y;\alpha)$, where y represents the actual data observed. The resulting exceedance probability, ϕ_k, can be interpreted as a measure of how likely it is that a given model, k, is the most prevalent model in a population (Eq. 4.8).

$$\forall j \in \{1...K|j \neq k\}: \quad \phi_k = P(r_k > r_j|y;\alpha) \tag{4.8}$$

where: ϕ_k = exceedance probability for the model, k,
K = total number of models.

Rigoux et al.[270] further extend this measure to produce the *protected* exceedance probability, $\tilde{\phi_k}$, which considers an error rate, or 'Bayesian omnibus risk', and calculates how likely a given model, k, is to be the most prevalent model, *over and above chance* (Eq. 4.9).

$$\forall j \in \{1...K|j \neq k\}: \quad \tilde{\phi_k} = P(r_k > r_j|y;\alpha,H_0,H_1) \tag{4.9}$$

where: $\tilde{\phi_k}$ = protected exceedance probability for the model, k,
H_0 = null hypothesis that all models are equally frequent,
H_1 = alternative hypothesis that all models are not equally frequent.

These were calculated using the `bms` function created by Gershman[279] and the protected exceedance probabilities are reported in Fig. 4.6B.
For consistency with the common literature, we also indicate the distribution of BIC and AIC values. Under an assumption that all members of a population utilise the same model, the fixed effect comparison method[191,278] expects the correct model to consistently have the lowest (here, most negative) BIC across all subjects (Fig. 4.6A).

The fixed effect analysis confirms that the non-TD-RL models have significantly lower BICs and AICs and, of these, TD-AL$_\gamma$ is the best fitting model on average, but this difference is negligible relative to TD-AL$_0$. As expected, TD-RL performs the worst.

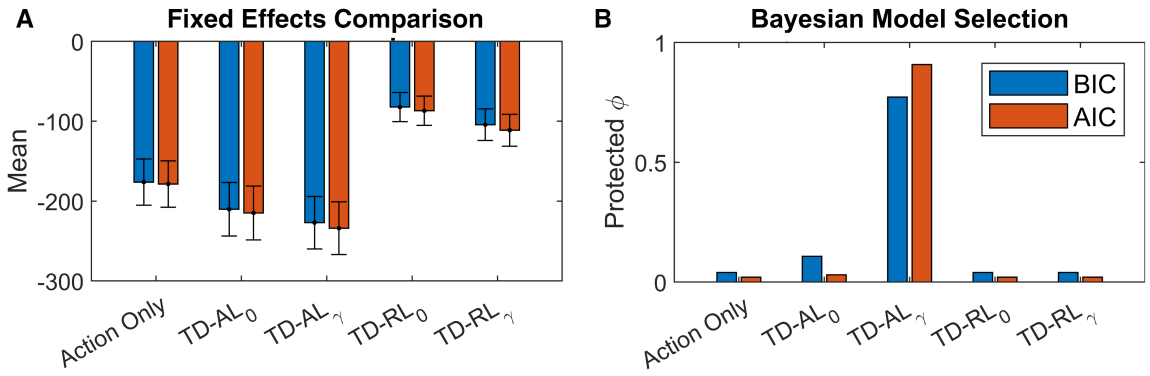


Figure 4.6: Group Analysis of Model Fits.
A: A fixed effects comparison of the mean BIC (blue) and AIC (red) values produced by each of the models. The standard error of the mean is reported in the error bars.
B: A comparison of the protected exceedance probabilities, $\tilde{\phi}_k$, for each model using BIC and AIC as approximations for model likelihood.

In contrast, the results from our BMS analysis are much clearer. Under the assumption that *some* mice may use alternative (potentially nested) models, the protected exceedance probability gives a 77.2% likelihood that TD-AL$_\gamma$ is the most prevalent model within the population, over and above chance. TD-AL$_0$ is the next most likely at 10.8%. This separation of $\tilde{\phi}_k$ between the two TD-AL models is exacerbated when the additional parameter is penalised less strongly, with probabilities of 90.8% and 3.0%, respectively. Thus, according to AIC, there is a 93.8% likelihood that, of our 5 models, TD-AL underlies the dopaminergic signal for most of the population, with a significant portion of these applying a $\gamma_h \neq 0$.
In all cases, TD-AL far exceeds the presence of action-only and TD-RL models, therefore providing strong evidence that, for this population of mice, the dopaminergic signal is more likely to contain an APE than an RPE.

4.4.3 Representing habits

Finally, although TD-AL makes no assumption about how H_a values may influence the selection and intensity of future actions, a complementary exploration of how well the learnt H_a variables were able to describe the real action in the world was undertaken.
Given that TD-AL provided the best fit for the majority of models and that, when $\gamma_h = 0$, H_a is a direct representation of current expected action, a comparison of the resulting H_a

values to the speed in each trial is possible. Fig. 4.7 demonstrates the results for mouse 17, which was the example mouse shown in Greenstreet et al.[3]. This was a visual analysis only and no quantitative tests were taken.

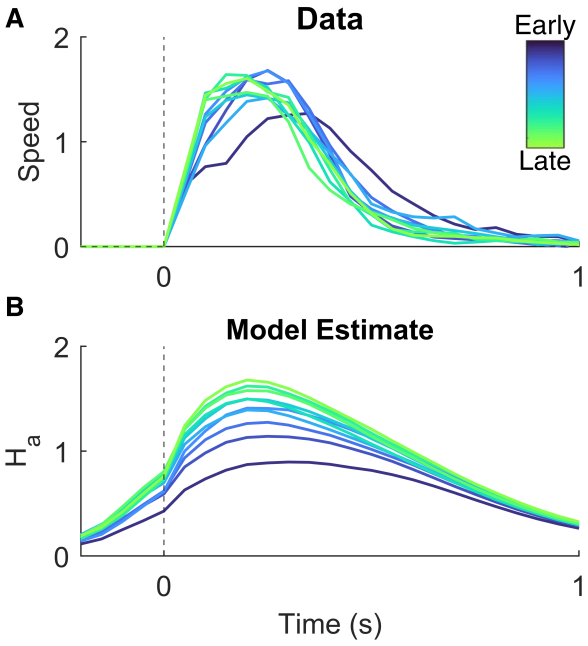


Figure 4.7: Speed and H_a Estimation.
A: The average speed of mouse 17 across 50 consecutive trials from early (blue) to late (green) in training, aligned to action initiation.
B: The average H_a estimate for the same trials as A, produced by TD-AL$_0$ using the best-fitting parameters.

The speed estimation on these trials is not perfect, as determined by the non-zero APE signal and the delay in H_a's decay following the peak speed. Nonetheless, Fig. 4.7 illustrates how, with training, H_a is able to expect action initiation and the time at which the peak speed will be reached.

The slower rate of decay for H_a is due to the microstimulus formulation as the model attempts to replicate a non-Gaussian speed with a series of Gaussian basis functions that have increasingly poor temporal accuracy (Fig. 3.1A) and without new microstimuli being produced at the peak speed. This could potentially be ameliorated by increasing the number of microstimuli produced by each event (currently, $m = 12$); however, this would rapidly become computationally expensive as the number of weights, and thus, calculations, increases proportionally.

Ultimately, it would be beneficial to record corresponding efferent signals from the striatum, to test whether the TS SPN themselves contain such a habit signal, just as value signals have been found in the VS[280].

4.5 Discussion

4.5.1 Summary

This chapter has described the application of the TD-AL model developed in Chapter 3 to real dopaminergic data collected in from the TS of six mice during a two-choice CoT task.

Five models were compared overall: TD-AL and TD-RL (both with and without γ) and an action-only model which did not learn to predict future actions. Across these models, strong preference was found for TD-AL over TD-RL and the action-only model.

The existence of a non-zero γ_h parameter was less apparent for two of the mice, whose BIC values are lower for TD-AL$_0$. However, the significance of this result is weakened when we consider the negligible difference between the TD-AL models' BIC and AIC values for these mice and the fact that TD-AL$_0$ is a nested model within TD-AL$_\gamma$.

In future, repeating this analysis with a greater number of mice and trials would be useful. Of the mice data provided by Greenstreet et al.[3], three were missing a significant number of trials due to a recording error, which rendered it difficult to extract a good-fitting learning model since the data during learning is absent. Further, to confirm that RPEs are present in other striatal regions, an additional control could be completed by applying the same fitting pipeline to data collected in the VS. Unfortunately, as this data was not provided, the proposed analysis could not be undertaken at this stage.

4.5.2 TD-AL performance and limitations

Overall, the results presented in this chapter provide evidence that it *is* possible to model APE-like signals using real neurophysiological data. The simulated APEs are categorically distinct from RPEs and are more than a pure kinematic signal.

With a minimal number of free parameters, TD-AL is able to capture a significant degree of the temporal dynamics in dopaminergic signalling. Even for the mice with a reduced number of recorded trials, Fig 4.5 shows how well this model can qualitatively match the collapse in peak magnitude as well as the forward shift in timing.

Moreover, the fitting procedure uses real behavioural data to represent action intensity, thus reducing the number of assumptions required when fitting the learning model. The simulations from Chapter 3 (Section 3.3) applied a fixed Gaussian curve in each trial and TD-AL was naturally able to learn this structure well, given the Gaussian basis functions which underlie the microstimuli. In contrast, the real speed of the mice is far less stereotyped, with the action duration and acceleration varying between every trial. Impressively, despite this variability, TD-AL was still able to capture the dopaminergic dynamics (Fig. 4.5) and produce a good estimate of future speed (Fig. 4.7).

One key limitation of using TD-AL, as presented in this chapter, is the absence of an observation function. Without the ability to predict how this learning model would influence real action, we are unable to (1) determine how the resultant habit influences subsequent speed and choice and (2) produce surrogate data with which we could assess model and parameter recovery.

This latter issue is particularly pertinent given the difference seen in our BIC and AIC results. Successful model recovery analysis would confirm the degree to which BIC overestimates the influence of additional parameters, while also introducing the possibility to quantitatively determine how the number of trials included alters the confidence with which we can recover the correct model. A deeper discussion regarding the use of penalised fit metrics is provided in Section 7.2.1.2.

However, several factors render the development of such a model intractable. Most pertinently, the continuous nature of this model across time- and action-space necessarily implies that an observation function would not only need to select which choice is made (left port or right), but also be capable of establishing both when an action should be initiated and the temporal dynamics of the resultant action intensity. Further, the goal-directed system must be defined and an assumption made on how the two processes interact to produce actions. The data interrogated in this chapter would be insufficient to extract the best fitting parameters for a goal-directed model and, behaviourally, the mice are too stereotyped to isolate the extent to which habitual learning controls actions. The introduction of additional constraints (for example, by

including reversal trials) would allow for clearer disentanglement of the two processes.
Alternatively, though behavioural surrogate data cannot be produced, a less extensive form of parameter and model recovery could be undertaken by creating simulated PE signals from the behavioural data of each individual mouse across a systematic range of parameters. This work would confirm whether the forward (leftward) shift in dopaminergic peaks across training is sufficient to constrain γ_h and, in concert with a comparison to VS data, would reduce our reliance on BIC and AIC when drawing conclusions regarding the nature of the underlying PE.

A second limitation in fitting models to this study arises from the inconsistent temporal separation between the auditory cues and action initiation.
As the fitting procedure compares TD-AL to a dLight signal which is centred at the start of movement and averaged across 200 trials, the temporal dynamics at cue presentation are lost. Regrettably, the variable delay between the go-cue and action initiation from one trial to another prevented the inclusion of any cue-evoked dopaminergic dynamics since doing so would mask the action-related information.
It would be particularly beneficial to repeat the fitting procedure described in Section 4.3.2 on a dataset which enforces a constant delay between the go-cue and action initiation. The regression analysis completed in the original paper by Greenstreet et al.[3] provides evidence supporting a lack of cue-response in the TS data and, thus, implies an absence of γ_h. This contradicts our results which indicate that most of the datasets prefer TD-AL$_\gamma$. Introducing the cue-responses as an additional constraint on γ_h would improve the reliability of our model recovery and confidence in our parameter estimations.
Despite this, it is encouraging that the γ_h values reported appear to correlate with real information within the signals, which suggests some success in the constraining of their final value. Though a peak at the predictive cue is the clearest indicator of a non-zero γ_h value, in Chapter 3, the simulations of models D and E (Fig. 3.2D and Fig. 3.2E, respectively) demonstrate that the δ_{H_a} dynamics following action initiation are also influenced by γ_h and so, it is possible that these differences may be detectable here.
It is also notable that our γ_h parameters remained lower than those usually reported for

RPE models. This aligns with our discussion in Section 3.3.1.4, which explains that the ability to calculate a continuous average of all expected future actions may confer less of an advantage than the knowledge of what action should be occurring immediately. A γ_h value of 0 represents this argument taken to the extreme. For similar reasons, in their implementation of an APE learning rule, Lee et al.[255] compromised between the two options and set the APE discount factor to 0.5. It is possible that this lowered γ may be sufficient to obscure any cue-associated dopaminergic peaks in the data collected by Greenstreet et al.[3].

4.5.3 Dopamine in the TS

Finally, this chapter aimed to establish whether APE signals exist in the striatum.
As briefly described in Section 4.2, the additional analysis completed by Greenstreet et al.[3] provided strong evidence supporting the existence of a *learning* signal that is associated with action kinematics.
The application of a biologically-plausible TD-AL model has further established that a realistic and mechanistic APE model *is* capable of reproducing these recorded dopaminergic dynamics, using an identical computational unit to that normally assumed for RPEs in the VS. Indeed, TD-AL provides a better fit for this data than an RPE-based model. Both value-based goals and value-free habits can be learnt over continuous time using the same neural pattern of connectivity, simply by altering the information provided to the dopaminergic system.

RPEs and APEs are not the only signals proposed to be contained within TS dopaminergic dynamics. The TS has also been strongly associated with *threat* prediction errors (TPEs) as part of the 'weal and woe' theory of dopamine[104], which posits that the caudal TS learns to respond to aversive stimuli with avoidant actions, in contrast to the approach response to rewards encoded in the more ventro-medial areas.
Greenstreet et al.[3] do report seeing such TPEs during a threat-based looming-stimulus experiment, within the *same* mice that were reported as showing an APE. However, these signals were distinct from the APEs and optogenetic stimulation during this task did not

increase aversive responses, where it had previously been shown to influence movement kinematics during action learning (Section 4.2).
Thus, these variables could feasibly be encoded in a separate subpopulation of dopaminergic cells[132] and their existence is not sufficient to disprove TS-located APE signals, nor do they contradict the use of TD-AL as an explanatory model.

Overall, the data provided by Greenstreet et al.[281] appears to show a real APE contained within dopaminergic data. This signal is described by the TD-AL model over and above both TD-RL algorithms and an action-only non-learning variable.
In the next chapter, we develop an observation model which accounts for the influence of habits on both choice *and* RT and, in Chapter 7, we discuss ways in which this observation model may be extended to utilise TD-AL.

5

Two-Drift Race Diffusion Model

Contents

5.1 Introduction

The third research question for this PhD addresses the influence of habits on both choices and RTs.

As discussed in Section 2.1.3.4, isolating the expression of habits in humans poses a significant challenge. So far, the primary solution has been to influence 'slips-in-action' by constraining the ability of the goal-directed system.

As S-R relationships are beneficial due to their efficiency in both time and cognitive resources required, reducing the availability of said resources should logically increase the expression of these habits while impairing the capacity of the goal-directed system. To use the terms provided by Damaso et al.[216], constraining the time available to react will increase the number of fast 'response-speed' errors. If an S-R association has developed during the task and is no longer appropriate, there should be a period of time for which

these fast errors will be more likely to express the habitual action over other chance errors (induced by a lack of processing time).

In Chapter 6, we will interrogate data from Hardwick et al.[4] who applied precisely this logic to expose human habits and posited that habits will have a stronger effect on actions with shorter RTs. However, Hardwick et al.[4] were unable to test their data with a mechanistic model since, to date, no RL-EAM has been developed which is able to describe a system with multiple (>2) alternative choices and within-trial changes to drift-rates.

The creation of an observation model imparts several advantages, as briefly discussed in the previous chapter (Section 4.5). Specifically, determining how a subject utilises information to guide future actions allows for this aspect to be considered during model fitting. Thus, we can assess the likelihood that a subject's behaviour was generated by the model and the probability that a given action will be made in the future. Further, the observation model can be used to produce surrogate data which enables the analysis of the fitting procedure's reliability.

Inspired by the proposal of a two-process time-dependency in habit expression and the increasing prevalence of RL-EAMs (Section 2.3.4.2), we elected to develop an observation model that could account for the influence of habits on RTs and replicate slips-of-action under constrained time conditions.

In this chapter, the mathematical equations underlying our novel two-drift race diffusion model (TD-RDM) are outlined and a simplified value-free habit learning algorithm is proposed. We then present a near-analytical cost function which can be used to efficiently fit the TD-RDM algorithm to real data. The chapter ends with a discussion of the predictions made by a value-free habit TD-RDM on choices and RTs, alongside potential future extensions to both the learning and observation models.

5.2 The Mathematical Model

This section begins with a description of the mathematical algorithm underlying the two-process race model. Subsequently, a potential value-free habit learning system is

outlined.

Fig. 5.1 provides a schematic framework of both the learning and observation models within a single trial.

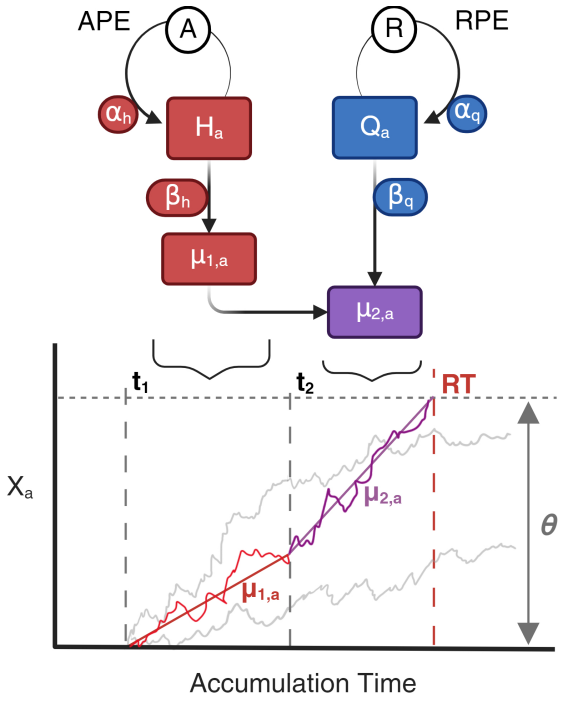


Figure 5.1: The Habit-Race TD-RDM. **Top**: A visualisation of the learning model described in Section 5.2.2. The two drift-rates in a given trial are extracted from approximations of APE-based habits and RPE-based 'action-value' variables. **Bottom**: A schematic representation of the evidence accumulation process in a single trial, using the formulation developed in Section 5.2.1. For each potential action, after a non-decision time, t_1, accumulation begins with a constant drift-rate, $\mu_{1,a}$. If an action has not been selected by t_2, the drift-rate changes to a sum of habit and goal-directed evidence, $\mu_{2,a}$. The first accumulator to reach a threshold value, θ, determines the choice and RT of the trial.

When developing the TD-RDM, we explored trial-by-trial observation models for two reasons.

Firstly, when working in continuous time (as TD-AL does) it rapidly becomes computationally intractable to use perpetually evolving variables (e.g., H_a and Q_a) to predict a subject's choices and RTs (Section 4.5.2). This is in large part due to the absence of any analytical solution which can describe the probability of the action's temporal dynamics alongside choice and RT. Thus, even if such an observation model existed, the processing power required to simulate each timestep within the data would far exceed what is currently available.

Moreover, the measurements that TD-RDM replicates (choice and RT) do not require continuous simulation, as the action intensity itself is not recorded. Each trial contains only two datapoints, so increasing the model complexity confers no advantage and trial-by-trial RL-EAMs are sufficient.

5.2.1 Two-drift observation model

Individually, the multi-alternate forced choice scenario and changing drift-rate requirements have both been solved by different models. The former was explicitly addressed by the RL-RDM[226], while the latter has been developed in an adaptation of the DDM framework[221,282]. However, to our knowledge, no cohesive combination of the two elements has previously been presented.

We elected to extend the RL-RDM framework to include time-dependent drift-rates, rather than working with the RL-DDM for several reasons. Specifically, (1) the within-trial noise is sufficient to account for both fast and slow errors[227], rendering the start-point variability unnecessary, (2) the RT distribution of a single RDM accumulator follows an inverse Gaussian distribution function[226,283], and (3) each accumulator evolves independently of all others. These factors vastly simplify the resulting analytical probability function (Section 5.3) of our TD-RDM.

Mathematically, within a trial, the activity of each accumulator, X_a, drifts according to a Wiener diffusion process. Accumulation with the first drift-rate, $\mu_{1,a}$, begins at a non-decision time, t_1, and switches to a second drift-rate, $\mu_{2,a}$, at a fixed time, t_2, as calculated in Eq. 5.1 and schematically shown in Fig. 5.1.

$$dX_a(T) = \begin{cases} \mu_{1,a}dT + \sigma dW(T) & \text{if } t_1 \leq T < t_2 \\ \mu_{2,a}dT + \sigma dW(T) & \text{if } T \geq t_2 \end{cases} \tag{5.1}$$

where:
- X_a = accumulator activity for action, a,
- $\mu_{1,a}$ = first drift-rate for action, a,
- $\mu_{2,a}$ = second drift-rate for action, a,
- t_1 = non-decision time,
- t_2 = time at which the second process starts,
- T = timestep within the trial,
- $dW(T)$ = Wiener process,
- σ = standard deviation of noise.

As with RDM, the choice and RT of a trial is determined by the earliest time at which any accumulator crosses a shared threshold value, θ.

5.2.2 Habit RL model

The TD-RDM observation function is generalisable and can flexibly be adapted to multiple RL frameworks. Fig. 5.1 portrays one such algorithm, henceforth termed the *Habit-Race* model. This model adapts the value-free system developed by Miller et al.[1] to a simpler, two-equation learning framework, while TD-RDM replaces the explicit arbiter used in the original algorithm (Section 2.4.2).

For this Habit-Race model, the action-value variable, Q_a, and value-free habit variable, H_a, learn from an RPE and APE, respectively, on any given trial (Eq. 5.2 and Eq. 5.3).

Note that an additional condition exists for the goal-directed learning rule (Eq. 5.2) which does not for habit (Eq. 5.3) - the Q_a values only update when the associated action, a, has occurred on that trial, such that $A_a(t) = 1$. This is equivalent to an assumption that the agent only learns from the consequences of the executed choice since no information has been provided regarding the potential value of the unselected action. In contrast, the habits associated with all available actions update on every trial, wherein $A_a(t) = 1$ for the executed choice and $A_a(t) = 0$ for all other potential actions. So the S-R relationship, and thus, H_a, is weakened when the associated response is not made in a given state.

$$Q_a(s, t+1) = Q_a(s, t) + \alpha_q\Big(R(t) - Q_a(s, t)\Big)A_a(t) \tag{5.2}$$

$$H_a(s, t+1) = H_a(s, t) + \alpha_h\Big(A_a(t) - H_a(s, t)\Big) \tag{5.3}$$

where:
- Q_a = estimated value for action, a,
- H_a = habit variable for action, a,
- t = trial number,
- s = state (or, stimulus seen),
- $R(t)$ = reward received on trial, t,
- $A_a(t)$ = action intensity for action, a, on trial, t,
- $\alpha_{q/h}$ = learning rate for Q_a and H_a, respectively.

These values are then translated into drift-rates for the subsequent trial (Eq. 5.4 and Eq. 5.5). In the Habit-Race model, the second drift-rate is the sum of the two variables, modulated by two 'temperature parameters', $\beta_{q/h}$. Thus, for the current state, s, at trial, t;

$$\mu_{1,a}(t) = \beta_h H_a(s,t) \tag{5.4}$$

$$\mu_{2,a}(t) = \beta_h H_a(s,t) + \beta_q Q_a(s,t) \tag{5.5}$$

where: $\beta_{q/h}$ = constant temperature parameter for Q_a and H_a, respectively.

This simple formulation is not an absolute requirement of the TD-RDM and the application of the learnt variables can be made more complicated; for example, through a constant decay rate or by introducing competition between accumulators. Some alternatives are discussed further in Section 5.4.

5.3 A Near-Analytical Cost Function

Simply establishing the trial-by-trial dynamics of the TD-RDM is only the first step in applying RL-EAMs to explain behavioural data. Though this can be used to produce surrogate data, it is also beneficial to be able to calculate the probability of both choice and RT, given the previous experience of an agent. Doing so allows for the model's parameters, and thus, predicted output, to be specifically adapted to match the empirical data of a given subject. This provides the possibility to compare (1) the relative likelihood of various model formulations and (2) quantifiable parameters across subjects and/or conditions.

To increase the efficiency of the fitting procedure, we have devised a near-analytical cost function through which the expected model behaviour can be compared to experimental results.

Specifically, this model can be exposed to two potential types of trial:

1. **Free-RT trials:** The agent is free to make an action with no explicit time constraints, i.e., when the accumulator passes a threshold value.

2. **Time-controlled trials:** The preparation time given to the agent is externally enforced such that the trial's RT is fixed. Thus, its action is determined by whichever accumulator is highest at this point.

The cost function is unique in both of these, as the former requires the calculation of $P(c, RT)$ and the latter, $P(c|RT)$.

5.3.1 Free-RT trials

The first trial type, wherein the agent is told to respond as fast as possible but is free to act whenever it chooses, is the framework classically modelled using EAMs - a stimulus is presented and cognitively processed, before the evidence for all available actions accumulates until the first to reach the threshold activity level is executed. Fig. 5.2A provides a schematic illustration of how this can be mathematically represented.

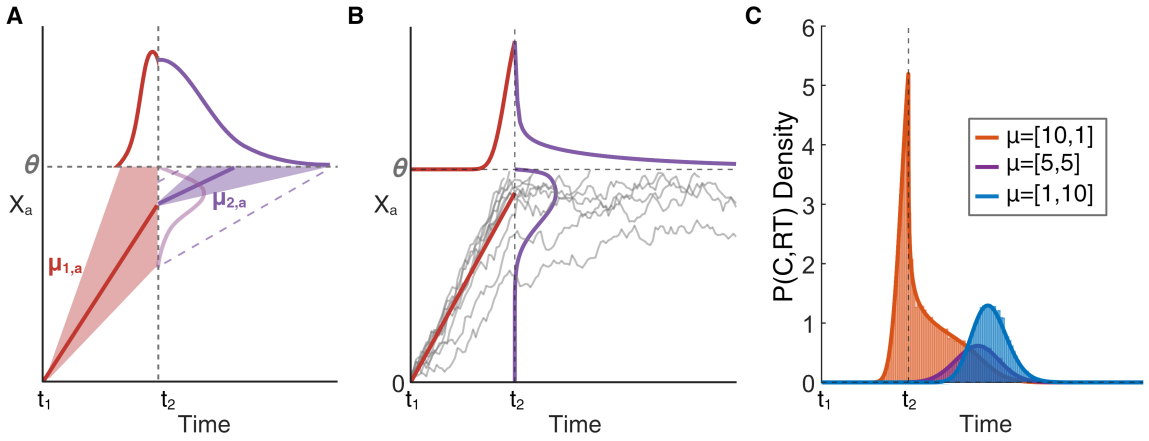


Figure 5.2: TD-RDM Free-RT Trials.
A: A schematic representation of the evidence accumulation process when the agent is free to react at any time. Between t_1 and t_2 (red), the accumulator drifts with a mean rate μ_1 towards a threshold θ. The associated RTs follow an inverse Gaussian distribution. In the second stage (purple), the accumulator drifts at a mean rate $\mu_{2,a}$. The new start point of each accumulator is determined by its activity level at t_2. Their distribution at this time follows the solution to the BVP (Section 5.3.1.2).
B: A simulation replicating a free-RT trial as in A, for 10 trials with the parameter values outlined in Table 5.1.
C: The RT distributions in a three choice context. The accumulator from B (red) competes against two others, purple ($\mu_1 = 5, \mu_2 = 5$) and blue ($\mu_1 = 1, \mu_2 = 10$). The RT distributions from 10,000 simulated trials are well approximated by the analytical solution (solid line).

The probability that accumulator, i, reaches threshold activity level, θ, at time T before all others can be calculated through Eq. 5.6:

$$P(A_i = 1, T = RT) = f_i(T) \prod_{j \neq i} (1 - F_j(T)) \tag{5.6}$$

where: f_i = RT distribution's probability density function (pdf) for action, i,
F_j = RT distribution's cumulative density function (cdf) for action, j.

This equation states that the likelihood of any action can be determined by the probability that the associated accumulator, i, has reached the threshold activity level ($f_i(T)$) while all other accumulators, j, have not ($1 - F_j(T)$).

At its core, this premise is simple enough. However, the complexity arises in providing an analytical solution of the RT distribution for any given accumulator which requires a consideration of the change in dynamics before and after t_2. Specifically, we can redefine $f_i(T)$ and $F_i(T)$ as two separate formulae, such that:

$$f_i(T) = \begin{cases} f_{i,1}(T) & \text{if } t_1 \leq T < t_2 \\ f_{i,2}(T) & \text{if } T \geq t_2 \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

$$F_i(T) = \begin{cases} F_{i,1}(T) & \text{if } t_1 \leq T < t_2 \\ F_{i,2}(T) & \text{if } T \geq t_2 \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

5.3.1.1 Before t_2

During the first stage of accumulation ($t_1 \leq T < t_2$), the TD-RDM distribution is the same as in the standard RDM framework. Anders et al.[283] demonstrated how the first-passage time of a Brownian accumulator diffusing towards a single threshold can be approximated by the 'inverse Gaussian' (or shifted Wald) distribution (Eq. 5.9).

$$T \sim IG\left(\frac{\Theta}{\mu}, \left(\frac{\Theta}{\sigma}\right)^2\right), \quad f_1(T; \mu, \sigma, \Theta) = \frac{\Theta}{\Theta\sqrt{2\pi T^3}} e^{-\frac{(\Theta - \mu T)^2}{2\sigma^2 T}} \tag{5.9}$$

where: T = time to reach threshold,
Θ = distance from start point to threshold ($\theta - z$),
μ = mean drift-rate,
σ = standard deviation of diffusion.

Similarly, the cdf of the inverse Gaussian distribution already has an established analytical solution (Eq. 5.10):

$$F_1(T;\mu,\sigma,\Theta) = \Phi\left(\sqrt{\frac{\lambda}{T}}\left(\frac{\mu T}{\Theta}-1\right)\right) + \exp\left(\frac{2\lambda\mu}{\Theta}\right)\cdot\Phi\left(-\sqrt{\frac{\lambda}{\nu}}\left(\frac{\mu T}{\Theta}+1\right)\right), \quad \lambda = \left(\frac{\Theta}{\sigma}\right)^2 \tag{5.10}$$

where: $\Phi(x)$ = cdf of the standard Normal distribution at x.

For Fig. 5.2A, this period corresponds to the first accumulation time (red) and the pdf function begins by following an inverse Gaussian distribution. Fig. 5.2B illustrates the same formulation, as produced by 10 trials of a single accumulator, using the parameters given in Table 5.1.

Parameter	μ_1	μ_2	θ	t_1	t_2	σ
Value	10	1	4.5	0	0.4	1

Table 5.1: Parameter values used to produce simulations in Fig. 5.2 and Fig. 5.3.

5.3.1.2 After t_2

Following the drift-rate change, calculating the pdf becomes more complex.

Theoretically, each accumulator should continue to produce an inverse Gaussian distribution with the new parameter settings. However, the distance to threshold, Θ, is no longer a fixed value. Instead, it is dependent on the distribution of the accumulator's potential activity at t_2, which is limited by the absorbing boundary at threshold.

Calculating the distribution of a Wiener accumulator's activity as it approaches a single, non-zero, absorbing boundary, θ, is also known as the 'boundary value problem (BVP)'. Thankfully, this already has an established solution[284–286] (Eq. 5.11).

$$\text{bvd}(z,T;\mu,\sigma,\Theta) = \frac{1}{\sqrt{2\pi\sigma^2 T}}\exp\left(\frac{-(z-\mu T)^2}{2\sigma^2 T}\right) - \exp\left(\frac{2\mu\Theta}{\sigma^2}\right)\exp\left(\frac{-(z-2\Theta-\mu T)^2}{2\sigma^2 T}\right) \tag{5.11}$$

where: z = start point of second accumulator,

bvd = boundary value distribution of z.

Thus, we have all the elements necessary to calculate the cdf of our RT distribution following t_2, using Eq. 5.9, Eq. 5.10 and Eq. 5.11, by combining the cdf of the first accumulation period with an integration of the second phase's cdf across all potential remaining distances to threshold (Eq. 5.12).

$$F_{i,2}(T;\mu_1,\mu_2,t_1,\sigma,\theta) = F_{i,1}(t_2-t_1;\mu_1,\sigma,\theta) + \int_{-\theta}^{\theta} F_{i,1}(T-t_2;\mu_2,\sigma,\theta-z)\cdot \mathrm{bvd}(z,t_2-t_1;\mu_1,\sigma,\theta)\,dz \tag{5.12}$$

Unfortunately, determining the pdf of these trials analytically is intractable, as Eq. 5.12 cannot be differentiated. However, the pdf at any RT can be rapidly approximated numerically as follows:

$$f_{i,2}(T;\mu_1,\mu_2,t_1,\sigma,\theta) = \frac{F_{i,2}(T+dt) - F_{i,2}(T-dt)}{2dt} \tag{5.13}$$

The requirement of this numerical approximation is what renders the whole solution presented here *near*-analytical. Regardless, it remains more efficient than the completely numerical method of approximating the RT distribution for every combination of parameters via simulation. Fig. 5.2C portrays the equivalency of these two methods. The near-analytical solution provides a good fit to the simulated RT distributions for each choice.

5.3.2 Time-controlled trials

Calculating the cost function for the second trial type is much simpler as it has an entirely analytical solution. The action selected in these trials is determined by whichever accumulator has the highest activity at the externally enforced response time, and thus, the lack of a threshold removes any need to consider the BVP. As such, the cost function is consistent regardless of RT and amounts to a summation of the Normal distributions for each potential accumulation step. Specifically, at any given RT, the activity of any individual accumulator, X_a, follows a Normal distribution (Eq. 5.14) whose mean and standard deviation are determined by the time elapsed and the drift-rates experienced.

$$X_a(T) \sim \begin{cases} 0 & \text{if } T < t_1 \\ N(\mu_{1,a}(T - t_1), \sigma\sqrt{T - t_1}) & \text{if } t_1 \leq T < t_2 \\ N(\mu_{1,a}(t_2 - t_1) + \mu_{2,a}(T - t_2), \sigma\sqrt{T - t_1}) & \text{if } T \geq t_2 \end{cases} \tag{5.14}$$

where: $N(\mu, \sigma)$ = Normally distributed variable with mean, μ, and s.d., σ.

Fig. 5.3A and 5.3B show the within-trial evolution of an accumulator and associated activity distribution schematically and during numerical simulations, respectively. The same parameter values (Table 5.1) are used for Fig. 5.3B as in Fig. 5.2B, with the clear exception of θ. Fig 5.3B confirms the Normal distribution of a single accumulator across 10 trials (shown), as predicted in Fig. 5.3A.

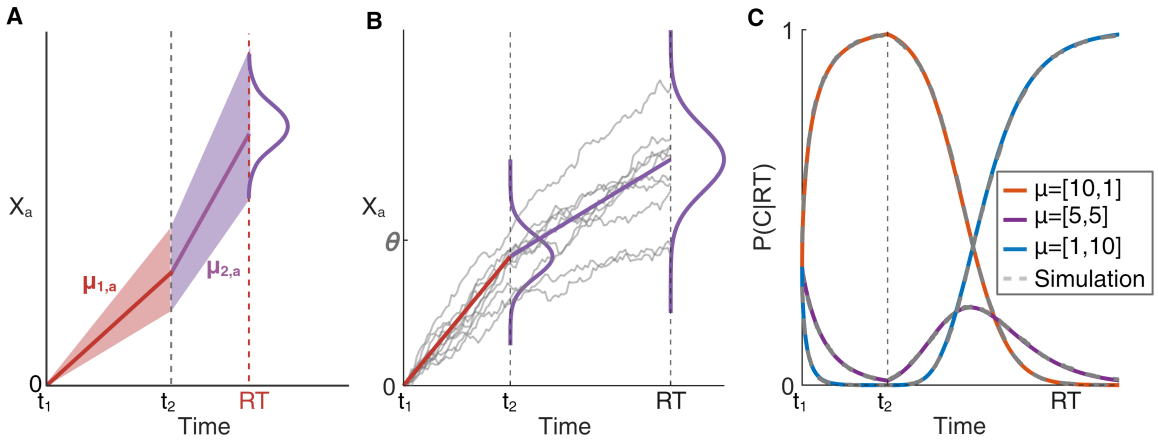


Figure 5.3: TD-RDM Time-Controlled Trials.
A: A schematic representation of the evidence accumulation process when the agent needs to hold responding until a predetermined response time. As there is no threshold, the accumulator's activity follows a Normal distribution throughout accumulation. The mean drift-rate of each accumulation step is $\mu_{1,a}$ during the first accumulation period (red) and switches to $\mu_{2,a}$ following t_2 (purple).
B: A simulation replicating the time-controlled trials as in **A**, for 10 trials with the parameter values given in Table 5.1. The θ value from Fig. 5.2B is included for comparison.
C: The probability distribution of choices across different RTs for the three accumulators in Fig. 5.2C. The numerical solution from 10,000 trials (dashed lines) is overlaid on the analytical solution (coloured lines).

Finally, to determine the likelihood of a given choice being made at the predetermined RT, we must establish the probability that one accumulator has a higher activity than all others at that point in time. This can be calculated through Eq. 5.15.

$$P(A_i = 1|RT = T) = P(X_i(T) > X_j(T)|T) = \int_{-\infty}^{\infty} \phi_i(x) \prod_{j \neq i} \Phi_j(x) dx \tag{5.15}$$

where: ϕ_i = Normal pdf of X_i at time T,

Φ_j = Normal cdf of X_j at time T.

Fig. 5.3C exhibits the probability dynamics in the same three choice scenario provided in Section 5.3.1.1, calculated analytically through Eq. 5.15 and numerically via the simulation of 10,000 trials.

Comparing the dynamics of the red ($\mu_1 > \mu_2$) and blue ($\mu_2 > \mu_1$) accumulators highlights how a two-stage drift system should theoretically replicate slips-of-action in time-constrained conditions. In a scenario where the habitual action and optimal goal-directed choice are different, the former accumulator will have the larger μ_1 and so will approach the threshold rapidly during the first time period. Thus, just as the probability of the red choice approaches 1, any actions made during this time will be biased towards those that have been repeated often in the past.

In contrast, the goal-directed system contributes to the drift-rate at a later timepoint, and so, the latter choice with a strong Q_a but weak H_a will require more time after t_2 to gain on and surpass the activity of the habitual 'error' accumulator. The larger the RT, the more likely the agent is to execute the optimal, goal-directed, action.

5.4 Discussion

5.4.1 Summary

In this chapter, our novel RL-EAM was introduced, a model capable of simulating many choices (> 2) whilst also providing a time-dependent drift-rate.

First, the mathematical equations underlying the EAM framework, extended from Miletić et al.'s RL-RDM[226] were presented. Then, we outlined how one potential RL algorithm with value-free habits (simplified from Miller et al.'s formulation[1]) can be used in conjunction.

The following section describes the near-analytical solution we developed to calculate the probability of any choice and RT for a given trial and parameter combination, regardless of whether (1) the action was made before or after the drift-rate changes and (2) the agent was free to react at any time or had an externally enforced preparation time. As part of this, we confirmed that the novel analytical pdf and cdf equations successfully replicate the numerical approximations produced through simulations. Importantly, these equations work on a trial-by-trial basis, such that the drift-rates can update between trials, as well as changing within, through various RL algorithms.

This chapter concludes with a brief overview of the predictions made for both the general TD-RDM and a more specific habit-based TD-RDM, followed by a discussion of potential alternative frameworks and future extensions.

5.4.2 Predictions of the TD-RDM

Several predictions can be made about the RT and choice distributions produced by a TD-RDM framework.

Some are shared with other RL-EAMs, such as the assumption that, if choices with a greater amount of evidence have larger drift-rates, then they are both more likely to be selected over the alternatives and executed at an earlier RT. Further, RL frameworks tell us that the RT distribution of an optimal choice should speed (left-shift) with training as the drift-rates update and increase.

Additionally, the TD-RDM algorithm treats each accumulator as independent from all others. When the diffusion process considers absolute evidence in this way, we predict that two accumulators with equal levels of evidence will have identical mean drift-rates, and thus, will produce identically speeded RT distributions regardless of whether the other choice is available to execute. This assumption can be contrasted with DDMs, where two actions are always placed in direct opposition and a single accumulator drifts according to the *relative* evidence ($\mu = \mu_a - \mu_b$). If the two options have equivalent evidence, the drift-rates cancel out and the accumulator will take much longer to reach either threshold, forcibly slowing the RTs.

Similarly, the two drift-rates of a given accumulator *within* a trial can be made independent (depending on how they form from the RL algorithm). As such, a strong μ_2 in a TD-RDM framework does not necessarily imply a faster RT if μ_1 remains low. Moreover, when the drift-rates are in agreement and positive, the RT distribution speeds and probability of action execution increases, as the distance to threshold is continuously reduced.

The Habit-Race TD-RDM framework provides two additional predictions. Most pertinently, subjects are more likely to make habitual errors at fast RTs, as posited by Hardwick et al.[4]. This assumption is an extension of the generalised speed-accuracy trade-off and introduces a temporary additional bias towards habitual errors over random errors between t_1 and t_2.

Further, as with all value-free habit models, we can also predict that execution of habitual actions will continually reinforce them. Thus, if an agent is made to react before t_2 repeatedly, we would expect that they will be more likely to make habitual errors in future trials than an agent that regularly acts after a long delay.

5.4.3 Potential alternative formulations

The Habit-Race framework presented in this chapter is vastly simplified and, when appropriate, many additional elements can be introduced to both the RL and EAM systems. Chapter 6 interrogates how adaptations to the learning rules influence the predicted outcomes to determine whether human behavioural data provides evidence of habits.

Despite our upcoming focus on the RL component of TD-RDM, a few key manipulations of the EAM are also available. The simplest of these may be a constant bias term whereby the drift-rate for any action is either increased or decreased by a fixed amount depending on the individual's preference for choosing that action. Equally, the internal motivation of the animal could be modelled using an additive *urgency* signal, U, which is constant across drift-rates but may adapt between trials[287].

Notably, the TD-RDM as presented here does not apply the advantage framework[288] used

by Miletić et al.[226]. This modification introduces direct competition between choices. Rather than each accumulator representing a single choice that drifts according to the absolute evidence, the 'advantage RDM' calculates the drift-rates as a combination of (1) the relative reward expectancy (Q-values) of a given pair of actions, (2) the sum of those Q-values, and (3) a baseline urgency signal.

However, this rapidly becomes complicated for multi-alternate choice scenarios since separate accumulators are required for each pair of available actions. As such, each choice is present in multiple accumulators and an additional assumption must be applied to determine how an action is triggered. Miletić et al.[226] resolved this by modelling action selection as the first choice for which *all* associated accumulators had crossed threshold, with the RT determined by the accumulator with the latest first-passage time. Although there are valid arguments supporting the application of this framework, we did not explore this further since our decision to extend the RDM was specifically motivated by the straightforward inclusion of multi-alternate choices. Despite this, the TD-RDM can easily be extended to apply this framework as the analytical solution is unaffected - the cost function must simply be based on the *slowest* accumulator for any given choice.

Briefly, adaptations may also be made to the way in which the RL algorithm is applied by the EAM. For example, $\beta_{q/h}$ does not need to be a constant term and could be permitted to change both within and across trials. For the former, we can imagine a scenario in which β_h is reduced after t_2 as the agent relies more strongly on the goal-directed evidence.

Adapting β parameters across trials would allow for the introduction of an 'arbiter'-like term[1]. Consistently across models of action selection and habits, *uncertainty* is believed to play a role in their expression (Section 2.1.2.3, Section 2.3.1). The value-free[1] and DopAct[2] models both include this explicitly. Alternatively, a simpler trial-by-trial formulation was proposed by Mikhael and Bogacz[48] wherein the degree to which Q_a (and H_a) fluctuates across trials, $S_{q/h}$, is calculated through biologically-plausible mechanisms which take advantage of the D1R- and D2R-expressing SPNs. This uncertainty term could either influence β directly or be introduced as an additional variable when determining drift-rates.

Finally, the RL update rules themselves can be changed; for example, by introducing a decay term to the Q-values so that they do not remain high when unselected for long periods of time or by employing separate learning rates for positive and negative PEs. The decision on whether to include any of the potential adaptations proposed above rests with the researcher's interest in investigating whether the added complexity would allow the model to better explain key features in the data.

In Chapter 6, we will explore and compare the extent to which four variations of the RL model can explain the real human data collected by Hardwick et al.[4] that inspired the development of this model.

6

Analysis of Human Behavioural Data

Contents

6.1 Introduction

Having developed the TD-RDM in Chapter 5, we are now in a position to tackle our final research question: is it possible to quantify the impact of habits in human behavioural data?

As previously discussed (Section 2.1.3.4, Section 5.1), humans have a larger degree of cognitive control over their actions and (are likely to) understand the experimental tasks better than their rodent counterparts. One consequence of this is that it is much harder to detect habits in humans through extinction tests, regardless of the method used (devaluation, reversal tests or contingency degradation)[82]. Thus, human behavioural research turned instead to manipulating participants' capacity for cognitive control and

interrogating whether habits emerge as a compensatory mechanism.

The novel TD-RDM provides a mechanistic structure with which we can test Hardwick et al.'s[4] proposal that, since the S-R relationship is quicker to process, habits can be used to prepare a response while the slower, optimal goal-directed choice is being evaluated. As a result, the effect of these habits would be entirely masked when a participant is able to wait for the goal-directed information to arrive, irrespective of the amount of training experienced.

In Chapter 5, we specifically demonstrated how introducing temporal constraints to an agent using the Habit-Race model would predict an RT-dependent increase in the number of slips-of-action produced, as was reported by Hardwick et al.[4] during their time-controlled trials.

Now, the TD-RDM is applied to investigate whether Hardwick et al.'s[4] behavioural data contains evidence of two processes with different temporal profiles. The mechanistic model's parameters can be further used to explore how individuals differ in their expression of habits and how this is influenced by factors such as habit strength, the relative weighting of S-R and A-O information, and the time required to process the goal-directed expectations.

This chapter begins with an overview of the experimental methods and results presented by Hardwick et al.[4] in their study investigating the effect of time constraints on habit expression. Following this, four RL-EAM algorithms are outlined, three of which are variations of the TD-RDM presented in the previous chapter. An analysis of the model and parameter recovery is provided before the models are fit to the human behavioural data and the best-fitting parameters are analysed. Finally, we conclude with a brief discussion of TD-RDM's performance and limitations.

6.2 Study by Hardwick et al., 2019

Hardwick et al.[4] interrogated the influence of time constraints on the expression of human habits in a visuomotor association task with a novel paradigm (Fig. 6.1), in 22 participants (initially 24 with two incomplete datasets). The experiment can be broken

down into three core components:

1. **Initial learning**

   In the initial instrumental association task, participants were asked to maximise reward by learning to respond to a visual stimulus with the correct computer key-press. In each trial, one of four potential cues was shown (letters of the Phoenician alphabet) and selection of the associated key would always produce a reward (a pleasant sound), while errors triggered a 'punishment' (1 second delay period with a buzzer noise) before the participant could attempt the trial again.

   The participant group underwent two counterbalanced rounds of (1) minimal and (2) extensive training, with distinct stimuli sets. During the extensive training condition, participants completed an additional 4000 trials over a period of 4 days prior to testing.

   Participant accuracy was determined in a 'criterion phase'; individuals could only progress to the next stage when they correctly responded to five consecutive trials

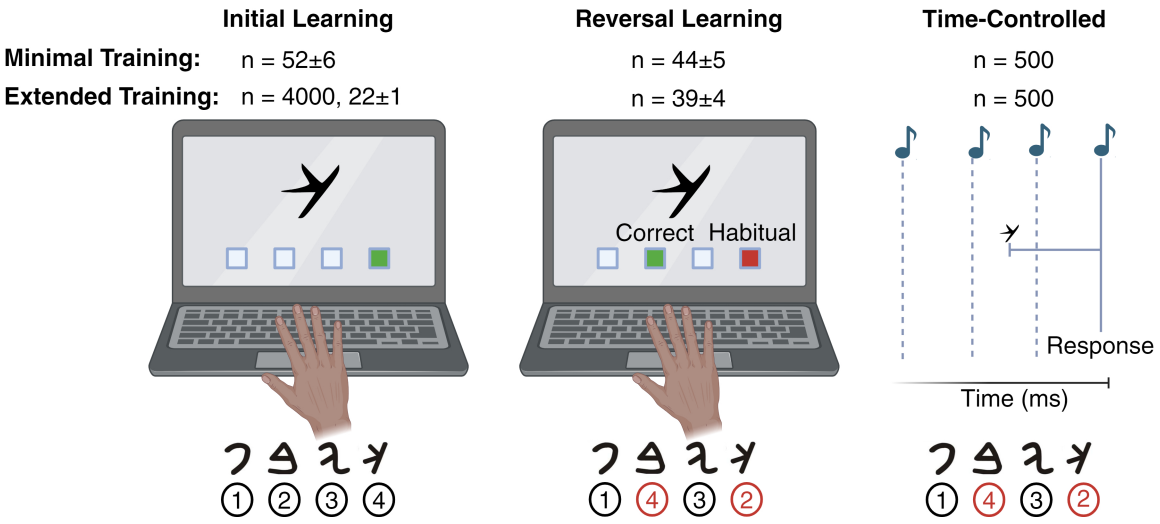


Figure 6.1: Hardwick et al. 2019, Experimental Design.
A schematic overview of Hardwick et al.'s[4] visuomotor association task. Participants learnt a deterministic cue-key response mapping for four stimuli, before the relationship was switched for two of the cues. Once participants achieved a high enough accuracy on the new mapping they completed 500 trials under a time-controlled paradigm. Here, the participant had to respond on the fourth tone, and the overall response time was externally determined by the interval between stimulus presentation and the final tone. All phases of the task were performed twice in a counterbalanced manner, once with minimal training and once with extended experience (an additional 4000 trials).

for each stimulus (thus, creating a minimum requirement of 20 trials). The minimal training condition began at this criterion phase and the task was learnt very quickly by most participants ($n = 52 \pm 6$ in the minimal training condition).

2. **Reversal learning**

   Once participants had reached the required accuracy, the mapping of stimulus to action was reversed for half of the cues. This resulted in two cues with 'consistent' mapping across all trials, and two with a distinguishable 'habitual error' and correct 'remapped' response.

   This stage is an identical criterion phase as for initial learning. Once participants had reached the required accuracy they proceeded to the time-controlled trials. Reversal learning required approximately 40 trials during the extended training condition ($n = 39 \pm 4$), which was statistically equivalent to their performance after minimal training ($n = 44 \pm 5$).

3. **Time-controlled trials**

   The explicit testing of habit expression was performed in the final phase of the experiment. Theoretically, since the reversal training was shorter than initial learning under the extended condition, the habitual S-R associations should not have changed substantially, even when their expression is masked by the updated goal-directed A-O mapping.

   In each trial, four tones were sounded out in a regular rhythm and participants were instructed to respond simultaneously with the fourth. To control the participants' preparation times for each trial, the presentation of the cue occurred at different time intervals in relation to this final tone. The accuracy of actions across the range of response durations was compared between (1) the stimuli with consistent S-R and A-O relationships, (2) the habitual errors made, and (3) the actions that were newly correct for the remapped cues.

As expected, Hardwick et al.[4] reported no statistically significant difference between the minimal and extended training conditions in either the RT measurements or the number of trials-to-accuracy during the reversal phase. They found no evidence supporting the

existence of habits following extended training when the participant is free to respond at an internally generated RT. This result is in line with much of the literature on human habits, which has rarely found a developed resistance to extinction (Section 2.1.3.4). However, measurements during the time-controlled paradigm were markedly different. As shown in Fig. 6.2, behaviour immediately following the initial processing period temporarily reverted to habitual errors (~ 300–600ms), before correcting to goal-directed responses as the preparation time increased. This effect does not occur in the minimal practice condition, nor for the stimuli with consistent cue-key associations.

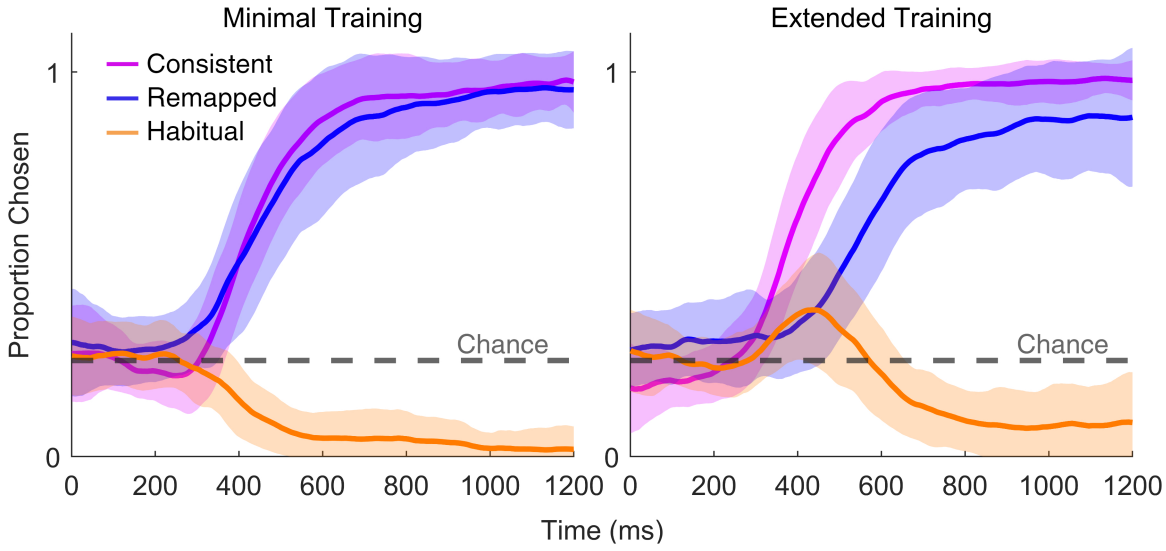


Figure 6.2: Hardwick et al. 2019, Results.
The average behaviour of participant responses during the time-controlled trials in the minimal and extended condition. Each line represents a moving average of how often the associated response was selected as a function of time elapsed following stimulus presentation. Random (non-habitual) errors are not shown.
Minimal Training: Following a non-decision time of ~ 300ms, the proportion of correct responses steadily increases for both the consistent (pink) and the remapped stimuli (blue). The habitual errors (orange) decay to 0 as accuracy increases.
Extended Training: The curves show the same non-decision time and correct responses for consistent stimuli (pink) as in the minimal condition. Between ~ 300 to 600ms, when the remapped stimuli are presented, there is a preference for selecting the habitual error (orange) which delays the increase in correct remapped responses (blue). The saturation point for the blue curve does not always reach 100%.

Hardwick et al.[4] concluded that their results likely arose as a product of two parallel processes with different non-decision times: a faster 'habitual' accumulator and a slower 'goal-directed' one. To test this proposal, they developed their 'response-selection' model (Fig. 6.3).

This probabilistic model makes no assumptions about the mechanistic process underlying the likelihood of an action being selected at a given timepoint. Instead, it posits that the habitual and goal-directed processes work to *prepare* a response and the time required for this preparation may vary between trials (Fig. 6.3A and Fig. 6.3C). In TD-RDM terms, this can be conceptualised as allowing t_1 and t_2 to vary with a Gaussian distribution (Eq. 6.1) whilst removing any drift-diffusion delays.

$$t_1 \sim N(\mu_1, \sigma_1), \qquad t_2 \sim N(\mu_2, \sigma_2) \tag{6.1}$$

Following presentation of a consistent stimulus, once the associated 'preparation time' has passed, the newly planned action has a set probability of occurring, q_1, which was set to 0.95 for Hardwick et al.'s[4] analyses. All other actions are equally likely to happen

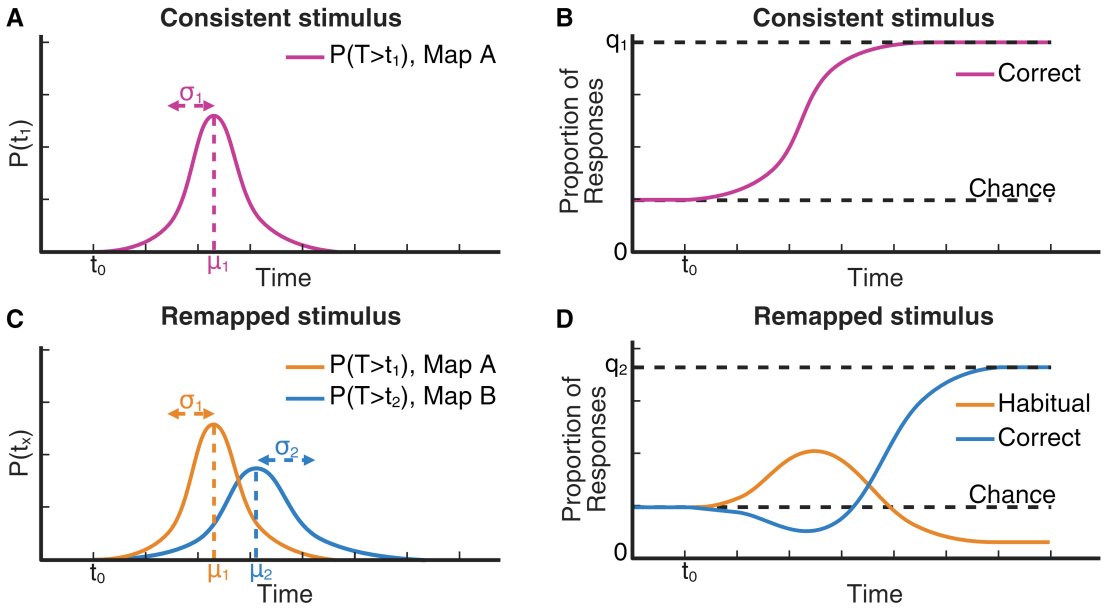


Figure 6.3: The Response-Selection Model.
A: The probability distribution of the response preparation time occurring when a consistent stimulus is shown.
B: The probability of the correct choice being made in response to a consistent stimulus during a time-controlled trial. Accuracy saturates at q_1.
C: As **A**, for remapped stimuli. The parametrisation of t_1 (orange), and thus the habitual error being prepared, is unchanged since it experienced the same training as the consistent mapping. The correct choice is prepared at time t_2, and has a delayed and extended distribution (blue).
D: As **B**, for remapped stimuli. There is an early period for which a habitual error is more likely to occur ($t_1 < t < t_2$), after which the correct response dominates and saturates at q_2. Habitual errors continue to occur as there is a probability, ρ, of the goal-directed action failing to replace habit.

(Table 6.1). Since the prepared actions are non-specific (correct/incorrect, rather than a fixed key), the Gaussian parameters are shared for all stimuli. Therefore, as RT lengthens, the probability that t_1 has occurred increases and so too does the likelihood of the correct response being executed, until it saturates at q_1 (Fig. 6.3B).

In the case of a remapped stimulus, two preparation time curves are produced: one for the habitual error (t_1) and the other for the remapped action (t_2) (Fig. 6.3C). The response-selection model assumes that, as both habitual errors and consistent stimulus actions share the same training experience, the t_1 distribution parameters and probability of action selection (q_1) are identical in both cases.

In contrast, the correct response to remapped stimuli is less established, and so, the preparation time is delayed such that $\mu_{t_1} < \mu_{t_2}$ and $\sigma_{t_1} < \sigma_{t_2}$. The new probability of executing the correct response, q_2, was allowed to vary between participants but not across trials. Hardwick et al.[4] further enforced that t_1 must be less than t_2 on any given trial, so these curves are not entirely independent. Consequently, the proportion of habitual actions being executed is determined by the probability that t_1 has occurred and t_2 has not, which produces the temporary expression of habits seen in Fig. 6.2D.

Interestingly, this model applies an 'all-or-nothing' policy - once the remapped action is prepared, the probability of making a habitual error becomes equal to all other actions (Table 6.1). Late habitual errors are accounted for using an additional probability parameter, ρ, which is superimposed over the response-selection model and determines whether the prepared goal-directed action fails to override the habitual error (i.e., a t_2

Stimulus Seen	**Action**	$\mathbf{t < t_1}$	$\mathbf{t \geq t_1}$	
Consistent	Correct	0.25	q_1	
	Error	0.75	$1 - q_1$	
Stimulus Seen	**Action**	$\mathbf{t < t_1}$	$\mathbf{t_1 \leq t < t_2}$	$\mathbf{t \geq t_2}$
Remapped	Correct	0.25	$(1 - q_1) \cdot 1/3$	q_2
	Habitual Error	0.25	q_1	$(1 - q_2) \cdot 1/3$
	Other Error	0.5	$(1 - q_1) \cdot 2/3$	$(1 - q_2) \cdot 2/3$

Table 6.1: The influence of stimulus and time on the probability of action selection in Hardwick et al.'s[4] response selection model.

failure rate).

Using this model, Hardwick et al.[4] showed that applying this two-curve setting for remapped stimuli after extended training provided a better fit than a single distribution alone for 17/22 participants.

However, though the response-selection framework provides a good visual representation of the data, several core elements are absent in such probabilistic models. Most importantly, they include no information regarding the individual participant's experience nor the process by which the non-response time distributions are formed. Similarly, the all-or-nothing approach implies that, once the goal-directed system has finished preparation, there is no bias towards habitual choices over completely unrewarded errors. This is known to be false due to the incomplete saturation of the correct choices on remapped trials and occasional late-RT habitual errors. The inclusion of the ρ term superimposes this effect with no mechanistic explanation.

With the same number of parameters, the TD-RDM algorithm can account for and quantify these effects. Thus, in this chapter, we extend upon Hardwick et al.'s[4] work to directly test the two-accumulator theory using our mechanistic TD-RDM (Section 5.2).

6.3 Methods

The TD-RDM described in Chapter 5 was developed in order to mathematically simulate the time-dependent accumulation process proposed by Hardwick et al.[4]. In this chapter, this new mechanistic model allows us to address two core questions:

1. Does Hardwick et al.'s[4] data provide evidence for a time-dependent two-process accumulation system?
2. If so, is one of these processes best explained by a value-free S-R association?

This section begins with a summary of the four models that we applied to the data collected by Hardwick et al.[4]. Then, the method used to estimate the best-fitting parameters for a given dataset is outlined. The section ends with a description of the model and parameter recovery measures that are used to refine the fitting procedure ahead of its application on real human data (Section 6.4).

6.3.1 Four potential models

Section 5.4 discussed some potential adaptations to the TD-RDM learning rules and a subset of these were tested here.

The first of these algorithms, the *RL-Race* model, is a simplified version of the RL-RDM framework developed by Miletic et al.[226] and acts as our control (Fig. 6.4A). Habits are not included and all accumulators experience a single drift-rate. This can alternatively be conceptualised as a special case of the TD-RDM where the drift-rate is unchanged at t_2.

Second, a TD-RDM without value-free habits is included (Fig. 6.4B). The *RL2-Race* model instead initially drifts according to a Q-value with a lower learning rate than the latter ($\alpha_{q_1} << \alpha_{q_2}$). This algorithm aligns with previous formulations of model-free habits, such that the secondary process learns from an RPE but is slower to update, thus producing an extended record of previously experienced A-O events.

Finally, we developed two versions of our APE-based Habit-Race model, presented in Chapter 5 (Section 5.2), which differ in their calculation of $\mu_{2,a}$. One, the Habit-Race$_{1\beta}$ model, applies the previously shown additive structure wherein the goal-directed information is superimposed on a constant habit-based drift (Fig. 6.4C). This provides the simplest conceptualisation of two independent processes. However, it brings the unintended consequence that $\mu_{2,a}$ can never be smaller than $\mu_{1,a}$, causing an action with a strong habit to continue to race towards the threshold at the same speed, even when the goal-directed information is in direct opposition.

To counteract this, we introduced the alternative *Habit-Race*$_{2\beta}$ algorithm with a time-dependent β_h value (Fig. 6.4D). Specifically, the degree to which H_a influences the drift-rate changes at t_2 from β_{h_1} to β_{h_2}, both of which are constant parameters. We make no assumptions regarding the relative values of these two temperature terms.

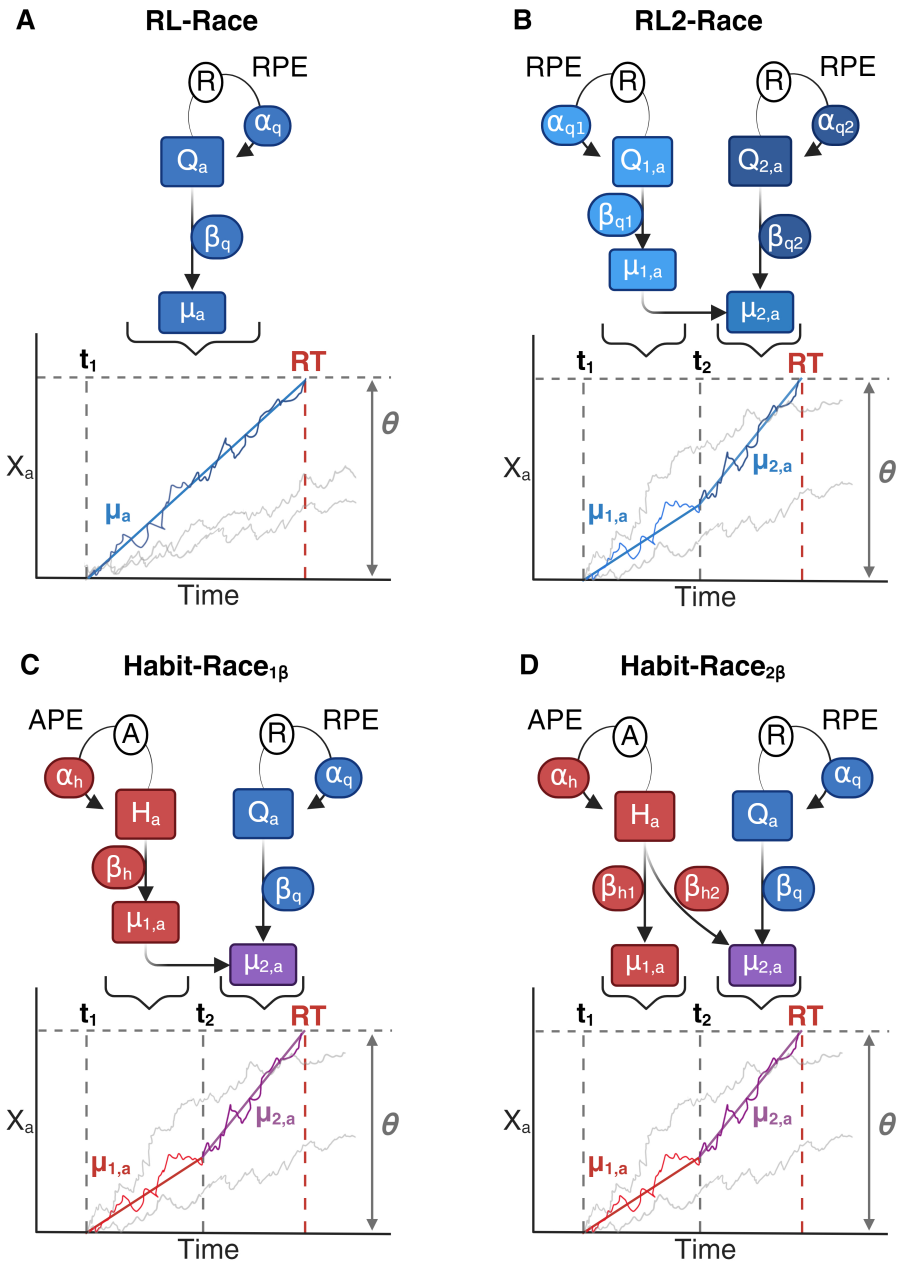


Figure 6.4: Four RL-EAM Models.
Each model is separated into the learning rule (top), and the EAM (bottom). The associated graph portrays a single free-RT trial.
A, RL-Race: a simplified RL-RDM with a single drift-rate. Between each trial, a Q-value of the previously selected action (blue) updates according to the reward received and is then used to determine the drift-rate on that subsequent trial (blue).
B, RL2-Race: a TD-RDM with no value-free habits. The first drift-rate is formed from $Q_{1,a}$ (light blue) which has a slower learning rate than $Q_{2,a}$ (dark blue). These are summed to form $\mu_{2,a}$ (blue).
C, Habit-Race$_{1\beta}$: as presented in Chapter 5, Fig. 5.1.
D, Habit-Race$_{2\beta}$: an extension of **C**, with two value-free habit temperature terms (red). The first, β_{h_1} is used to calculate the first drift-rate, $\mu_{1,a}$ (red), while β_{h_2} controls H_a's contribution to $\mu_{2,a}$ (purple).

For ease of reference, the equations used by all four models are repeated below.

Evidence Accumulation

$$dX_a(T) = \begin{cases} \mu_{1,a}dT + \sigma dW(T) & \text{if } t_1 \leq T < t_2 \\ \mu_{2,a}dT + \sigma dW(T) & \text{if } T \geq t_2 \end{cases} \tag{5.1}$$

Learning:

$$Q_a(s,t+1) = Q_a(s,t) + \alpha_q\Big(R(t) - Q_a(s,t)\Big)A_a(t) \tag{5.2}$$

$$H_a(s,t+1) = H_a(s,t) + \alpha_h\Big(A_a(t) - H_a(s,t)\Big) \tag{5.3}$$

Drift-rates:

	$\mathbf{t_1 \leq T < t_2}$	$\mathbf{T \geq t_2}$
RL-Race	$\beta_q Q_a$	$\beta_q Q_a$
RL2-Race	$\beta_{q_1} Q_{1,a}$	$\beta_{q_1} Q_{1,a} + \beta_{q_2} Q_{2,a}$
Habit-Race$_{1\beta}$	$\beta_h H_a$	$\beta_h H_a + \beta_q Q_a$
Habit-Race$_{2\beta}$	$\beta_{h_1} H_a$	$\beta_{h_2} H_a + \beta_q Q_a$

6.3.2 Fitting procedure

The fitting procedure applied in this chapter calculates the best-fitting parameters of a given model to each participant's dataset through maximum likelihood estimation (MLE) and the use of MATLAB's `fmincon` function (as in Chapter 4). The likelihood of any given trial was calculated using the appropriate TD-RDM cost functions developed in Section 5.3.

We begin by specifying which trials are included in the fitting procedure, before discussing the model parameterisation. Finally, we describe the mathematical details of calculating the negative loglikelihood (NLL) when cost functions differ between trials.

6.3.2.1 Data preprocessing

When developing the fitting procedure, we initially had to decide which trials were relevant and necessary to include.

Trials where the participant either responded before t_1, after 2 seconds, or did not respond

at all were omitted. This exclusion criteria arises from an assumption that such trials are produced by a separate cognitive process or that the participant was not engaging with the task.

The trials with long latencies or no response suggest that the participant was either allowing a much greater degree of cognitive consideration or was not attempting to respond as quickly as they could. In either case, EAMs are no longer appropriate to explain the data. Hardwick et al.[4] applied the same criteria in their analysis.

Removing trials prior to t_1 is less common but is justified specifically when working with time-controlled trials and the TD-RDM. Excluding very rapid trials (e.g., <100ms) is often done under the assumption that these are aberrations, either due to mechanical errors in recording or because the participant had already prepared a response before the stimulus was seen (hence, as above, was not applying an EAM)[289,290]. Classically, the non-decision time can then be constrained by the fastest of the remaining trials so no further exclusion is necessary. However, as we will explore further below (Section 6.3.2.2), Hardwick et al.'s[4] time-controlled paradigm introduces multiple trials with responses before t_1, and so, before any accumulation can occur, thus obscuring parameter recovery.

While all remaining trials are included during the learning process, they do not necessarily confer equal amounts of information regarding the underlying model. Specifically, the extended condition of the Hardwick task comprises of roughly 4000 trials of very deterministic data that vastly outnumbers all other trial types while providing very little information with which the parameters can be constrained.

Across all participants and conditions, 50 trials were sufficient for an individual to fully learn the task. Therefore, to reduce the chance of over-representing this deterministic subset of data, we elected to omit all but these first 50 trials from the initial learning phase of the extended condition. No further trials were excluded from fitting in the reversal or time-controlled phases. With this approach, we are still able to determine which parameters will best replicate the *learning* behaviour of both the initial and reversal mapping, as well as the responses to time-controlled trials.

Lastly, during the free-RT trials, when participants made a mistake, they were allowed

to try again following a punishment delay period. These further attempts provided the participants with additional information regarding the stimulus-key mapping and so were included during the learning process. However, since the stimulus remained consistent, we assume that preparation for the subsequent response occurred during the punishment phase and could not be used for fitting the EAM.

6.3.2.2 Model parameterisation

All but two of the parameters included in this model are allowed to freely vary during fitting.

The first fixed parameter is the standard deviation of noise in the Weiner process, σ, which is typical when fitting diffusion models[227]. The absolute values of threshold, drift-rates and noise only hold value in relation to each other, since it is their ratios which determines the final behaviour of the model. As such, one of these values must be held constant in order for any fitting procedure to be successful and σ is classically fixed at 1[227]. This has the added benefit of simplifying many analytical equations.

The second fixed parameter, t_1, requires a heuristic estimation due to its latent nature during free-RT trials and our trial exclusion criteria affecting the time-controlled responses. As briefly touched on, the estimation of a non-decision time is often constrained by a participant's earliest RT. However, humans rarely act before t_2 during the free-RT trials, instead waiting for the goal-directed processing to complete before acting. Equally, during the time-controlled trials, participants are *required* to act before t_1 and TD-RDM assumes that actions are selected with equal probability during these trials. As a consequence, if t_1 is allowed to freely vary and these trials are included during fitting, then a strong local minima is produced when t_1 collapses to the largest possible value. This is further exacerbated when these trials are excluded as it will always be optimal for the final solution to make this value as large as possible at the expense of the overall model accuracy. Thus, t_1 must be held constant and the earlier trials can be removed without consequence.

We tested several potential methods to estimate t_1 as the nature of TD-RDM's masked

habit process impairs our ability to constrain this parameter's value.

Theoretically, t_1 represents the earliest moment in which a participant uses information to guide their actions, i.e., when the time-controlled responses deviate from random selection. Ideally, this could be extracted from the standard deviation of the response curves shown in Fig. 6.2, or in the mutual information between stimulus and response at a given timepoint[291]. Unfortunately, for this dataset, the number of trials during the time-controlled phase is insufficient to calculate either of these measures with any accuracy. Instead, we must rely on a moving average when plotting the time-controlled data. As a result, the value of these measures at any timepoint is strongly dependent on the window used: too large and the time-controlled curves are too generalised to pinpoint an exact t_1, while if it is too small, the resultant plots are pure noise.

Ultimately, we estimate t_1 thanks to a specific feature of participant's responses that emerges during the extended condition only. As behaviour remains deterministic after 3000+ repetitions, participants increasingly express occasional trials with very rapid responses. These trials are still within the range of realistic EAM responses (RT > 0.2s)[289,290] and largely match the participant's accuracy in surrounding trials, which implies that visual processing of the stimulus is still occurring and that these aren't periods of random fast presses.

Under an assumption that t_1 represents the insurmountable limitation in visual processing and mechanistic execution of key-presses, then the shortest of these rapid RTs can be used as a proxy heuristic. Thus, t_1 is fixed to the fastest RT with a correct response in the final 1000 trials of early learning. We further specify that only the first action after presentation of a stimulus is included, for the same reasons outlined above when excluding post-error trials from fitting. Additional analyses confirm that model and parameter recovery are both robust to errors in t_1 (Appendix A).

Our heuristic is specifically appropriate to the exact Hardwick et al.[4] set-up. Other experimental paradigms may be able to estimate t_1 differently, particularly if there are a sufficient number of trials to accurately determine when the mutual information between stimulus and response first increases.

The limits of the remaining free parameters (α_x, β_x, θ and t_2) are summarised in Table 6.2. Note that α_h has very strict constraints because an APE-based H variable is strongly self-reinforcing; i.e., if the habit rapidly becomes very strong, the agent never explores other options and is unable to learn the task. We further enforce that α_{q_1} must be smaller than α_{q_2} for RL2-Race, to ensure that the Q_1 variable is slower to update.

Parameter	α_q/α_{q_x}	α_h	β_x	t_1	t_2	θ	σ
Limits	0-1	0-0.005	0-100	Est.	t_1 - 0.6	0.1-100	1

Table 6.2: Limits placed on the value of parameters during MLE parameter estimation.

6.3.2.3 Calculating goodness-of-fit

The trial-by-trial nature of the TD-RDM ensures a simplistic learning process. For all datapoints included in the learning simulation, the agent is provided with the stimulus seen, choice made, RT and outcome experienced. From this, the model's learning rules extract all the information required to update the Q_a and (when included) H_a variables for that trial.

Once learning is complete, these variables can be used to calculate the associated drift-rates and, thus, the likelihood of an action occurring at a given timepoint for each relevant trial. Section 5.3 described the development of the TD-RDM cost functions in detail, but their final equations are repeated here for reference. The likelihood for either a free-RT trial, L_f, or a time-controlled trial, L_c, are given by Eq. 5.6 and Eq. 5.15, respectively.

$$L_f(t) = P(A_i = 1, T = RT) = f_i(T) \prod_{j \neq i} (1 - F_j(T)) \tag{5.6}$$

$$L_c(t) = P(A_i = 1 | RT = T) = \int_{-\infty}^{\infty} \phi_i(x) \prod_{j \neq i} \Phi_j(x) dx \tag{5.15}$$

The calculation of the *overall* likelihood that a complete dataset was produced by a given model and set of parameters is complicated by the inclusion of two trial types. More specifically, not only do the cost functions for the free-RT and time-controlled trials differ, but they also operate on a different number of trials.

To account for these differences when calculating NLL, the free and time-controlled trials are summed separately, before being weighted by an additional parameter, w_c (Eq. 6.2).

$$NLL = -\left((1 - w_c)\sum_{t=0}^{n_f}\log(L_f(t)) + w_c\sum_{t=0}^{n_c}\log(L_c(t))\right) \tag{6.2}$$

where: NLL = negative loglikelihood of data given the model and parameters,
w_c = weighting towards time-controlled trials,
n_f = number of free-RT trials included in fitting,
n_c = number of time-controlled trials included in fitting,
L_f = likelihood of a free-RT trial,
L_c = likelihood of a time-controlled trial.

The fixed weight, w_c, can be optimised through systematic testing in order to maximise model recovery (Section 6.3.3.3). In the next section, we outline several quantifiable measures that describe the efficacy of the fitting procedure.

Finally, as in Chapter 4, once the minimal NLL has been extracted, BIC analysis can be used to assess which model best describes the data. However, as NLL is an explicit likelihood, rather than an SSE approximation (Section 4.3.2), we return to the generalised BIC and AIC equations[275] (Eq. 6.3 and Eq. 6.4). Thereafter, group BMS analysis (Section 4.4.2) is completed to determine the probability that a given model is the most prevalent in the population.

$$BIC = 2 \cdot NLL + p\log(t) \tag{6.3}$$

$$AIC = 2 \cdot NLL + 2\log(p) \tag{6.4}$$

where: p = number of free parameters in the model,
t = number of trials included in fitting.

6.3.3 Recovery analysis

This subproject includes a focussed analysis of our model and parameter recovery methods. By assessing the reliability and accuracy of the fitting procedure, not only can

we better establish our confidence in the results, but we are also able to optimise the w_c weighting parameter rather than using an arbitrary assumption.

6.3.3.1 Producing surrogate data

Performing parameter and model recovery analysis first requires the creation of *surrogate data* via simulation. These datasets are created using the same experimental paradigm as was experienced by the participants.
Since the true underlying model and parameters are known, we are able to gain a measure of how reliably the parameter estimation and model selection provided by our fitting procedure can be trusted.

For each potential model, 100 agents were randomly assigned a set of parameter values from a fixed range (Table 6.3) which were confirmed to produce realistic results. Specifically, simulations run with parameters outside this range were either incapable of (1) learning the task at all, (2) completing either criterion phase within a realistic number of trials or (3) producing RT distributions similar to those recorded by Hardwick et al.[4]. Again, we further specified that α_{q_1} be smaller than α_{q_2} for the RL2-Race model datasets.

Parameter	α_q/α_{q_2}	α_{q_1}	α_h	β_q/β_{q_2}	β_{h_x}/β_{q_1}	t_1	t_2	θ	σ
Minimum	0.1	0.001	0.001	5	1	0.2	t_1	2	1
Maximum	0.5	0.25	0.005	13	5	0.4	0.6	5	1

Table 6.3: The parameter values used to produce surrogate data of the Hardwick task.

The datasets themselves were simulated in two stages to replicate the minimal and extended condition datasets following the sequence described in Section 6.2 and Fig. 6.1. The free-RT and time-controlled trials differed in how the choices and RTs were generated, but not in their learning rules. For any given agent, the simulation of each trial follows a set series of steps:

1. A stimulus, s, (1-4) is presented.
2. The drift-rates of four accumulators are calculated using the appropriate model (as above) and the associated $Q_{a,s}/H_{a,s}$ values for the stimulus, s.
3. The accumulators drift for 2 seconds (dt = 0.001) according to Eq. 5.1.

4. A choice is made ($A_a(t)$ is set):
   (a) In a free-RT trial, the first accumulator to pass threshold is determined and the associated action and RT is returned.
   (b) In a time-controlled trial, the trial duration is selected from a uniform distribution ($T \sim U(0, 1.8)$) and the accumulator with the highest value at that timepoint determines the action returned.
5. If the selected action is correct for the current cue-key mapping, a reward is given ($R(t) = 1$), otherwise, $R(t) = 0$.
6. The model's $Q_{a,s}/H_{a,s}$ values for that stimulus update according to Eq. 5.2 or Eq. 5.3.

Once the time-controlled trials for the minimal condition are completed, the RL variables for all actions and stimuli are reset to $H_{a,s} = 0$ and $Q_{a,s} = 0.5$ before the simulation of extended condition trials begins. The start values for Q_a ensure exploration at an early stage and represent an assumption that, when an agent is in a new environment, it believes all actions have an equal potential to produce a reward.

6.3.3.2 Measuring recovery

Parameter estimates and NLL values can be extracted for each of the surrogate datasets by applying the fitting procedure as though we were working with real behavioural data. Using these results, parameter recovery is straightforward to assess - the correlation between true and estimated parameter values should approach a perfect positive correlation as the accuracy increases.

It is worth noting that there is a known difficulty in constraining associated α and β terms when fitting RL-EAMs, as they largely trade-off against one another[191]. Instead, it is standard to assess recovery of their product, since it is their *relative* values that are detectable within the data.

Model recovery is usually quantified in two ways, through a *confusion matrix* and/or an *inverse matrix*[191] (Fig 6.5A). For each surrogate dataset, BIC analysis is used to select the best-fitting model. By comparing the true model, M_T, with the recovered model, M_R, we can calculate both the $P(M_R|M_T)$ and the $P(M_T|M_R)$, for every model. The former

of these is the most commonly reported in a confusion matrix, while the latter inverts the confusion matrix to provide a measure of our confidence that the recovered model truly is underlying the data.

As the accuracy and reliability of model selection increases, the mean value of the matrices' diagonals will also increase, thereby allowing us to compare a single measure when discussing recovery efficiency.

6.3.3.3 Estimating w_c

We applied this surrogate data recovery to determine the optimal value for w_c, the relative weight parameter that controls the bias towards time-controlled trials in the NLL function

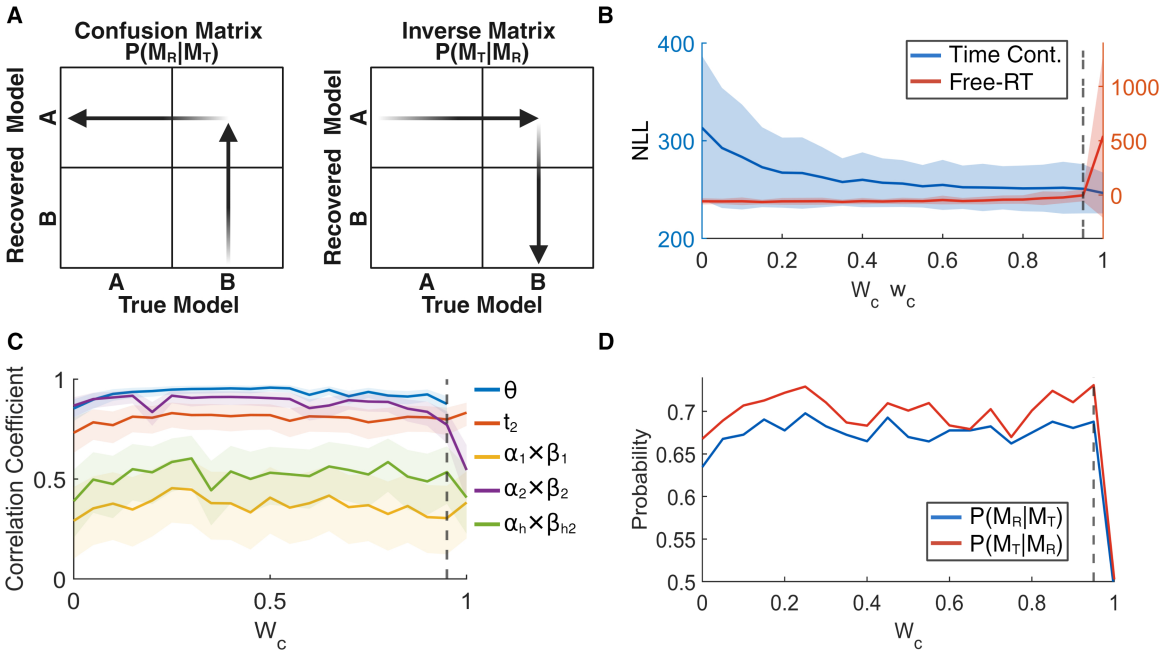


Figure 6.5: Optimising w_c.
A: A diagrammatic representation of the information contained within confusion and inverse matrices. The arrows begin at the information given and point to the model in question. The co-ordinates at which they meet is where the respective probability is recorded (e.g., $P(A_{\text{rec}}|B_{\text{true}})$ and $P(B_{\text{true}}|A_{\text{rec}})$, for the confusion and inverse matrix respectively).
B: The varying impact of w_c on recovery of the free-RT trials (red) or time-controlled trials (blue). The plots provide the average NLL value across models ($\pm s.d.$). A w_c of 0.95 is indicated by a dashed line.
C: As **B** for the average correlation coefficients for each type of parameter. The product of α and β has been subdivided into those used for variables in the first accumulation period ($\alpha_1 \times \beta_1$, H_a and $Q_{1,a}$) and the variables that are introduced at t_2 ($\alpha_2 \times \beta_2$, Q_a and $Q_{2,a}$). The β_{h_2} term used by Habit-Race$_{2\beta}$ is further separated.
D: As **B**, for the diagonal sum of the confusion matrix (blue) and inverse matrix (red) across w_c.

(Eq. 6.2). By running the fitting procedure many times with w_c ranging from 0 (free-RT trials only) to 1 (time-controlled trials only), in steps of 0.05, we are able to directly compare improvements to both parameter and model recovery. As these simulated datasets did not include the subset of rapid trials we will use to estimate t_1 in the human data, we assumed a perfect recovery and set t_1 to its true value for the purposes of selecting w_c. The results of this analysis, and the justification for our w_c value, are described in this section.

First, we assessed whether inclusion of both trial types was appropriate. Comparison of the raw NLL data (Fig. 6.5B) shows that, unsurprisingly, w_c = 0 dramatically worsens the model's ability to explain behaviour in the time-controlled trials, and vice-versa for w_c = 1. Therefore, looking at the time-controlled data alone, as in Hardwick et al.[4], is insufficient to describe habitual behaviour when realistic learning processes are considered. Equally, parameter recovery largely saturates for w_c values between 0.05 and 0.95 with impairments at the extremes, particularly for $\alpha_2 \times \beta_2$, which encompasses $\alpha_q \times \beta_q$ and $\alpha_{q_2} \times \beta_{q_2}$ (Fig. 6.5C).

Model recovery also saturates across the central w_c values, though two peaks in $P(M_T|M_R)$, our inverted confidence metric, emerge at w_c = 0.25 and 0.95 (Fig. 6.5D). The latter of these is minimally larger at a mean recovery of 0.73 across all four models. Therefore, as parameter recovery at w_c = 0.95 is acceptable, this setting shall be used for all future analyses. It is worth noting that this does not necessarily represent a prioritisation of the time-controlled data, but rather a compromise to directly compare probabilities and probability densities, the latter of which can reach much greater values given that it is not constrained between 0-1.

6.4 Results

6.4.1 Surrogate data

This section provides analysis of the parameter and model recovery produced by applying the optimised fitting procedure. These results will then be used to explore the relationship

between the expression of habits and the best-fitting parameters extracted for the Habit-Race$_{1\beta}$ model under Hardwick et al.'s[4] experimental paradigm.

6.4.1.1 Parameter and model recovery

Across all four models, parameter estimation is largely successful. As shown in Fig. 6.6, the correlation between the true parameter and the best-fitting parameter returned by our fitting procedure is strong for θ, t_2 and the Q-learning parameters applied during the second accumulation period. This effect is weaker for RL2-Race, but we deemed the correlation strength to still be acceptable for the analyses completed in this chapter.

Unfortunately, recovery is not as successful for the RL parameters which influence the strength of $Q_{1,a}$ and H_a. The reasons for this are twofold. First, the absolute values of these parameters are very small, and `fmincon` is often less accurate at this scale. Second, these parameters most strongly affect a very small subset of the data - the time-controlled

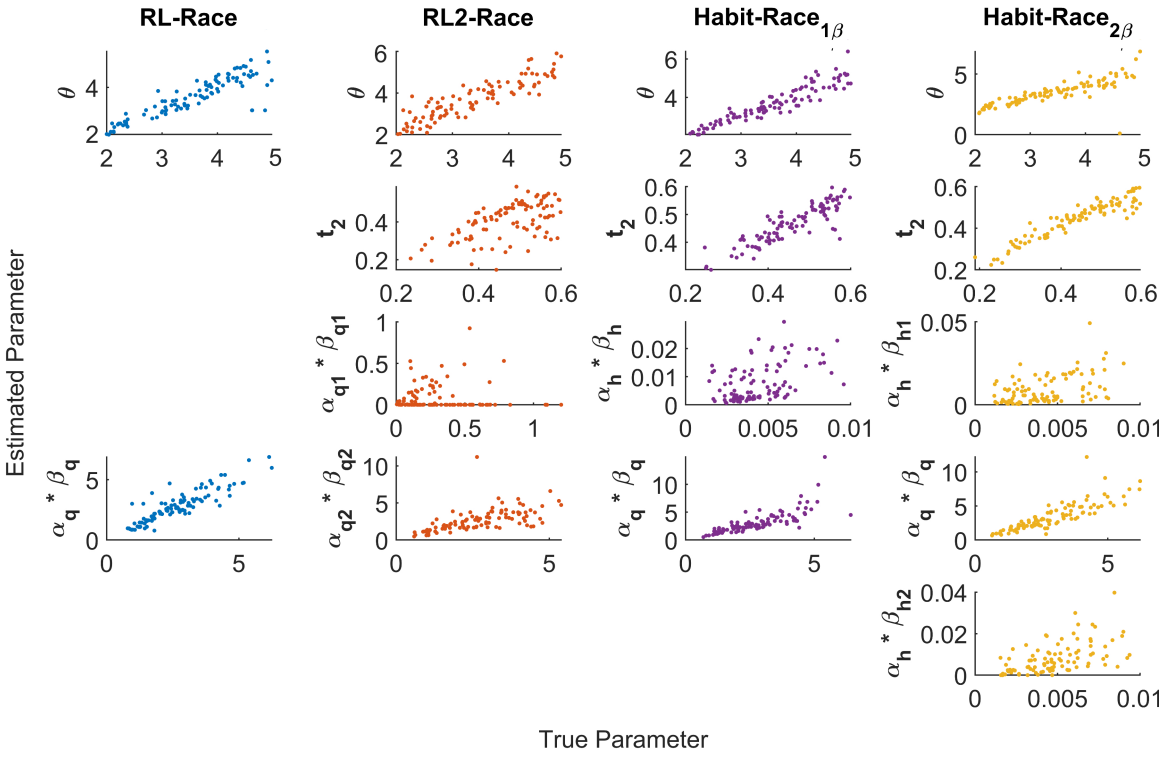


Figure 6.6: Parameter Recovery Analysis.
The correlation of true parameters (x-axis) against the recovered parameters (y-axis) for all free parameters. Each column represents data produced by one of the four separate models, from left to right: RL-Race, RL2-Race, Habit-Race$_{1\beta}$, Habit-Race$_{2\beta}$.

trials with an RT between t_1 and t_2. Any trials with a longer RT can largely compensate for small differences in the poorly recovered parameters with equally small changes in the later, larger RL parameters (α_{q/q_2} and β_{q/q_2}). This is similar to the known redundancies between α and β for RL-EAMs[191].

Overall, we deemed our parameter recovery to be sufficient to continue with the current fitting procedure, though the accuracy of a given parameter will be considered when discussing results on human data in later sections.

The confusion and inverse matrices (Fig. 6.7) provide further insight into our confidence that the fitting procedure and BIC analysis selects the correct model.

Perhaps unsurprisingly, the two habit models are difficult to distinguish from each other.

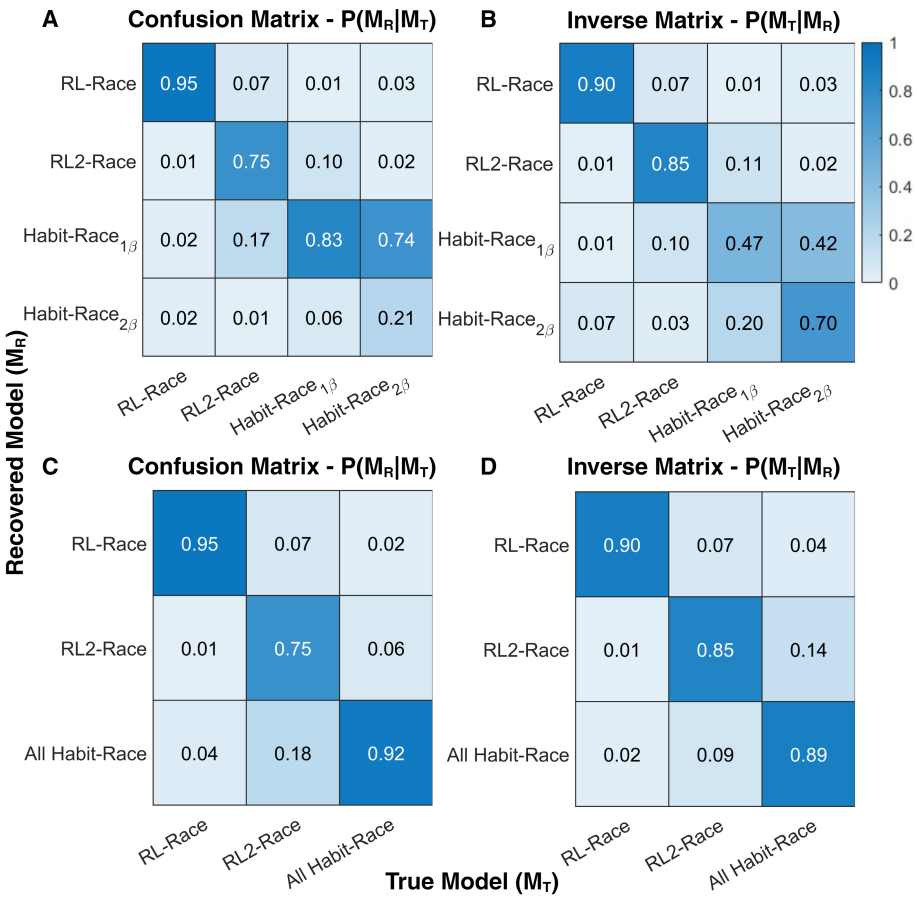


Figure 6.7: Model Recovery Analysis.
A: Confusion matrix for recovery of all four models when $w_c = 0.95$, organised as in Fig. 6.5A.
B: Inverse matrix for recovery of all four models when $w_c = 0.95$, organised as in Fig. 6.5A.
C: As **A**, but the two Habit-Race models have been grouped together.
D: As **B**, but the two Habit-Race models have been grouped together.

They differ by a single parameter and Habit-Race$_{1\beta}$ is nested within Habit-Race$_{2\beta}$ - if β_{h_2} approaches β_{h_1}, then the two systems become equivalent. Despite this, Fig. 6.7A and 6.7B show that it is possible to detect the existence of multiple drift-rates and to determine whether the former drift-rate uses an APE.
This separation of habit-based drift from an RPE variable is made more apparent if the two habit models are grouped together, as in Fig. 6.7C and 6.7D. The probability that the recovered model is accurate is over 85% for all three model iterations.

As with parameter estimation, RL2-Race is the least-well classified of the models, with 18% being recovered as one of the two Habit-Race systems. More importantly, the inverse matrix in Fig. 6.7D shows that 14% of datasets recovered as RL2-Race are actually produced using an APE-based drift-rate. The total confidence of the fitting procedure is high, but these values will be considered when drawing future conclusions regarding participant data.

6.4.1.2 Estimating habits with mechanistic parameters

Finally, this surrogate data can be applied to assess the degree to which different parameters influence habit expression in time-controlled trials. Specifically, we can produce a linear regressor that predicts 'habit strength' using the estimated parameters extracted during fitting. This exploratory analysis (1) allows us to confirm that the best-fitting parameters are capturing the behaviours of interest and (2) indicates which parameters are of particular interest when quantifying human habitual behaviour.

As the 'habit strength' is a latent property, we must rely on proxy measures. Two interrelated and quantifiable consequences of habit expression were extracted from the time-controlled trials following extended training: (1) the maximum height of the habitual error curve, R_h, and (2) the maximum separation between the consistent and remapped curves, R_s (Fig. 6.8A). In combination, these observable measures encompass the relative strength of H_a and Q_a whilst accounting for influence of other drift parameters. For conciseness, rather than producing two regression models for these highly correlated values, a principal component analysis (PCA) was completed to extract

their latent shared relationship. In accounting for both measures, the final regression model allows us to interrogate a generalised influence of 'habits' overall, which may be further advantageous for any future work that can apply the Habit-Race model but is unable to specifically record R_h and R_s.

From the 200 datasets produced by either Habit-Race model, we excluded 15 whose β_h parameter exceeded 30 as outliers before the PCA components were calculated. The first of these, PC1, explained over 91% of the variance in the proxy measures, with coefficients of 0.629 and 0.778 for R_h and R_s, respectively. The resultant relationship is shown in Fig. 6.8B. Thus, we interpret PC1 as a measure of the latent habit strength in the data.

Next, α_q, α_h, β_q and β_h were selected as predictors together with an additional term, t_h, that is formed from the difference between t_1 and t_2 to provide the duration of the first accumulation period. Since the habit strength proxies were extracted from the time-controlled data, θ was omitted. These variables were supplemented with five interaction terms, representing the relationship within the Q_a or H_a parameters ($\alpha_x : \beta_x$), the innate duration of drift ($\beta_x : t_h$) and the relative contributions of Q_a and H_a ($\beta_q : \beta_h$). The Habit-Race$_{1\beta}$ parameters were chosen for this investigation as the majority of the Habit-Race datasets were recovered as this simpler model and the introduction of β_{h1} and β_{h2} would increase the number of terms by 50%.

The regressor's final variables are summarised in Eq. 6.5 and Table 6.4.

$$\text{PC1} \sim k + \alpha_q + \alpha_h + \beta_q + \beta_h + t_h + \alpha_q : \beta_q + \alpha_h : \beta_h + \beta_q : t_h + \beta_h : t_h + \beta_q : \beta_h \tag{6.5}$$

where: PC1 = latent habit strength in extended training time-controlled trials,
t_h = duration of the first accumulation period $(t_2 - t_1)$,
k = intercept.

The final regression model was statistically significant ($R^2 = .64$, $F(10, 174) = 31.5$, $p < 0.000$) and capable of predicting PC1 (Fig. 6.8C). Of the eleven terms in Eq. 6.5, only three were found to significantly affect habit strength: α_h, t_h and $\alpha_h : \beta_h$. The significance of these three parameters is intuitively predictable as they have the greatest impact on H_a

Term	Coefficient	Standard Error	t-statistic	p-value
k	-0.149	0.158	-0.942	0.348
α_q	-0.0368	0.197	-0.187	0.852
β_q	-0.0127	0.0139	-0.911	0.363
$\boldsymbol{\alpha_h}$	**-74.9**	**11.8**	**-6.36**	**1.74×10^{9}**
β_h	-0.0103	0.0182	-0.563	0.574
$\mathbf{t_h}$	**1.44**	**0.531**	**2.71**	**0.00741**
$\alpha_q : \beta_q$	-0.0109	0.0196	-0.558	0.578
$\boldsymbol{\alpha_h : \beta_h}$	**25.7**	**3.23**	**7.96**	**2.13×10^{-13}**
$\beta_q : t_h$	-0.00362	0.0490	-0.0738	0.941
$\beta_h : t_h$	0.0808	0.0511	1.58	0.116
$\beta_q : \beta_h$	-7.05×10^{-4}	0.00128	-0.549	0.584

Table 6.4: Linear regressor coefficients when estimating habit strength. Significant terms are highlighted in bold.

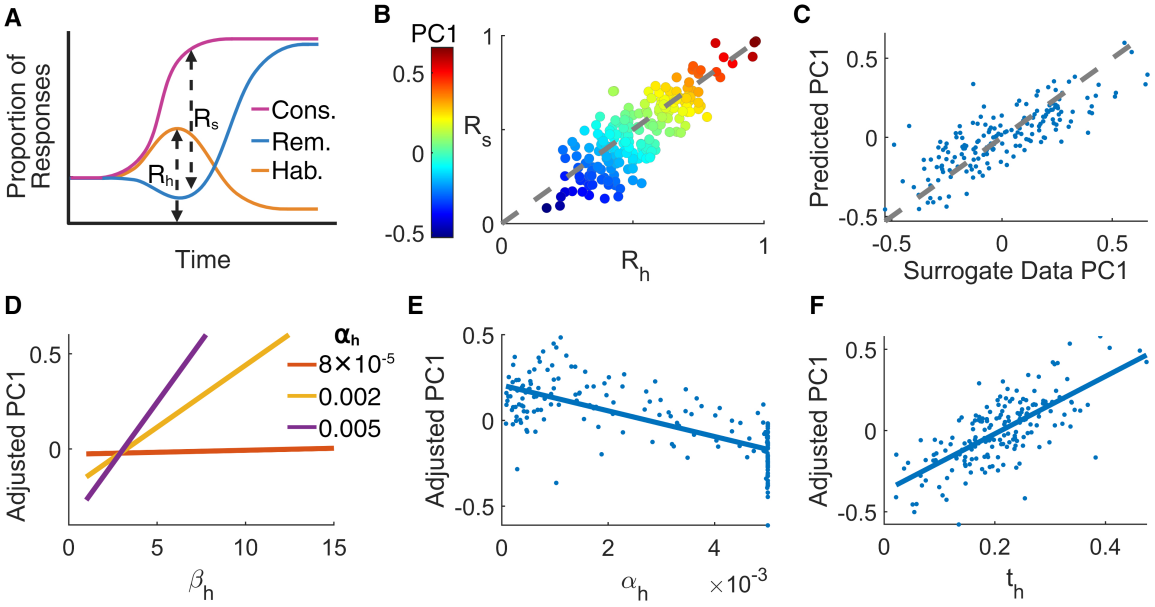


Figure 6.8: The Habit-Race$_{1\beta}$ Linear Regressor.
A: A schematic representation of how R_h and R_s are calculated from the time-controlled trials in the extended training condition, as in Fig. 6.2.
B: A scatter-plot showing the positive correlation between two measures of habit expression, R_h and R_s. The latent habit-strength component, PC1, is shown through the overlaid colour-map. An $x = y$ reference line is provided (grey dashed).
C: The accuracy of the PC1 predictions made by the linear regressor. An $x = y$ reference line is provided (grey dashed).
D: A visualisation of how α_h influences the relationship between β_h and the PC1 estimate, adjusted to average out the influence of all other terms.
E: The α_h values plotted against adjusted PC1 estimates when the influence of all other terms are accounted for, including $\alpha_h : \beta_h$. The least-squares line of best fit is provided (blue).
F: The t_h values plotted against adjusted PC1 estimates with the least-squares line of best fit (blue).

expression. The interaction of α_h and β_h determines the initial habitual drift-rate, while a larger t_h delays the influence of Q_a.

Next, we interrogate the significance of the interaction term more closely. The positive regression co-efficient states that, as α_h increases, so too does the impact of β_h on PC1, and vice-versa (Fig. 6.8D). This follows rationally from our understanding of the model - a larger α_h increases the speed at which H_a changes. Thus, when habitual actions are regularly executed as a result of a large β_h, a positive feedback loop forms and the H_a variable gains strength proportionally to α_h. Similarly, an accumulator with a very low β_h will struggle to reach threshold, regardless of habit strength, and so, habitual errors would become rare. In this case, a large α_h will promote extinction and the PC1 estimate will decrease.

The negative main effect of α_h alone can be seen in Fig. 6.8E. The regression coefficient of the α_h term denotes the influence of this parameter when the effect of all other terms are accounted for, including $\alpha_h : \beta_h$. This suggests that, in isolation, this parameter promotes extinction rather than habit strength.

The final term discussed here is t_h, whose value has an expected positive impact on PC1 (Fig. 6.8F). When time spent accumulating at an average drift of μ_1 lengthens, the interval in which a participant is likely to make a habitual error increases, together with the absolute number of time-controlled trials within this period.

Having developed this deeper understanding of how the fitted parameters influence expression of habits in simulated behaviour, we can now interrogate Hardwick et al.'s[4] participant data.

6.4.2 Human data

6.4.2.1 Qualitative Assessment

As in Chapter 4, visual analysis of the models' capacity to replicate the true data is key to ensuring that the recovered parameters are capturing real patterns in the participant's recorded behaviour. Later sections will then interrogate BIC recovery across the whole population and explore the quantifiable measures of habit strength in human data.

For clarity and conciseness, this chapter focusses on a subset of four participants rather than the entire cohort and explores the degree to which they successfully recover behaviour in both the free-RT and time-controlled trials. Though the free-RT distributions generally vary less between participants and models, they are also included for their utility in (1) confirming that the RT structure and accuracy are well-captured and (2) identifying when recovery of the unselected models is especially poor. Supplementary figures containing the full dataset are included in Appendix B.

We begin with a discussion of two individuals who exemplify the expected behaviour of their recovered model (S15 and S23, Fig. 6.9). Then, a participant with a particularly strong R_s signature in the absence of a clear R_h peak is examined (S18, Fig. 6.10A). The final case study (S24, Fig. 6.10B) demonstrates an exceptionally poor qualitative recovery of R_h and is included to highlight a key limitation when fitting the same parameter values to disparate datasets.

The first participant, S15, behaves exactly as a classical single drift-rate RDM would predict (Fig. 6.9A).

During the time-controlled trials, after an initial period of random action selection, a swift and steady increase in accuracy can be observed. The gradient of this improvement is highly similar between the two training conditions. Indeed, the most salient difference after extended training is a increased propensity to make occasional habitual errors, as the associated error and remapped curves saturate above zero and below one, respectively.

The RL-Race simulation provides an adequate replication of this time-controlled behaviour, though it is unable to maintain habitual errors without the inclusion of an additional process, as with ρ in the response-selection model[4] (Section 6.2). The more complex models provide no improvement to this fit and RL-Race is sufficient to explain the data. Similarly, the free-RT distribution is well captured by all models, though RL-Race has the lowest proportion of habitual errors.

Our next example participant, S23, is the first to demonstrate a transient increase in habitual errors after extended training (Fig. 6.9B), which is further associated with a

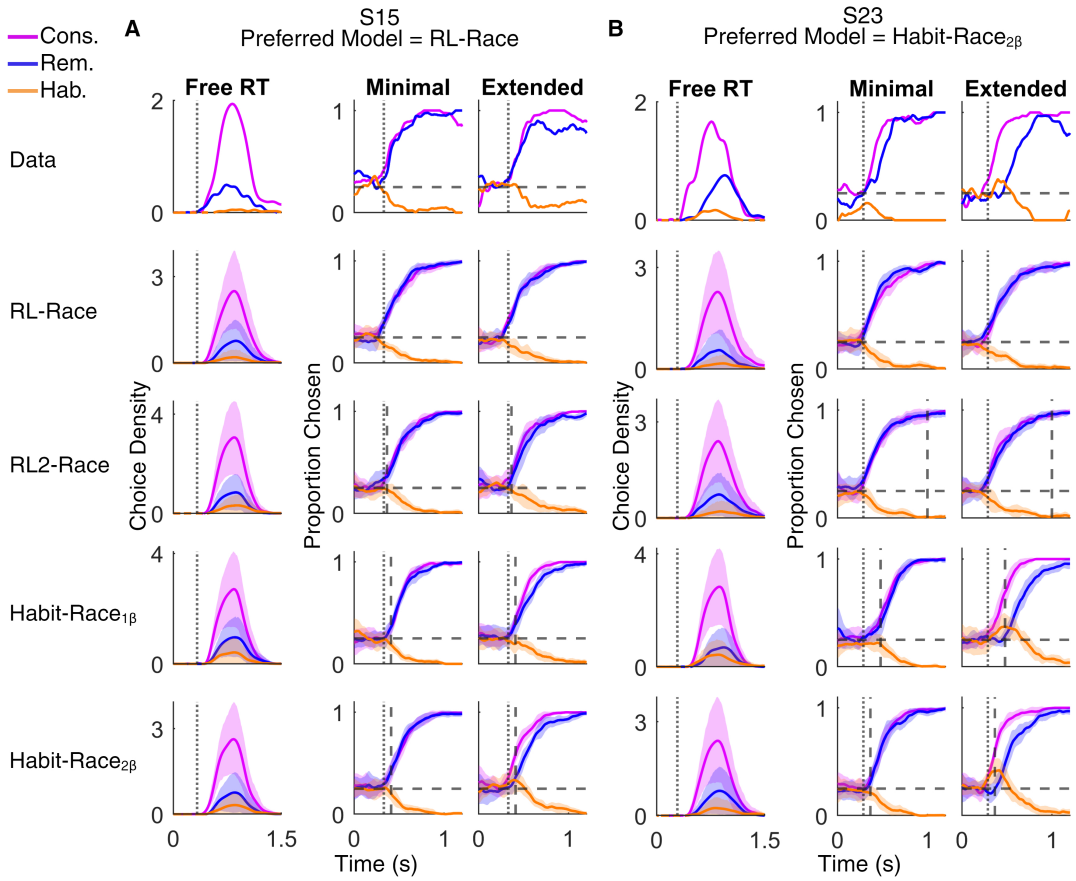


Figure 6.9: Example Participants.
Comparison of participant behaviour and the best-fitting model simulations. Two subjects are included, S15 (**A**) and S23 (**B**), who were recovered as RL-Race and Habit-Race$_{2\beta}$, respectively. Each row illustrates behavioural data (row 1) or a given model (rows 2-5). The simulations are averaged over 10 iterations and the $\pm s.d.$ is provided. The t_1 and t_2 estimates are indicated (dotted and dashed lines, respectively).
Free RT (left): The RT distribution of the free-RT trials included during fitting. Correct responses during initial learning and consistent trials during reversal learning are combined (pink). The remapped trials during reversal learning are divided between the correct responses (blue) and habitual errors (orange).
Time-Controlled (centre and right): As in Fig. 6.2, the moving average of choice proportions is provided for correct responses to consistent (pink) and remapped stimuli (blue), alongside habitual errors (orange).

much larger R_s than S15. Both habitual models were able to capture this behaviour well. Although Habit-Race$_{2\beta}$ mildly overestimates R_h, it is the preferred model.

The explanation for this recovery can be found by considering the other two subsets of data. First, Habit-Race$_{1\beta}$ produced a greater number of habitual errors during the remapping free-RT trials. More importantly, the recovered t_2 value is much larger for the simpler model, which results in an overlong period of random action selection during the

minimal time-controlled trials.

This difference illustrates the impact of H_a's mutable influence within trials. Including β_{h_2} allows the balance of habits and goal-directed evidence to shift following t_2 and thus, for μ_2 to actively correct earlier accumulation. As such, early habitual expression can be much larger without impairing the rate of reversal and increased accuracy during the second accumulation period. Habit-Race$_{1\beta}$ must perform the same compromise with a constant $\beta_h H_a$ term. This can only be achieved by drastically reducing the habit strength while increasing the duration of t_h, such that that the associated accumulators can reach the appropriate activity level.

Together, the above participants provide clear examples of successful qualitative simulations of human behaviour. Now, we complement this analysis with the exploration of two case studies who illustrate two interesting latent features of the data that were highlighted during recovery. Specifically, S18 represents the small subset of participants whose habitual behaviour continues into late RTs (Fig. 6.10A), while S24 demonstrates a particularly strong disconnect between the data's visual representation of habits and the model that is recovered (Fig. 6.10B).

To begin, S18 selected a Habit-Race model as its preferred fit despite no visual evidence of a transient habit R_h peak (Fig. 6.10A). This is rendered possible by the recovery of R_s and the continued influence of habits on action selection. Specifically, following extended training, this participant exhibits a distinct and prolonged period in which habitual actions occur at a consistently high rate as accuracy fails to reach saturation for the remapped trials. This protracted expression of habits without a peak requires H_a to influence the drift-rate following t_2 without any prior accumulation, which is evidenced by the collapse between t_1 and t_2. In so doing, a difference is maintained between the pattern of action selection in consistent and remapped trials.

This participant's behaviour is particularly extreme, but exemplifies how habits continue to exist and influence action selection beyond the initial preparation period. Interestingly, with the loss of the first accumulation period, this participant has essentially returned to a single-drift RL-RDM which utilises additional APE-based variables. Only

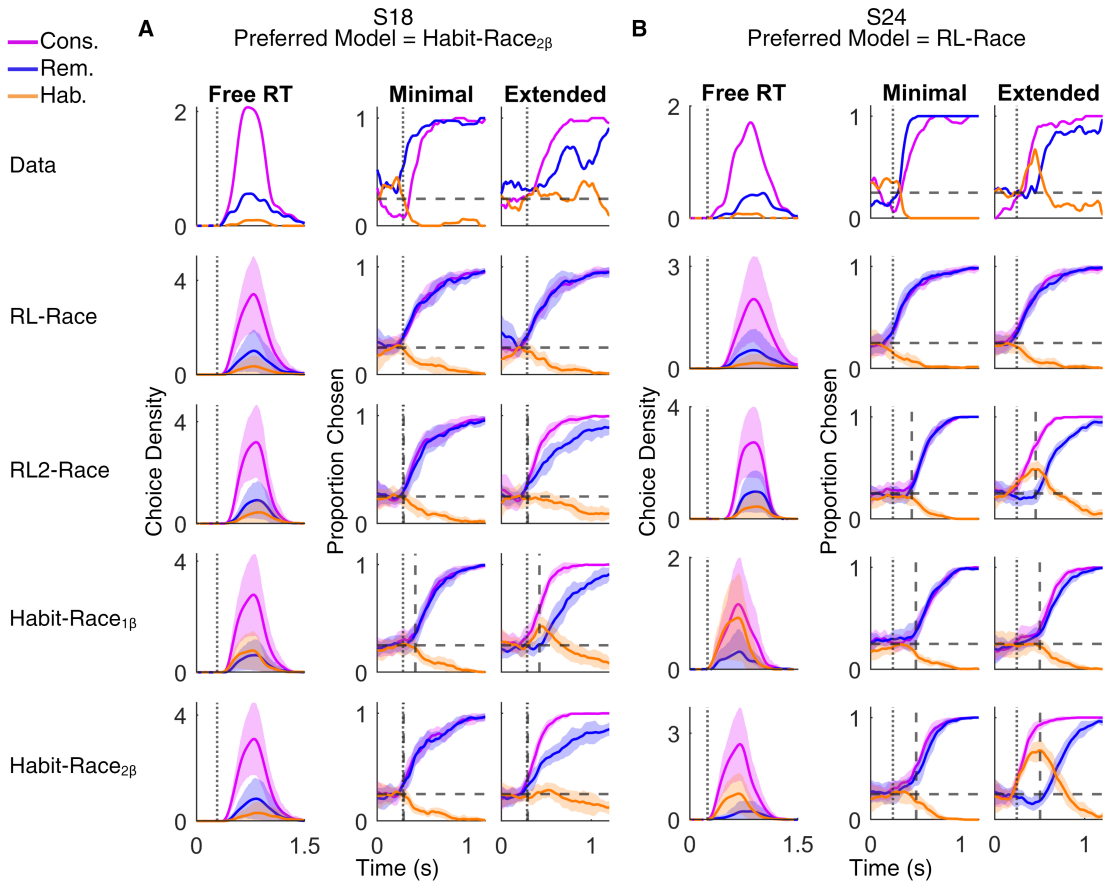


Figure 6.10: Two Case Studies.
Comparison of participant behaviour and the best-fitting model simulations, as in Fig. 6.9. Two subjects are included, S18 (**A**) and S24 (**B**), who were recovered as Habit-Race$_{2\beta}$ and RL-Race, respectively. Each row illustrates behavioural data (row 1) or a given model (rows 2-5). The simulations are averaged over 10 iterations and the $\pm s.d.$ is provided. The t_1 and t_2 estimates are indicated (dotted and dashed lines, respectively). The free-RT (left) and time-controlled (centre, right) behaviour are provided.

one other participant, S21 (Appendix B), demonstrated a similar collapse in t_2.

In contrast to the other participants shown here, the replication of S24's behaviour is qualitatively poor (Fig. 6.10B). This individual shows a very clear peak in habitual errors (R_h) following extended training with a corresponding delay in accuracy. Despite this, RL-Race was selected as the preferred model. This is an especially striking outcome given that, qualitatively, the R_h value for S24 is better captured by both RL2-Race and Habit-Race$_{2\beta}$.

Instead, model preference seems to have been more strongly determined by the best-fitting habit parameters' failure to replicate behaviour during the free-RT and

minimal training datasets. In the latter of these, S24's data shows an immediate increase in accuracy that very quickly saturates at 1 - an effect which cannot currently be balanced against the large R_h after extended training. Achieving such a rapid increase during the first accumulation period would require either (1) habits to have been completely relearnt during remapping - thus placing H_a and Q_a in agreement, or (2) Q_a to influence drift-rates at an earlier timepoint. The former of these options is possible, but we would then expect the same absence of habits after extended training. In future work, this forced compromise could be resolved through extensions to the model. Some potential adaptions will be discussed further in Section 7.1.3.2.

Similarly, in attempting to recover time-controlled trials under both training conditions with a single set of parameters, replication of the free-RT trials in S24 has failed significantly for both habit models. As briefly mentioned above, there is generally very little difference in the precise replication of the free-RT trials, particularly as habitual errors are rare during this period. S24 is one of the exceptions to this. For all subjects, the best-fitting model provides a good replication of free-RT behaviour.

Overall, key patterns in human behaviour are reproduced by their best-fitting models. However, the accuracy and magnitude of this recovery appears to be partially impaired by the inclusion of multiple conditions and trial-types for the same parameter set. The replication of R_h and R_s is explored further in Section 6.4.2.3.

6.4.2.2 BIC analysis and group BMS

Having confirmed that the best-fitting parameters are capturing real patterns in the data, we can now quantitatively analyse model recovery at a group level.

Model	Number of participants	
	BIC Model Selection	AIC Model Selection
RL-Race	5	5
RL2-Race	2	2
Habit-Race$_{1\beta}$	7	6
Habit-Race$_{2\beta}$	8	9

Table 6.5: The frequency distribution of model selection across Hardwick et al.'s[4] participants.

Under both BIC and AIC analysis, 15/22 participant datasets were best-fit by a TD-RDM model with an initial value-free habitual drift-rate (Table 6.5). No further conclusions can be drawn regarding preference between the two Habit-Race models as it was confirmed that these are indistinguishable with the data available to us. However, it is interesting to note that ~50% were recovered as the more complex 2β model whereas Fig. 6.7 shows that only 6% of Habit-Race$_{1\beta}$ datasets were mis-recovered in this way. Another 2 participants were fit best by RL2-Race.

Reducing the impact of the parameter number with AIC analysis only altered preference for one model (S21, Appendix B) which shifted between the Habit-Race models. As we have already established that these models are difficult to correctly recover, this change was not interrogated further.

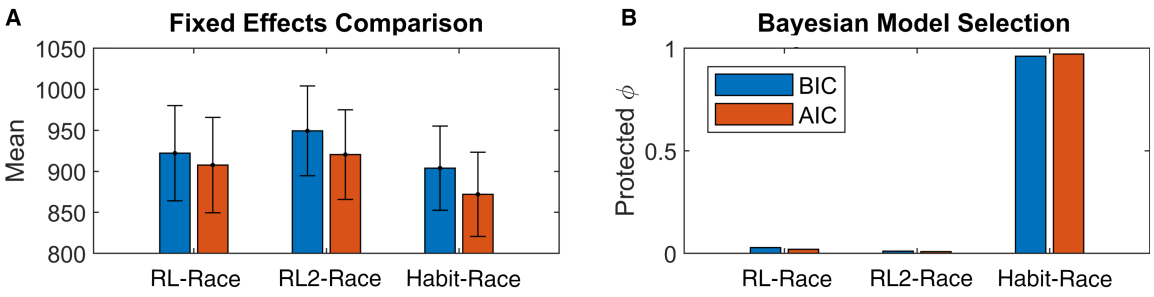


Figure 6.11: Population Analysis of TD-RDM Recovery.
A: A fixed effects comparison of the mean BIC (blue) and AIC (red) values across each of the RL-EAM frameworks. The standard error of the mean is reported in the error bars.
B: The protected exceedance probabilities, $\tilde{\phi}_k$, calculated using the `bms` function created by Gershman[279] with BIC (blue) and AIC (red) used as approximations for model likelihood.

At the group level, BMS analysis suggests that, according to the BIC values, there is a 96% chance that the majority of the population utilises value-free habits in a multi-alternate forced choice task (Fig. 6.11). The fixed effects comparison is less clear, but on average the best-fitting Habit-Race model has a lower BIC than the two pure RPE models. This pattern holds for the AIC values.

6.4.2.3 Recovering habits using TD-RDM

One key advantage of using mechanistic models that parametrise the underlying processes is that the parameter values can reveal latent relationships within the data, as

discussed in Section 6.4.1.2.

This work can be further extended to determine whether the theoretical associations found within the surrogate models still hold true when considering real participants.

In Section 6.4.1.2, a regression model (Eq. 6.5) confirmed that the latent association between two measures of habit strength, R_h and R_s, was most strongly impacted by the duration of the first accumulation period, t_h, and the interaction between the learning rate and temperature parameter of H_a.

The relationship between these measures can be similarly extracted from both the human data and the best-fitting simulations, thereby enabling a quantitative examination of the extent to which habits are captured by the Habit-Race TD-RDM and interrogation of how the linear regressor's significant terms relate to recovery of these measures.

First, Fig. 6.12A confirms that a correlation exists between R_s and R_h in both the real data and in the simulations created with the best-fitting parameters. Further, the relationship between these measures and the latent PC1 is equivalent to that seen in Fig. 6.8. It is worth noting that R_h and R_s are less tightly correlated in the real data than in model simulations. This effect is likely due to both behavioural noise and additional processes that influence either R_h or R_s independently which are not captured by TD-RDM.

Next, the recovery of habit measures is visualised in Fig. 6.12B. The simulated behaviour is significantly and positively correlated with human data for all measures of habit. However, the correlation coefficients remain between 0.45 and 0.6. Thus, while the preferred models do capture real information regarding a participant's latent habit strength, recovery is not perfect and there is a trend towards underestimation of R_h and R_s.

Lastly, Fig. 6.12C also illustrates how successfully the linear regressor is able to predict PC1 and the relationship between latent habits and the significant regression terms. Again, all correlations are positive and significant.

In combination, this analysis confirms that the surrogate data used in model recovery (Section 6.3.3) closely approximated the latent habit behaviour expressed by the participants and that the Habit-Race$_{1\beta}$ best-fitting parameters allow us to quantify the

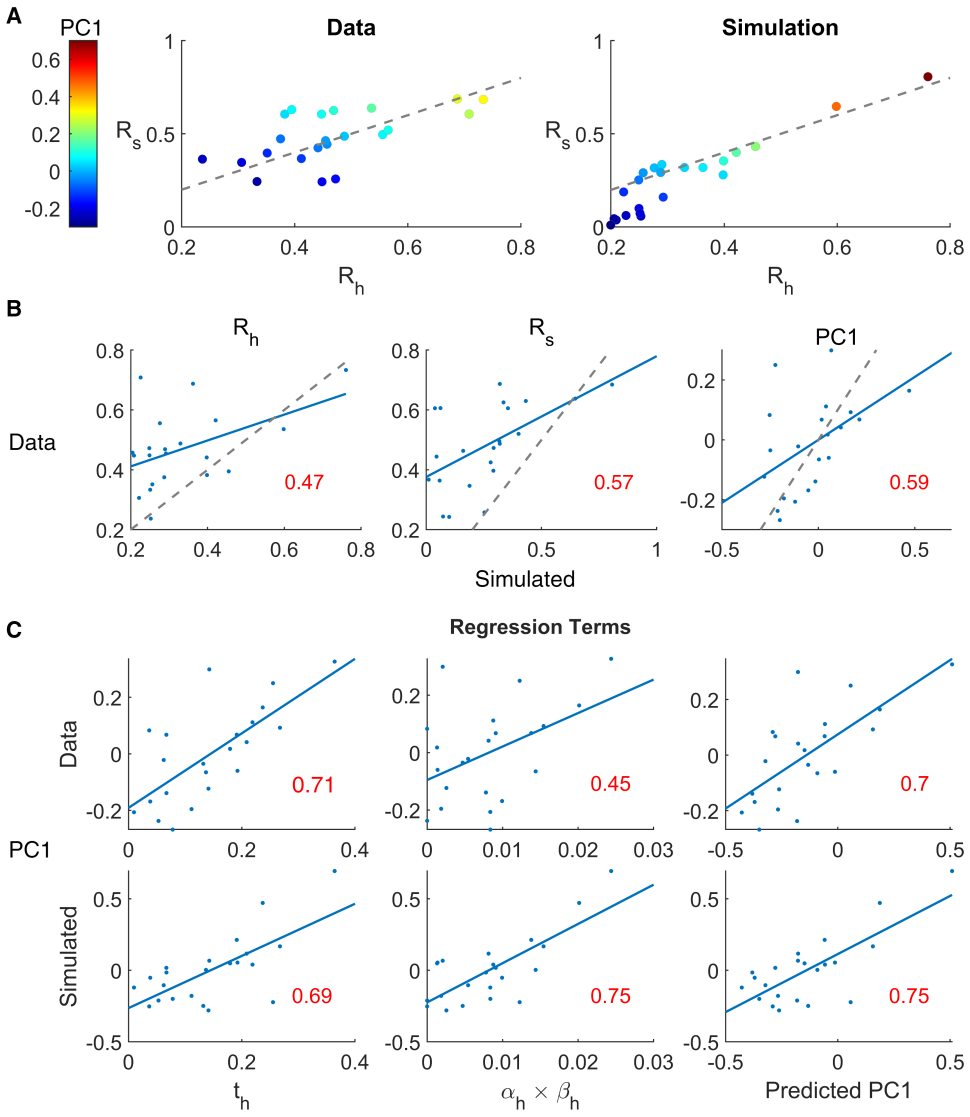


Figure 6.12: Parameters, PC1 and Habit Strength.
A: As in Fig. 6.8A, a scatter-plot showing the positive relationship between R_h, R_s and PC1, calculated in the participant data (left) and model simulations (right). An $x = y$ reference line is provided (grey, dashed).
B: The correlation of R_h, R_s and PC1 between the data and simulated values. The least-squares line of best fit (blue) is provided alongside an $x = y$ reference line (grey, dashed) and the correlation co-efficient is indicated (red when significant, black otherwise).
C: A comparison between PC1 as estimated in the data and in simulations when correlated with the significant regression terms and the linear regressor's estimation of PC1. The least-squares line of best fit is provided (blue) and the correlation co-efficient is indicated (red when significant, black otherwise).

specific subprocesses that are likely to be responsible for the individual variation between participants.

6.5 Discussion

6.5.1 Summary

Over the course of this chapter, the novel TD-RDM was applied to the behavioural data of 22 human participants in a multi-alternate forced choice paradigm with additional time constraints, provided by Hardwick et al.[4].

Four models were tested, two of which used an APE-based value-free H_a variable alongside the standard RPE-based Q_a. These Habit-Race models were compared against two controls that exclusively learnt from the outcome-based RPE but differed in the number of drift-rates used during evidence accumulation.

Analysis of surrogate data provided three key insights:

1. While the time-controlled trials provide the clearest visual expression of habits, inclusion of the free-RT trials is necessary in order to constrain parameters and successfully distinguish the models.
2. The data provided by this experimental paradigm is sufficient to identify (1) which participants used two drift-rates rather than one and (2) whether an APE learning rule influenced time-constrained behaviour, although the Habit-Race models were often confounded.
3. The surrogate data's best-fitting Habit-Race$_{1\beta}$ parameters can be used to estimate a participant's habit strength following extended training. Specifically, the expression of habits is influenced by two factors: the delay between the non-decision times and the interaction of α_h and β_h. These effects were significant despite the relatively poor recovery of the latter terms, which implies that, although the H_a parameters are only minimally constrained, they still capture important latent information within the data.

Across the four models, the human data strongly supported the existence of a two drift-rate evidence accumulation process and an underlying APE-based habit, which was distinct from a slow-learning Q_a variable.

Qualitative analysis of the model behaviour revealed that accurate recovery is limited by

the application of the same fixed parameters across all trial types. Repeating Hardwick et al.'s[4] experimental paradigm with a greater number of time-controlled trials or repeated reversals may allow for more complex versions of the TD-RDM framework to be studied.

6.5.2 The response-selection model and TD-RDM

In their original study, Hardwick et al.[4] found that, in 17 of the 22 participants, the time-controlled behaviour following extended training was better replicated by a probabilistic model that assumes two time-dependent preparation processes are occurring. Though they labelled this model as a 'habit-model', they did not specify how these habits would form, nor did they attempt to replicate the behaviour during free-RT trials.

The model selection presented in this chapter concurs with their results, in that the TD-RDM framework was preferred in 17 datasets over the RL-Race model's single drift-rate. We have additionally established that 15 of these participants likely used a value-free APE-based habit to control the initial drift-rate.

Though the Habit-Race TD-RDM and the response-selection model were both designed to interrogate the existence of two time-dependent drift-rates under Hardwick et al.'s experimental paradigm, they apply two very different approaches.

The response-selection model was explicitly parameterised to reproduce the time-controlled figure exactly, through probabilistic action selection rates across the RT distribution. As such, the Gaussian distributions of habitual errors and responses to consistent stimuli are treated as equivalent, whilst also maintaining full independence between the minimal and extended training conditions. Consequently, this model must forcibly disregard an individual's learning experience and assume that all learning halts before the time-controlled phase begins.

TD-RDM uses approximately the same number of parameters as Hardwick et al.'s[4] response-selection model and provides a deeper insight into the mechanistic learning processes underlying human behaviour. By considering the likelihood of an action occurring in a given trial, the ability to replicate time-controlled behaviour arises as

a consequence of the final parameters rather than their primary function. Further, the application of TD-RDM allows for the habitual errors and consistent mapping to be treated independently and for action selection to be influenced by an individual's previous experience - even within the time-constrained trials.

This is particularly relevant given that participants continued to be provided with feedback during the final stage of testing, which had ten-fold more trials than the reversal learning phase. As the participants were able to adjust to the new mapping within ~40 trials, logically, this learning should continue during the time-controlled trials and thus, could potentially influence their expression of habits. More concretely, if the majority of a participant's trials with RTs before t_2 occurred during the later time-constrained trials, their expression of habits would be weaker than if they were tested in the first trials immediately following remapping. By considering the impact of learning during all trials, the TD-RDM is able to account for the influence of the trial-sequence experienced by the participant.

The clearest difference between this study and Hardwick et al.[4] is the inclusion of free-RT trials. Through the analysis of surrogate data, we confirmed that model recovery was markedly improved when these experimental stages were also considered.

During the exploration of individual case studies, these trials were not discussed to the same extent as the time-controlled figures for two reasons. First, Hardwick et al.[4] confirmed that there was no statistically significant evidence of habits within this data, which aligns with previous research into human slips-of-action (Section 2.1.3.4). Second, there tended to be very little variation across models in the recovery of the free-RT behaviour. When clear collapses did occur, such as with S24 (Fig. 6.10B), the associated models were never selected during BIC analysis.

Since certain parameter combinations can only accurately mimic the time-controlled trials at the cost of realistic EAM behaviours, it is likely that such impairments in RT distributions are at least partially responsible for the positive impact that including both trial-types has on model recovery. For example, in S24 the free-RT errors were caused by θ values that had collapsed to <1 in order to ensure that habitual choices were reinforced

throughout the first experimental phases and so, H_a could remain artificially strong ahead of the time-controlled trials.

6.5.3 TD-RDM performance and limitations

Using both surrogate and human data, the analysis presented in this chapter strongly supports the TD-RDM's capacity to model behaviour in a multi-alternate forced choice paradigm, across trials where either the response time was externally constrained or the participants were free to respond at an internally generated time.

The best-fitting simulations discussed in Section 6.4 provide confirmation that single drift-rate EAMs are not only unable to produce the transient peak in habitual errors without applying additional processes, but also cannot separate the different temporal accuracy curves in consistent and remapped stimuli.

RL2-Race models were the most difficult to reliably recover, as demonstrated by the model recovery analysis (Section 6.4.1.1). When t_1 is accurate, roughly 18% of the RL2-Race datasets were recovered as Habit-Race and they had the lowest $P(M_T|M_R)$ at 85% (Fig. 6.7).

To understand how RL2-Race and Habit-Race models can be distinguished at all, it is important to consider the impact of applying either a reward or action PE. More precisely, regardless of the action selected, the outcome received or the difference in learning rate, $Q_{1,a}$ and $Q_{2,a}$ will *always* be updated in the same direction due to their shared RPE. Consequently, they can only contradict each other for a fixed number of trials, as determined by their relative learning weights.

Conversely, the self-reinforcing property of H_a means that habits will continue to exist for as long as the associated action is selected. Therefore, if the outcome on a given trial is negative, Q_a will weaken while the habit is strengthened. This effect could potentially be behaviourally disentangled by comparing the habitual error rate following the execution of many error trials such that the Q_a values for these habitual errors would approach 0 while H_a would be unimpaired.

In the context of the time-controlled trials, this difference is important as it necessitates

that RL2-Race can only separate remapped and consistent stimuli if the associated trials occur early enough during training, while APE-based habits are likely to influence action selection for much longer. This impact of trial number is similarly responsible for the common absence of error peaks within RL2-Race simulations, despite the fact that these models have the capacity to produce them, as is shown by S24.
Overall, if RL2-Race was more prevalent in the data, we would expect 75% of them to be accurately selected (Section 6.4.1.1). Thus, given that only two participants were recovered as RL2-Race, we conclude that it is unlikely that any such mis-classifications will have strongly impacted our population-level results.

The limitations faced by TD-RDM, already largely discussed throughout the chapter, are briefly summarised here.
The majority of the constraints on TD-RDM performance are caused by the latent nature of the habitual variables and parameters. The most blatant of these is the imperfect recovery of any parameters associated primarily with the initial drift-rate. The inaccessibility of t_1 has been described in some detail, but a similar impairment occurs with the estimation of α_h and the β_h terms, for the same reason that human habits are difficult to detect in free-RT studies (Section 2.1.3.4). Specifically, there are very few trials in which the H_a variable is unmasked to a degree that the other parameters cannot compensate for noise in the α_h and β_h terms. Thus, the effect of these errors on the overall likelihood is limited and accuracy suffers as a result.
Thankfully, through the analysis of surrogate data, we were able to determine that the reliability of model selection was not significantly impacted by these inaccuracies. Further, the influence of the α_h:β_h interaction term in the linear regressor suggests that some redundancy in the exact values of these parameters is likely, as they still capture important information regarding the expression of habits. This is supported by the known collinearity between α and β terms in RL-EAMs[191].

The second limitation which has most significantly affected the results of this project arises from the implicit assumption that all behaviours use the same mechanistic observation model and that the parameters are constant across conditions. In truth, this

is an oversimplification as it is unrealistic to assume that humans do not apply some form of parameter adaptation in response to different contexts. Indeed, our removal of θ during the time-controlled trials represents a necessary modification without which the participants would be unable to hold their response until the fourth tone and thus could not correctly complete the task.

The example participants examined throughout Section 6.4.2 highlight the direct effect this assumption has on accurate replication of the data. Specifically, the inflexibility of Q_a's temporal relationship with action selection resulted in a prioritisation of either (1) the immediate increase in accuracy following t_1 after minimal training or (2) the appearance of a transient habit peak after extended training, which were often balanced one at the expense of the other. The necessity of this compromise could be ameliorated by extensions to the current Habit-Race framework, some potential options are discussed further in Section 7.1.3.2.

Overall, the results presented in this chapter for both the surrogate and human data support the claim that human habits can be quantified from behavioural RT data despite potentially being masked by a goal-directed process. The mechanistic nature of TD-RDM provides strong evidence in favour of Hardwick et al.'s[4] hypothesis that (1) introducing time-constraints to a multi-alternate forced choice paradigm will increase expression of habitual behaviour which (2) is produced by a separate process from goal-directed learning that influences action selection at a faster timescale.

Importantly, we were further able to conclude that habitual associations developed using value-free APEs provide a better explanation for most human behaviour relative to variables produced using the standard model-free RPE.

The next and final chapter summarises our answers to the four research questions posed in Chapter 1, discusses further extensions to TD-AL and TD-RDM and explores their wider impact on our understanding of how habitual S-R relationships form in the brain.

7
Discussion

Contents

Over the course of this thesis, four pieces of research have been presented which address separate requirements needed to establish the plausibility of APE-based value-free computational habits.

In this final chapter, the questions raised in Chapter 1 are directly answered and the new avenues of work which arise from our conclusions are highlighted, with a particular focus on potential theoretical extensions to the two novel models of computational habits. Following this, we briefly assess the key methodological limitations faced throughout this research project before finally discussing three outstanding questions from the wider field.

7.1 Four Research Questions

7.1.1 Can habit learning be generalised to continuous and scalar actions?

Yes, we were able to adapt the trial-by-trial APE-based habits presented by Miller et al.[1] and Bogacz[2] to learn across continuous time and from time-dependent action intensities.

7.1.1.1 Key findings

Our first body of work, presented in Chapter 3, provided the mathematical description of our novel TD-AL model. Just as APEs follow the computational structure of RPEs with an adaptation of the information received, so too was TD-AL developed from Ludvig et al.'s[175,176] microstimulus TD-RL model. This algorithm was selected for its biological plausibility and accurate replication of dopaminergic dynamics.

Through simulations of classical instrumental association and omission experiments, we confirmed that TD-AL behaved as expected and in accordance with the key axioms that define an APE. Specifically:

1. The APE spikes in response to unexpected actions and the magnitude of the APE peak decreases in size as the action becomes expected.
2. There is a negative APE when expected actions are omitted.
3. The APE is entirely outcome (and so, reward) insensitive.

TD-AL creates several predictions that can be used to determine whether a given dopaminergic dataset contains APE signals.

Firstly, the absence of a response to predictive cues is insufficient to disprove the existence of APEs, as this can be explained through adaptations of the discount parameter, γ_h. More precisely, when future actions are entirely discounted and H_a calculates only the action it expects at that specific time-point, then, unless an action occurs simultaneously to the cue, no prediction errors will be produced.

Secondly, if there *is* a non-zero γ_h then action initiation should act as the most immediate predictive cue, since it is responsible for producing its own microstimuli.

Most importantly, a TD-AL-produced APE will always dip below tonic levels following

the omission of an expected action. This final prediction will hold true regardless of the specific parameter settings.

7.1.1.2 Future TD-AL extensions

The greatest limitation faced by TD-AL is the absence of an associated observation model. Without this, the impact of time-continuous habits on behaviour cannot be tested and the quality of model recovery cannot be assessed ahead of parameter fitting. The three barriers to producing an observation model are (1) the absence of Q-variable measure of action-value, (2) a decision of how such a Q_a and H_a would interact, and (3) a model of how these striatal values influence the release of action at the thalamus.

One simple solution exists for developing an appropriate Q_a learning algorithm. Just as TD-RDM's Q_a only learnt from an RPE when the action had been executed, so too could a continuous Q_a only update its weights when the microstimuli of its associated action are active. Currently, the microstimulus model calculates the value of a context, V, according to all events. This formulation would produce Q_a variables which calculate the expected value specifically *if* the associated action is executed. Action omission or selection of another action would then reduce the activity of the unselected Q_a.
In this way, the 'action-value' system would form an intermediary balance of V and H_a, such that actions which regularly precede a reward will have a greater response to earlier predictive cues. The primary limitation to this approach is that it requires the action microstimuli to overlap with the subsequent reward.
Such 'gating' of learning still maintains biological plausibility as mechanisms are increasingly proposed to modulate the impact of dopamine in the striatum. For example, while it could be as simple as Hebbian requirements of co-firing from the action microstimuli, learning could also be impacted by inhibitory GABA interneurons or through cholinergic modulation of dopamine release[292].

The other two barriers to the creation of an observation model likely need to be resolved in tandem, as the solution to one will influence the other. Here, we primarily focus on the question of whether the TD-AL continuous variables can influence action.

In Chapter 5, the Habit-Race formulation of the TD-RDM was outlined. This EAM assumed that H_a and Q_a are both fixed to a single value in every trial and that the average drift-rate is constant between non-decision times.

In contrast, our TD-AL simulations have shown that, before expected actions are initiated, H_a will linearly increase as predictive cues arrive. In particular, Fig. 3.2 demonstrates that, if $\gamma_h > 0$, H_a spikes to a high value following the cue and remains high until the action begins. Despite learning at every time-point, α_h is often so low that the adaptations remain small from one moment to the next, especially if the action is consistent in its temporal dynamics between trials.

It is therefore plausible that this variable could be taken as the source of 'evidence' for a given action and used in the calculation of a thalamic drift-rate. Then, similar to DopAct's separation of action planning and action learning[2], once the action is triggered, H_a would promote the expected intensity and learn from its accuracy.

The complication arises in applying such a function to a computational model. Simulations are surely executable, but the fitting procedure employed in Chapter 6 was only feasible due to the efficient near-analytical cost function. Introducing further drift-rate changes rapidly becomes intractable as each one requires the calculation of the new BVP and integration over all possible activity levels. Consequently, a continually evolving drift-rate cannot, currently, be fit to any data.

However, as in all models, simplifications can be made; for example, the drift-rate could be set to the average H_a value or be fixed to H_a's activity level at a pre-determined t_1 following the predictive cue.

7.1.2 Can evidence of action prediction errors be found in striatal dopamine?

Yes, analysis of dopaminergic data from the TS during a CoT instrumental association task was better replicated by TD-AL than both TD-RL and a pure motor signal.

7.1.2.1 Key findings

In Chapter 4, the novel TD-AL algorithm was tested on dopaminergic data provided by Greenstreet et al.[3]. Overall, for each of the six mice investigated, BIC analysis found that TD-AL was the preferred model over the alternatives.

The best-fitting parameters of four mice strongly supported the existence of a γ_h signal within this data, though its magnitude varied greatly between individuals. This was largely unexpected as Greenstreet et al.[3] found no evidence of a dopaminergic response to the auditory cue, which we would expect TD-AL$_\gamma$ to produce.

However, this cue-response was not the only discriminatory factor since the inclusion of a discount factor will produce a left-shift in dopaminergic responses with training, which could be seen within the data and was captured by the best-fitting parameters.

The absence of an observation model precluded us from assessing the reliability of our parameter and model recovery but, given that the simulations of other models produced qualitatively distinct signals (Fig. 4.5), we determined that an APE-like signal was detected within this data.

BMS analysis was in agreement with these results, as it suggested that there was a 88% chance that one of the two TD-AL formulations was the most prominent within the population. This number climbed to 93.8% when AIC was used instead of BIC.

These analyses were particularly encouraging as the dynamics within the dopaminergic data are quite complex and TD-AL$_\gamma$ was able to reproduce many key features with only 3 free parameters, one of which was an artificial scaling of unitless data and did not impact the structure itself.

7.1.2.2 Future research questions

Some additional studies have already been proposed in Section 4.5.2, such as running a control repeat using signals from the VS, analysing parameter recovery using simulated datasets produced from the animals' behavioural data or testing for γ_h by explicitly fixing the temporal separation between cue and action.

Another evident direction through which this work could be further expanded would be to interrogate whether APEs could be found in a greater sample size, using experimental paradigms designed to target TD-AL specifically.

The sample size is simple enough to address, as the Greenstreet et al.[3] task can be scaled to a greater number of mice and our fitting procedure repeated.

TD-AL can be explicitly tested for using the model predictions established in Chapter 3. We saw that omission of an expected action will produce a negative APE and cause dopaminergic levels to dip, irrespective of γ_h's value. Unfortunately, the data used in this thesis was extremely stereotyped in the final stages of the experiment, causing the number of incorrect/omission trials to be few and far between. In future, this could be resolved by the addition of a reversal task, such that the mice learn to change their response to a new mapping and execute different choices while the habitual system continues to expect the old action.

It should alternatively be possible to identify TD-AL in the existent data through action grouping. Specifically, if the dopaminergic signals in the later trials were classed according to the overall speed of the trial and averaged within these groups rather than over consecutive trials, then we hypothesise that the APE behaviour would differ between the slowest and fastest trials. When the animal moves the same distance over less time, the action intensity must initially spike above what was expected and produce a positive APE. In tandem, the action will also end earlier than usual, which, theoretically, would behave the same as an omission trial since the expected action is absent, causing a negative APE to follow. The reverse would be true in actions that were slower than the typical learnt movement, as the intensity remains low and continues for longer than expected.

In contrast, TD-RL algorithms would not show this bidirectional PE response, as they would instead respond only to the timing of reinforcer, which may arrive sooner or later than expected but should still be temporally distinct from the execution of action. Ludvig et al.[176] have previously demonstrated that, late in learning, the microstimulus model will be partially surprised by an early reward, but the omission signals at the usual reward timing are negated and absent.

Finally, the existence of APE-based habits would be further supported by the discovery of H_a-like signals within the striatal SPNs.

Just as studies have searched for value and action-value signals within the striatum[280,293,294], the value-free habit hypothesis supposes that action-expectation activity should be triggered during the preparation and execution of habitual behaviours. This signal should further be entirely unmodulated by outcome and reward.

7.1.3 Can the process by which habits affect choices and reaction times be mechanistically described?

Yes, under the assumption that habits influence action selection on a shorter time-frame than goal-directed information, we developed the novel TD-RDM algorithm, which predicts how the presence of an initial habitual drift would impact behaviour in a multi-alternate forced choice task.

7.1.3.1 Key findings

In Chapter 5, we extended a simplified form of the RL-RDM to include a within-trial change in drift-rate for multiple accumulators racing towards a shared threshold.

Most importantly, an analytical solution was developed that calculates the probability of a specific choice being made and the RT of that action. This cost function allowed us to efficiently run parameter-fitting analyses under a RL framework where the distribution must be recalculated in every trial.

The application of our model confers several advantages. For example, TD-RDM requires no implicit assumption of competition between drift-rates or accumulators and it is able to simulate multi-alternate forced choice tasks.

It is worth noting that, while TD-RDM is placed in the context of habits for the purposes of this thesis, the algorithm itself is general and can be applied across a wide variety of applications. For instance, TD-RDM would be an appropriate model in multi-alternate tasks where additional information is received mid-trial, such as stop-change experiments[295]. As the individual accumulators adapt their rate, the initial

drift can be accounted for in the likelihood of action execution.

7.1.3.2 Future TD-RDM extensions

In Chapter 5, several adaptations to TD-RDM were proposed. As mentioned above (Section 7.1.1.2), introducing additional within-trial dependency in the calculation of μ_1 or μ_2 rapidly becomes intractable for any investigation beyond pure simulation.
However, since the model itself only requires μ_1 and μ_2 to be constant within a trial, any adaptation of the algorithm which produces these drift-rates is possible, whether that arrives in the RL framework or in the mechanism through which these variables influence drift.

To focus particularly on the APE-based habit formulation, the potential extensions can be classed into two groups, depending on whether they (1) introduce additional variables or (2) alter the implementation of Q_a and H_a.
The most natural additional variables to include are an innate bias towards a given action, an urgency signal, or a measure of confidence (e.g., the S term proposed by Mikhael and Bogacz[48]). The latter of these is particularly relevant given the common association of certainty with both expression of habits (Section 2.1.2.3) and RT distributions (Section 2.3.4.2).
Two of the models tested in Chapter 6 alter the influence of Q_a and H_a on behaviour, as both the introduction of $Q_{a,1}$ in μ_1 and of β_{h_2} in μ_2 have no impact on the structure of the learning rules, but instead adapt how these variables are used by the EAM.
Further, the introduction of direct competition between actions is possible by combining TD-RDM and the advantage framework proposed by Miletić et al.[226]. However, it is important to consider that, in this algorithm, each accumulator represents a pair of actions, and thus, the number of calculations required rapidly increases as the number of available actions grow.

The primary limitation we encountered when using TD-RDM in Chapter 6 was the inflexibility of parameters between conditions. Though many extensions to the current TD-RDM framework could ameliorate this, the increased model complexity must be

balanced against the data's ability to constrain parameter values. As we saw with Hardwick et al.'s[4] data, the ability to constrain parameters can be highly limited by the latent nature of the first drift-rate and this must be taken into account when designing future experiments.

In particular, introducing flexibility in the non-decision times appears to be paramount in allowing the model to adapt under different conditions. The fixed values and distance between t_1 and t_2 could alternatively be conceptualised as functions of trial-number, task complexity, stimulus novelty or participant certainty.

Lastly, the current TD-RDM is limited to a single change between drift-rates. As addressed in Section 7.1.1.2, this likely represents an upper limit for an applicable analytical function in the free-RT trials, as each additional drift-change requires integration across all potential activity levels of all accumulators at the new t_x. If focus is only given to the time-controlled trials, then the number of drift-changes is irrelevant to the model complexity as, without a threshold, the analytical solution is produced through a simple summation of normal distributions.

Similarly, though one could envision a model which combines the response-selection model's distribution of non-decision times with the TD-RDM, the variability in start-time introduces further complications to the BVP, particularly if t_2 can vary between accumulators. Once again, simulation would be feasible but the advantages conferred by an analytical solution are lost. It is also unclear whether this variation in start-point would significantly impact the results, as Tillman et al.[227] established that such between-trial variability was unnecessary for the RDM.

7.1.4 Can the impact of habits in human behavioural data be quantified?

Yes, the majority of human behavioural data was better replicated by a TD-RDM agent which used a APE learning signal. Further, the best-fitting t_h, α_h and β_h parameter values were significantly correlated with habitual expression in both surrogate and real datasets.

7.1.4.1 Key findings

In Chapter 6, four iterations of the TD-RDM were tested on the human behavioural data provided by Hardwick et al.[4]. The learning behaviour during free-RT trials and the time-controlled responses under both training conditions were fit using the same fixed set of parameters.

Overall, of the 22 participants evaluated, 17 were best-fit by a two-drift RDM and a further 15 of these were specifically recovered by one of the two APE models. Model recovery had previously established that confidence in this classification was high.

Qualitative analysis confirmed that the recovered behaviour replicated the key behaviours of interest. Further, the BMS analysis of the human data revealed that there is a 96% chance that, of the models tested, the majority of the population will use one of the Habit-Race formulations.

We completed additional surrogate data analysis which determined that (1) inclusion of the free-RT trials was necessary to correctly recover the underlying model and (2) the two Habit-Race models could not be accurately distinguished.

The association between the recovered parameter values themselves and the strength of expressed habits was interrogated via linear regression analysis. Through this work, we found that the latent shared component between R_h and R_s was significantly affected by the duration of the first accumulation period, t_h, and the interaction of α_h and β_h.

The precise parameter values could therefore be used to analyse the qualitative differences between the individual subject's reliance on H_a.

7.1.4.2 Future research questions

Once again, the primary limitation for this analysis was the inflexibility of our fixed parameters across different conditions. In attempting to replicate time-controlled responses after minimal and extended training without parameter adaptation, the model often had to compromise between realistic accuracy dynamics before habits were over-strengthened and the magnitude of R_h and R_s in the latter training condition.

The previous section (Section 7.1.3.2) addressed potential solutions to this issue in future

work, though increasing the complexity of the model would require a much richer dataset. This is particularly important given that the current data was already insufficient to correctly identify whether β_{h_1} and β_{h_2} had different values.
The simplest experimental approaches to improve detailed recovery would likely be to either (1) increase the number of reversals or (2) increase the number of time-controlled trials within the first ~600ms.

Having established that it is possible to quantify habits through a mechanistic RL-EAM, future studies will now be able to investigate how the specific parameters are influenced through different experimental manipulations, such as interrogating which parameters adapt to allow subjects to compensate for increased task complexity or the impact of attention on the application of habits.
Understanding which precise features of habit expression are affected by context and physiological differences will be especially valuable when identifying the underlying causes of maladaptive habit formation and isolating what may cause a given individual to be particularly susceptible to compulsive habit pathologies.

7.2 Methodological Limitations

It is important to acknowledge and assess the methodological limitations that were encountered during this research and how these may impact our conclusions. This next section briefly outlines the implicit assumptions inherent to our use of MLE, Bayesian statistical analysis and the adjustments required when working with suboptimal datasets.

7.2.1 Model selection

7.2.1.1 MLE

In both Chapter 4 and Chapter 6, MLE was applied to determine the parameter values which provided the best replication of an individual dataset for each tested model. This method is well-established and especially useful when working with models which update between trials.

While MLE recovery is sensitive to outliers and small datasets, this was not an issue faced in our work as each individual (both mouse and human) provided many datapoints. Additionally, though overfitting can often be a risk, the number of parameters for all models was relatively small given the complexity of the recovered information. When fitting TD-RDM, qualitative analysis confirmed that simulations were able to replicate behaviour following different sequences of stimuli and responses.

However, MLE is time- and resource-expensive and suffers from the potential local minima that can impair recovery. The latter of these issues was addressed by repeatedly running our fitting procedures from different starting values, but, unfortunately, this approach only worsens the time and resource requirements. Indeed, we have regularly highlighted that further extensions to these models will be increasingly computationally expensive and thus, be limited in their capacity to be fit to data.

Fortunately, the cost of these resource-requirements was partially alleviated by our access to the Advanced Research Computing (ARC) services provided by the University of Oxford, through which we could run the fitting procedures and parallelise all iterations across many computational nodes. This was particularly impactful when completing the w_c optimisation and t_1 heuristic analysis described in Section 6.3.3.2 and Appendix A, during which best-fitting parameters were found for 4 models, across 400 datasets at 21 different w_c and t_1 values. Even with the considerable efficiency of the near-analytical cost function, running either of these analyses required between 6-12 hours per dataset. These time-limitations regrettably precluded us from continuing this work to explore the proposed variety of TD-RDM formulations.

7.2.1.2 BIC and AIC

Once the best-fitting parameters and model likelihood were extracted, BIC analysis was used for model selection and AIC was reported for comparison. The application of information criteria such as these constitutes the industry standard during similar model recovery analyses as they both provide varying degrees of protection against overfitting caused by the inclusion of superfluous parameters.

However, the use of these measures requires an assumption that the penalisation term sufficiently punishes the flexibility provided by additional free parameters without losing model features which may truly reflect the real processes, and thus, improve the fit. Neither BIC nor AIC is likely to be correct in their calculations of such a term as they are, fundamentally, approximations. It is therefore essential to supplement these statistical tests with qualitative analysis and predictions regarding the influence of each parameter on the final data.

Although, of the two, BIC is the more conservative test, it is particularly sensitive to the number of datapoints included since the penalty term includes this value. For this reason, both values were reported.

In Chapter 4, the 3-parameter nature of the most complicated model and the vast datasets for each mouse led us to consider both AIC and BIC in our conclusions. When these measures disagreed, additional qualitative analysis was undertaken to assess why this difference existed. Thankfully, the preferred learning model was consistent between these two measures and the disagreement was mostly caused by the nested nature of TD-AL$_0$ and TD-AL$_\gamma$ combined with markedly low γ_h values in the latter model. In future, these conclusions could be further tested through recovery analysis and direct comparison to the VS data from the same experiment.

In Chapter 6, all conclusions drawn were founded in the BIC values since the models contained a greater number of parameters than TD-AL and there were fewer datapoints.

It must further be acknowledged that these statistical tests (including the subsequent fixed-effects comparison and BMS) share a fundamental assumption that the true model is contained within those tested. Therefore, all conclusions drawn can only be placed relative to each other - they provide evidence that TD-AL and the Habit-Race algorithms produce better explanations for the data tested than the alternatives and not that these are, in fact, the true models used.

We already know that, as with all neurocomputational models, this belief is false since the data cannot be entirely explained by the best-fitting agents. However, the results can be used to support hypotheses and guide further research, i.e., TD-AL is (unsurprisingly)

unlikely to be a complete model of the basal ganglia, but that does not negate that the work presented in Chapter 4 can safely conclude that the dopaminergic data tested is more likely to contain an APE-like signal than a TD-RL RPE or a pure motor response.

7.2.2 Data collection

Lastly, I would like to address the limitation which has the largest impact on these projects - the data was originally collected to answer different questions to those tested here.
More precisely, while the data provided by Greenstreet et al.[3] and Hardwick et al.[4] was designed to assess the existence of dopaminergic APE and expression of human habits, respectively, they were both undertaken prior to the development of these models. As a result, they were not created to specifically interrogate the predictions made by TD-AL or a Habit-Race TD-RDM.
The clearest consequences of this are (1) the inability to test for γ_h at the presentation of predictive cues, (2) the fitting of a learning model to data with missing training sessions, (3) the application of a sub-optimal t_1 heuristic, and (4) the attempt to establish habitual behaviour in datasets with, at most, one reversal of mapping. Throughout this thesis, we have regularly proposed how future work could improve on this and provide a more specific test of continuous APEs and value-free habits in RT data.
However, it must be acknowledged that the datasets themselves were rich and it is highly encouraging that the results were so conclusive despite this non-specificity.

7.3 Outstanding Questions

Before concluding the thesis, this section considers some of the wider questions that arose during our work and which remain, at this stage, unanswered.

7.3.1 What is an action?

In Chapter 3, we briefly touched on how an action may be defined. Throughout the projects presented here, actions and choices have been treated as synonymous - whether a mouse went left or right, or a human pressed a given key, each had its own

corresponding H_a value.

This definition is formed through a specific assumption regarding the mechanism through which the BG affects behaviour, which states that the cortex produces multiple action plans and presents them to the striatum. The BG is then hypothesised to filter through these plans according to prior experience until a single action is dis-inhibited and 'released' by the thalamus. In contrast, the precise details of the action plan, such as the combination of muscle engagement between different groups and sequence of movements required, are instead attributed to the brainstem and the cerebellum, in particular. As always, this is a simplification which allows us to produce tractable and intuitive models of action selection.

In the development of TD-AL, we also assumed that the striatum learnt from action *intensity* with the implication that it was, to some degree, responsible for the vigour with which an action plan is executed. This aligns with the historical studies associating striatal dopamine with motivation and execution of planned movement (Section 2.2.3), but requires the striatum to not only act as a 'gate' for action initiation but also to take into account factors other than just the intended movement and outcome.

Consideration of how an action is defined is particularly relevant in the context of habits, as maladaptive expression is rarely associated with the execution of a single key-press.

Instead, realistic habitual behaviours combine a sequence of movements that occur over a greater length of time. Consider the example of driving home from Chapter 1, the habitual movement requires the involvement of all limbs and the visual system, and it may take several moments before the cognitive system re-engages and remembers that a different goal had been set.

Dezfouli and Balleine[233] described these habitual action-sequences in their presentation of a model-based habit. They argue that the strongest S-R relationship will be the termination of one action and the execution of the action that follows it.

Interestingly, these action-sequences would be similarly applicable under TD-AL. In fact, these associations arise as a natural consequence of action-produced microstimuli. Just as these microstimuli will act as the closest predictor of the remaining action intensity in

TD-AL and as predictive cues for subsequent rewards in TD-RL, so too can these actions be used to build expectation for those that regularly follow.

Further, under the assumption of a non-zero γ_h, Fig. 3.2 illustrates how the APE will transmit to the earliest predictive cue following training. Thus, for an extremely regular series of actions, the APE dynamics will all converge to respond at the same moment in time - at the presentation of whichever cue initiates the sequence of movements.

Taken to its natural conclusion, this means that when habits are completed as expected, there should be no detectable difference in the dopaminergic dynamics between a singular action plan or a series of associated muscular movements. Indeed, we would once again have to rely on omission studies to attempt to disentangle these two, such that a singular movement is skipped without impacting the rest of the sequence.

7.3.2 What is a state?

Under all computational models of action selection, one shared concept is that of a 'state' which guides the outcome according to the agent's context (Section 2.3). These states can be explicit, such as the presentation of a given stimulus in TD-RDM, or they can be a more latent understanding regarding when a specific relationship is appropriate to the environment.

In Chapter 2, we briefly mentioned that habits appear to be suppressed during extinction learning, rather than being completely unlearnt (Section 2.1.2). For this to be true, it inherently requires the habit to be context-specific even when the state is a hidden change in the rules of the task.

This is a particularly significant property underlying the APE hypothesis. Every iteration of an APE-based habit model has required a very small learning rate because these habits are strongly self-reinforcing - if execution of an action promotes future expression then the feedback loop will always be positive.

The adaptation of H_a between states has already been explicitly included in the Habit-Race TD-RDM, as these were further subscripted according to the stimulus seen, s. More precisely, 16 values of H_a existed since each action had a different habitual response

to every cue, though only 4 competed at any given trial. Furthermore, the H_a values were only updated for the accumulators that were racing, the habitual actions in response to other stimuli were unchanged.

This is simple enough when the state changes are obvious, but additional processes must be included to consider how animals respond to latent states. To remain with the multi-alternate forced choice task as an example, we would imagine that if the same reversal occurs multiple times (such that the consistent and habitual errors switch after a set number of trials), eventually the human participants should no longer be surprised by the first error they get in the new state and be able to quickly adapt by reusing the information they previously learnt. In the context of our Habit-Race TD-RDM, 32 values of H_a would therefore be split between these two latent states.

This question therefore requires us to ascertain how these states are created and at what point the agent determines that there are two hidden rules and divides the Q_a and H_a values to learn independently across these conditions. It is highly unlikely that this complex process would be centralised within the BG and, instead, would probably require the introduction of a top-down controller.

7.3.3 Are actions and rewards unique?

Finally, we come to a question which, if answered, could redefine the paradigm of BG computational models.

Every application of APEs across all projects in this thesis has included a core axiom, that the BG variables all share identical computational units that differ only in the information they receive and the precise parameter values used. We have primarily focussed on how this may be applied in the context of rewards and actions, given the wealth of evidence that the BG and associated dopaminergic systems are associated with both.

However, the question remains - are the basal ganglia uniquely focussed on these two features or are more elements encoded with the same unit?

It is becoming increasingly apparent that many heterogeneous PEs are detectable within dopaminergic data.

Historically, early iterations of this work searched, with mixed success, for an 'aversive' prediction error that acted in counterpart with the positive reinforcement of RPEs[296]. In the past decade, signals have been reported which respond to other environmental contexts, including Uchida et al.'s[104] TPE in their dopaminergic theory of weal and woe. Currently and in parallel, many studies are beginning to explore how the BG determines that something is rewarding. More specifically, these projects investigate how dopaminergic signals may vary in response to different reward 'features' according to an animal's internal homeostatic state. For example, a starving animal who has had unlimited access to drinking water is less likely to produce RPEs following a water-drop reward when compared to a water-deprived one. Determining how these varied signals and computational units interact is one of the great questions facing BG research today.

Interestingly, a natural classification emerges for the heterogeneous PEs, one of state (threat, novelty), outcome (reward, aversion), and response (action). Given the absolute wealth of evidence supporting the association of the BG with context (or state) dependent goal-directed actions (A-O) and habits (S-R), accounting for these disparate signals may not require as large a paradigm shift as it first may appear.

7.4 Conclusion

Overall, this thesis has interrogated and expanded upon the computational models of value-free habits created by Miller et al.[1] and Bogacz[2].

I developed two novel computational models which could address some of the remaining questions regarding the concept of an action prediction error. These models were then compared against equivalent value-based systems for their capacity to explain neural and behavioural data.

Specifically, I found that the microstimulus model[176] provided a better fit to dopaminergic data when it was adapted to learn both from which choices an animal had previously made, but also the intensity with which the action was carried out. Following this, I extended the reinforcement-learning race diffusion model[226] to create a generalisable evidence accumulation system which could adapt the drift-rate of an accumulator mid-trial. In doing so, I produced an associated (near-) analytical likelihood function which allowed the fitting procedures to remain tractable. Finally, I found evidence supporting that, under specific time-constraints, humans apply two action-selection processes which differ in their non-decision times and established that the faster of these systems is likely to learn using an action prediction error, rather than a Q-learning variable.

In combination, these results provide a strong proof-of-concept that habitual behaviours can be explained through value-free learning systems. Further, I have shown that the discrepancy between the experimental and computational definition of habits can be solved through the introduction of action prediction errors.

References

1. Miller, K. J., Shenhav, A. & Ludvig, E. A. Habits without values. *Psychological Review* **126.** Place: US Publisher: American Psychological Association, 292–311. ISSN: 1939-1471 (2019).
2. Bogacz, R. Dopamine role in learning and action inference. *eLife* **9** (eds Kahnt, T. & Wassum, K. M.) Publisher: eLife Sciences Publications, Ltd, e53262. ISSN: 2050-084X. https://doi.org/10.7554/eLife.53262 (2021) (July 2020).
3. Greenstreet, F. *et al.* Dopaminergic action prediction errors serve as a value-free teaching signal. en. *Nature*. Publisher: Nature Publishing Group, 1–10. ISSN: 1476-4687. https://www.nature.com/articles/s41586-025-09008-9 (2025) (May 2025).
4. Hardwick, R. M., Forrence, A. D., Krakauer, J. W. & Haith, A. M. Time-dependent competition between goal-directed and habitual response preparation. en. *Nature Human Behaviour* **3.** Number: 12 Publisher: Nature Publishing Group, 1252–1262. ISSN: 2397-3374. https://www.nature.com/articles/s41562-019-0725-0 (2023) (Dec. 2019).
5. Rescorla, R. A. & Solomon, R. L. Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review* **74.** Place: US Publisher: American Psychological Association, 151–182. ISSN: 1939-1471 (1967).
6. Belin, D., Belin-Rauscent, A., Murray, J. E. & Everitt, B. J. Addiction: failure of control over maladaptive incentive habits. en. *Current Opinion in Neurobiology. 23/4 Addiction* **23,** 564–572. ISSN: 0959-4388. http://www.sciencedirect.com/science/article/pii/S0959438813000445 (2021) (Aug. 2013).

7. Zack, M., St George, R. & Clark, L. Dopaminergic signaling of uncertainty and the aetiology of gambling addiction. eng. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* **99,** 109853. ISSN: 1878-4216 (Apr. 2020).
8. Marques, A., Durif, F. & Fernagut, P.-O. Impulse control disorders in Parkinson's disease. en. *Journal of Neural Transmission* **125,** 1299–1312. ISSN: 1435-1463. https://doi.org/10.1007/s00702-018-1870-8 (2021) (Aug. 2018).
9. Muresanu, D. F., Stan, A. & Buzoianu, A. Neuroplasticity and impulse control disorders. *Journal of the Neurological Sciences* **316,** 15–20. ISSN: 0022-510X. https://www.sciencedirect.com/science/article/pii/S0022510X12000172 (2025) (May 2012).
10. Watkins, E. R. & Nolen-Hoeksema, S. A habit-goal framework of depressive rumination. *Journal of Abnormal Psychology* **123.** Place: US Publisher: American Psychological Association, 24–34. ISSN: 1939-1846 (2014).
11. Burguière, E., Monteiro, P., Mallet, L., Feng, G. & Graybiel, A. M. Striatal circuits, habits, and implications for obsessive–compulsive disorder. *Current Opinion in Neurobiology. SI: Neuropsychiatry* **30,** 59–65. ISSN: 0959-4388. https://www.sciencedirect.com/science/article/pii/S0959438814001706 (2025) (Feb. 2015).
12. Leckman, J. F. & Riddle, M. A. Tourette's Syndrome: When Habit-Forming Systems Form Habits of Their Own? English. *Neuron* **28.** Publisher: Elsevier, 349–354. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(00)00114-8 (2025) (Nov. 2000).
13. Erdman, A. *et al.* Ruminative Tendency Relates to Ventral Striatum Functionality: Evidence From Task and Resting-State fMRI. English. *Frontiers in Psychiatry* **11.** Publisher: Frontiers. ISSN: 1664-0640. https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2020.00067/full (2025) (Feb. 2020).
14. Belin, D., Jonkman, S., Dickinson, A., Robbins, T. W. & Everitt, B. J. Parallel and interactive learning processes within the basal ganglia: Relevance for the understanding of addiction. en. *Behavioural Brain Research. Special issue on the role of the basal ganglia in learning and memory* **199,** 89–102. ISSN: 0166-4328. https:

//www.sciencedirect.com/science/article/pii/S0166432808005354 (2021) (Apr. 2009).

15. Robbins, T. W. & Costa, R. M. Habits. en. *Current Biology* **27,** R1200–R1206. ISSN: 0960-9822. https://www.sciencedirect.com/science/article/pii/S0960982217312587 (2022) (Nov. 2017).
16. Smith, K. S. & Graybiel, A. M. Investigating habits: strategies, technologies and models. English. *Frontiers in Behavioral Neuroscience* **8.** Publisher: Frontiers. ISSN: 1662-5153. https://www.frontiersin.org/journals/behavioral-neuroscience/articles/10.3389/fnbeh.2014.00039/full (2025) (Feb. 2014).
17. Daw, N. D. Are we of two minds? en. *Nature Neuroscience* **21.** Publisher: Nature Publishing Group, 1497–1499. ISSN: 1546-1726. https://www.nature.com/articles/s41593-018-0258-2 (2025) (Nov. 2018).
18. Huang, Q. *Model-Based or Model-Free, a Review of Approaches in Reinforcement Learning* in *2020 International Conference on Computing and Data Science (CDS)* (Aug. 2020), 219–221. https://ieeexplore.ieee.org/abstract/document/9275964?casa_token=Sm3yD7Ue_0sAAAAA:XUzURH6PmEJYHg3FK4siLNoSGuTSm7uX_hapZsUNwoBVd9-rhFsl40g48MclFhZVaxDD-wg (2025).
19. Orbell, S. & Verplanken, B. The automatic component of habit in health behavior: Habit as cue-contingent automaticity. *Health Psychology* **29.** Place: US Publisher: American Psychological Association, 374–383. ISSN: 1930-7810 (2010).
20. Dickinson, A. & Balleine, B. Motivational control of goal-directed action. en. *Animal Learning & Behavior* **22,** 1–18. ISSN: 1532-5830. https://doi.org/10.3758/BF03199951 (2025) (Mar. 1994).
21. De Wit, S. & Dickinson, A. Associative theories of goal-directed behaviour: a case for animal–human translational models. en. *Psychological Research PRPF* **73,** 463–476. ISSN: 1430-2772. https://doi.org/10.1007/s00426-009-0230-6 (2025) (July 2009).
22. Norman, D. A. & Shallice, T. en. in *Consciousness and Self-Regulation: Advances in Research and Theory Volume 4* (eds Davidson, R. J., Schwartz, G. E. & Shapiro, D.) 1–18 (Springer US, Boston, MA, 1986). ISBN: 978-1-4757-0629-1. https://doi.org/10.1007/978-1-4757-0629-1_1 (2025).

23. Bouton, M. E. Context, attention, and the switch between habit and goal-direction in behavior. en. *Learning & Behavior* **49,** 349–362. ISSN: 1543-4508. https://doi.org/10.3758/s13420-021-00488-z (2025) (Dec. 2021).

24. Evans, J. S. B. T. & Stanovich, K. E. Dual-Process Theories of Higher Cognition: Advancing the Debate. EN. *Perspectives on Psychological Science* **8.** Publisher: SAGE Publications Inc, 223–241. ISSN: 1745-6916. https://doi.org/10.1177/1745691612460685 (2025) (May 2013).

25. Decker, J. H., Otto, A. R., Daw, N. D. & Hartley, C. A. From Creatures of Habit to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based Reinforcement Learning. EN. *Psychological Science* **27.** Publisher: SAGE Publications Inc, 848–858. ISSN: 0956-7976. https://doi.org/10.1177/0956797616639301 (2025) (June 2016).

26. .Corbit, L. H. Understanding the balance between goal-directed and habitual behavioral control. *Current Opinion in Behavioral Sciences. Habits and Skills* **20,** 161–168. ISSN: 2352-1546. https://www.sciencedirect.com/science/article/pii/S2352154617302371 (2025) (Apr. 2018).

27. Schwabe, L. & Wolf, O. T. Stress-induced modulation of instrumental behavior: From goal-directed to habitual control of action. *Behavioural Brain Research* **219,** 321–328. ISSN: 0166-4328. https://www.sciencedirect.com/science/article/pii/S0166432811000258 (2025) (June 2011).

28. Skinner, B. F. Operant behavior. *American Psychologist* **18.** Place: US Publisher: American Psychological Association, 503–515. ISSN: 1935-990X (1963).

29. SKINNER, B. F. THE EXPERIMENTAL ANALYSIS OF BEHAVIOR. *American Scientist* **45.** Publisher: Sigma Xi, The Scientific Research Society, 343–371. ISSN: 0003-0996. https://www.jstor.org/stable/27826953 (2022) (1957).

30. Postman, L. The history and present status of the law of effect. *Psychological Bulletin* **44.** Place: US Publisher: American Psychological Association, 489–563. ISSN: 1939-1455 (1947).

31. Urcelay, G. P. & Jonkman, S. Delayed rewards facilitate habit formation. *Journal of Experimental Psychology: Animal Learning and Cognition* **45.** Place: US Publisher: American Psychological Association, 413–421. ISSN: 2329-8464 (2019).

32. Wiltgen, B. J. *et al.* The Effect of Ratio and Interval Training on Pavlovian-Instrumental Transfer in Mice. en. *PLOS ONE* **7.** Publisher: Public Library of Science, e48227. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0048227 (2025) (Oct. 2012).

33. Adams, C. D. & Dickinson, A. Instrumental Responding following Reinforcer Devaluation. en. *The Quarterly Journal of Experimental Psychology Section B* **33.** Publisher: SAGE Publications, 109–121. ISSN: 0272-4995. https://doi.org/10.1080/14640748108400816 (2025) (May 1981).

34. Rescorla, R. A. Response-outcome versus outcome-response associations in instrumental learning. en. *Animal Learning & Behavior* **20,** 223–232. ISSN: 1532-5830. https://doi.org/10.3758/BF03213376 (2025) (Sept. 1992).

35. Suzuki, A. *et al.* Memory Reconsolidation and Extinction Have Distinct Temporal and Biochemical Signatures. en. *Journal of Neuroscience* **24,** 4787–4795. ISSN: 0270-6474, 1529-2401. http://www.jneurosci.org/content/24/20/4787 (2019) (May 2004).

36. Dunsmoor, J. E., Niv, Y., Daw, N. & Phelps, E. A. Rethinking Extinction. English. *Neuron* **88.** Publisher: Elsevier, 47–63. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(15)00817-X (2025) (Oct. 2015).

37. Campos, R. C., Dias, C., Darlot, F. & Cador, M. Double dissociation between actions of dopamine D1 and D2 receptors of the ventral and dorsolateral striatum to produce reinstatement of cocaine seeking behavior. en. *Neuropharmacology* **172,** 108113. ISSN: 0028-3908. https://www.sciencedirect.com/science/article/pii/S0028390820301817 (2022) (Aug. 2020).

38. Acosta, J. I., Thiel, K. J., Sanabria, F., Browning, J. R. & Neisewander, J. L. Effect of schedule of reinforcement on cue-elicited reinstatement of cocaine-seeking behavior. en-US. *Behavioural Pharmacology* **19,** 129. ISSN: 0955-8810. https://

journals . lww . com / behaviouralpharm / abstract / 2008 / 03000 / effect _ of _ schedule_of_reinforcement_on.4.aspx (2025) (Mar. 2008).

39. Dickinson, A., Balleine, B., Watt, A., Gonzalez, F. & Boakes, R. A. Motivational control after extended instrumental training. en. *Animal Learning & Behavior* **23,** 197–206. ISSN: 1532-5830. https://doi.org/10.3758/BF03199935 (2025) (June 1995).
40. Ferster, C. B. & Skinner, B. F. *Schedules of reinforcement* Pages: vii, 744 (Appleton-Century-Crofts, East Norwalk, CT, US, 1957).
41. Morgan, D. L. Schedules of Reinforcement at 50: A Retrospective Appreciation. en. *The Psychological Record* **60,** 151–172. ISSN: 2163-3452. https://doi.org/10.1007/BF03395699 (2022) (Jan. 2010).
42. Dickinson, A., Nicholas, D. J. & Adams, C. D. The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B* **35.** Publisher: Routledge _eprint: https://doi.org/10.1080/14640748308400912, 35–51. ISSN: 0272-4995. https://doi.org/10.1080/14640748308400912 (2022) (Feb. 1983).
43. DeRusso, A. L. *et al.* Instrumental Uncertainty as a Determinant of Behavior Under Interval Schedules of Reinforcement. *Frontiers in Integrative Neuroscience* **4,** 17. ISSN: 1662-5145. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2901087/ (2022) (May 2010).
44. Dickinson, A. & Weiskrantz, L. Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **308.** Publisher: Royal Society, 67–78. https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1985.0010 (2022) (Feb. 1985).
45. De Villiers, P. A. & Herrnstein, R. J. Toward a law of response strength. *Psychological Bulletin* **83.** Place: US Publisher: American Psychological Association, 1131–1153. ISSN: 1939-1455 (1976).
46. Garr, E., Bushra, B., Tu, N. & Delamater, A. R. Goal-directed control on interval schedules does not depend on the action–outcome correlation. *Journal of Experimental Psychology: Animal Learning and Cognition* **46.** Place: US Publisher: American Psychological Association, 47–64. ISSN: 2329-8464 (2020).

47. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. en. *Nature Neuroscience* **8.** Publisher: Nature Publishing Group, 1704–1711. ISSN: 1546-1726. https://www.nature.com/articles/nn1560 (2025) (Dec. 2005).

48. Mikhael, J. G. & Bogacz, R. Learning Reward Uncertainty in the Basal Ganglia. en. *PLOS Computational Biology* **12.** Publisher: Public Library of Science, e1005062. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005062 (2022) (Sept. 2016).

49. Thrailkill, E. A., Trask, S., Vidal, P., Alcalá, J. A. & Bouton, M. E. Stimulus control of actions and habits: A role for reinforcer predictability and attention in the development of habitual behavior. *Journal of Experimental Psychology: Animal Learning and Cognition* **44.** Place: US Publisher: American Psychological Association, 370–384. ISSN: 2329-8464 (2018).

50. Wingard, J. C. & Packard, M. G. The amygdala and emotional modulation of competition between cognitive and habit memory. *Behavioural Brain Research* **193,** 126–131. ISSN: 0166-4328. http://www.sciencedirect.com/science/article/pii/S016643280800243X (2018) (Nov. 2008).

51. Dias-Ferreira, E. *et al.* Chronic Stress Causes Frontostriatal Reorganization and Affects Decision-Making. en. *Science* **325,** 621–625. ISSN: 0036-8075, 1095-9203. http://science.sciencemag.org/content/325/5940/621 (2018) (July 2009).

52. Schwabe, L. & Wolf, O. T. Stress Prompts Habit Behavior in Humans. en. *Journal of Neuroscience* **29.** Publisher: Society for Neuroscience Section: Articles, 7191–7198. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/29/22/7191 (2025) (June 2009).

53. Dolan, R. J. & Dayan, P. Goals and Habits in the Brain. English. *Neuron* **80.** Publisher: Elsevier, 312–325. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(13)00805-2 (2025) (Oct. 2013).

54. Yin, H. H. & Knowlton, B. J. The role of the basal ganglia in habit formation. en. *Nature Reviews Neuroscience* **7.** Number: 6 Publisher: Nature Publishing Group,

464–476. ISSN: 1471-0048. https://www.nature.com/articles/nrn1919 (2022) (June 2006).

55. Graybiel, A. M. & Grafton, S. T. The Striatum: Where Skills and Habits Meet. en. *Cold Spring Harbor Perspectives in Biology* **7.** Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, a021691. ISSN: , 1943-0264. http://cshperspectives.cshlp.org/content/7/8/a021691 (2025) (Aug. 2015).

56. Balleine, B. W. & Dickinson, A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37,** 407–419. ISSN: 0028-3908. https://www.sciencedirect.com/science/article/pii/S0028390898000331 (2025) (Apr. 1998).

57. Rocha, G. S. *et al.* Basal ganglia for beginners: the basic concepts you need to know and their role in movement control. English. *Frontiers in Systems Neuroscience* **17.** Publisher: Frontiers. ISSN: 1662-5137. https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/fnsys.2023.1242929/full (2025) (Aug. 2023).

58. Balleine, B. W. & O'Doherty, J. P. Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. en. *Neuropsychopharmacology* **35.** Publisher: Nature Publishing Group, 48–69. ISSN: 1740-634X. https://www.nature.com/articles/npp2009131 (2025) (Jan. 2010).

59. Alexander, G. E., DeLong, M. R. & Strick, P. L. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience* **9.** Place: US Publisher: Annual Reviews, 357–381. ISSN: 1545-4126 (1986).

60. Chevalier, G. & Deniau, J. M. Disinhibition as a basic process in the expression of striatal functions. English. *Trends in Neurosciences* **13.** Publisher: Elsevier, 277–280. ISSN: 0166-2236, 1878-108X. https://www.cell.com/trends/neurosciences/abstract/0166-2236(90)90109-N (2025) (July 1990).

61. Alexander, G. E. & Crutcher, M. D. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. English. *Trends in Neurosciences* **13.**

Publisher: Elsevier, 266–271. ISSN: 0166-2236, 1878-108X. https://www.cell.com/trends/neurosciences/abstract/0166-2236(90)90107-L (2025) (July 1990).

62. Guttman, M. Receptors in the Basal Ganglia. en. *Canadian Journal of Neurological Sciences* **14,** 395–401. ISSN: 0317-1671, 2057-0155. https://www.cambridge.org/core/journals/canadian-journal-of-neurological-sciences/article/receptors-in-the-basal-ganglia/DE4C9F17F12D7FB07DB8310BB5BA17B2 (2025) (Aug. 1987).
63. Bolam, J. P., Hanley, J. J., Booth, P. a. C. & Bevan, M. D. Synaptic organisation of the basal ganglia. en. *The Journal of Anatomy* **196,** 527–542. ISSN: 1553-0795, 0002-9106. https://www.cambridge.org/core/journals/journal-of-anatomy/article/abs/synaptic-organisation-of-the-basal-ganglia/97F5E3C65B4D25D01C79A1183768E70E (2025) (May 2000).
64. Nauta, W. J. K. en. in *Neuroanatomy* (ed Nauta, W. J. H.) 598–618 (Birkhäuser, Boston, MA, 1993). ISBN: 978-1-4684-7920-1. https://doi.org/10.1007/978-1-4684-7920-1_30 (2025).
65. Haber, S. N., Fudge, J. L. & McFarland, N. R. Striatonigrostriatal Pathways in Primates Form an Ascending Spiral from the Shell to the Dorsolateral Striatum. en. *Journal of Neuroscience* **20.** Publisher: Society for Neuroscience Section: ARTICLE, 2369–2382. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/20/6/2369 (2025) (Mar. 2000).
66. Olds, J. & Milner, P. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology* **47.** Place: US Publisher: American Psychological Association, 419–427. ISSN: 0021-9940 (1954).
67. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. en. *Science* **275.** Publisher: American Association for the Advancement of Science Section: Articles, 1593–1599. ISSN: 0036-8075, 1095-9203. https://science.sciencemag.org/content/275/5306/1593 (2021) (Mar. 1997).
68. Corbit, L. H. & Balleine, B. W. Double Dissociation of Basolateral and Central Amygdala Lesions on the General and Outcome-Specific Forms of Pavlovian-Instrumental Transfer. en. *Journal of Neuroscience* **25.** Publisher:

Society for Neuroscience Section: Behavioral/Systems/Cognitive, 962–970. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/25/4/962 (2025) (Jan. 2005).

69. Setlow, B., Holland, P. C. & Gallagher, M. Disconnection of the basolateral amygdala complex and nucleus accumbens impairs appetitive Pavlovian second-order conditioned responses. *Behavioral Neuroscience* **116.** Place: US Publisher: American Psychological Association, 267–275. ISSN: 1939-0084 (2002).
70. Killcross, S. & Coutureau, E. Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. *Cerebral Cortex* **13,** 400–408. ISSN: 1047-3211. https://doi.org/10.1093/cercor/13.4.400 (2025) (Apr. 2003).
71. Packard, M. G. & McGaugh, J. L. Inactivation of Hippocampus or Caudate Nucleus with Lidocaine Differentially Affects Expression of Place and Response Learning. *Neurobiology of Learning and Memory* **65,** 65–72. ISSN: 1074-7427. https://www.sciencedirect.com/science/article/pii/S1074742796900076 (2025) (Jan. 1996).
72. Guida, P., Michiels, M., Redgrave, P., Luque, D. & Obeso, I. An fMRI meta-analysis of the role of the striatum in everyday-life vs laboratory-developed habits. *Neuroscience & Biobehavioral Reviews* **141,** 104826. ISSN: 0149-7634. https://www.sciencedirect.com/science/article/pii/S0149763422003153 (2025) (Oct. 2022).
73. Yin, H. H., Ostlund, S. B., Knowlton, B. J. & Balleine, B. W. The role of the dorsomedial striatum in instrumental conditioning. en. *European Journal of Neuroscience* **22.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2005.04218.x, 513–523. ISSN: 1460-9568. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2005.04218.x (2025) (2005).
74. Ragozzino, K. E., Leutgeb, S. & Mizumori, S. J. Dorsal striatal head direction and hippocampal place representations during spatial navigation. en. *Experimental Brain Research* **139,** 372–376. ISSN: 1432-1106. https://doi.org/10.1007/s002210100795 (2025) (Aug. 2001).
75. Miyachi, S., Hikosaka, O. & Lu, X. Differential activation of monkey striatal neurons in the early and late stages of procedural learning. en. *Experimental Brain Research*

146, 122–126. ISSN: 1432-1106. https://doi.org/10.1007/s00221-002-1213-7 (2025) (Sept. 2002).

76. Bender, B. N., Stringfield, S. J. & Torregrossa, M. M. Changes in dorsomedial striatum activity during expression of goal-directed vs. habit-like cue-induced cocaine seeking. *Addiction Neuroscience* **11,** 100149. ISSN: 2772-3925. https://www.sciencedirect.com/science/article/pii/S2772392524000087 (2025) (June 2024).

77. Smith, K. S. & Graybiel, A. M. Using optogenetics to study habits. *Brain Research. Optogenetics and Pharmacogenetics in Neuronal Function and Dysfunction* **1511,** 102–114. ISSN: 0006-8993. https://www.sciencedirect.com/science/article/pii/S0006899313000516 (2025) (May 2013).

78. Crego, A. C. G. *et al.* Complementary Control over Habits and Behavioral Vigor by Phasic Activity in the Dorsolateral Striatum. en. *Journal of Neuroscience* **40.** Publisher: Society for Neuroscience Section: Research Articles, 2139–2153. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/40/10/2139 (2025) (Mar. 2020).

79. Thorn, C. A., Atallah, H., Howe, M. & Graybiel, A. M. Differential Dynamics of Activity Changes in Dorsolateral and Dorsomedial Striatal Loops during Learning. English. *Neuron* **66.** Publisher: Elsevier, 781–795. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(10)00329-6 (2025) (June 2010).

80. Turner, K. M., Svegborn, A., Langguth, M., McKenzie, C. & Robbins, T. W. Opposing Roles of the Dorsolateral and Dorsomedial Striatum in the Acquisition of Skilled Action Sequencing in Rats. *The Journal of Neuroscience* **42,** 2039–2051. ISSN: 0270-6474. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8916752/ (2025) (Mar. 2022).

81. Whishaw, I. Q., Mittleman, G., Bunch, S. T. & Dunnett, S. B. Impairments in the acquisition, retention and selection of spatial navigation strategies after medial caudate-putamen lesions in rats. *Behavioural Brain Research* **24,** 125–138. ISSN: 0166-4328. https://www.sciencedirect.com/science/article/pii/0166432887902506 (2025) (May 1987).

82. De Wit, S. *et al.* Shifting the balance between goals and habits: Five failures in experimental habit induction. eng. *Journal of Experimental Psychology. General* **147,** 1043–1065. ISSN: 1939-2222 (July 2018).

83. Gillan, C. M. *et al.* Disruption in the Balance Between Goal-Directed Behavior and Habit Learning in Obsessive-Compulsive Disorder. *American Journal of Psychiatry* **168.** Publisher: American Psychiatric Publishing, 718–726. ISSN: 0002-953X. https://psychiatryonline.org/doi/full/10.1176/appi.ajp.2011.10071062 (2025) (July 2011).

84. De Wit, S. *et al.* Reliance on habits at the expense of goal-directed control following dopamine precursor depletion. en. *Psychopharmacology* **219,** 621–631. ISSN: 1432-2072. https://doi.org/10.1007/s00213-011-2563-2 (2025) (Jan. 2012).

85. Tricomi, E., Balleine, B. W. & O'Doherty, J. P. A specific role for posterior dorsolateral striatum in human habit learning. en. *European Journal of Neuroscience* **29.** Publisher: John Wiley & Sons, Ltd, 2225–2232. ISSN: 1460-9568. https://onlinelibrary.wiley.com/doi/10.1111/j.1460-9568.2009.06796.x (2025) (June 2009).

86. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. English. *Neuron* **66.** Publisher: Elsevier, 585–595. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(10)00287-4 (2025) (May 2010).

87. Valentin, V. V., Dickinson, A. & O'Doherty, J. P. Determining the Neural Substrates of Goal-Directed Learning in the Human Brain. en. *Journal of Neuroscience* **27.** Publisher: Society for Neuroscience Section: Articles, 4019–4026. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/27/15/4019 (2025) (Apr. 2007).

88. Tanaka, S. C. *et al.* en. in *Behavioral Economics of Preferences, Choices, and Happiness* (eds Ikeda, S., Kato, H. K., Ohtake, F. & Tsutsui, Y.) 593–616 (Springer Japan, Tokyo, 2016). ISBN: 978-4-431-55402-8. https://doi.org/10.1007/978-4-431-55402-8_22 (2025).

89. Tanaka, S. C. *et al.* Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. en. *Nature Neuroscience* **7.** Publisher: Nature Publishing Group, 887–893. ISSN: 1546-1726. https://www.nature.com/articles/nn1279 (2025) (Aug. 2004).

90. Schwabe, L. & Wolf, O. T. Socially evaluated cold pressor stress after instrumental learning favors habits over goal-directed action. *Psychoneuroendocrinology* **35,** 977–986. ISSN: 0306-4530. https://www.sciencedirect.com/science/article/pii/S0306453009003722 (2025) (Aug. 2010).

91. Pool, E. R. *et al.* Determining the effects of training duration on the behavioral expression of habitual control in humans: a multilaboratory investigation. en. *Learning & Memory* **29.** Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, 16–28. ISSN: 1072-0502, 1549-5485. http://learnmem.cshlp.org/content/29/1/16 (2025) (Jan. 2022).

92. Schultz, W. Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience* **18,** 23–32. ISSN: 1294-8322. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4826767/ (2021) (Mar. 2016).

93. Chinta, S. J. & Andersen, J. K. Dopaminergic neurons. *The International Journal of Biochemistry & Cell Biology. Cancer and Aging at the Crossroads* **37,** 942–946. ISSN: 1357-2725. https://www.sciencedirect.com/science/article/pii/S1357272504003711 (2025) (May 2005).

94. Klein, M. O. *et al.* Dopamine: Functions, Signaling, and Association with Neurological Diseases. eng. *Cellular and Molecular Neurobiology* **39,** 31–59. ISSN: 1573-6830 (Jan. 2019).

95. Watabe-Uchida, M., Zhu, L., Ogawa, S. K., Vamanrao, A. & Uchida, N. Whole-Brain Mapping of Direct Inputs to Midbrain Dopamine Neurons. English. *Neuron* **74.** Publisher: Elsevier, 858–873. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(12)00281-4 (2025) (June 2012).

96. Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. en. *Trends in Neurosciences* **12,** 366–375. ISSN: 0166-2236. https://www.sciencedirect.com/science/article/pii/016622368990074X (2022) (Jan. 1989).

97. Matsuda, W. *et al.* Single Nigrostriatal Dopaminergic Neurons Form Widely Spread and Highly Dense Axonal Arborizations in the Neostriatum. en. *Journal of Neuroscience* **29.** Publisher: Society for Neuroscience Section: Articles, 444–453. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/29/2/444 (2025) (Jan. 2009).

98. Gerfen, C. R., Keefe, K. & Gauda, E. D1 and D2 dopamine receptor function in the striatum: coactivation of D1- and D2-dopamine receptors on separate populations of neurons results in potentiated immediate early gene response in D1-containing neurons. *The Journal of Neuroscience* **15,** 8167–8176. ISSN: 0270-6474. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6577952/ (2025) (Dec. 1995).

99. Jones-Tabah, J., Mohammad, H., Paulus, E. G., Clarke, P. B. S. & Hébert, T. E. The Signaling and Pharmacology of the Dopamine D1 Receptor. *Frontiers in Cellular Neuroscience* **15,** 806618. ISSN: 1662-5102. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8801442/ (2025) (Jan. 2022).

100. Neve, K. A., Seamans, J. K. & Trantham-Davidson, H. Dopamine Receptor Signaling. EN. *Journal of Receptors and Signal Transduction.* Publisher: Taylor & Francis. https://www.tandfonline.com/doi/abs/10.1081/RRS-200029981 (2025) (Jan. 2004).

101. Valjent, E. & Gangarossa, G. The Tail of the Striatum: From Anatomy to Connectivity and Function. en. *Trends in Neurosciences* **44,** 203–214. ISSN: 0166-2236. https://www.sciencedirect.com/science/article/pii/S0166223620302496 (2022) (Mar. 2021).

102. Xiong, Q., Znamenskiy, P. & Zador, A. M. Selective corticostriatal plasticity during acquisition of an auditory discrimination task. en. *Nature* **521.** Publisher: Nature Publishing Group, 348–351. ISSN: 1476-4687. https://www.nature.com/articles/nature14225 (2025) (May 2015).

103. Menegas, W., Babayan, B. M., Uchida, N. & Watabe-Uchida, M. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife* **6** (ed Westbrook, G. L.) Publisher: eLife Sciences Publications, Ltd, e21886. ISSN: 2050-084X. https://doi.org/10.7554/eLife.21886 (2025) (Jan. 2017).

104. Watabe-Uchida, M. & Uchida, N. Multiple Dopamine Systems: Weal and Woe of Dopamine. en. *Cold Spring Harbor Symposia on Quantitative Biology* **83.** Publisher: Cold Spring Harbor Laboratory Press, 83–95. ISSN: 0091-7451, 1943-4456. http://symposium.cshlp.org/content/83/83 (2022) (Jan. 2018).

105. Yokel, R. A. & Wise, R. A. Increased Lever Pressing for Amphetamine After Pimozide in Rats: Implications for a Dopamine Theory of Reward. *Science* **187.** Publisher: American Association for the Advancement of Science, 547–549. https://www.science.org/doi/10.1126/science.1114313 (2025) (Feb. 1975).

106. Sutton, R. S. Learning to predict by the methods of temporal differences. en. *Machine Learning* **3,** 9–44. ISSN: 1573-0565. https://doi.org/10.1007/BF00115009 (2025) (Aug. 1988).

107. Van Elzelingen, W. *et al.* Striatal dopamine signals are region specific and temporally stable across action-sequence habit formation. *Current Biology* **32,** 1163–1174.E6. ISSN: 0960-9822 (Feb. 2022).

108. Van Elzelingen, W. *et al.* A unidirectional but not uniform striatal landscape of dopamine signaling for motivational stimuli. *Proceedings of the National Academy of Sciences* **119.** Publisher: Proceedings of the National Academy of Sciences, e2117270119. https://www.pnas.org/doi/full/10.1073/pnas.2117270119 (2022) (May 2022).

109. Brown, H. D., McCutcheon, J. E., Cone, J. J., Ragozzino, M. E. & Roitman, M. F. Primary food reward and reward-predictive stimuli evoke different patterns of phasic dopamine signaling throughout the striatum. en. *European Journal of Neuroscience* **34.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2011.07914.x, 1997–2006. ISSN: 1460-9568. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2011.07914.x (2025) (2011).

110. Matsumoto, M. & Hikosaka, O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. en. *Nature* **459.** Publisher: Nature Publishing Group, 837–841. ISSN: 1476-4687. https://www.nature.com/articles/nature08028 (2025) (June 2009).

111. Jong, J. W. d. *et al.* A Neural Circuit Mechanism for Encoding Aversive Stimuli in the Mesolimbic Dopamine System. English. *Neuron* **101.** Publisher: Elsevier, 133–151.e7. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(18)30996-6 (2025) (Jan. 2019).

112. Matsumoto, H., Tian, J., Uchida, N. & Watabe-Uchida, M. Midbrain dopamine neurons signal aversion in a reward-context-dependent manner. *eLife* **5** (ed Costa, R. M.) Publisher: eLife Sciences Publications, Ltd, e17328. ISSN: 2050-084X. https://doi.org/10.7554/eLife.17328 (2025) (Oct. 2016).

113. Cone, J. J., McCutcheon, J. E. & Roitman, M. F. Ghrelin Acts as an Interface between Physiological State and Phasic Dopamine Signaling. en. *Journal of Neuroscience* **34.** Publisher: Society for Neuroscience Section: Articles, 4905–4913. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/34/14/4905 (2025) (Apr. 2014).

114. Schultz, W. Recent advances in understanding the role of phasic dopamine activity. *F1000Research* **8,** F1000 Faculty Rev–1680. ISSN: 2046-1402. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6760455/ (2025) (Sept. 2019).

115. Diuk, C., Tsai, K., Wallis, J., Botvinick, M. & Niv, Y. Hierarchical Learning Induces Two Simultaneous, But Separable, Prediction Errors in Human Basal Ganglia. en. *Journal of Neuroscience* **33.** Publisher: Society for Neuroscience Section: Articles, 5797–5805. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/33/13/5797 (2025) (Mar. 2013).

116. Millidge, B., Song, Y., Lak, A., Walton, M. E. & Bogacz, R. Reward Bases: A simple mechanism for adaptive acquisition of multiple reward types. en. *PLOS Computational Biology* **20.** Publisher: Public Library of Science, e1012580. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012580 (2025) (Nov. 2024).

117. Fahn, S. The history of dopamine and levodopa in the treatment of Parkinson's disease. en. *Movement Disorders* **23.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mds.22028, S497–S508. ISSN: 1531-8257. https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.22028 (2025) (2008).

118. Muñoz, J. M., Williams, J. T. & Lebowitz, J. J. Morphological and functional decline of the SNc in a model of progressive parkinsonism. en. *npj Parkinson's Disease* **11.** Publisher: Nature Publishing Group, 1–9. ISSN: 2373-8057. https://www.nature.com/articles/s41531-025-00873-9 (2025) (Jan. 2025).

119. Ungerstedt, U. Adipsia and Aphagia after 6-Hydroxydopamine Induced Degeneration of the Nigro-striatal Dopamine System. en. *Acta Physiologica Scandinavica* **82.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-201X.1971.tb11001.x, 95–122. ISSN: 1365-201X. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-201X.1971.tb11001.x (2025) (1971).

120. Cenci, M. A. Dopamine dysregulation of movement control in l-DOPA-induced dyskinesia. English. *Trends in Neurosciences* **30.** Publisher: Elsevier, 236–243. ISSN: 0166-2236, 1878-108X. https://www.cell.com/trends/neurosciences/abstract/S0166-2236(07)00066-5 (2025) (May 2007).

121. Schultz, W., Ruffieux, A. & Aebischer, P. The activity of pars compacta neurons of the monkey substantia nigra in relation to motor activation. en. *Experimental Brain Research* **51,** 377–387. ISSN: 1432-1106. https://doi.org/10.1007/BF00237874 (2025) (Aug. 1983).

122. Baunez, C. & Robbins, T. W. Effects of dopamine depletion of the dorsal striatum and further interaction with subthalamic nucleus lesions in an attentional task in the rat. *Neuroscience* **92,** 1343–1356. ISSN: 0306-4522. https://www.sciencedirect.com/science/article/pii/S0306452299000652 (2025) (June 1999).

123. Salamone, J. D., Correa, M., Mingote, S. & Weber, S. M. Nucleus Accumbens Dopamine and the Regulation of Effort in Food-Seeking Behavior: Implications for Studies of Natural Motivation, Psychiatry, and Drug Abuse. *The Journal of*

Pharmacology and Experimental Therapeutics **305,** 1–8. ISSN: 0022-3565. https://www.sciencedirect.com/science/article/pii/S0022356524362445 (2025) (Apr. 2003).

124. Ishiwari, K., Weber, S. M., Mingote, S., Correa, M. & Salamone, J. D. Accumbens dopamine and the regulation of effort in food-seeking behavior: modulation of work output by different ratio or force requirements. *Behavioural Brain Research* **151,** 83–91. ISSN: 0166-4328. https://www.sciencedirect.com/science/article/pii/S0166432803002924 (2025) (May 2004).

125. Cagniard, B., Balsam, P. D., Brunner, D. & Zhuang, X. Mice with Chronically Elevated Dopamine Exhibit Enhanced Motivation, but not Learning, for a Food Reward. en. *Neuropsychopharmacology* **31.** Publisher: Nature Publishing Group, 1362–1370. ISSN: 1740-634X. https://www.nature.com/articles/1300966 (2025) (July 2006).

126. Wang, W. *et al.* Motor Preparation Disrupts Proactive Control in the Stop Signal Task. English. *Frontiers in Human Neuroscience* **0.** Publisher: Frontiers. ISSN: 1662-5161. https://www.frontiersin.org/articles/10.3389/fnhum.2018.00151/full (2021) (2018).

127. Berke, J. D. What does dopamine mean? en. *Nature Neuroscience* **21.** Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Motivation;Operant learning;Psychology;Reward Subject_term_id: motivation;operant-learning;psychology;reward, 787–793. ISSN: 1546-1726. https://www.nature.com/articles/s41593-018-0152-y (2021) (June 2018).

128. Hamid, A. A. *et al.* Mesolimbic dopamine signals the value of work. en. *Nature Neuroscience* **19.** Publisher: Nature Publishing Group, 117–126. ISSN: 1546-1726. https://www.nature.com/articles/nn.4173 (2025) (Jan. 2016).

129. Barter, J. W. *et al.* Beyond reward prediction errors: the role of dopamine in movement kinematics. English. *Frontiers in Integrative Neuroscience* **9.** Publisher: Frontiers. ISSN: 1662-5145. https://www.frontiersin.org/journals/integrative-neuroscience/articles/10.3389/fnint.2015.00039/full (2025) (May 2015).

130. Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during locomotion and reward. en. *Nature* **535.** Publisher: Nature Publishing Group, 505–510. ISSN: 1476-4687. https://www.nature.com/articles/nature18942 (2025) (July 2016).

131. Da Silva, J. A., Tecuapetla, F., Paixão, V. & Costa, R. M. Dopamine neuron activity before action initiation gates and invigorates future movements. en. *Nature* **554.** Publisher: Nature Publishing Group, 244–248. ISSN: 1476-4687. https://www.nature.com/articles/nature25457 (2025) (Feb. 2018).

132. Azcorra, M. *et al.* Unique functional responses differentially map onto genetic subtypes of dopamine neurons. en. *Nature Neuroscience* **26.** Publisher: Nature Publishing Group, 1762–1774. ISSN: 1546-1726. https://www.nature.com/articles/s41593-023-01401-9 (2025) (Oct. 2023).

133. Fan, D., Rossi, M. A. & Yin, H. H. Mechanisms of Action Selection and Timing in Substantia Nigra Neurons. en. *Journal of Neuroscience* **32.** Publisher: Society for Neuroscience Section: Articles, 5534–5548. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/32/16/5534 (2025) (Apr. 2012).

134. Syed, E. C. J. *et al.* Action initiation shapes mesolimbic dopamine encoding of future rewards. en. *Nature Neuroscience* **19.** Publisher: Nature Publishing Group, 34–36. ISSN: 1546-1726. https://www.nature.com/articles/nn.4187 (2025) (Jan. 2016).

135. Hollon, N. G. *et al.* Nigrostriatal dopamine signals sequence-specific action-outcome prediction errors. English. *Current Biology* **31.** Publisher: Elsevier, 5350–5363.e5. ISSN: 0960-9822. https://www.cell.com/current-biology/abstract/S0960-9822(21)01280-X (2025) (Dec. 2021).

136. Coddington, L. T. & Dudman, J. T. Learning from Action: Reconsidering Movement Signaling in Midbrain Dopamine Neuron Activity. en. *Neuron* **104,** 63–77. ISSN: 0896-6273. https://www.sciencedirect.com/science/article/pii/S0896627319307421 (2022) (Oct. 2019).

137. Lex, B. & Hauber, W. The Role of Dopamine in the Prelimbic Cortex and the Dorsomedial Striatum in Instrumental Conditioning. *Cerebral Cortex* **20,** 873–883. ISSN: 1047-3211. https://doi.org/10.1093/cercor/bhp151 (2025) (Apr. 2010).

138. Moss, M. M., Zatka-Haas, P., Harris, K. D., Carandini, M. & Lak, A. Dopamine Axons in Dorsal Striatum Encode Contralateral Visual Stimuli and Choices. en. *Journal of Neuroscience* **41.** Publisher: Society for Neuroscience Section: Research Articles, 7197–7205. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/41/34/7197 (2025) (Aug. 2021).

139. Klanker, M., Fellinger, L., Feenstra, M., Willuhn, I. & Denys, D. Regionally distinct phasic dopamine release patterns in the striatum during reversal learning. *Neuroscience. Cognitive Flexibility: Development, Disease, and Treatment* **345,** 110–123. ISSN: 0306-4522. https://www.sciencedirect.com/science/article/pii/S0306452216301609 (2025) (Mar. 2017).

140. Wise, R. A. & Jordan, C. J. Dopamine, behavior, and addiction. en. *Journal of Biomedical Science* **28,** 83. ISSN: 1423-0127. https://doi.org/10.1186/s12929-021-00779-7 (2025) (Dec. 2021).

141. Belin, D., Mar, A. C., Dalley, J. W., Robbins, T. W. & Everitt, B. J. High Impulsivity Predicts the Switch to Compulsive Cocaine-Taking. *Science* **320.** Publisher: American Association for the Advancement of Science, 1352–1355. https://www.science.org/doi/full/10.1126/science.1158136 (2021) (June 2008).

142. Belin, D., Belin-Rauscent, A., Everitt, B. J. & Dalley, J. W. In search of predictive endophenotypes in addiction: insights from preclinical research. en. *Genes, Brain and Behavior* **15.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gbb.12265, 74–88. ISSN: 1601-183X. https://onlinelibrary.wiley.com/doi/abs/10.1111/gbb.12265 (2021) (2016).

143. Grall-Bronnec, M. *et al.* Dopamine Agonists and Impulse Control Disorders: A Complex Association. en. *Drug Safety* **41,** 19–75. ISSN: 1179-1942. https://doi.org/10.1007/s40264-017-0590-6 (2025) (Jan. 2018).

144. Nelson, A. & Killcross, S. Amphetamine Exposure Enhances Habit Formation. en. *Journal of Neuroscience* **26.** Publisher: Society for Neuroscience Section: Articles, 3805–3812. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/26/14/3805 (2025) (Apr. 2006).

145. Faure, A., Haberland, U., Condé, F. & Massioui, N. E. Lesion to the Nigrostriatal Dopamine System Disrupts Stimulus-Response Habit Formation. en. *Journal of Neuroscience* **25.** Publisher: Society for Neuroscience Section: Behavioral/Systems/Cognitive, 2771–2780. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/25/11/2771 (2025) (Mar. 2005).

146. Becchi, S., Buson, A. & Balleine, B. W. Inhibition of vascular adhesion protein 1 protects dopamine neurons from the effects of acute inflammation and restores habit learning in the striatum. en. *Journal of Neuroinflammation* **18,** 233. ISSN: 1742-2094. https://doi.org/10.1186/s12974-021-02288-8 (2025) (Oct. 2021).

147. Wang, L. P. *et al.* NMDA Receptors in Dopaminergic Neurons Are Crucial for Habit Learning. English. *Neuron* **72.** Publisher: Elsevier, 1055–1066. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(11)00931-7 (2025) (Dec. 2011).

148. Marr, D. A theory of cerebellar cortex. *The Journal of Physiology* **202,** 437–470.1. ISSN: 0022-3751. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1351491/ (2021) (June 1969).

149. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory* ISBN: 978-1-4106-1240-3 (Psychology Press, New York, Apr. 2005).

150. Cooper, S. J. Donald O. Hebb's synapse and learning rule: a history and commentary. *Neuroscience & Biobehavioral Reviews* **28,** 851–874. ISSN: 0149-7634. https://www.sciencedirect.com/science/article/pii/S0149763404000995 (2025) (Jan. 2005).

151. Bliss, T. V. & Lomo, T. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. eng. *The Journal of Physiology* **232,** 331–356. ISSN: 0022-3751 (July 1973).

152. Perrin, E. & Venance, L. Bridging the gap between striatal plasticity and learning. *Current Opinion in Neurobiology. Neurobiology of Learning and Plasticity* **54,** 104–112. ISSN: 0959-4388. https://www.sciencedirect.com/science/article/pii/S0959438818300497 (2025) (Feb. 2019).

153. Wilson, C. & Kawaguchi, Y. The origins of two-state spontaneous membrane potential fluctuations of neostriatal spiny neurons. *The Journal of Neuroscience* **16,** 2397–2410. ISSN: 0270-6474. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6578540/ (2025) (Apr. 1996).

154. Gerfen, C. R. & Surmeier, D. J. Modulation of Striatal Projection Systems by Dopamine. en. *Annual Review of Neuroscience* **34.** Publisher: Annual Reviews, 441–466. ISSN: 0147-006X, 1545-4126. https://www.annualreviews.org/content/journals/10.1146/annurev-neuro-061010-113641 (2025) (July 2011).

155. Yagishita, S. *et al.* A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345.** Publisher: American Association for the Advancement of Science, 1616–1620. https://www.science.org/doi/full/10.1126/science.1255514 (2025) (Sept. 2014).

156. Littman, M. L. A tutorial on partially observable Markov decision processes. *Journal of Mathematical Psychology. Special Issue: Dynamic Decision Making* **53,** 119–125. ISSN: 0022-2496. https://www.sciencedirect.com/science/article/pii/S0022249609000042 (2025) (June 2009).

157. Doll, B. B., Simon, D. A. & Daw, N. D. The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology. Decision making* **22,** 1075–1081. ISSN: 0959-4388. https://www.sciencedirect.com/science/article/pii/S0959438812001316 (2025) (Dec. 2012).

158. Jensen, K. T. An introduction to reinforcement learning for neuroscience. en. *Neurons, Behavior, Data analysis, and Theory.* Publisher: The neurons, behavior, data analysis and theory collective, 1–30. https://nbdt.scholasticahq.com/article/127771-an-introduction-to-reinforcement-learning-for-neuroscience (2025) (Dec. 2024).

159. Watkins, C. J. C. H. & Dayan, P. Q-learning. en. *Machine Learning* **8,** 279–292. ISSN: 1573-0565. https://doi.org/10.1007/BF00992698 (2022) (May 1992).

160. Haith, A. M. & Krakauer, J. W. *Model-Based and Model-Free Mechanisms of Human Motor Learning* en. in *Progress in Motor Control* (eds Richardson, M. J., Riley, M. A. & Shockley, K.) (Springer, New York, NY, 2013), 1–21. ISBN: 978-1-4614-5465-6.

161. Wunderlich, K., Smittenaar, P. & Dolan, R. J. Dopamine Enhances Model-Based over Model-Free Choice Behavior. English. *Neuron* **75.** Publisher: Elsevier, 418–424. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(12)00527-2 (2025) (Aug. 2012).

162. Botvinick, M. & Weinstein, A. Model-based hierarchical reinforcement learning and human action control. EN. *Philosophical Transactions of the Royal Society B: Biological Sciences*. Publisher: The Royal Society. https://royalsocietypublishing.org/doi/10.1098/rstb.2013.0480 (2025) (Nov. 2014).

163. Dayan, P. & Berridge, K. C. Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. en. *Cognitive, Affective, & Behavioral Neuroscience* **14,** 473–492. ISSN: 1531-135X. https://doi.org/10.3758/s13415-014-0277-8 (2025) (June 2014).

164. Lee, S. W., Shimojo, S. & O'Doherty, J. P. Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. English. *Neuron* **81.** Publisher: Elsevier, 687–699. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(13)01125-2 (2025) (Feb. 2014).

165. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. English. *Neuron* **69.** Publisher: Elsevier, 1204–1215. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(11)00125-5 (2025) (Mar. 2011).

166. Daw, N. D. Of goals and habits. *Proceedings of the National Academy of Sciences* **112.** Publisher: Proceedings of the National Academy of Sciences, 13749–13750. https://www.pnas.org/doi/abs/10.1073/pnas.1518488112 (2022) (Nov. 2015).

167. Soltani, A. & Koechlin, E. Computational models of adaptive behavior and prefrontal cortex. en. *Neuropsychopharmacology* **47.** Publisher: Nature Publishing Group, 58–71. ISSN: 1740-634X. https://www.nature.com/articles/s41386-021-01123-1 (2025) (Jan. 2022).

168. Schultz, W. Subjective neuronal coding of reward: temporal value discounting and risk. en. *European Journal of Neuroscience* **31.** _eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2010.07282.x, 2124–2135. ISSN: 1460-9568. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2010.07282.x (2025) (2010).

169. Critchfield, T. S. & Kollins, S. H. Temporal discounting: Basic research and the analysis of socially important behavior. *Journal of Applied Behavior Analysis* **34.** Place: US Publisher: Journal of Applied Behavior Analysis, 101–122. ISSN: 1938-3703 (2001).

170. Nguyen, T. N., McDonald, C. & Gonzalez, C. Credit Assignment: Challenges and Opportunities in Developing Human-like Learning Agents. en. *Proceedings of the AAAI Symposium Series* **3.** Number: 1, 54–57. ISSN: 2994-4317. https://ojs.aaai.org/index.php/AAAI-SS/article/view/31180 (2025) (May 2024).

171. Zhang, Z., Costa, K. M., Langdon, A. J. & Schoenbaum, G. The devilish details affecting TDRL models in dopamine research. English. *Trends in Cognitive Sciences* **0.** Publisher: Elsevier. ISSN: 1364-6613, 1879-307X. https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(25)00033-6 (2025) (Feb. 2025).

172. SUTTON, R. S. *Temporal Credit Assignment in Reinforcement Learning* English. ISBN: 9798204592971. Ph.D. (University of Massachusetts Amherst, United States – Massachusetts, 1984). https://www.proquest.com/docview/303321395/abstract/E4D428D9CC9A49ABPQ/1 (2025).

173. Cichosz, P. Truncating Temporal Differences: On the Efficient Implementation of TD(lambda) for Reinforcement Learning. en. *Journal of Artificial Intelligence Research* **2,** 287–318. ISSN: 1076-9757. https://www.jair.org/index.php/jair/article/view/10128 (2025) (1994).

174. Sutton, R. S. & Barto, A. G. in *Learning and computational neuroscience: Foundations of adaptive networks* 497–537 (The MIT Press, Cambridge, MA, US, 1990). ISBN: 978-0-262-07102-4.

175. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Evaluating the TD model of classical conditioning. en. *Learning & Behavior* **40,** 305–319. ISSN: 1543-4508. https://doi.org/10.3758/s13420-012-0082-6 (2022) (Sept. 2012).

176. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation* **20.** Place: US Publisher: MIT Press, 3034–3054. ISSN: 1530-888X (2008).

177. Roesch, M. R., Esber, G. R., Li, J., Daw, N. D. & Schoenbaum, G. Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. en. *European Journal of Neuroscience* **35.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2011.07986.x, 1190–1200. ISSN: 1460-9568. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2011.07986.x (2025) (2012).

178. Luksys, G. & Sandi, C. Neural mechanisms and computations underlying stress effects on learning and memory. *Current Opinion in Neurobiology. Behavioural and cognitive neuroscience* **21,** 502–508. ISSN: 0959-4388. https://www.sciencedirect.com/science/article/pii/S0959438811000432 (2025) (June 2011).

179. Van den Ende, M. W. J. *et al.* A review of mathematical modeling of addiction regarding both (neuro-) psychological processes and the social contagion perspectives. *Addictive Behaviors* **127,** 107201. ISSN: 0306-4603. https://www.sciencedirect.com/science/article/pii/S0306460321003865 (2025) (Apr. 2022).

180. Chen, Z. S. & Wang, J. Pain, from perception to action: A computational perspective. English. *iScience* **26.** Publisher: Elsevier. ISSN: 2589-0042. https://www.cell.com/iscience/abstract/S2589-0042(22)01980-0 (2025) (Jan. 2023).

181. Barto, A. G., Sutton, R. S. & Anderson, C. W. Looking Back on the Actor–Critic Architecture. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **51.** Conference Name: IEEE Transactions on Systems, Man, and Cybernetics: Systems, 40–50. ISSN: 2168-2232. https://ieeexplore.ieee.org/abstract/document/9306925 (2025) (Jan. 2021).

182. Barto, A. G. in *Models of information processing in the basal ganglia* 215–232 (The MIT Press, Cambridge, MA, US, 1995). ISBN: 978-0-262-08234-1.

183. Houk, J. C., Buckingham, J. T. & Barto, A. G. *Models of the cerebellum and motor learning* en. 1996. https://philpapers.org/rec/houmot-2 (2025).

184. Joel, D., Niv, Y. & Ruppin, E. Actor–critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks* **15,** 535–547. ISSN: 0893-6080. https://www.sciencedirect.com/science/article/pii/S0893608002000473 (2025) (June 2002).

185. Peters, J., Vijayakumar, S. & Schaal, S. *Natural Actor-Critic* en. in *Machine Learning: ECML 2005* (eds Gama, J., Camacho, R., Brazdil, P. B., Jorge, A. M. & Torgo, L.) (Springer, Berlin, Heidelberg, 2005), 280–291. ISBN: 978-3-540-31692-3.

186. Bornstein, A. M. & Daw, N. D. Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology. Behavioural and cognitive neuroscience* **21,** 374–380. ISSN: 0959-4388. https://www.sciencedirect.com/science/article/pii/S0959438811000365 (2025) (June 2011).

187. Haarnoja, T. *et al. Soft Actor-Critic Algorithms and Applications* arXiv:1812.05905 [cs]. Jan. 2019. http://arxiv.org/abs/1812.05905 (2025).

188. Sewak, M. en. in *Deep Reinforcement Learning: Frontiers of Artificial Intelligence* (ed Sewak, M.) 141–152 (Springer, Singapore, 2019). ISBN: 9789811382857. https://doi.org/10.1007/978-981-13-8285-7_11 (2025).

189. Takahashi, Y., Schoenbaum, G. & Niv, Y. Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an Actor/Critic model. English. *Frontiers in Neuroscience* **2.** Publisher: Frontiers. ISSN: 1662-453X. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/neuro.01.014.2008/full (2025) (July 2008).

190. Rummery, G. A. & Niranjan, M. *On-line Q-learning using connectionist systems* (University of Cambridge, Department of Engineering Cambridge, UK, 1994).

191. Daw, N. Trial-by-trial data analysis using computational models. *Affect, Learning and Decision Making, Attention and Performance XXIII* **23.** ISSN: 9780199600434 (Mar. 2011).

192. Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L. & Platt, M. L. A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research. en. *Neuropsychopharmacology* **42.** Publisher: Nature Publishing Group, 1931–1939.

ISSN: 1740-634X. https://www.nature.com/articles/npp2017108 (2025) (Sept. 2017).

193. Cohen, J. D., McClure, S. M. & Yu, A. J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. EN. *Philosophical Transactions of the Royal Society B: Biological Sciences*. Publisher: The Royal SocietyLondon. https://royalsocietypublishing.org/doi/10.1098/rstb.2007.2098 (2025) (Mar. 2007).

194. Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *eLife* **8** (ed Behrens, T. E.) Publisher: eLife Sciences Publications, Ltd, e49547. ISSN: 2050-084X. https://doi.org/10.7554/eLife.49547 (2025) (Nov. 2019).

195. Cohen, M. X. & Frank, M. J. Neurocomputational models of basal ganglia function in learning, memory and choice. *Behavioural Brain Research. Special issue on the role of the basal ganglia in learning and memory* **199,** 141–156. ISSN: 0166-4328. https://www.sciencedirect.com/science/article/pii/S0166432808005421 (2025) (Apr. 2009).

196. Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T. & Hutchison, K. E. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences* **104.** Publisher: Proceedings of the National Academy of Sciences, 16311–16316. https://www.pnas.org/doi/full/10.1073/pnas.0706111104 (2025) (Oct. 2007).

197. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. en. *Nature* **441.** Publisher: Nature Publishing Group, 876–879. ISSN: 1476-4687. https://www.nature.com/articles/nature04766 (2025) (June 2006).

198. Bossaerts, P. & Murawski, C. From behavioural economics to neuroeconomics to decision neuroscience: the ascent of biology in research on human decision making. *Current Opinion in Behavioral Sciences. Neuroeconomics* **5,** 37–42. ISSN: 2352-1546. https://www.sciencedirect.com/science/article/pii/S2352154615000881 (2025) (Oct. 2015).

199. Rutledge, R. B., Skandali, N., Dayan, P. & Dolan, R. J. Dopaminergic Modulation of Decision Making and Subjective Well-Being. en. *Journal of Neuroscience* **35.** Publisher: Society for Neuroscience Section: Articles, 9811–9822. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/35/27/9811 (2025) (July 2015).

200. Antelo, C., Martinho, D. & Marreiros, G. *A Review on Supervised Learning Methodologies for Detecting Eating Habits of Diabetic Patients* en. in *Progress in Artificial Intelligence* (eds Marreiros, G., Martins, B., Paiva, A., Ribeiro, B. & Sardinha, A.) (Springer International Publishing, Cham, 2022), 374–386. ISBN: 978-3-031-16474-3.

201. Reverdy, P. & Leonard, N. E. Parameter Estimation in Softmax Decision-Making Models With Linear Objective Functions. *IEEE Transactions on Automation Science and Engineering* **13,** 54–67. ISSN: 1558-3783. https://ieeexplore.ieee.org/document/7336571 (2025) (Jan. 2016).

202. Beilock, S. L., Bertenthal, B. I., Hoerger, M. & Carr, T. H. When does haste make waste? Speed-accuracy tradeoff, skill level, and the tools of the trade. *Journal of Experimental Psychology: Applied* **14.** Place: US Publisher: American Psychological Association, 340–352. ISSN: 1939-2192 (2008).

203. Mickevičienė, D., Motiejūnaitė, K., Skurvydas, A., Darbutas, T. & Karanauskienė, D. How do Reaction Time and Movement Speed Depend on the Complexity of the Task? en. *Baltic Journal of Sport and Health Sciences* **2.** Number: 69. ISSN: 2538-8347. https://journals.lsu.lt/baltic-journal-of-sport-health/article/view/504 (2025) (2008).

204. Ma, H.-i. & Trombly, C. A. Effects of Task Complexity on Reaction Time and Movement Kinematics in Elderly People. en. *The American Journal of Occupational Therapy* **58.** Publisher: American Occupational Therapy Association, 150–158. ISSN: 0272-9490. https://research.aota.org/ajot/article/58/2/150/4797/Effects-of-Task-Complexity-on-Reaction-Time-and (2025) (Mar. 2004).

205. Laszlo, J. I. & Livesey, J. P. Task Complexity, Accuracy, and Reaction Time. *Journal of Motor Behavior* **9.** Publisher: Routledge _eprint:

https://doi.org/10.1080/00222895.1977.10735107, 171–177. ISSN: 0022-2895. https://doi.org/10.1080/00222895.1977.10735107 (2025) (June 1977).

206. Mills, C. B. Effects of context on reaction time to phonemes. *Journal of Verbal Learning and Verbal Behavior* **19,** 75–83. ISSN: 0022-5371. https://www.sciencedirect.com/science/article/pii/S0022537180905368 (2025) (Feb. 1980).

207. Martin, K. *et al.* The Impact of Environmental Stress on Cognitive Performance: A Systematic Review. EN. *Human Factors* **61.** Publisher: SAGE Publications Inc, 1205–1246. ISSN: 0018-7208. https://doi.org/10.1177/0018720819839817 (2025) (Dec. 2019).

208. Molloy, M. F., Galdo, M., Bahg, G., Liu, Q. & Turner, B. M. What's in a response time?: On the importance of response time measures in constraining models of context effects. *Decision* **6.** Place: US Publisher: Educational Publishing Foundation, 171–200. ISSN: 2325-9973 (2019).

209. Zhang, B. *et al.* Reaction time and physiological signals for stress recognition. *Biomedical Signal Processing and Control* **38,** 100–107. ISSN: 1746-8094. https://www.sciencedirect.com/science/article/pii/S1746809417300885 (2025) (Sept. 2017).

210. Pawar, N. M. & Velaga, N. R. Modelling the influence of time pressure on reaction time of drivers. *Transportation Research Part F: Traffic Psychology and Behaviour* **72,** 1–22. ISSN: 1369-8478. https://www.sciencedirect.com/science/article/pii/S1369847820304149 (2025) (July 2020).

211. Chittka, L., Skorupski, P. & Raine, N. E. Speed–accuracy tradeoffs in animal decision making. English. *Trends in Ecology & Evolution* **24.** Publisher: Elsevier, 400–407. ISSN: 0169-5347. https://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(09)00122-0 (2025) (July 2009).

212. Heitz, R. P. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. English. *Frontiers in Neuroscience* **8.** Publisher: Frontiers. ISSN: 1662-453X. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2014.00150/full (2025) (June 2014).

213. Du, Y., Krakauer, J. W. & Haith, A. M. The relationship between habits and motor skills in humans. English. *Trends in Cognitive Sciences* **26.** Publisher: Elsevier, 371–387. ISSN: 1364-6613, 1879-307X. https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(22)00038-9 (2025) (May 2022).

214. Keramati, M., Dezfouli, A. & Piray, P. Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. en. *PLOS Computational Biology* **7.** Publisher: Public Library of Science, e1002055. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002055 (2025) (May 2011).

215. Ratcliff, R. & Rouder, J. N. Modeling Response Times for Two-Choice Decisions. EN. *Psychological Science* **9.** Publisher: SAGE Publications Inc, 347–356. ISSN: 0956-7976. https://doi.org/10.1111/1467-9280.00067 (2025) (Sept. 1998).

216. Damaso, K., Williams, P. & Heathcote, A. Evidence for different types of errors being associated with different types of post-error changes. en. *Psychonomic Bulletin & Review* **27,** 435–440. ISSN: 1531-5320. https://doi.org/10.3758/s13423-019-01675-w (2025) (June 2020).

217. Ratcliff, R. A theory of memory retrieval. *Psychological Review* **85.** Place: US Publisher: American Psychological Association, 59–108. ISSN: 1939-1471 (1978).

218. Ratcliff, R., Smith, P. L., Brown, S. D. & McKoon, G. Diffusion Decision Model: Current Issues and History. English. *Trends in Cognitive Sciences* **20.** Publisher: Elsevier, 260–281. ISSN: 1364-6613, 1879-307X. https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(16)00025-5 (2025) (Apr. 2016).

219. Voskuilen, C., Ratcliff, R. & Smith, P. L. Comparing fixed and collapsing boundary versions of the diffusion model. *Journal of mathematical psychology* **73,** 59–79. ISSN: 0022-2496. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5450920/ (2024) (Aug. 2016).

220. Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R. & Brown, S. D. Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. eng. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **35,** 2476–2484. ISSN: 1529-2401 (Feb. 2015).

221. Diederich, A. A Dynamic Dual Process Model for Binary Choices: Serial Versus Parallel Architecture. en. *Computational Brain & Behavior* **7,** 37–64. ISSN: 2522-087X. https://doi.org/10.1007/s42113-023-00186-1 (2024) (Mar. 2024).

222. Alós-Ferrer, C. A Dual-Process Diffusion Model. en. *Journal of Behavioral Decision Making* **31.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.1960, 203–218. ISSN: 1099-0771. https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.1960 (2025) (2018).

223. Hübner, R., Steinhauser, M. & Lehle, C. A dual-stage two-phase model of selective attention. *Psychological Review* **117.** Place: US Publisher: American Psychological Association, 759–784. ISSN: 1939-1471 (2010).

224. Usher, M. & McClelland, J. L. The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review* **108.** Place: US Publisher: American Psychological Association, 550–592. ISSN: 1939-1471(Electronic),0033-295X(Print) (2001).

225. Boucher, L., Palmeri, T. J., Logan, G. D. & Schall, J. D. Inhibitory control in mind and brain: An interactive race model of countermanding saccades. *Psychological Review* **114.** Place: US Publisher: American Psychological Association, 376–397. ISSN: 1939-1471 (2007).

226. Miletić, S. *et al.* A new model of decision processing in instrumental learning tasks. *eLife* **10** (eds Wyart, V., Gold, J. I. & de Gee, J. W.) Publisher: eLife Sciences Publications, Ltd, e63055. ISSN: 2050-084X. https://doi.org/10.7554/eLife.63055 (2024) (Jan. 2021).

227. Tillman, G., Van Zandt, T. & Logan, G. D. Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. en. *Psychonomic Bulletin & Review* **27,** 911–936. ISSN: 1531-5320. https://doi.org/10.3758/s13423-020-01719-6 (2024) (Oct. 2020).

228. Miletić, S., Boag, R. J. & Forstmann, B. U. Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia* **136,** 107261. ISSN: 0028-3932. https://www.sciencedirect.com/science/article/pii/S0028393219303033 (2024) (Jan. 2020).

229. Tuerlinckx, F. & Boeck, P. D. Two Interpretations of the Discrimination Parameter. en. *Psychometrika* **70,** 629–650. ISSN: 0033-3123, 1860-0980. https://www.cambridge.org/core/journals/psychometrika/article/abs/two-interpretations-of-the-discrimination-parameter/44500E3DE45642A2C46C657B6A57BBE4 (2025) (Dec. 2005).

230. Doody, M., Van Swieten, M. M. H. & Manohar, S. G. Model-based learning retrospectively updates model-free values. en. *Scientific Reports* **12.** Publisher: Nature Publishing Group, 2358. ISSN: 2045-2322. https://www.nature.com/articles/s41598-022-05567-3 (2025) (Feb. 2022).

231. Feher da Silva, C., Lombardi, G., Edelson, M. & Hare, T. A. Rethinking model-based and model-free influences on mental effort and striatal prediction errors. en. *Nature Human Behaviour* **7.** Publisher: Nature Publishing Group, 956–969. ISSN: 2397-3374. https://www.nature.com/articles/s41562-023-01573-1 (2025) (June 2023).

232. Gillan, C. M., Otto, A. R., Phelps, E. A. & Daw, N. D. Model-based learning protects against forming habits. en. *Cognitive, Affective, & Behavioral Neuroscience* **15,** 523–536. ISSN: 1531-135X. https://doi.org/10.3758/s13415-015-0347-6 (2025) (Sept. 2015).

233. Dezfouli, A. & Balleine, B. W. Habits, action sequences and reinforcement learning. *European Journal of Neuroscience* **35.** Publisher: John Wiley & Sons, Ltd, 1036–1051. ISSN: 0953-816X. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1460-9568.2012.08050.x (2025) (Apr. 2012).

234. Friston, K. *et al.* Active inference and learning. *Neuroscience & Biobehavioral Reviews* **68,** 862–879. ISSN: 0149-7634. https://www.sciencedirect.com/science/article/pii/S0149763416301336 (2025) (Sept. 2016).

235. Bootsma, R. J., Marteniuk, R. G., MacKenzie, C. L. & Zaal, F. T. J. M. The speed-accuracy trade-off in manual prehension: effects of movement amplitude, object size and object width on kinematic characteristics. en. *Experimental Brain Research* **98,** 535–541. ISSN: 1432-1106. https://doi.org/10.1007/BF00233990 (2025) (Apr. 1994).

236. Sutton, R. S. & Barto, A. G. *Reinforcement learning: an introduction* Second edition. en. ISBN: 978-0-262-03924-6 (The MIT Press, Cambridge, Massachusetts, 2018).

237. Hollerman, J. R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. en. *Nature Neuroscience* **1.** Publisher: Nature Publishing Group, 304–309. ISSN: 1546-1726. https://www.nature.com/articles/nn0898_304 (2025) (Aug. 1998).

238. Adler, A. *et al.* Temporal Convergence of Dynamic Cell Assemblies in the Striato-Pallidal Network. en. *Journal of Neuroscience* **32.** Publisher: Society for Neuroscience Section: Articles, 2473–2484. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/32/7/2473 (2022) (Feb. 2012).

239. Matell, M. S., Meck, W. H. & Nicolelis, M. A. L. Interval timing and the encoding of signal duration by ensembles of cortical and striatal neurons. *Behavioral Neuroscience* **117.** Place: US Publisher: American Psychological Association, 760–773. ISSN: 1939-0084 (2003).

240. Matell, M. S. & Meck, W. H. Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Cognitive Brain Research. Neuroimaging of Interval Timing* **21,** 139–170. ISSN: 0926-6410. https://www.sciencedirect.com/science/article/pii/S0926641004001697 (2025) (Oct. 2004).

241. Lau, B., Monteiro, T. & Paton, J. J. The many worlds hypothesis of dopamine prediction error: implications of a parallel circuit architecture in the basal ganglia. *Current Opinion in Neurobiology. Computational Neuroscience* **46,** 241–247. ISSN: 0959-4388. https://www.sciencedirect.com/science/article/pii/S0959438817301587 (2025) (Oct. 2017).

242. Schroll, H., Vitay, J. & Hamker, F. H. Working memory and response selection: A computational account of interactions among cortico-basalganglio-thalamic loops. *Neural Networks* **26,** 59–74. ISSN: 0893-6080. https://www.sciencedirect.com/science/article/pii/S089360801100270X (2025) (Feb. 2012).

243. Berns, G. S. & Sejnowski, T. J. en. in *Neurobiology of Decision-Making* (eds Damasio, A. R., Damasio, H. & Christen, Y.) 101–113 (Springer, Berlin, Heidelberg, 1996). ISBN: 978-3-642-79928-0. https://doi.org/10.1007/978-3-642-79928-0_6 (2025).

244. Klaus, A., Silva, J. A. d. & Costa, R. M. What, If, and When to Move: Basal Ganglia Circuits and Self-Paced Action Initiation. en. *Annual Review of Neuroscience* **42.** Publisher: Annual Reviews, 459–483. ISSN: 0147-006X, 1545-4126. https://www.annualreviews.org/content/journals/10.1146/annurev-neuro-072116-031033 (2025) (July 2019).

245. Albus, J. S. A theory of cerebellar function. en. *Mathematical Biosciences* **10,** 25–61. ISSN: 0025-5564. https://www.sciencedirect.com/science/article/pii/0025556471900514 (2021) (Feb. 1971).

246. Swain, R. A., Kerr, A. L. & Thompson, R. F. The Cerebellum: A Neural System for the Study of Reinforcement Learning. English. *Frontiers in Behavioral Neuroscience* **5.** Publisher: Frontiers. ISSN: 1662-5153. https://www.frontiersin.org/journals/behavioral-neuroscience/articles/10.3389/fnbeh.2011.00008/full (2025) (Mar. 2011).

247. Sendhilnathan, N., Semework, M., Goldberg, M. E. & Ipata, A. E. Neural Correlates of Reinforcement Learning in Mid-lateral Cerebellum. English. *Neuron* **106.** Publisher: Elsevier, 188–198.e5. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(19)31098-0 (2025) (Apr. 2020).

248. Thompson, R. F., Thompson, J. K., Kim, J. J., Krupa, D. J. & Shinkman, P. G. The Nature of Reinforcement in Cerebellar Learning. *Neurobiology of Learning and Memory* **70,** 150–176. ISSN: 1074-7427. https://www.sciencedirect.com/science/article/pii/S107474279893845X (2025) (July 1998).

249. Hart, G., Burton, T. J. & Balleine, B. W. What Role Does Striatal Dopamine Play in Goal-directed Action? *Neuroscience* **546,** 20–32. ISSN: 0306-4522. https://www.sciencedirect.com/science/article/pii/S0306452224001337 (2025) (May 2024).

250. Miller, S., Konorski, J. & Skinner, B. F. ON A PARTICULAR FORM OF CONDITIONED REFLEX. en. *Journal of the Experimental Analysis of Behavior* **12.** Publisher: John Wiley & Sons, Ltd, 187–189. ISSN: 1938-3711. https://onlinelibrary.wiley.com/doi/10.1901/jeab.1969.12-187 (2025) (Jan. 1969).

251. Balleine, B. W. Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology & Behavior. Purdue University Ingestive*

Behavior Research Center Symposium. Dietary Influences on Obesity: Environment, Behavior and Biology **86,** 717–730. ISSN: 0031-9384. https://www.sciencedirect.com/science/article/pii/S0031938405004002 (2025) (Dec. 2005).

252. Dickinson, A. Omission Learning after Instrumental Pretraining. *The Quarterly Journal of Experimental Psychology Section B* **51.** Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1080/713932679, 271–286. ISSN: 0272-4995. https://www.tandfonline.com/doi/abs/10.1080/713932679 (2022) (Aug. 1998).

253. Collingwood, C., Stephenson-Jones, M. & Bogacz, R. *Habits Through Temporal-Difference Action Learning* English. in *Proceedings of the 2023 Conference on Cognitive Computational Neuroscience* (Cognitive Computational Neuroscience, Oxford, England, Aug. 2023), P–3A.53. https://2023.ccneuro.org/view_paper6b76.html?PaperNum=1156.

254. Collingwood, C. & Stephenson-Jones, M. *Habits Through Temporal-Difference Action Learning* Highlighted Poster Presenation. Stockholm, June 2023. https://2023.ibags.global/final-programme/.

255. Lee, R. S., Sagiv, Y., Engelhard, B., Witten, I. B. & Daw, N. D. A feature-specific prediction error model explains dopaminergic heterogeneity. en. *Nature Neuroscience*. Publisher: Nature Publishing Group, 1–13. ISSN: 1546-1726. https://www.nature.com/articles/s41593-024-01689-1 (2024) (July 2024).

256. Guttman, N. Operant conditioning, extinction, and periodic reinforcement in relation to concentration of sucrose used as reinforcing agent. *Journal of Experimental Psychology* **46.** Place: US Publisher: American Psychological Association, 213–224. ISSN: 0022-1015 (1953).

257. Schaeffer, R. W. & Hanna, B. Effects of Quality and Quantity of Reinforcement Upon Response Rate in Acquisition and Extinction. en. *Psychological Reports.* Publisher: SAGE PublicationsSage CA: Los Angeles, CA. https://journals.sagepub.com/doi/abs/10.2466/pr0.1966.18.3.819?casa_token=ZDfWC6qF9_QAAAAA%3Ag0x63szkuLT1YXYQDiK8qnPqGtYQGZCqNZ8QELwdl0EOO_uKB1DjAQIRFztbe7ldV6BTMS1uyaY (2025) (June 1966).

258. Summerside, E. M., Shadmehr, R. & Ahmed, A. A. Vigor of reaching movements: reward discounts the cost of effort. en. *Journal of Neurophysiology*. Publisher: American Physiological Society Bethesda, MD. https://journals.physiology.org/doi/10.1152/jn.00872.2017 (2025) (June 2018).

259. Jackson, R. E., Waran, N. K. & Cockram, M. S. Methods for Measuring Feeding Motivation in Sheep. en. *Animal Welfare* **8,** 53–63. ISSN: 0962-7286, 2054-1538. https://www.cambridge.org/core/journals/animal-welfare/article/abs/methods-for-measuring-feeding-motivation-in-sheep/0DF74EBCEEDC061A67E8632ECC7DBCD2 (2025) (Feb. 1999).

260. Shen, Y. J. & Chun, M. M. Increases in rewards promote flexible behavior. en. *Attention, Perception, & Psychophysics* **73,** 938–952. ISSN: 1943-393X. https://doi.org/10.3758/s13414-010-0065-7 (2025) (Apr. 2011).

261. Zoratto, F., Laviola, G. & Adriani, W. The subjective value of probabilistic outcomes: Impact of reward magnitude on choice with uncertain rewards in rats. *Neuroscience Letters* **617,** 225–231. ISSN: 0304-3940. https://www.sciencedirect.com/science/article/pii/S0304394016300921 (2025) (Mar. 2016).

262. Stauffer, W. R., Lak, A. & Schultz, W. Dopamine Reward Prediction Error Responses Reflect Marginal Utility. English. *Current Biology* **24.** Publisher: Elsevier, 2491–2500. ISSN: 0960-9822. https://www.cell.com/current-biology/abstract/S0960-9822(14)01128-2 (2025) (Nov. 2014).

263. Tobler, P. N., Fiorillo, C. D. & Schultz, W. Adaptive Coding of Reward Value by Dopamine Neurons. *Science* **307.** Publisher: American Association for the Advancement of Science, 1642–1645. https://www.science.org/doi/full/10.1126/science.1105370 (2025) (Mar. 2005).

264. Heath, C. J., Phillips, B. U., Bussey, T. J. & Saksida, L. M. Measuring Motivation and Reward-Related Decision Making in the Rodent Operant Touchscreen System. en. *Current Protocols in Neuroscience* **74.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471142301.ns0834s74, 8.34.1–8.34.20. ISSN: 1934-8576. https://onlinelibrary.wiley.com/doi/abs/10.1002/0471142301.ns0834s74 (2025) (2016).

265. Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. E. M. & Rushworth, M. F. S. Weighing up the benefits of work: Behavioral and neural analyses of effort-related decision making. *Neural Networks. Neurobiology of Decision Making* **19,** 1302–1314. ISSN: 0893-6080. https://www.sciencedirect.com/science/article/pii/S0893608006001602 (2025) (Oct. 2006).

266. Treadway, M. T., Buckholtz, J. W., Schwartzman, A. N., Lambert, W. E. & Zald, D. H. Worth the 'EEfRT'? The Effort Expenditure for Rewards Task as an Objective Measure of Motivation and Anhedonia. en. *PLOS ONE* **4.** Publisher: Public Library of Science, e6598. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006598 (2025) (Aug. 2009).

267. Wittmann, B. C., Daw, N. D., Seymour, B. & Dolan, R. J. Striatal Activity Underlies Novelty-Based Choice in Humans. English. *Neuron* **58.** Publisher: Elsevier, 967–973. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(08)00381-4 (2025) (June 2008).

268. Nussenbaum, K. *et al.* Novelty and uncertainty differentially drive exploration across development. *eLife* **12** (eds Badre, D., Behrens, T. E. & Wu, C. M.) Publisher: eLife Sciences Publications, Ltd, e84260. ISSN: 2050-084X. https://doi.org/10.7554/eLife.84260 (2025) (Aug. 2023).

269. Wang, Y., Lak, A., Manohar, S. G. & Bogacz, R. Dopamine encoding of novelty facilitates efficient uncertainty-driven exploration. en. *PLOS Computational Biology* **20.** Publisher: Public Library of Science, e1011516. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011516 (2025) (Apr. 2024).

270. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies — Revisited. *NeuroImage* **84,** 971–985. ISSN: 1053-8119. https://www.sciencedirect.com/science/article/pii/S1053811913009300 (2025) (Jan. 2014).

271. Znamenskiy, P. & Zador, A. M. Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. en. *Nature* **497.** Publisher: Nature Publishing Group, 482–485. ISSN: 1476-4687. https://www.nature.com/articles/nature12077 (2025) (May 2013).

272. Patriarchi, T. *et al.* Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science (New York, N.Y.)* **360,** eaat4422. ISSN: 0036-8075. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6287765/ (2025) (June 2018).

273. Bernklau, T. W., Righetti, B., Mehrke, L. S. & Jacob, S. N. Striatal dopamine signals reflect perceived cue–action–outcome associations in mice. en. *Nature Neuroscience* **27.** Publisher: Nature Publishing Group, 747–757. ISSN: 1546-1726. https://www.nature.com/articles/s41593-023-01567-2 (2025) (Apr. 2024).

274. Suri, R. E. TD models of reward predictive responses in dopamine neurons. *Neural Networks* **15,** 523–533. ISSN: 0893-6080. https://www.sciencedirect.com/science/article/pii/S0893608002000461 (2025) (June 2002).

275. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. en. *WIREs Computational Statistics* **4.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.199, 199–203. ISSN: 1939-0068. https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.199 (2025) (2012).

276. Cavanaugh, J. E. & Neath, A. A. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. en. *WIREs Computational Statistics* **11.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1460, e1460. ISSN: 1939-0068. https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1460 (2025) (2019).

277. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* en. ISBN: 978-0-387-95364-9 (Springer Science & Business Media, Dec. 2003).

278. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *NeuroImage* **46,** 1004–1017. ISSN: 1053-8119. https://www.sciencedirect.com/science/article/pii/S1053811909002638 (2025) (July 2009).

279. Gershman, S. *MFIT: simple model-fitting tools* original-date: 2015-06-29T01:13:26Z. Apr. 2025. https://github.com/sjgershm/mfit (2025).

280. Lau, B. & Glimcher, P. W. Value Representations in the Primate Striatum during Matching Behavior. English. *Neuron* **58.** Publisher: Elsevier, 451–463. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(08)00175-X (2025) (May 2008).

281. Green, L. & Myerson, J. A Discounting Framework for Choice With Delayed and Probabilistic Rewards. *Psychological Bulletin* **130.** Place: US Publisher: American Psychological Association, 769–792. ISSN: 1939-1455 (2004).

282. Diederich, A. & Zhao, W. J. A Dynamic Dual Process Model of Intertemporal Choice. en. *The Spanish Journal of Psychology* **22,** E54. ISSN: 1138-7416, 1988-2904. https://www.cambridge.org/core/journals/spanish-journal-of-psychology/article/dynamic-dual-process-model-of-intertemporal-choice/461344A77C72DB280F6BD624D0BE244E (2025) (Jan. 2019).

283. Anders, R., Alario, F.-X. & Van Maanen, L. The shifted Wald distribution for response time data analysis. *Psychological Methods* **21.** Place: US Publisher: American Psychological Association, 309–327. ISSN: 1939-1463 (2016).

284. Gardiner, C. *Stochastic Methods: A Handbook for the Natural and Social Sciences* 4th ed. en. ISBN: 978-3-540-70712-7. https://link.springer.com/book/9783540707127 (2025) (Springer Berlin, Heidelberg, Jan. 2009).

285. Schuss, Z. *Brownian Dynamics at Boundaries and Interfaces: In Physics, Chemistry, and Biology* en. ISSN: 0066-5452, 2196-968X. ISBN: 978-1-4614-7686-3 978-1-4614-7687-0. https://link.springer.com/10.1007/978-1-4614-7687-0 (2025) (Springer, New York, NY, 2013).

286. Seshadri, V. & Seshadri, V. *The Inverse Gaussian Distribution: A Case Study in Exponential Families* ISBN: 978-0-19-852243-0 (Oxford University Press, Oxford, New York, Jan. 1994).

287. Thura, D., Beauregard-Racine, J., Fradet, C.-W. & Cisek, P. Decision making by urgency gating: theory and experimental support. *Journal of Neurophysiology* **108.** Publisher: American Physiological Society, 2912–2930. ISSN: 0022-3077. https://journals.physiology.org/doi/full/10.1152/jn.01071.2011 (2025) (Dec. 2012).

288. Van Ravenzwaaij, D., Brown, S. D., Marley, A. A. J. & Heathcote, A. Accumulating advantages: A new conceptualization of rapid multiple choice. *Psychological Review* **127.** Place: US Publisher: American Psychological Association, 186–215. ISSN: 1939-1471 (2020).

289. Whelan, R. Effective Analysis of Reaction Time Data. en. *The Psychological Record* **58,** 475–482. ISSN: 2163-3452. https://doi.org/10.1007/BF03395630 (2025) (July 2008).

290. Luce, R. D. *Response Times: Their Role in Inferring Elementary Mental Organization* en. Google-Books-ID: WSmpNN5WCw0C. ISBN: 978-0-19-536146-9 (Oxford University Press, May 1991).

291. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* en. Google-Books-ID: VWq5GG6ycxMC. ISBN: 978-1-118-58577-1 (John Wiley & Sons, Nov. 2012).

292. Cragg, S. J. Meaningful silences: how dopamine listens to the ACh pause. en. *Trends in Neurosciences* **29,** 125–131. ISSN: 0166-2236. https://www.sciencedirect.com/science/article/pii/S016622360600004X (2022) (Mar. 2006).

293. Elber-Dorozko, L. & Loewenstein, Y. Striatal action-value neurons reconsidered. *eLife* **7** (ed Behrens, T. E.) Publisher: eLife Sciences Publications, Ltd, e34248. ISSN: 2050-084X. https://doi.org/10.7554/eLife.34248 (2025) (May 2018).

294. Kim, H., Sul, J. H., Huh, N., Lee, D. & Jung, M. W. Role of Striatum in Updating Values of Chosen Actions. en. *Journal of Neuroscience* **29.** Publisher: Society for Neuroscience Section: Articles, 14701–14712. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/29/47/14701 (2025) (Nov. 2009).

295. Verbruggen, F., Schneider, D. W. & Logan, G. D. How to stop and change a response: The role of goal activation in multitasking. *Journal of Experimental Psychology: Human Perception and Performance* **34.** Place: US Publisher: American Psychological Association, 1212–1228. ISSN: 1939-1277 (2008).

296. Goedhoop, J. N. *et al.* Nucleus accumbens dopamine tracks aversive stimulus duration and prediction but not value or prediction error. *eLife* **11** (eds Iordanova, M. D., Wassum, K. M. & Oleson, E.) Publisher: eLife Sciences Publications, Ltd,

e82711. ISSN: 2050-084X. https://doi.org/10.7554/eLife.82711 (2025) (Nov. 2022).

Figures produced using BioRender.com

Appendices

A Estimating t_1

In Chapter 6, the surrogate data recovery analysis assumed perfect estimation of t_1. By nature, heuristic estimation methods are unlikely to be this accurate. Therefore, we completed a systematic exploration of how errors in t_1 estimation influence our recovery metrics.

This process was similar to the optimisation of w_c; the fitting procedure was repeated on each surrogate dataset across a series of t_1 errors from -0.1s to +0.1s in steps of 0.01. The results of this work are shown in Fig. A.1. This 200ms window covered the range of realistic t_1 values and is consistent with the spread of heuristic t_1 estimates calculated for the participants.

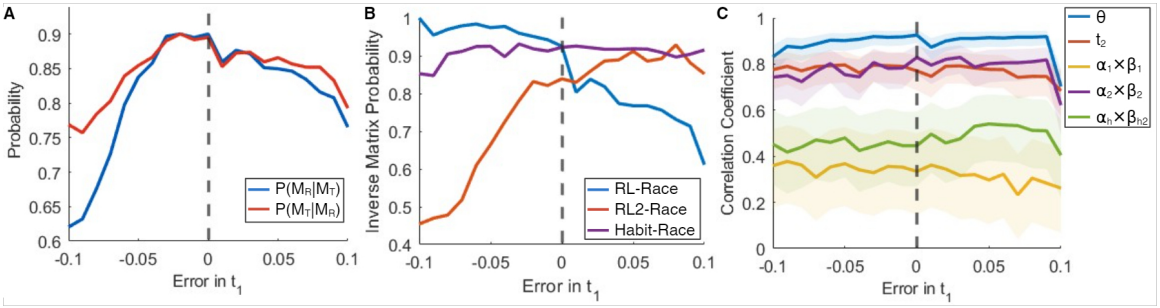


Figure A.1: Assessing t_1.
A: The influence of errors in t_1 on the diagonal sum of the confusion matrix (blue) and inverse matrix (red). **B:** As with A for the inverse matrix ($P(M_T|M_R)$), separated into the three model classes; RL-Race (blue), RL2-Race (red) and the two Habit-Race models (yellow). Recovery of Habit-Race is largely unimpaired by noise in the t_1 heuristic. **C:** The varying impact of errors in t_1 on the average correlation coefficients of each parameter group. The plots provide the average correlation across models ($\pm s.d.$). As in Fig. 6.5C, the product of α and β has been subdivided into those used for variables in the first accumulation period ($\alpha_1 \times \beta_1$, H_a and $Q_{1,a}$) and the variables that are introduced at t_2 ($\alpha_2 \times \beta_2$, Q_a and $Q_{2,a}$). The β_{h_2} term used by Habit-Race$_{2\beta}$ is further separated.

Encouragingly, the reliability of the model recovery process remains above 80% as the

accuracy of t_1 jitters between -60ms to +90ms from its true value (Fig. A.1A). Further insight into the impact of this error on recovery is provided in Fig. A.1B. In particular, the likelihood of accurately having detected a TD-RDM that uses APE-based habits when either Habit-Race model is selected consistently remains at ~ 90% across across the entire 200ms range. This suggests that the factors used to distinguish the Habit-Race models from the controls are largely unaffected by noise in t_1, potentially through compensatory mechanisms such as equivalent changes to t_2.

Instead, the drop in reliability results from a shift between the two reward-based models. An erroneously low t_1 causes many RL-Race models to instead be recovered as RL2-Race, and vice-versa for a positive t_1 errors. This results from the nested nature of these two models. Although RL-Race is presented as a single-drift RDM, it is also a specific case of RL2-Race where μ_1 and μ_2 are equal. As such, certain parameter settings can produce equivalent behaviour by either agent.

For example, in datasets created from the RL-Race model, the RL2-Race framework is able to compensate for an underestimation in t_1 by reducing t_2 towards the true t_1 value and maintaining low α_{q_1} and β_{q_1} parameter values. This results in minimal accumulation until the second drift-rate begins, which is functionally equivalent to RL-Race with a non-decision time of t_2 rather than the erroneously estimated t_1.

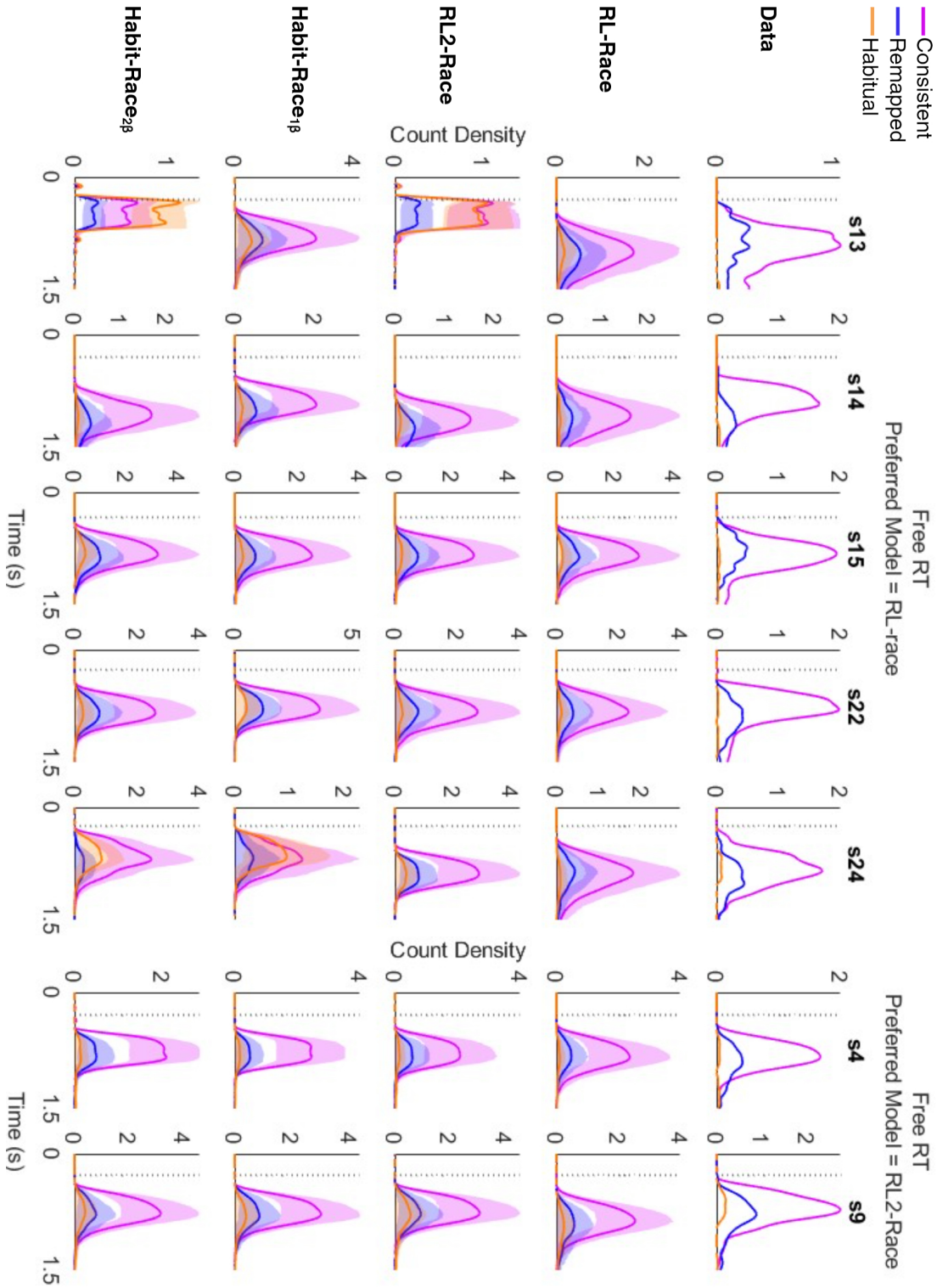
Conversely, due to the exclusion criterion for trials with RTs shorter than t_1, if this parameter is mistakenly large then much of the behaviour prior to t_2 is lost and the two systems are harder to distinguish from one another. This causes the more complex RL2-Race model to be rejected in favour of RL-Race during BIC analysis.

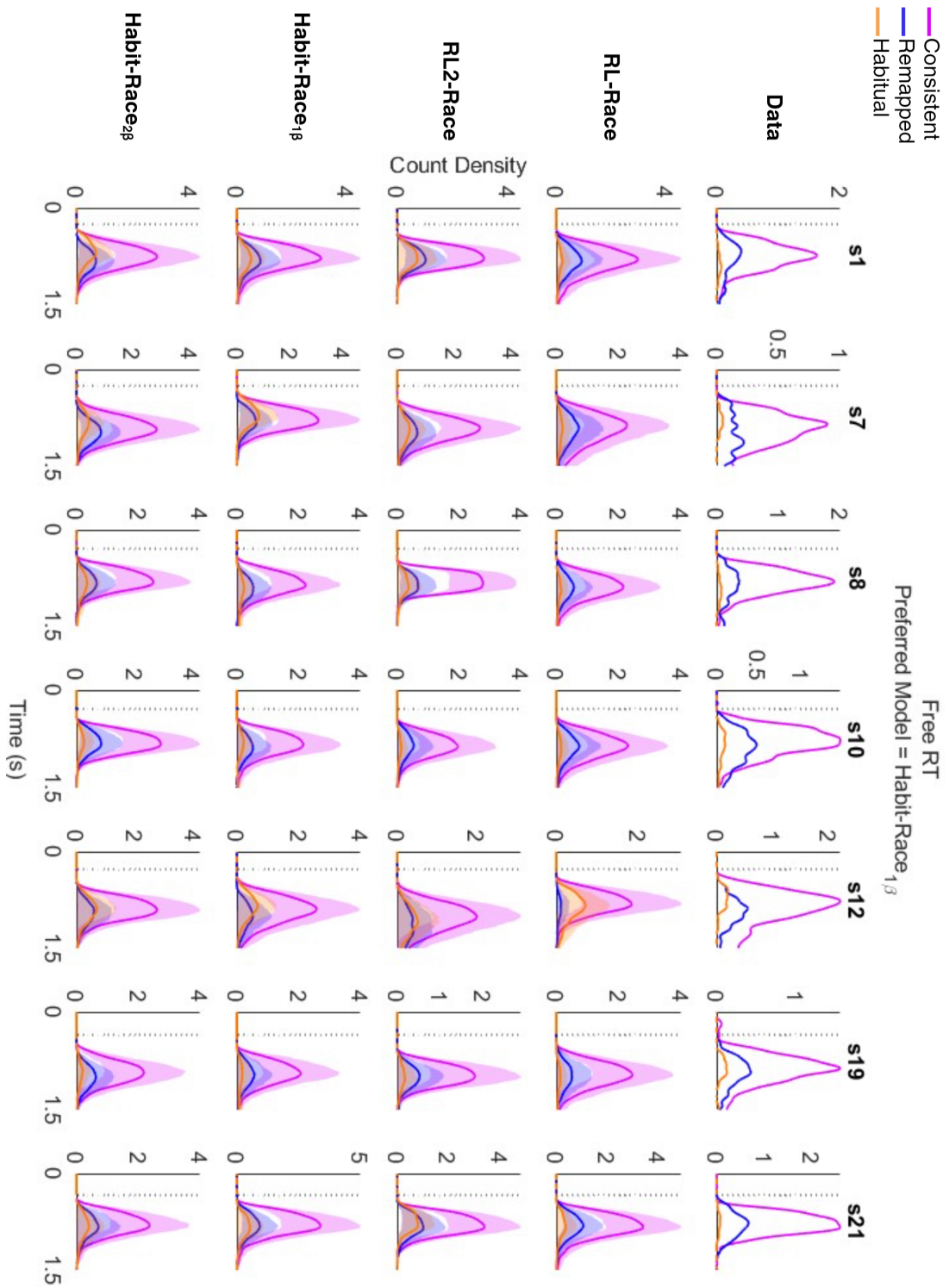
Additionally, Fig. A.1C reveals that the recovery of the remaining parameters is primarily unaffected by deviations in t_1.

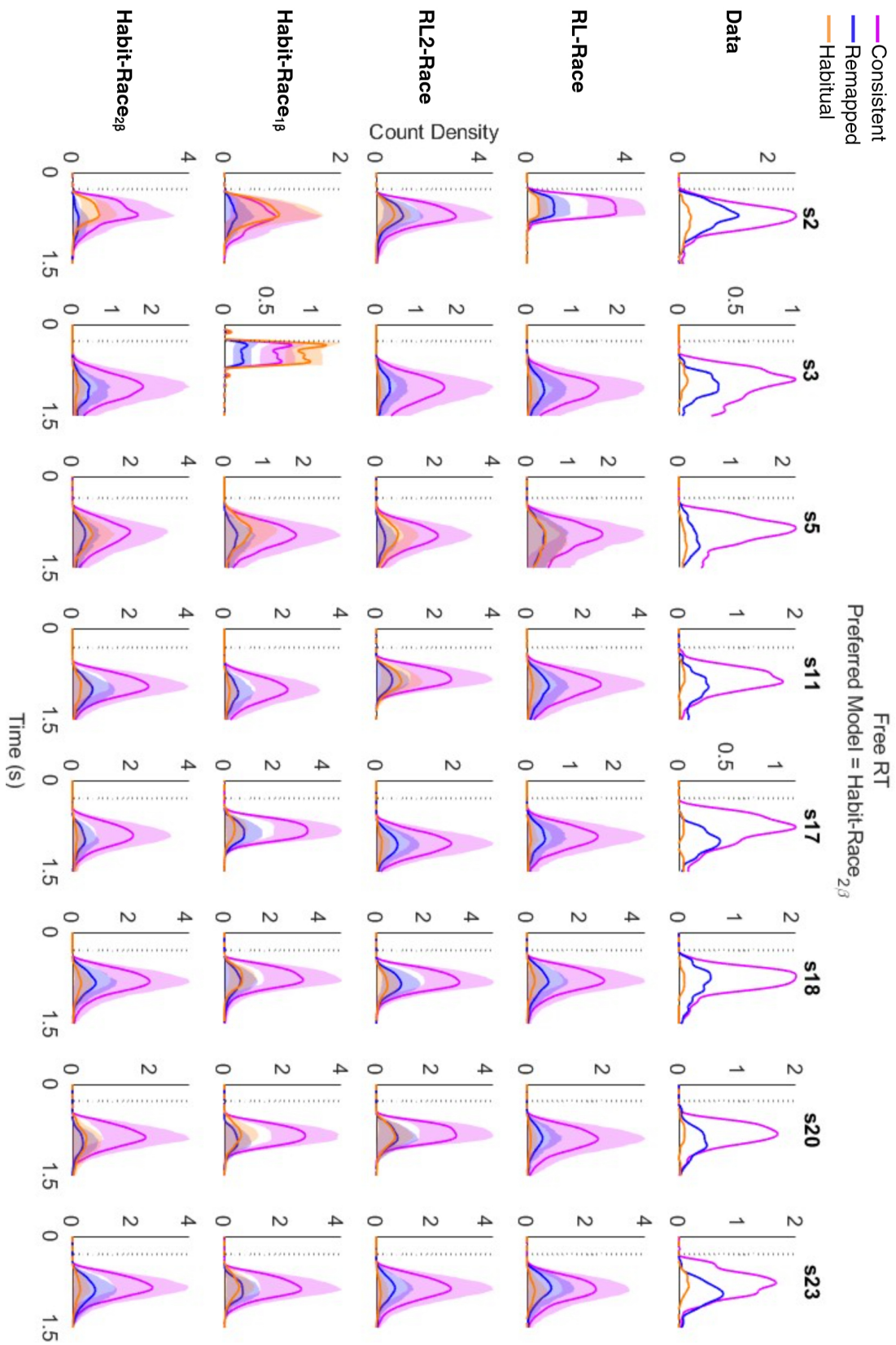
Overall, it is unlikely that the expected inaccuracies in the t_1 heuristic will dramatically effect the accuracy of the final estimated parameters or our confidence in the recovery of the habit-based TD-RDM. Some consideration must, however, be given when comparing the two pure RPE systems.

B Supplementary Figures for Chapter 6

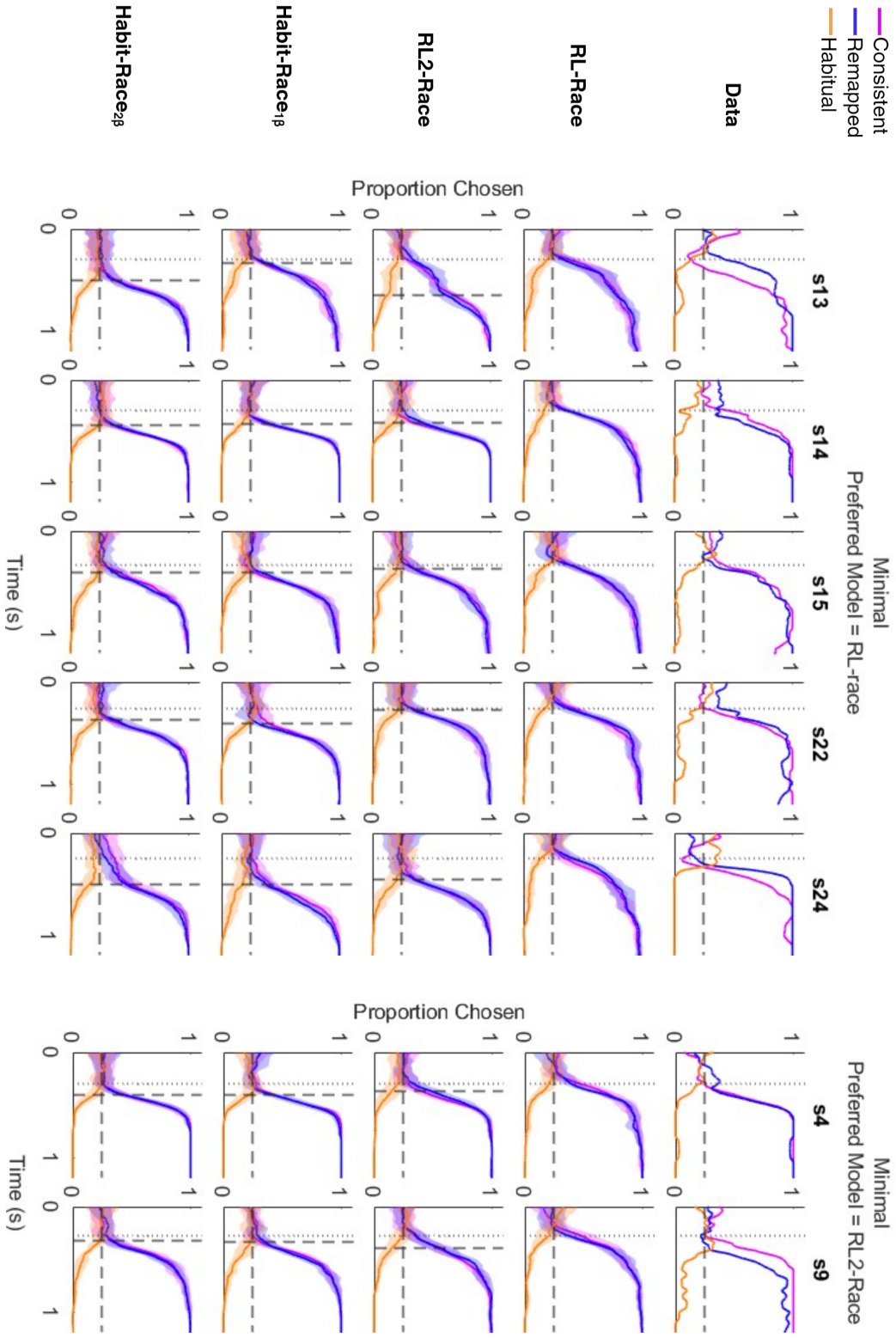
B.1 Free-RT trials

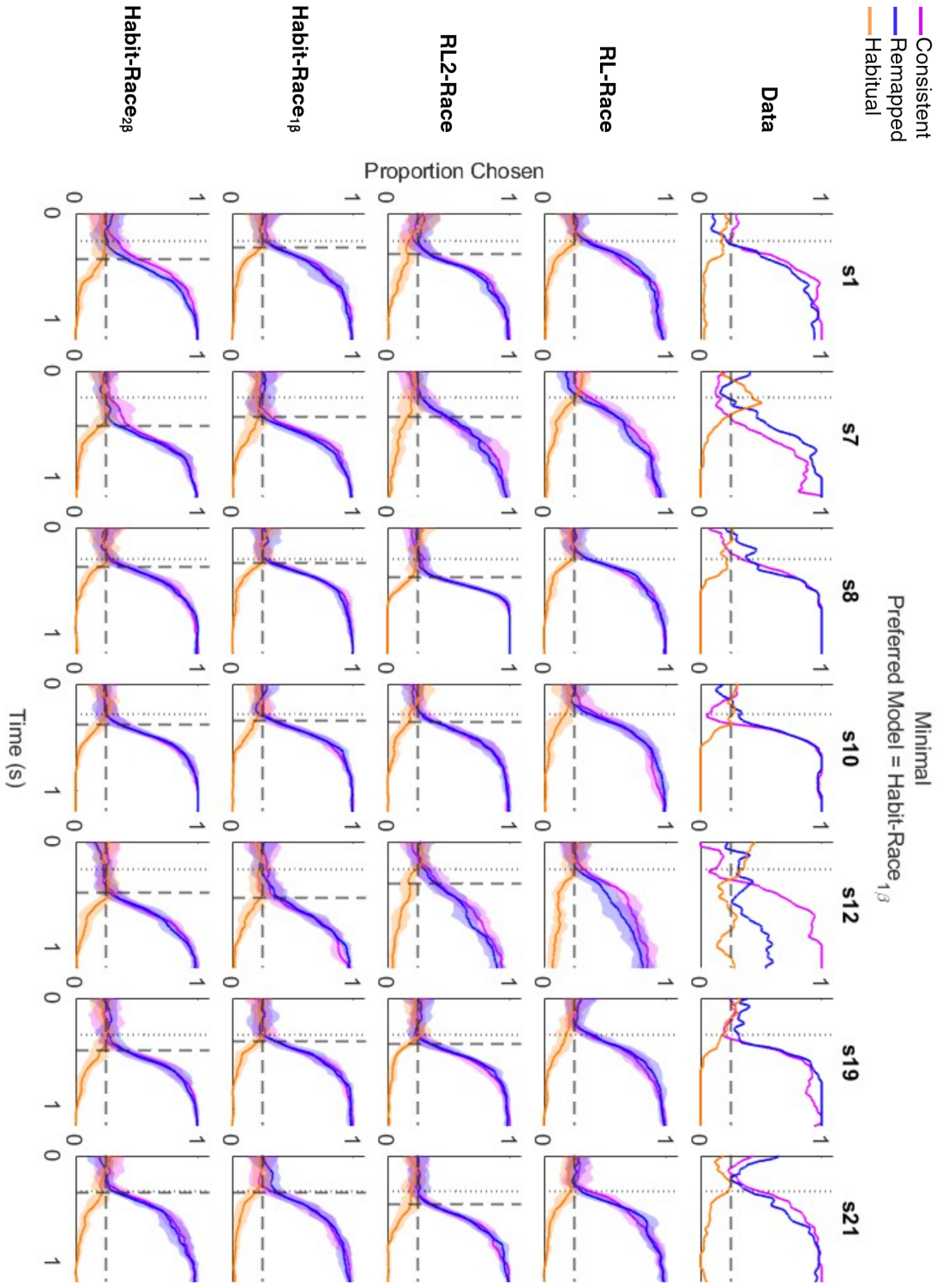


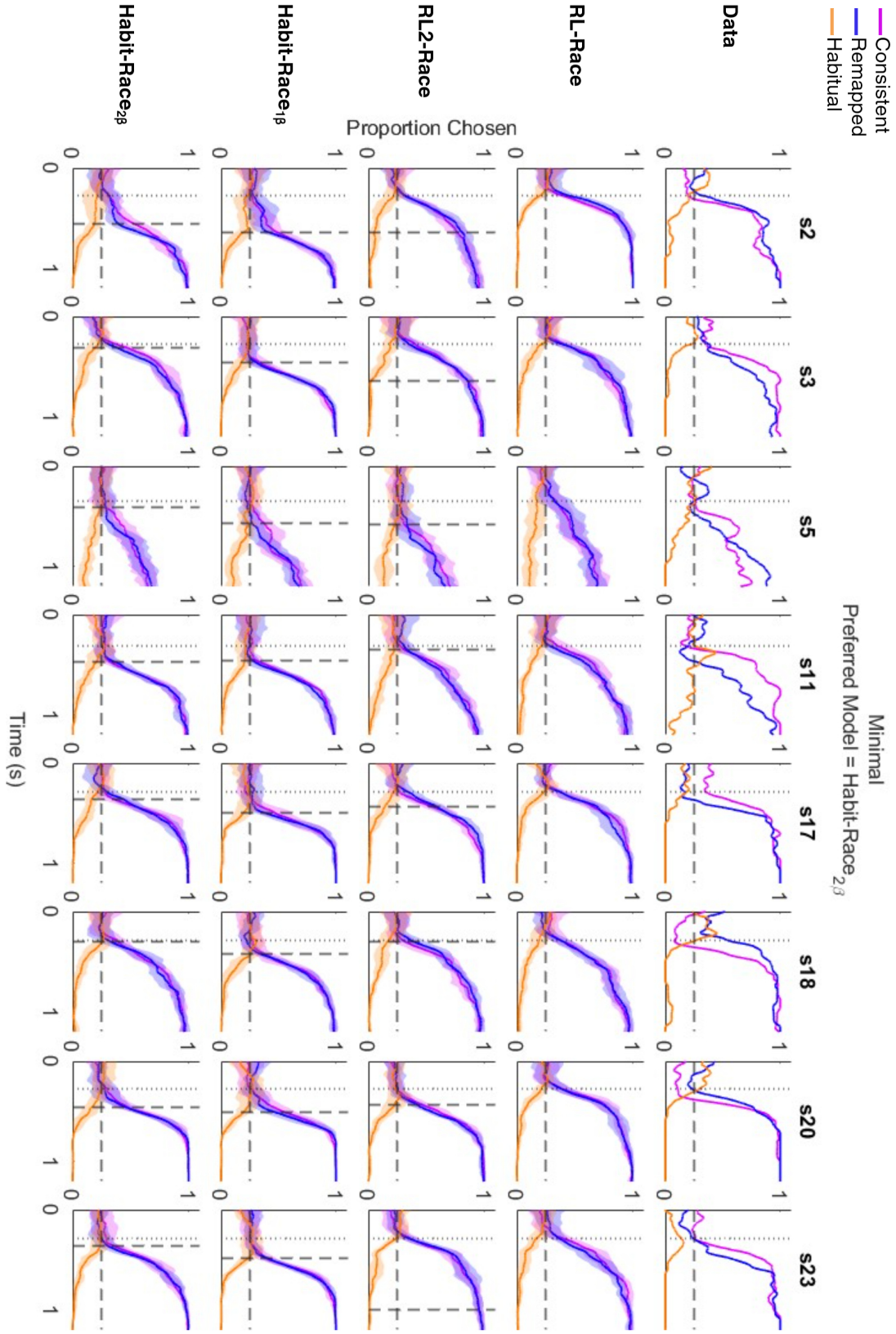

Consistent
Remapped
Habitual
Free RT
Preferred Model = Habit-Race1β
Data
RL-Race
RL2-Race
Habit-Race1β
Habit-Race2β
s1
s7
s8
s10
s12
s19
s21
Count Density
Time (s)


Consistent
Remapped
Habitual
Free RT
Preferred Model = Habit-Race2β
Data
RL-Race
RL2-Race
Habit-Race1β
Habit-Race2β
s2
s3
s5
s11
s17
s18
s20
s23
Count Density
Time (s)

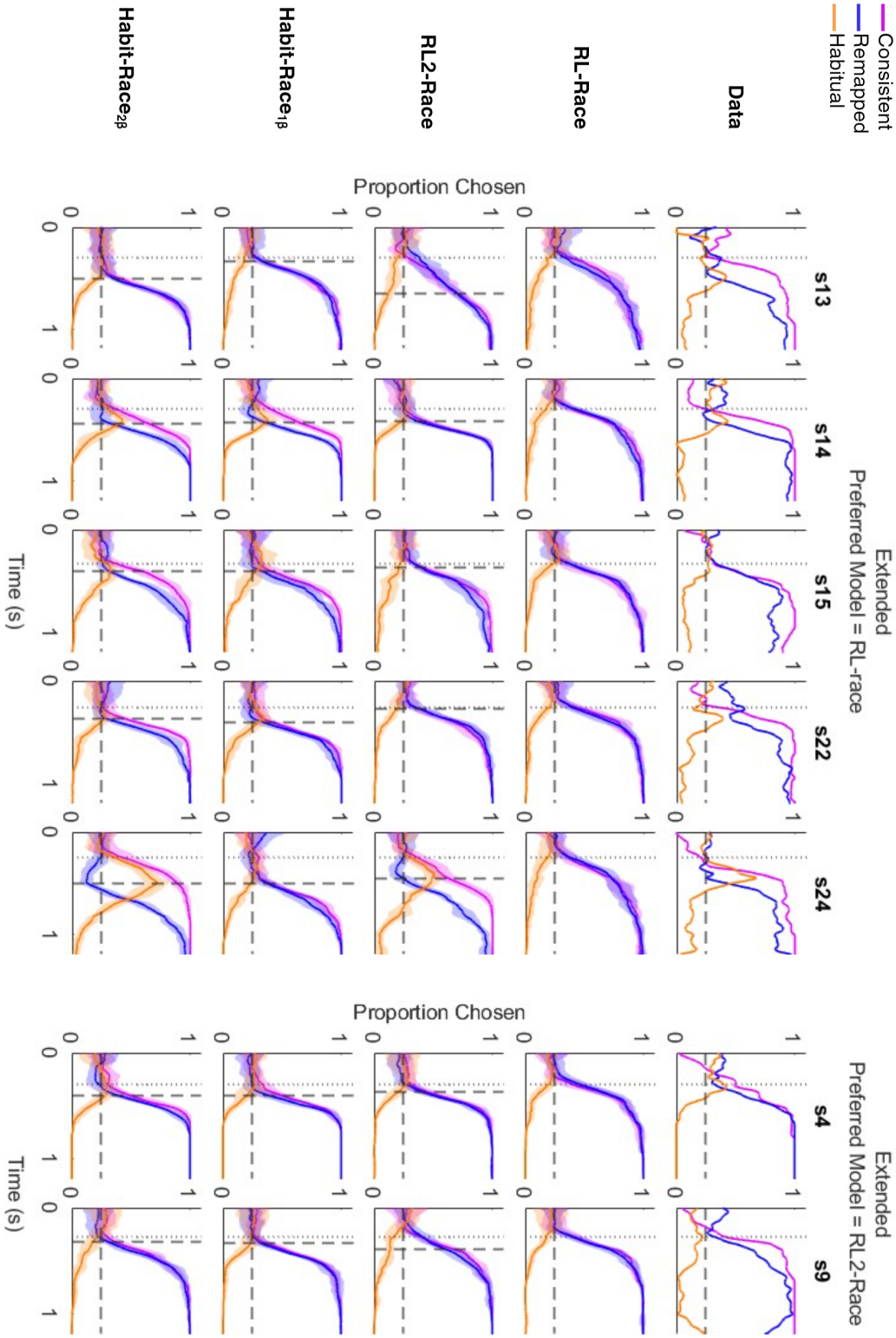
B.2 Minimal time-controlled trials

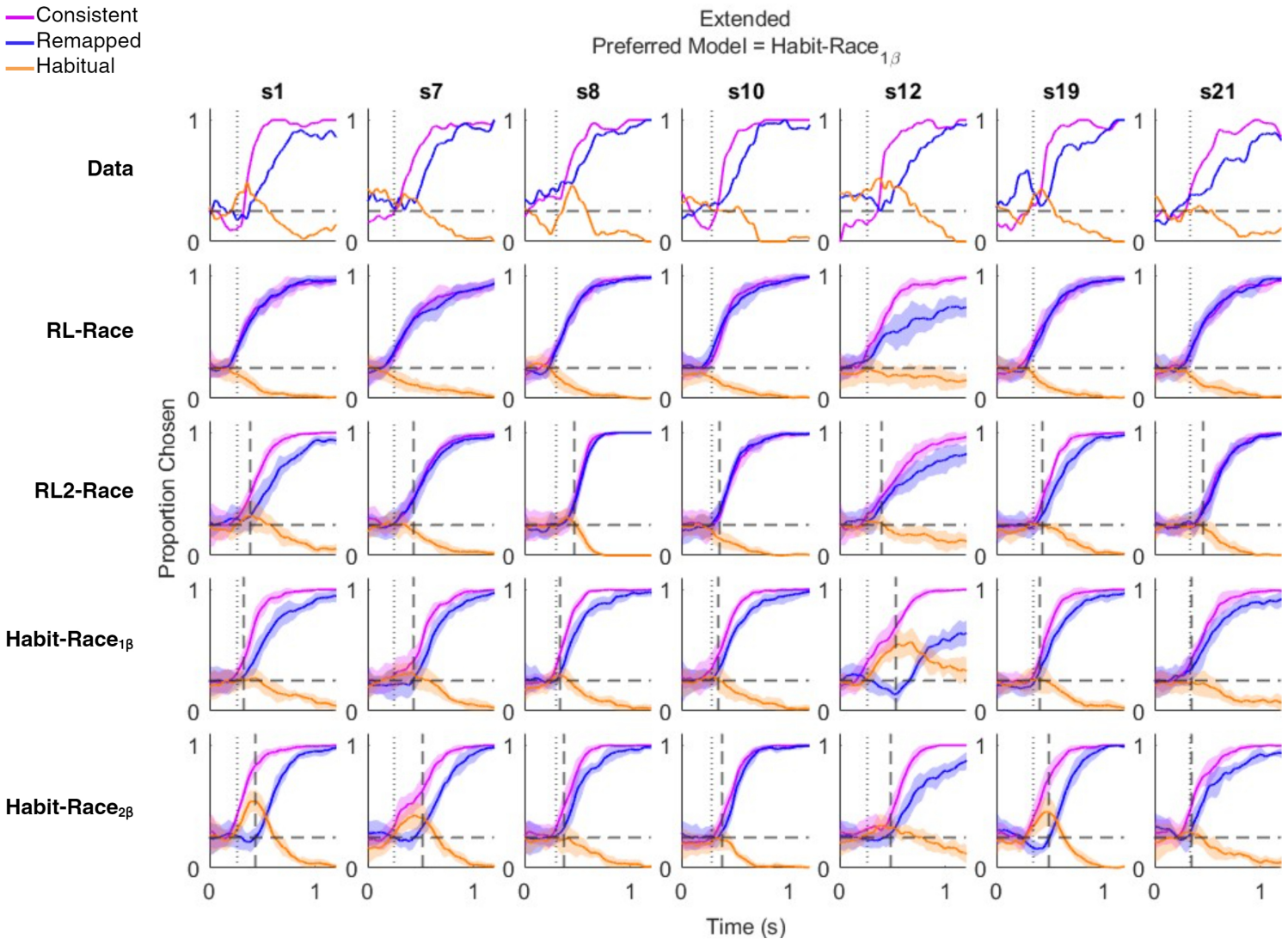


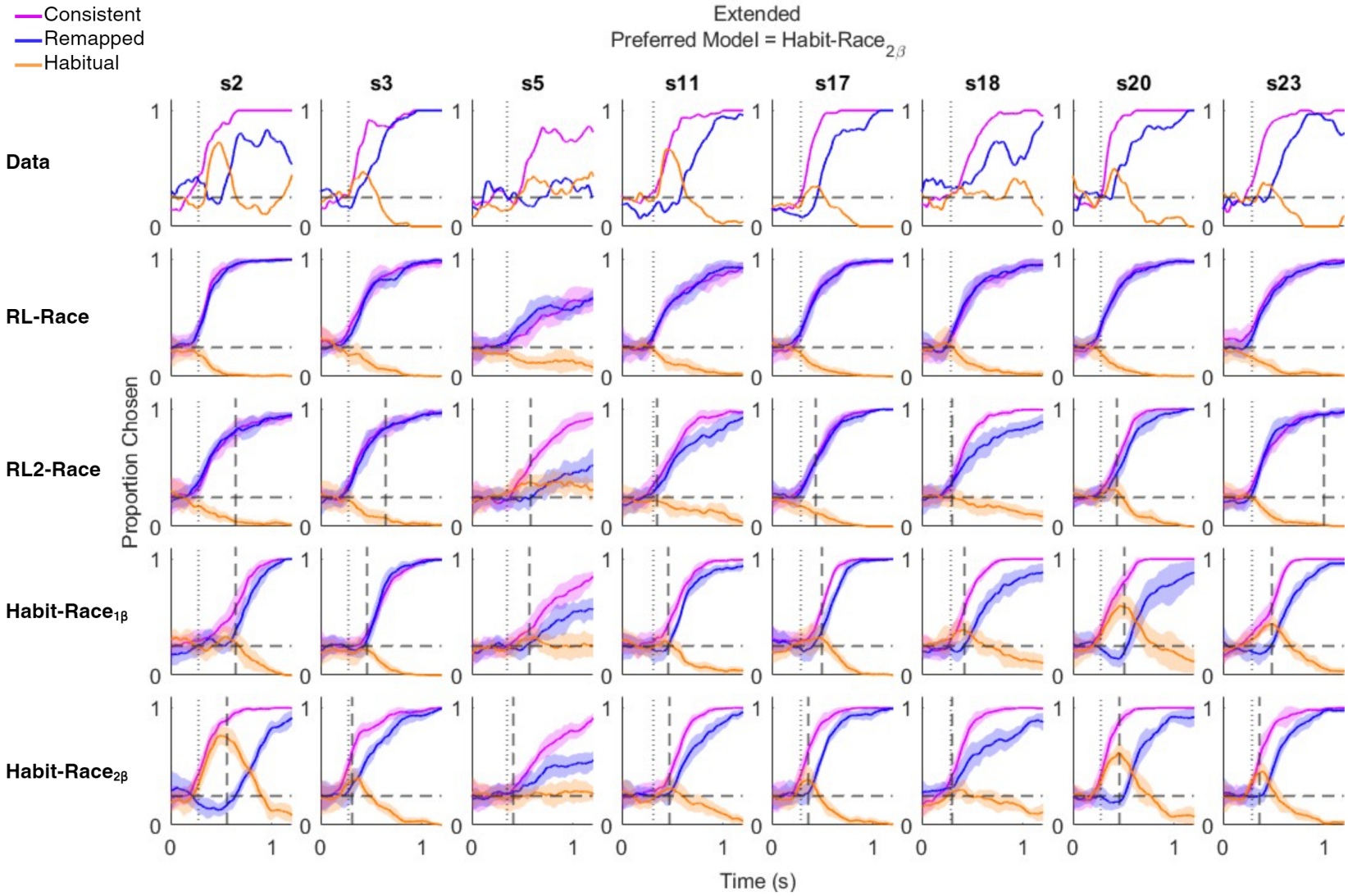

Consistent
Remapped
Habitual
Minimal
Preferred Model = Habit-Race1,β
Data
RL-Race
RL2-Race
Habit-Race1,β
Habit-Race2,β
s1
s7
s8
s10
s12
s19
s21
Proportion Chosen
Time (s)


Consistent
Remapped
Habitual
Minimal
Preferred Model = Habit-Race2β
Data
RL-Race
RL2-Race
Habit-Race1β
Habit-Race2β
s2
s3
s5
s11
s17
s18
s20
s23
Proportion Chosen
Time (s)

B.3 Extended time-controlled trials




Consistent
Remapped
Habitual
Extended
Preferred Model = Habit-Race1β
s1
s7
s8
s10
s12
s19
s21
Data
RL-Race
RL2-Race
Habit-Race1β
Habit-Race2β
Proportion Chosen
Time (s)


Consistent
Remapped
Habitual
Extended
Preferred Model = Habit-Race2β
s2
s3
s5
s11
s17
s18
s20
s23
Data
RL-Race
RL2-Race
Habit-Race1β
Habit-Race2β
Proportion Chosen
Time (s)

B.4 Individual Participant Recovery

Interestingly, there is minimal overlap between our results and Hardwick's AIC analysis regarding the specific datasets that were best replicated with a single drift-rate. More precisely, only one of our five RL-Race participants, S15, was also found by Hardwick et al.[4] to prefer a 'no-habit' model after both minimal and extended training.

Though on the surface this discrepancy may be surprising, further consideration provides a few explanations. First, the 17 participants quoted for Hardwick et al.[4] is specific to the extended training condition whereas only one participant, S11, provided evidence for multiple drift-rates after minimal training. In contrast, the TD-RDM analysis was fit equally to both conditions and is largely able to replicate both sets of accuracy curves, though there are some limitations to this method (Section 6.5.3). Qualitatively, if the extended time-controlled trials alone were considered, even with the current best-fitting parameters, it is likely that at least one of the RL-Race participants (i.e., S24) would be reclassified as either RL2-Race or Habit-Race$_{2\beta}$.

Second, the assumption of constant parameter values is likely to be an oversimplification and recovery of R_h appears to be particularly compromised as the model attempts to account for both minimal and extended data. Similarly, the inclusion of the free-RT trials in combination with fixed non-decision times appears to disadvantage the TD-RDMs to a greater degree than RL-Race. Future studies could potentially ameliorate these effects by allowing the observation model to vary between trial types while holding the RL parameters constant.

In combination, precise classifications for the response-selection and TD-RDM are likely to differ as their cost functions apply to different types and subsets of data. Though the exact participant fits may be inconsistent, the same conclusion can be drawn - the majority of participants' behaviour could be better explained by an observation model that applies two drift-rates.