



Neural circuit mechanisms underlying contingency learning

Citation

Qian, Lechen. 2024. Neural circuit mechanisms underlying contingency learning. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37378696>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Department of **Molecular and Cellular Biology**
have examined a dissertation entitled
Neural circuit mechanisms underlying contingency learning

presented by **Lechen Qian**

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature *Joshua R Sanes*
Joshua R Sanes (Feb 8, 2024 13:16 EST)

Typed name: Prof. **Joshua Sanes**

Signature *Adam Cohen*
Adam Cohen (Feb 8, 2024 13:42 EST)

Typed name: Prof. **Adam Cohen**

Signature *Samuel Gershman*
Samuel Gershman (Feb 8, 2024 13:56 EST)

Typed name: Prof. **Sam Gershman**

Date: **February 8, 2024**

Neural circuit mechanisms underlying contingency learning

A dissertation presented

by

Lechen Qian

to

The Department of Molecular & Cellular Biology

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the subject of

System Neuroscience

Harvard University

Cambridge, Massachusetts

Feb 2024

© 2024 Lechen Qian

All rights reserved

Neural circuit mechanisms underlying contingency learning

Abstract

Studies on animal learning have shown that the efficacy of both associative learning and responding depends on contingency, the extent to which the likelihood of an outcome changes by a presentation of a stimulus. However, how neural representations in the mesolimbic dopamine system and ventral striatum are modulated by contingency remain unclear. Chapter I of my dissertation focuses on this by examining mice trained in a Pavlovian conditioning degradation task. I observed that anticipatory licking for odors previously associated with rewards significantly diminished when unpredicted rewards were introduced but remained stable with predicted additional rewards. Furthermore, dopamine (DA) axonal signals in the Olfactory Tubercle (OT) and Nucleus Accumbens (NAc), measured using photometry, mirrored these behavioral changes. While dopamine signals are commonly thought to resemble the temporal difference (TD) error in TD learning models, previous studies suggested these models couldn't explain the contingency degradation phenomenon. However, our findings indicated that DA responses in various experimental conditions align closely with a TD learning model incorporating state transitions reflective of task structure. Furthermore, we showed that recurrent neural networks (RNNs) trained in the task using a TD learning framework recapitulate dopamine responses and develop state representations consistent with the task states, merely from observations. Based on these results, we provided a theoretical framework linking TD errors to contingency and causal learning

Many studies of reinforcement learning in rodents have involved the olfactory system. Previous studies have indicated that neural activity related to the odor valence have been observed in the OT, a relatively understudied structure situated in both the olfactory pathway and the ventral striatum. In **Chapter II**, I examined how valence representations of odor cues evolve over time in the two distinct neuronal populations in the OT, and how stable these representations are under different conditions in contingency

learning using 2-photon microscopy. As the idea of representational stability and drift are vigorously debated, this study will add valuable biological data to inform conceptual ideas of neural representations in the brain.

Together, these results provided new insights into the neural circuits involved in the representation of contingency in the mammalian brain.

Acknowledgements

I want to express my heartfelt gratitude to my Ph.D. advisors, Nao and Venki. Their guidance has been invaluable throughout my journey. Nao is an incredible advisor. Nao's dedication and work ethic have inspired me, driving all of us in the lab to push our limits. His readiness to assist, with his patience and constructive approach, has been instrumental in helping me navigate through multiple challenges during my Ph.D. training. His insightful vision and our dynamic back-and-forth discussions on pinpointing promising research areas significantly catalyzed the breakthroughs presented in Chapter I of this dissertation. Venki is an exceptional mentor who has consistently gone the extra mile in every aspect of my scientific training. His unwavering support and optimism have been a beacon of hope, especially when I was deeply frustrated by the two-photon experiments. Venki is impressively knowledgeable but humble and passionate about science, jazz, and life, which also deeply inspires me. Moreover, Venki's encouragement and support in exploring career opportunities have been significant to me.

My gratitude also extends to my colleagues in both labs. Their assistance and friendship have made the lab a vibrant and enjoyable workplace. I want to express my special thanks to Mitsuko for her support and guidance in the literature review, presentation preparation, and data visualization. Her dedication to scientific truth and rigor has provided a guiding example for my work and a shining role model for female scientific workers. Ding has been an important mentor since my undergraduate internship at the Dulac Lab. His passion for science and the brain ignited my initial curiosity and love for neuroscience. His continuous support and engaging scientific discussions have been invaluable during my Ph.D. journey. Nune's generosity in sharing her expertise in two-photon imaging was crucial in establishing my first functional pipeline. Mark, my closest collaborator and coauthor, is incredibly amazing and has been a fountain of brilliant ideas, significantly enriching our dopamine project. My dear desk mate and late-night buddy, Isobel, has been a steadfast companion, sharing joys and challenges. Also, despite Vic's busy schedule, his willingness to help with laser issues is greatly appreciated.

To Hao, Abby, Ningjing, Bo, Siddarth, and Luis in the Murthy lab, thank you for creating a sense of community and belonging and I really enjoyed working with you. Iku, Ryu, Malcolm, and other lab mates have contributed to a harmonious environment in the Uchida lab, making my time there truly enjoyable. My experience in these labs has been a professional journey and a personal growth adventure, thanks to each of you.

I also deeply thank my dissertation advisory committee: Catherine Dulac, Florian Engert, and Samuel Gershman. Your invaluable advice during our committee meetings has kept my project focused and on track. The insights and guidance you have provided have significantly shaped my research journey. I also wish to express my sincere appreciation to my exam committee: Samuel Gershman, Adam Cohen, and the chair, Joshua Sanes. Thank you for dedicating time from your busy schedules to review my dissertation thoroughly and for attending my defense. Having such esteemed professionals contributing to my academic journey is truly an honor.

Last but certainly not least, I am profoundly thankful to my family. My parents, in particular, have been a constant source of support in my educational and career endeavors. Their encouragement has been a cornerstone of my success. Additionally, I am blessed to have friends like Yulu, Yuhan, Chris, Zhengjia, and Hanrui. Your accompany and friendship during challenging times have been invaluable. I am truly grateful for your presence in my life and your unwavering support.

To all of you, thank you for being part of this significant chapter in my life. Your support and guidance have been immeasurable.

Table of contents

Title page	i
Copyright	ii
Abstract	iii
Acknowledgements	v
Table of contents	vii
List of Figures	viii
Contributions	x
Introduction	1
Animal associative learning: classical conditioning and instrumental conditioning.....	2
Contingency and contingency learning theory.....	4
Contingency degradation	5
Temporal difference (TD) learning algorithm and TD errors	6
Dopamine as an important teaching signal	7
Ventral Striatum as a hub for reward learning	9
Olfactory Tubercle: anatomical structure and function in reward processing	10
Summary	12
Chapter I. Dopaminergic responses are sensitive to stimulus-outcome contingencies: computational accounts based on temporal difference learning models	13
<i>Introduction</i>	13
<i>Results</i>	16
Contingency degradation attenuates dopaminergic cue responses	21
TD learning models can explain dopamine responses in contingency degradation.....	23
Other aspects of dopamine responses and model predictions	31
Recurrent neural networks that learn to predict values through TD learning can explain dopamine responses	35
A retrospective learning model, ANCCR, cannot explain the dopamine responses.....	38
<i>Discussion</i>	41
TD learning model as a model of associative learning	41
State representations as population activity dynamics	43
Limitations of the ANCCR model as a model of associative learning and dopamine.....	44
<i>Methods</i>	46
Chapter II. Odor Valence Representation in Olfactory Tubercle during contingency learning	55
<i>Introduction</i>	55
<i>Results</i>	57
OT neuronal activity robustly and bidirectionally reflects odor-outcome contingency.....	59
Longitudinal analysis of individual neurons during contingency learning.....	62
<i>Discussion</i>	64
Representational stability and drift	65
<i>Methods</i>	67
Conclusion	72
References	78

List of Figures

Figure 1 Dynamic changes in lick response to olfactory cues across different phases of Pavlovian contingency learning task.	17
Figure 2 Population average behavior per session	20
Figure 3 Dopamine axonal activity recordings show different responses to rewarding cues in Degradation and Cued Reward condition	22
Figure 4 Population Average Dopamine Response per session	23
Figure 5 TD learning models can explain dopamine responses in contingency degradation with appropriate ITI representation	26
Figure 6 Discount factor determines modeled contingency degradation effect size	27
Figure 7 Belief-State model, but not Cue-Context model, explains variance in behavior and dopamine responses	29
Figure 8 Relative value explains decreased anticipatory licking during ISI during contingency degradation ·	30
Figure 9 Belief-State model's prediction recapitulate the other experimental data	32
Figure 10 Comparison of reward and omission responses between experimental data, Belief-State model and value RNN predictions	34
Figure 11 Value RNNs recapitulate experimental results using state-spaces akin to hand-crafted belief state model	37
Figure 12 ANCCR does not explain the experimental results.....	40
Figure 13 Outcome-specific contingency degradation explained by belief state model and value RNN model.	45
Figure 14 Task design and behavioral responses	58
Figure 15 Two- photon calcium imaging of D1- and D2- type neurons in the olfactory tubercle (OT) during odor- outcome association learning.	60
Figure 16 D1- type neurons in the olfactory tubercle	61

Figure 17 Population level quantification of D1/D2 neural responses	61
Figure 18 Neural response dynamics in D1 and D2 Neurons during contingency degradation and longitudinal tracking across Training Sessions	63

Contributions

The work presented in this thesis has been significantly bolstered by a collaborative effort. The data and methodologies detailed herein benefited from direct contributions as follows:

In Chapter I:

Mark Burrell has been my primary collaborator. He constructed all the TD learning models and conducted the analysis for both the TD learning models and the RNN model. He also provided critical feedback/discussions on other analyses.

Sara Matias established the multifiber photometry system and has been a constant source of support in troubleshooting and guidance. She also generously shared her analysis code with me and supplied the transgenic mice.

Jay A. Hennig constructed RNN models and has provided feedback and advice as we delved deeper into the analysis and interpretation of the models.

In Chapter II:

Nuné Martiros has been my primary mentor for this project. She meticulously taught me the fabrication of custom-made cannulae, their implantation, and the nuances of two-photon calcium imaging. Her mentorship enabled me to conduct the experiments independently and efficiently in a relatively short timeframe.

Vikrant Kapoor helped with early rig and olfactometer building. His advice on laser usage has been invaluable, as was his LabVIEW code which facilitated the control of the imaging process.

Hao Wu provided a Python toolbox essential for conducting the behavioral experiments. His contribution significantly streamlined the experimental process.

Introduction

The ability of animals to assign valence - positive or negative - to environmental cues is essential for flexible behaviors and survival. Such associations endow the animal with the ability to swiftly prepare appropriate behavioral responses to familiar stimuli and to update valence of sensory stimuli as environmental contingencies change. Neuronal circuits involved in assigning and recalibrating valence to sensory stimuli based on experience and environment remain a focal point of neuroscience research. Understanding these processes is essential, because certain neural associations can culminate in maladaptive behaviors¹. However, deciphering these mechanisms has been challenging due to the distributed and intricate circuits underlying the interplay of sensory perception, associative learning, and reward processing in mammalian brains²⁻⁶.

Over the last five decades, animal learning research has elucidated fundamental components of simple associative learning, which often involved reward processing by presenting cues followed by reward. A seminal neurophysiological discovery is that dopamine neurons in the midbrain appear to encode reward prediction errors⁷⁻⁹ (RPEs), aligning with the theoretical error signals posited in models of animal learning. Concurrently, the field of machine learning has advanced algorithms, e.g., temporal difference learning (TD learning) in the reinforcement learning framework¹⁰⁻¹³, detailing how an agent might optimize reward acquisition through environmental interactions, thereby offering computational frameworks for associative learning. Activities of the ventral striatum (VS)-projecting dopaminergic neurons have been consistently observed to reflect TD error¹⁴. Yet, how these neurons behave in various conditions in associative learning and whether their activity aligns with TD error theory remain unclear.

Understanding how sensory valence is assigned and represented in the brain during associative learning and the stability of this representation is another important neuroscientific inquiry. The rodent olfactory system is ideal for studying this, as mice and rats can quickly learn tasks based on odor cues, forming associations between odors, motor actions, and reward¹⁵⁻¹⁷. Therefore, my research utilized mice, with a particular focus on an olfactory area in the brain, olfactory tubercle (OT), a region considered to be

part of the ventral striatum and implicated in reward learning and motivated behaviors^{18–23}. The OT is predominantly made up of medium spiny neurons expressing D1 or D2-type dopamine receptors^{24–26}. These neurons, similar in cellular architecture to those in other striatal regions, have widespread connections with the basal ganglia and midbrain dopaminergic system^{26–28}. This extensive interconnectivity positions the OT as an ideal site for investigating the neural dynamics of reward processing and reinforcement learning^{17,19,29–33}. Investigations into the stability of D1 and D2-type neuron responses in the OT after performance stabilization, as well as their adaptive changes during the degradation of stimulus-outcome contingencies, could deepen our comprehension of the neural underpinnings of associative learning.

In this dissertation, I will focus mostly on Pavlovian conditioning tasks, which have consistently served as a robust paradigm for investigating associative learning models and their neural substrates. Through a combination of neurophysiological recording techniques and computational models, my findings will elucidate 1) the response patterns of the VS-projecting midbrain dopaminergic system during contingency learning, 2) the alignment of dopaminergic signaling with computational models within the TD learning framework, and 3) the acquisition, representation, and modulation of valence signals within OT. I will synthesize seminal studies from these related fields in the following sections.

Animal associative learning: classical conditioning and instrumental conditioning

Two primary animal learning frameworks – classical/Pavlovian conditioning and instrumental conditioning - have shaped our understanding of associative learning. In Pavlovian conditioning, organisms experience a sequence of conditioned stimulus (CS) and unconditioned stimulus (US) pairings, and the delivery of the US doesn't depend on the organism's actions between these stimuli. A US is a certain outcome that triggers a biological response without learning, such as salivation following a bite of tasty food. During Pavlovian conditioning, animals learn to predict a US, following a neutral sensory stimulus that does not evoke an automatic biological response. For example, in the seminal experiment, Pavlov^{34,35} paired the sound of a bell (CS) with the provision of food (US). After several repetitions, the dog started salivating - a typical conditioned responding - in response to the bell sound, even without the food. This

behavior demonstrates that the dog had learned to associate the bell sound with the reward, anticipating the forthcoming food.

The other type of conditioning is instrumental conditioning, where an organism must act to secure an outcome. An illustrative case is Thorndike's puzzle box experiment in which cats learn to escape to obtain fish³⁶. After repeating this procedure many times, Thorndike found that behaviors leading to food became more frequent in later attempts. Even over time, when cats were reintroduced to the box's similar context, these cats swiftly executed escape behaviors. In this dissertation, the primary focus will be on classical conditioning. Therefore, I will not extensively delve into the principles of instrumental conditioning.

Since the seminal demonstration of Pavlov, Pavlovian conditioning has been studied extensively. In 1955, Bush and Mosteller introduced a mathematical explanation for how association between CS and the US become learned in an animal's behavior. They proposed that the spatiotemporal contiguity of the CS and US is crucial for this link to form. Then, however, Rescorla and Wagner (1972) offered a different theory by proposing that conditioning is driven by the discrepancy between the actual and predicted outcome (prediction error)³⁷, not just the spatiotemporal contiguity. Their model focused on the strength of the association between CS and the US and considered the collective impact of all stimuli present. Rescorla and Wagner proposed that the change in associative strength between a CS and the US is proportional to the discrepancy between the maximum possible responding the US can produce and the current associative strength of all the CS's present. One intuitive way to understand this is that the US must be somewhat surprising, and surprise is defined as how much an outcome deviates from the expected outcome. Suppose a CS already predicts a reward fully, in that case, introducing another CS won't change the learning, because the prediction isn't violated. Thus, this model suggests that learning only happens when the prediction is incorrect - an error between what's expected and what occurs is necessary for learning to take place. The Rescorla-Wagner model has since been successful in explaining a wealth of experimental observations in animals' associative learning studies³⁸⁻⁴¹.

Contingency and contingency learning theory

As described above, behavioral and psychological research into animal learning revealed that mere spatiotemporal '*contiguity*' between two events is insufficient for establishing enduring associations. Instead, Rescorla further postulated that '*contingency*' – the degree to which the CS indicates an increase or decrease in the likelihood of the US occurring – is the critical factor in forming successful association. In 1968, Rescorla conducted two experiments to investigate the role of CS-US contingency in fear conditioning⁴². In experiment 1, he compared the conditioned response (CR) in the presence and absence of a tone (CS) and a shock (US). When the shock was given only during the tone, substantial conditioning occurred. However, he found that when the probability of the shock was equal in the presence and absence of the tone, there was no conditioned response to the tone. Then in experiment 2, he further explored four different probabilities of the shock in the presence and absence of the tone. He demonstrated that the amount of conditioning was higher when the probability of the shock was greater during the tone and lower when the probability of the shock was greater in the absence of the tone. When the two probabilities were equal, no conditioning resulted. This study was the first to provide evidence that the contingency between the CS and the US plays a critical role in conditioning and the effect of contingency has been well-documented later in many Pavlovian conditioning paradigms.

Rescorla proposed that the concept of contingency should be described as a comparison between two probabilities: the likelihood of an US when a CS is present, and the likelihood of the US in the absence of the CS. However, he did not provide a specific definition for this concept. So, what is a good measure of contingency remains to be clarified^{43–46}. Among the various mathematical definitions available for contingency, the most prevalent one assesses the change in outcome probabilities in the presence or absence of a specific stimulus^{6,47,48}. In Pavlovian settings, the measure of contingency, denoted as ΔP , is quantified by the formula $\Delta P = P(US|CS +) - P(US|CS -)$, where 'CS+' represents the presence and 'CS-' the absence of a conditioned stimulus. Although association alone does not establish causation, such associations often lead to the perception of causal links^{45,49}. In a seminal contribution to the field of causal inference, Judea Pearl demonstrated that, under certain conditions, contingency ΔP aligns with the key

metrics that define necessary and sufficient causes—namely, the "probability of necessity and sufficiency (PNS)" and the "probability of sufficiency (PS)." This establishes a robust mathematical foundation for both contingency and causal power⁵⁰. However, there are a couple of issues with the ΔP definition of contingency, as Gallistel pointed out in his review: First, counting events that didn't occur (CS-) was a challenge; Second, it's highly dependent on the structure of imaginary trial to determine the time intervals within which the probabilities of events such as CS and US^{46,51} can be defined; Lastly, some behavioral phenomena do not align with the classic ΔP definition of contingency, which led to some critiques on the usefulness of contingency in accounting for behavioral patterns⁵². As a result, there have been initiatives to refine the definition of contingency to address these discrepancies⁴³⁻⁴⁶.

Contingency degradation

Contingency degradation was first observed and formalized in Pavlovian conditioning by Rescorla⁵³⁻⁵⁵. Yet, contingency degradation has been more extensively studied in the instrumental conditioning⁵⁶⁻⁶⁰. In instrumental conditioning, the sensitivity of animals to the contingency between action and outcome has been used to distinguish goal-directed from habitual behavior⁶¹. Defined by Hammond as the difference in probability of reinforcer delivery following a response versus its absence, the contingency degradation paradigm is thought to examine animals' perception of the causal relation between response and reward^{62,63}. Hammond trained hungry rats to press a lever, where the first press per second consistently resulted in food pellet delivery. Then, he performed the contingency degradation by increasing the likelihood of food delivery without lever pressing, eventually equalizing the probabilities, and nullified the impact of pressing on reward likelihood. Then Hammond noted that this increase in reward probability without response depresses instrumental responding, a hallmark of contingency degradation.

Numerous subsequent studies have investigated animals' learning and sensitivity to contingency, observing their behavioral responses to conditioned stimuli signaling reward or punishment when contingency are altered. These studies demonstrate that mice^{64,65}, rats^{4,59,61}, dogs^{37,55}, non-human primates^{3,57}, and human⁶⁶⁻⁶⁸ all change their behavioral response in contingency degradation.

Temporal difference (TD) learning algorithm and TD errors

The Rescorla-Wagner model, a cornerstone in associative learning research, is constrained by its reliance on a trial-based approach largely ignoring the timing of events within and between trials. To represent the temporal dynamics of stimuli more accurately in classical conditioning, it is necessary to propose models that operate continuously on a moment-to-moment basis, rather than on a trial-by-trial framework. Addressing the gap of lack of temporal structure between stimuli, Sutton and Barto developed the TD learning algorithm, conceptualized as a prediction error-based model of associative learning^{69,70}. This approach has since become foundational in the study of value learning within the realm of reinforcement learning.

In TD learning models, TD errors is denoted by δ and defined by the equation:

$$\delta = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (1)$$

where r_t represents reward at time t , s_t denotes the state at that time t , $V(s_t)$ is value corresponding to state s_t , and γ , known as the temporal discount factor, falls between 0 and 1 ($0 < \gamma < 1$). The value $V(s_t)$ is defined as the expected sum of all future rewards starting from time t , with each future reward discounted by the factor γ per time step.

$$V(s_t) = E\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r(\tau)\right] \quad (2)$$

The TD error plays a pivotal role in learning by continuously refining the estimate of this value at each time point, unlike the Rescorla-Wagner model which limits learning to moments of reward delivery. A crucial component in TD learning models is the representation of “states”. In the original TD learning model⁷¹, states were represented using a ‘complete serial compound’ (CSC), sometimes known as a tapped delay line^{71,72}. In this construction, stimulus presentation triggers a sequential activation of sub-states $x(t) = \{x_1, x_2, \dots\}$, each representing a time step after the stimulus. At any given time after the stimulus, only one sub-state is active. The value estimate $\hat{V}(s_t)$ is then computed as the weighted sum of these substates $x(t)$, which in CSC is reduced to be the weight of the substate.

$$\hat{V}(s_t) = \sum_i w_i \cdot x_i(t) \quad (3)$$

where w_i is the weight associated with sub-state i . Weights are updated following the TD-learning rule:

$$\Delta w_i = \alpha \cdot x_i(t) \cdot \delta \quad (4)$$

where α is a learning rate ($0 \leq \alpha \leq 1$) and $\delta(t)$ is the prediction error in the value signal.

Initially, in the TD model, weights are unlearned, so the value estimate $\hat{V}(s_t)$ is zero everywhere. When a US, such as a reward, is presented, it generates a positive prediction error. The prediction error adjusts the weight of the corresponding CSC state, which influences future value estimates. This mechanism leads to a ripple effect of positive prediction errors, which propagate back in time, modifying the weights of the CSC state across the interstimulus interval (ISI) until the weight of the state that corresponds to the CS onset is modified. Consequently, the TD model with CSC state representation (CSC-TD model) learns to anticipate the future US in response to the CS.

TD learning models have been effective in capturing and forecasting the characteristics of animals' behaviors in a wealth of classical conditioning paradigms. These models could potentially offer insightful mechanisms that could underlie the contingency learning observed in animals. Within this structured framework, the features of learning and behavioral adaptation can potentially be systematically understood and anticipated.

Dopamine as an important teaching signal

In the prior section, it was noted that the TD error plays a pivotal role in TD learning algorithms. Nonetheless, it raises an intriguing query: does an animal's brain utilize a process similar to the TD learning algorithm, and if so, are there specific neural populations that mirror the TD error signals? This has been a subject of considerable research interest, as identifying these neural correlates could advance our understanding of how animals learn and adapt to their environment through experience. Seminal works by Schultz and colleagues revealed a strong link between the TD error and the response of midbrain dopamine neurons, and later, more evidence has revealed the similarity between the two^{73–77}. In Schultz's study, they

performed a Pavlovian conditioning paradigm with monkeys, teaching them to associate a visual cue (CS) with a drop of juice (US)⁷⁸. Initially, the putative dopamine neurons fire only when the juice is delivered. However, after the monkeys learned the association, these dopamine neurons fired at the onset of the visual cue instead of at the time of reward. The other important observation is the reward omission dip, which showed that if the predicted reward is omitted unexpectedly, dopamine neurons showed a phasic decrease in activity, forming a “dip” at the time when the reward was supposed to be delivered. These findings demonstrated that dopamine encodes reward prediction error (‘RPE’), a type of error that closely matches TD errors proposed in the TD learning models. In Uchida lab’s recent work, we have shown that when animals were learning an association, dopamine neuron firing activities showed a gradual shift from reward receipt time to cue onset, which aligns with the critical prediction of TD learning models⁷⁹. Another work⁸⁰ from the Uchida lab has demonstrated that dopamine neurons’ activities reflect the moment-by-moment changes in value using teleportation and manipulation of track movement speed in virtual reality. These findings further supported that dopamine signals match the essential property of TD error, rather than other variables such as value.

Despite these findings supporting the TD error hypothesis of dopamine, this hypothesis has been contested by some recent studies. One study, for instance, proposed an alternative model, **adjusted net contingency for causal relations** (ANCCR), where researchers proposed that dopamine conveys the learned meaningfulness of events and guides causal learning of reward event through a retrospective mechanism⁸¹. In their retrospective mechanism, animal backtracks in time to pinpoint probable causes of a "meaningful causal target" (MCT), such as a reward, after its occurrence. This model suggests that the significance of events is conveyed through mesolimbic dopamine signaling, which is pivotal in guiding the causal learning of current events. Contrasting with the forward-looking perspective of TD learning models, the ANCCR theory emphasizes the importance of the likelihood of a cue given a reward ($P(\text{cue}|\text{R})$) - essentially asking, "Is there a precedent cue for the reward?" Conversely, the TD model concentrates on the likelihood of a reward following a cue ($P(\text{R}|\text{cue})$)—asking, "Is a cue predicting a future reward?". The authors claimed that ANCCR model could successfully explain the dopamine responses in various

experiments that they performed, including a contingency degradation paradigm, while TD learning models failed.

Testing these competing models could refine our understanding of dopamine's role and stimulate fresh research. In Chapter I of my dissertation, I will aim to delineate the core properties and assumptions of these models, devising targeted experiments to test these theories and models about dopamine's functions in contingency learning and to address recent debates. This will enhance our comprehension of the fundamental principles of associative learning in the brain.

Ventral Striatum as a hub for reward learning

The striatum, a subcortical part of the forebrain, is a complex and multifaceted structure integral to the basal ganglia network. It plays a pivotal role in a myriad of functions, including motor control, learning, and decision-making. Characterized by a rich mosaic of cellular and cytoarchitectural composition, the striatum's heterogeneous nature is crucial for its diverse functional capabilities.

At the cellular level, the striatum is predominantly composed of medium spiny neurons (MSNs), which account for approximately 95% of its neuronal population. These GABAergic neurons are the principal projection neurons of the striatum and are distinguished by their dense, spine-covered dendrites from which the name was derived. MSNs are further categorized based on the type of dopamine receptor that they express: those expressing dopamine D1 receptors are part of the direct pathway, which projects mainly to the basal ganglia output structure such as globus pallidus internal segment and substantia nigra pars reticulata⁸², while those expressing D2 receptors are part of the indirect pathway, which project to intermediate structures rather than directly to the basal ganglia output structures⁸³. These direct and indirect pathways form the basis of major theories about striatal function: the direct pathway facilitates certain behaviors by disinhibiting their output structures such as the thalamus and the substantia nigra pars reticulata, while the indirect pathway inhibits certain behaviors by inhibiting these output structures, though this model has faced revisions in recent years⁸⁴. The striatum receives the densest dopamine input in the brain and acts as a junction for various sensory and reward processing regions⁸⁵⁻⁸⁷. These characteristics

underscore the striatum's pivotal role in the processes of learning and the selection of actions, which are informed by sensory/contextual cues and the history of reinforcement^{88–90}.

The ventral striatum (VS) receives inputs from dopaminergic neurons in the midbrain, which are crucial for the reward processing, motivation, and reinforcement learning^{18,91–95}. It comprises of two major components, Olfactory Tubercle and Nucleus Accumbens. It is widely believed that dopamine neurons in the ventral tegmental area (VTA) convey RPE signals to the VS for value learning. However, more evidence has shown that VS-projecting dopamine neurons are diverse and exhibit heterogeneous responses to rewarding or aversive events. Some old voltammetry studies on dopamine inputs to the ventral striatum found significant distinctions within this structure. For instance, Aragona et al. found that cocaine sensitization of dopamine signals was evident in the shell of the NAc during drug exposure, whereas the cue-related dopamine signal was more pronounced in the core^{96–98}. Recent studies have found that the dopamine terminals in the ventral medial shell of NAc are activated by unexpected aversive outcomes and to cues that predict them, whereas dopamine terminals in other NAc subregions are mainly responsive to reward and reward-predicting cues¹⁴. These observations suggest that the nature of the dopamine signal within the ventral striatum's subregions is not homogenous, pointing to a more intricate interplay within the striatum than previously recognized. In Chapter I of my dissertation, I monitored dopaminergic axonal activities in response to cues predicting rewards and the rewards themselves in multiple subregions in the VS using multi-fiber photometry. This was aimed at exploring the potential heterogeneity within the subregions in the VS-projecting dopamine neurons.

Olfactory Tubercle: anatomical structure and function in reward processing

The olfactory tubercle (OT), a long and tubular structure within the VS, is increasingly recognized for its involvement in the convergence of smell and reward pathways^{15,25,99}. This unique positioning of the OT, receiving extensive projections from both the olfactory areas and midbrain dopaminergic neurons^{15,20,100}, underscores its potential role in associative learning.

OT belongs to the olfactory system. The evolutionary importance of the olfactory system is profound. The olfactory system is potentially the earliest sensory modality to evolve. Unlike the more complex pathways of visual, auditory, and somatosensory systems, the olfactory pathways have remained relatively conserved^{87,101}. The OT is an example of this evolutionary conservation, as it continues to be a direct recipient of olfactory input, particularly from tufted cells of the olfactory bulb¹⁰². In addition, as an important part of the ventral striatal reward circuitry, OT shares cellular features with other regions of the striatum^{24,103}, which primarily comprises of D1- or D2-type dopamine receptors expressing spiny projection neurons and heavily innervated by projections from dopaminergic neuron in ventral tegmental areas^{18,24,103,104}. Therefore, anatomically, the OT could serve as an ideal system to study the formation and modulation of stimulus-outcome associations due to the uniquely direct interface of sensory information and reward signaling^{18,31,33,102,104,105}.

Functionally, the OT is thought to be crucial in transforming olfactory information into behaviorally relevant neural codes. Research has indicated that the firing rates of OT neurons correlate with the valence of conditioned odors^{21,30}, suggesting their potential role in guiding valence-dependent behaviors. The OT has been highly implicated in motivation and reward processing. Administering cocaine directly into the OT can induce rewarding effects that surpass those seen in regions traditionally associated with reward, like the NAc and the ventral pallidum^{19,95}. Colleagues in the Murthy lab have identified that odor responses in the OT are modulated by the prediction of reward or punishment regardless of odor identity^{22,23}, and many literatures reported that neurons in OT encode odor valence^{16,23,28,105,106}. Furthermore, the odor responses are highly plastic¹⁰⁵, allowing the valence representation to be rapidly updated in different circumstances.

Despite evidence suggesting that OT neurons encode odor valence and demonstrate plasticity in their responses, the precise mechanisms by which these neurons encode odor-reward contingency or pairing are not yet fully understood. This gap in knowledge formed the basis of the current study in Chapter 2, which aimed to investigate the role of OT neurons in associative learning, particularly during the contingency degradation process. The hypothesis driving this research is that OT neurons are sensitive to

reward contingency and will bidirectionally alter their responses to reward-predicting odor cues following contingency degradation. In addition, as D1- and D2-type SPNs have been found to differ in many aspects (e.g., ligand affinity, signaling pathways, projection patterns, and neural plasticity rules), it is important to understand the in vivo dynamics of how these two distinct neural populations exhibit plastic changes in contingency learning tasks.

Summary

In **Chapter I**, I examined the behavior and dopamine signals in the VS in mice performing Pavlovian conditioning tasks while manipulating stimulus-outcome contingencies. I will show that TD learning models can comprehensively explain dopamine signals. Furthermore, I will demonstrate that dopamine signals primarily reflected prospective stimulus-outcome relationships and strongly violated predictions of the recently proposed ANCCR model. Our results advanced our understanding of fundamental processes underlying associative learning and dopamine signals in the ventral striatum.

In **Chapter II**, I employed two-photon microscopy to monitor how OT D1- and D2-type spiny neurons' activities evolve in reward-based learning, in particular, during Pavlovian contingency degradation tasks used in Chapter I. This study aimed to better understand the OT's role in sensory valence encoding during contingency learning. By investigating the neural activity of OT neurons in a contingency degradation task, this research seeks to uncover the intricate mechanisms through which the OT contributes to the encoding and updating of odor-reward associations.

Chapter I. Dopaminergic responses are sensitive to stimulus-outcome contingencies: computational accounts based on temporal difference learning models

Introduction

The ability to discern predictive relationships between different events is crucial for adaptive behaviors. Early investigations into animal learning revealed that mere contiguity between two events (“pairing”) is insufficient for establishing enduring associations. To understand this, consider Pavlovian conditioning, where an initially neutral cue (conditioned stimulus, CS) is paired with an outcome (unconditioned stimulus, US), such as an electrical shock. Through repeated pairings, animals learn to anticipate the

outcome in response to the presentation of just the CS, leading to heightened conditioned responses (e.g., freezing). Now, consider a scenario where the same number of pairings takes place, yet additional shocks occur in the absence of the CS, such that shocks happen with equal likelihood whether or not the CS is present. In such conditions, animals fail to display conditioned responses^{55,107}. Moreover, when a CS predicts a decrease in the likelihood of the US, conditioned responses are reduced. Based on these experiments, Rescorla postulated that conditioning depends not on the contiguity between the CS and the US but rather on *contingency* – the degree to which the CS indicates an increase or decrease in the likelihood of the US occurring.

Contingency indicates conditional relationships between different events and is thought to be an important quantity not only in conditioning, but also in causal inference in statistics and artificial intelligence. What is a good measure of contingency, however, remains to be clarified^{43–46}. One commonly adopted definition in psychology and causal inference is ΔP , the difference in the probability of one event occurring in the presence or absence of another^{6,47–49}. In Pavlovian settings with trial-like structures, such as the present study, ΔP can be expressed as $\Delta P = P(US|CS+) - P(US|CS-)$, where 'CS+' and 'CS-' signify the presence and absence of a CS, respectively. While mere association does not inherently imply causality, these associations can give rise to perceived causal relationships, and it has been shown that the contingency (ΔP) correlates with its strength^{45,49,108}. Although ΔP provides a simple definition, its application necessitates a trial-like structure or defined time intervals within which the probabilities of events such as CS and US can be determined^{46,109,110}. Likewise, some behavioral observations cannot be explained by ΔP , leading some to argue against the usefulness of contingency in explaining behavior⁵². As a result, efforts have been made to better define contingency^{43–46}.

Following Rescorla's experiments discussed above, further experiments highlighted the crucial role of surprise in the establishment of associations. To account for this, Rescorla and Wagner (1972) postulated that conditioning is driven by the discrepancy between the actual and predicted outcome (prediction error)³⁷. Importantly, this contiguity-based model can explain the contingency degradation experiments described above, assuming that the context acts as another CS, which competes with the primary CS³⁷. While this

“cue-competition” account is attractive, and potentially replaces the classic contingency-based account, the validity of the cue-competition model remains contested^{37,65,111–113}.

Like ΔP , the Rescorla-Wagner model also assumes a trial-based structure, as it does not consider the timing of events either within or outside a trial. To address this limitation, Sutton and Barto developed the temporal difference (TD) learning algorithm, now a fundamental algorithm in reinforcement learning (RL)^{13,114}, as a prediction error-based model of associative learning⁶⁹. TD learning as a model of associative learning in animals finds support in the striking resemblance observed between the activity of midbrain dopamine neurons and the prediction error (TD error) used in TD learning algorithms^{80,115–117}.

Despite the success of TD learning models in accounting for both associative learning and dopamine signals^{115,118}, TD models have received various challenges from alternative models. For instance, a recent study⁸¹ proposed an alternative model for associative learning and dopamine, called an adjusted net contingency for causal relations (ANCCR) model. As the name implies, the ANCCR model posits contingency as a key driver of associative learning and causal inference. Conventional definitions of contingency as well as TD learning models rely on “prospective” predictive relationships between cues and outcomes, i.e. $P(US|CS)$. By contrast in the ANCCR model, learning is driven by “retrospective” relationships, that is the probability of a stimulus (CS) given the outcome (US), or $P(CS|US)$. The authors argued that ANCCR implements causal inference, and that dopamine signals convey a signal for causal learning (the “adjusted net contingency”), not TD errors. Evidence supporting these ideas came from their experiments in examining dopamine signals in mice⁸¹ and rats¹¹⁹ during Pavlovian tasks in which contingency was manipulated. The validity of ANCCR, as well as interpretations of the data presented in these studies, await further examination.

The concept of contingency lies at the heart of learning predictive relationships. Recent work^{81,120} has raised the novel question of whether associations are learned looking forward (prospectively) or looking backward (retrospectively), and how dopamine is involved in these processes^{46,81}. Yet how contingency affects dopamine signals and behavior, as well as how dopamine signals relate to causal inference, remains to be determined. To address these questions, we examined behavior and dopamine signals in the ventral striatum

(VS) in mice performing Pavlovian conditioning tasks while manipulating stimulus-outcome contingencies. We show that, contrary to previous claims^{81,119}, dopamine signals could be comprehensively explained by TD learning models. Furthermore, we found that dopamine signals primarily reflected prospective stimulus-outcome relationships, and strongly violated predictions of the ANCCR model. We then discuss a conceptual framework for how dopamine signals can be related to contingency and causal inference.

Results

Contingency degradation attenuates Pavlovian conditioned responding

To study the effects of contingency in a Pavlovian setting, we developed a task for head-fixed mice in which odor cues predicted a stochastic reward (Fig. 1a, b, c). All mice ($n = 29$), after being water restricted, were first trained on one reward-predicting odor (Odor A) that predicted a reward (9 μ L water) with 75% probability and one odor (Odor B) that indicated no reward. In this phase (Phase 1), Odor A trials accounted for 40% of trials, Odor B for 20%, with the remaining 40% being blank trials, in which neither odor nor reward was delivered.

In Phase 1, Odor A has positive stimulus outcome contingency, being predictive of reward (R; Fig. 1c). This can be quantified using the commonly applied ΔP definition of contingency: e.g., $\Delta P(A) = P(R|A+) - P(R|A-) = 0.75$ in Phase 1. Conversely, Odor B has a negative stimulus outcome contingency: $\Delta P(B) = P(R|B+) - P(R|B-) = -0.375$. Consistent with these contingencies, all animals developed anticipatory licking to Odor A, but not Odor B, within five training sessions (Fig. 1e).

In Phase 2, animals were split into groups (Fig. 1a). The first group ('Cond', $n = 6$) continued being trained on the identical conditioning task from Phase 1. With no change in contingency, the behavior did not significantly change in a further five sessions of training (Fig. 1d, e).

The second group ('Deg', $n = 11$) experience contingency degradation. To reduce the contingency of A, either $P(R|A+)$ can be decreased or $P(R|A-)$ can be increased. The effects of reward probability on responding and dopamine have been well-studied and explained previously. Thus, we increased $P(R|A-)$ by

introducing uncued rewards, an experimental design termed ‘contingency degradation’. From Phase 1, blank trials were replaced with ‘background water’ trials in which a reward was delivered on 75% of these trials. In this condition, $P(R|A+)$ remains unchanged at 0.75, while $P(R|A-)$ increases to 0.5 (2 out of every 3 non-odor A trials are background water trials, of which 75% are rewarded thus $P(R|A-)= 2/3 \times 0.75=0.5$). As a result, $\Delta P(A)$ is reduced to 0.25. Concomitant with this decreased contingency, the anticipatory licking to Odor A decreased across the five sessions of Phase 2 in the Deg group ($t_{11} = -4.78$, $P = 0.00074$, paired t-test). Moreover, Deg group animals increased licking during the inter-trial intervals (ITIs, $t_{11} = 3.34$, $P = 0.0074$, paired t-test), potentially reflecting an increased baseline reward expectation. Additionally, the Deg group exhibited longer latencies to initiate licking and an increase in trials where mice did not lick before water delivery in Odor A trials (Fig. 2c, d, g, h).

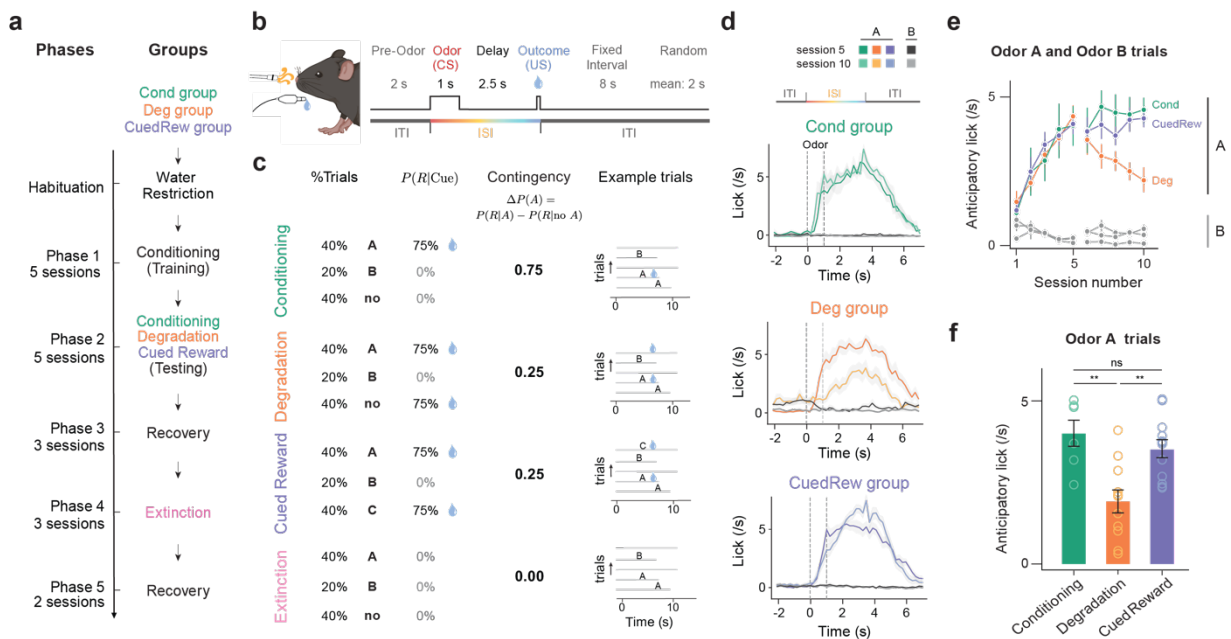


Figure 1 | Dynamic changes in lick response to olfactory cues across different phases of Pavlovian contingency learning task.

(a) Experimental design. Three groups of mice were subjected to four unique conditions of contingency learning. All animals underwent Phases 1 and 2. Deg group additionally Phases 3-5.

(b) Trial timing.

(c) Trial parameters by condition. In Conditioning, Degradation, and Cued Reward, Odor A predicts a 75% chance of reward (9 μ L water) delivery, and Odor B indicates no reward. In Degradation, blank trials were replaced with uncued rewards (75% reward probability). In Cued Reward, these additional rewards were cued by Odor C. In Extinction, no rewards were delivered.

- (d) PSTH of the average licking response of mice in three groups to the onset of Odor A and Odor B from the last session of Phase 1 (session 5) and Phase 2 (session 10). The shaded area is the standard error of the mean (SEM). Notably, the decreased licking response during ISI and increased during ITI in the Deg group. (green, Cond group, $n = 6$; orange, Deg group, $n = 11$; purple, CuedRew group, $n = 12$ mice).
- (e) Average lick rate in 3s post-cue (Odor A or B) by session. Error bars represent SEM.
- (f) Average lick rate in 3s post Odor A in the final session of each condition. Asterisks denote statistical significance: ns, $P > 0.05$; **, $P < 0.01$, Student's t -test, indicating a significant change in licking behavior to Odor A in Deg group across sessions.

The decrease in anticipatory licking, rather than reflecting the decreased contingency, could reflect satiety effects as animals in the Deg group receive twice as many rewards per session as the Cond group. We do not believe satiety explains this effect for at least two reasons: (1) all animals still received and drank about 1 ml supplementary water after each session, and (2) in all but the first degradation session, anticipatory licking was diminished compared to Cond controls in early trials (Fig. 2e). Nevertheless, a third group ('CuedRew') was included as a control for satiety effects. This group received identical rewards to the Deg group, but rather than delivering uncued rewards during the previously blank trials, these rewards were delivered following a third odor (Odor C). Unlike animals in the Deg group, animals in the CuedRew group did not decrease anticipatory licking to A. Furthermore, anticipatory licking, background licking and licking latency were similar to the Cond group (Fig. 1d, e; Fig. 2).

Using the ΔP definition of contingency, $\Delta P(A)$ is 0.25 in this condition, for identical reasoning as the Deg group. This indicates that the ΔP definition of contingency cannot be the sole determinant of conditioned responding (Fig. 1c). This phenomenon has been previously noted in the behavioral responses in conditioning tasks during contingency degradation. This is not resolved by considering a retrospective definition of contingency. Consider $\Delta P_{retro}(A) = P(A + |R) - P(A - |R)$ in both the CuedRew and Deg groups, this quantity is identical, with Odor A preceding the reward 50% of the time in both conditions.

In the subsequent stage of our investigation (Phase 3; 'Recovery 1'), we reinstated the original conditioning parameters for the Deg group, which increased the contingency back to 0.75 for Odor A, yielding an immediate recovery of the level of anticipatory licking (Fig. 2i).

To compare the behavior and neural correlates of contingency, we also introduced an Extinction phase (Phase 4) to the Deg group. In this phase, no reward was ever delivered following either odor cue.

Over three sessions, anticipatory licking to Odor A gradually waned. Finally, during a second recovery phase (Phase 5; Recovery 2), the anticipatory response to Odor A was effectively reinstated (Figure. 2i).

Notably, apart from the Extinction phase, the probability of a reward following Odor A was constant at $P(R|A) = 0.75$ throughout the experiment while behavior changes notably. Clearly, the contrast against the probability of reward in the absence of cue is an important consideration with the marked changes during contingency degradation. However, the Cued Reward controls showed it is not as straightforward as the contrast between the absence and presence of a cue.

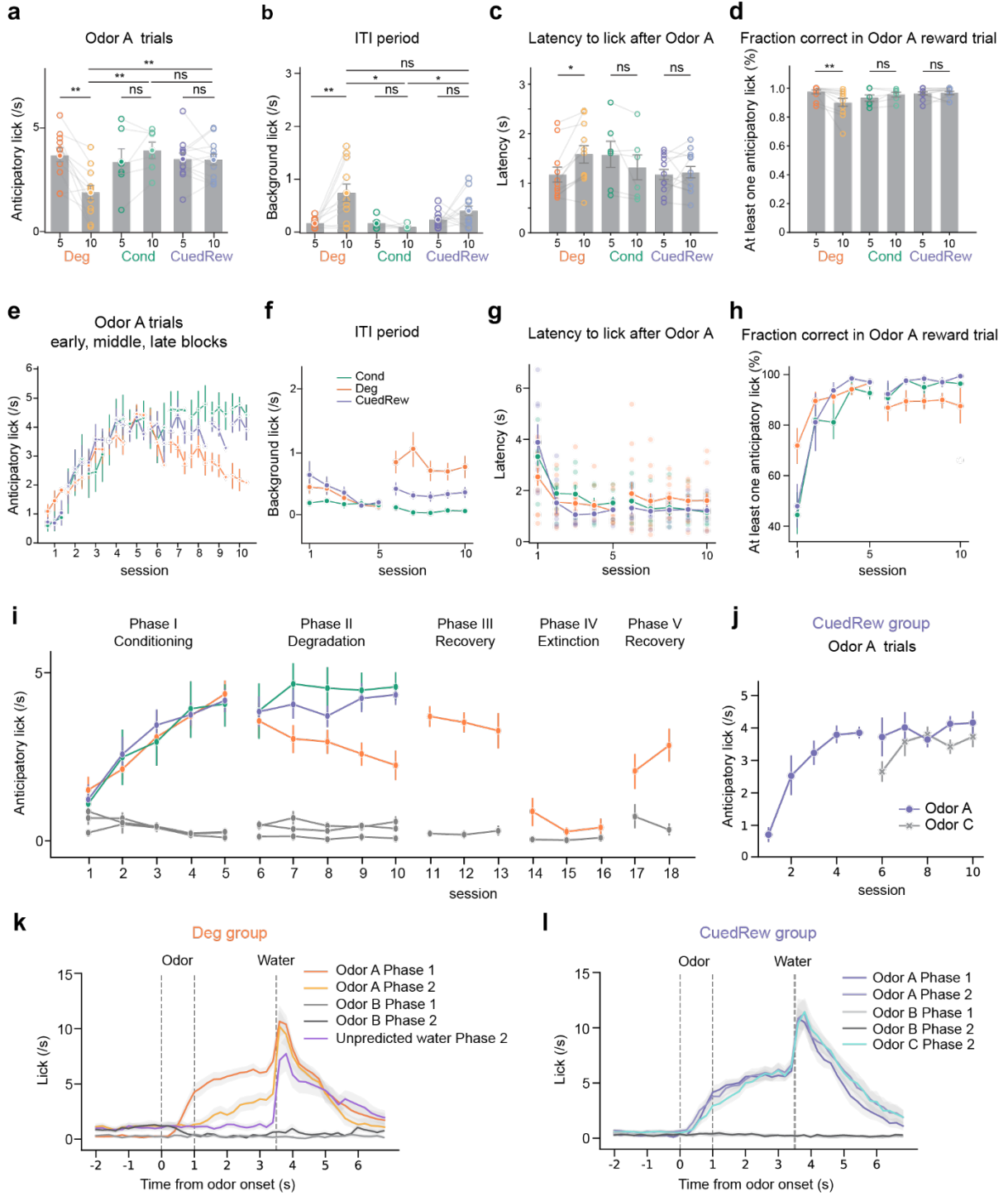


Figure 2 | Population average behavior per session

(a, b, c, d) Comparative bar graphs comparing the average number of licks to Odor A during the first 3s post-stimulus (a) and during ITI (b), latency to lick (c), and fraction correct (d) in the final sessions of phase 1 and phase 2 for Deg, Cond, and CuedRew groups. Error bars represent SEM. Asterisks denote statistical significance: ns $p > 0.05$, ** $p < 0.01$, paired Student's t-test

(e) Session-wise variation in anticipatory licking for Odor A trials, broken down into early, middle, and late blocks, for all groups.

(f, g, h) Line graphs showing the average number of licks to Odor A (colored) during ITI (g), latency to lick after Odor A and fraction correct in Odor A trials for each session in the Conditioning, Degradation, and Cued Reward phase (Deg group – orange, $n = 11$; Cond group – green, $n = 6$; CuedRew – purple, $n=12$ mice).

(i) Anticipatory licking rate in Odor A trials (colored) and in Odor B trials (grey) across multiple phases: Conditioning (Phase I), Degradation (Phase II), Recovery (Phase III), Extinction (Phase IV), and post-Extinction Recovery (Phase V).

(j) Anticipatory licking to Odor C develops quickly compared to Odor A, potentially reflecting generalization(j)

(k, l) PSTH shows the average licking response of mice in Deg group (k) and CuedRew group (l) to the various events. The response is time-locked to the odor presentation (time 0). The shaded area indicates the standard error of the mean (SEM).

Contingency degradation attenuates dopaminergic cue responses

Given the well-documented role of dopamine in associative learning, we sought to characterize the activity of dopamine neurons in our Pavlovian contingency manipulation task. We monitored axonal calcium signals of dopamine neurons using a multi-fiber fluorometry system¹²¹ with optical fibers targeting 6 locations within the ventral striatum (VS), including the nucleus accumbens (NAc, medial and lateral) and the olfactory tubercle (OT, 4 locations; Fig. 3a, b). Recordings were made only in the Deg and CuedRew groups, with the final session of Phase 1 used as the within animal conditioning control. To ensure similar levels of calcium sensor expression across the six recording locations, we employed a transgenic approach by crossing a transgenic mouse line expressing the Cre recombinase in dopamine neurons (DAT-Cre)¹²² and a reporter line that expresses a calcium sensor GCaMP6f in a Cre-dependent manner (Ai148)¹²³. Fiber locations were verified using post-mortem histology (see Methods for exclusion criteria, Fig. 3b). All results presented in the main text are from the lateral nucleus accumbens (INAc), where TD error-like dopamine signals have been observed most consistently¹⁴, though the main findings are consistent across all locations (minimum cosine similarity to INAc: 0.92).

During Phase 1 (initial conditioning) dopamine axons in INAc initially responded strongly to water and weakly to Odor A (Fig. 3c, d). As learning progressed, the response to water gradually decreased, while the response to Odor A increased over the course of 5 sessions ($t_{13} = 4.81$, $P = 0.0004$, paired t-test, cue response first vs. last session Phase 1), broadly consistent with previous reports of odor-conditioning on stochastic rewards^{79,124}.

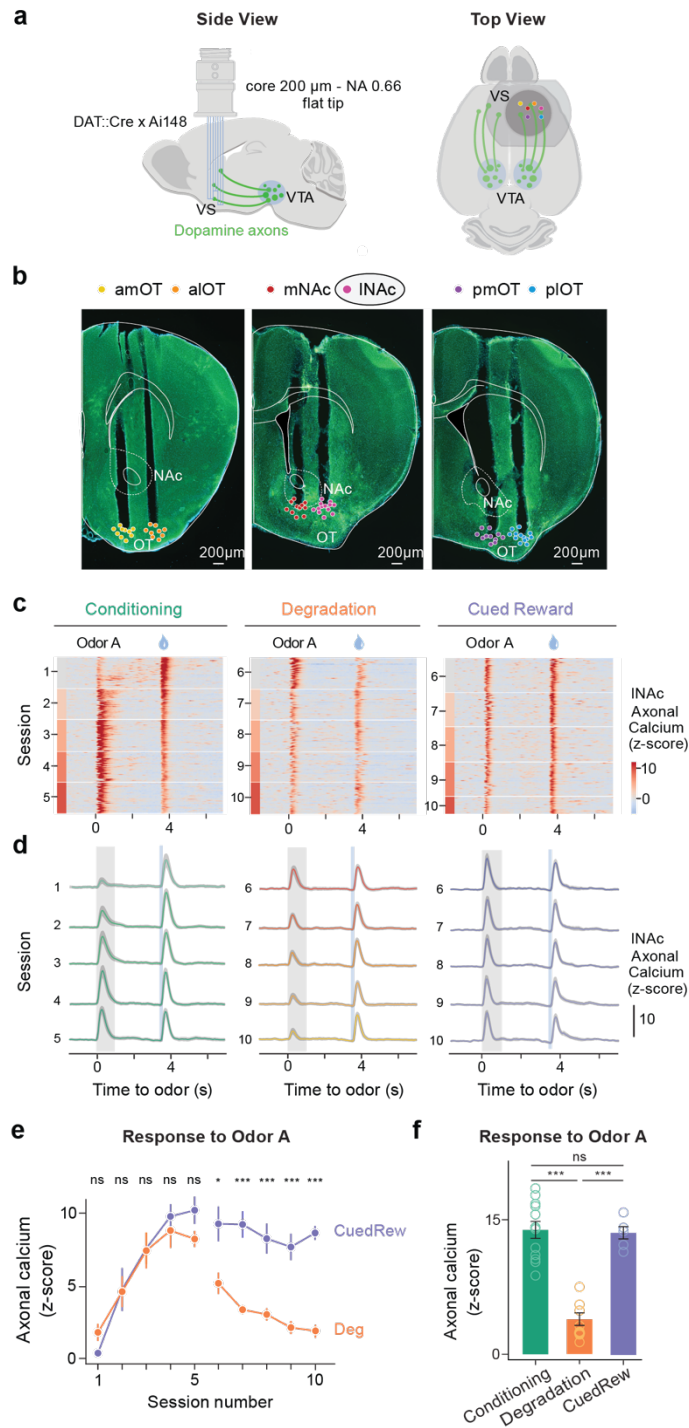


Figure 3 | Dopamine axonal activity recordings

show different responses to rewarding cues in

Degradation and Cued Reward condition

a) Configuration of multifiber photometry recordings. Coronal section from one DAT::cre x Ai148 mouse showing tracts for multiple fibers in the VS. Data recorded from INAc is used in the following analysis. INAc, Lateral nucleus accumbens; mNAc, Medial NAc; a lot, anterior lateral olfactory tubercle; plot, posterior lateral OT; amOT, anterior medial OT; pmOT, posterior medial OT.

b) Heatmap from two example mice (mouse 1, left two panels, mouse 2, right panel) illustrating the z-scored dopamine axonal signals in Odor A rewarded trials (rows), aligned to the onset of Odor A for three conditions.

c) Population average z-scored dopamine axonal signals in response to Odor A and water delivery. Shaded areas represent SEM.

d) Mean peak dopamine axonal signal (z-scored) of Odor A response by sessions for the Deg group (orange) and the CuedRew group (purple). Error bars are SEM. *, $P < 0.05$; ***, $P < 0.001$, Welch's t -test.

e) Mean peak dopamine axonal signal (z-scored) for the last session in Phase 1 (Conditioning) and 2 (Degradation and Cued Reward) for both Deg and CuedRew groups. Error bars represent SEM. ns, $P > 0.05$; ***, $P < 0.001$, Welch's t -test.

During contingency degradation (Deg group,

Phase 2), the response to Odor A decreased

across sessions ($t_8 = -11.50$, $P = 8.4 \times 10^{-6}$,

paired t -test, cue response, session 6 versus

10) consistent with the changes in

anticipatory licking and other recent reports of dopamine during contingency degradation^{6,81,119} (Fig. 3e, f).

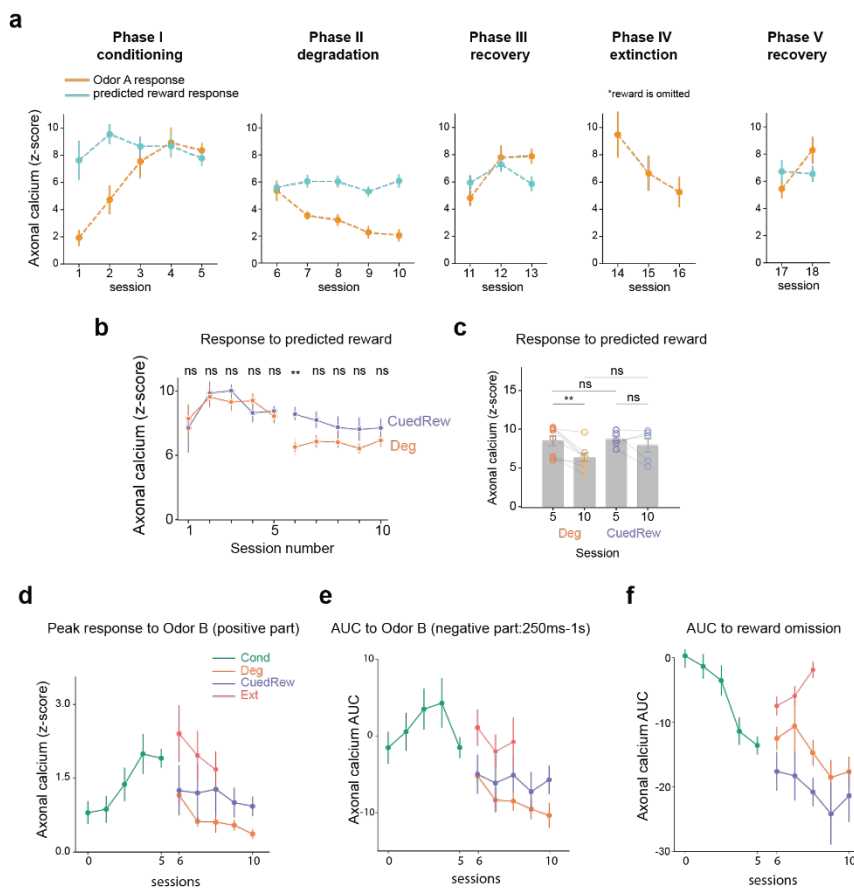
However, in the Cued Reward condition (CuedRew group, Phase 2), the response to Odor A did not

decrease compared to the Phase 1 response ($t_5 = -1.12$, $P = 0.32$, paired t -test, cue response, session 6 versus

10), aligning with the behavioral results but conflicting with the idea that dopamine neurons encode contingency, at least so far as defined by ΔP .

In the additional phases (3-5) in the Deg group, dopamine also mirrored behavior: the response to Odor A quickly recovered in Recovery 1 (Phase 3), decreased during Extinction (Phase 4) and recovered again during Recovery 2 (Phase 5; Fig. 4a). These results showed that dopamine cue responses track the stimulus-outcome contingency in our Pavlovian contingency degradation and extinction paradigms although they deviated from the contingency in the CuedRew group, yet in all groups and phases, dopamine tracked the anticipatory licking.

**Figure 4 | Population Average
Dopamine Response per session**



(a) Mean peak dopamine axonal signal (z-scored) of cue response (orange) and reward response (cyan) in Odor A rewarded trial by sessions for the Deg group across multiple phases: Conditioning (Phase I), Degradation (Phase II), Recovery (Phase III), Extinction (Phase IV), and post-Extinction Recovery (Phase V). Error bars are SEM.

(b) Mean peak dopamine axonal signal (z-scored) of reward response in Odor A trials by sessions for the Deg group (orange) and the CuedRew group (purple). Error bars are SEM. ns, $P > 0.05$; **, $P < 0.001$, Student's t -test.

(c) Mean peak dopamine axonal signal (z-scored) for the last session in Phase 1 and 2 for both Deg and CuedRew groups. Error bars represent SEM. ns, $P > 0.05$; ***, $P < 0.001$, paired t -test.

(d, e, f) Mean peak dopamine axonal signal (z-scored) across sessions for

four distinct conditions, represented for various events.

TD learning models can explain dopamine responses in contingency degradation

In both behavior and dopamine, the responses are not fully explained by contingency: there were diminished responses during contingency degradation, but not when the additional rewards are cued. Given

the match between dopamine responses and behavior, rather than consider new definitions of contingency, we sought to test if temporal difference (TD) models, which so far have been highly successful in accounting for dopamine activity, are able to explain the discrepancies from the contingency account.

Dopamine neurons are thought to convey TD errors, denoted by δ and defined by the equation: $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$, with r_t representing reward at time t , s_t representing the state at time t , $V(s_t)$ is the value at state s_t , and γ is the temporal discount factor ($0 < \gamma < 1$). Value $V(s_t)$ is defined as the expected sum of all future rewards starting from time t , with each future reward discounted by the factor γ at each time step. The role of the TD error in learning is to iteratively refine the value estimate (Fig. 5a), ultimately guiding behavior.

The response to Odor A differed most between our three test conditions (Conditioning, Degradation, Cued Reward) and thus our modeling efforts initially focused on explaining these changes. Noting there was no reward at the time of Odor A, by the definition of TD error, the response to Odor A is $\gamma V(s_{t+1}) - V(s_t)$, the difference between the value in the state immediately after Odor A (ISI) and the value in the state immediately before Odor A (pre-Odor ITI).

Previous studies have indicated whether dopamine responses as well as behaviors can be explained by TD learning models critically depends on what types of state representations are used^{125–128}. We therefore tested TD(0) learning models (Fig 5a) equipped with four different types of state representations (Fig 5c).

In the original application of TD models to dopamine activity, only the interval between a CS and a US, i.e. inter-stimulus interval (ISI), was considered and was represented using a ‘complete serial compound’ (CSC) representation, sometimes known as a tapped delay line^{10,72,129}. In this construction, a presentation of a stimulus triggers a sequential activation of sub-states $x_t = \{x_t(1), x_t(2), \dots\}$, each of which represents a time step after the stimulus (Fig. 5c). At any given time after the stimulus, only one sub-state is active. The value estimate $\hat{V}(s_t)$ is then computed as the weighted sum of these substates $\hat{V}(s_t) = \sum_j w_j x_t(j)$, which in CSC reduces to be the weight of the active substate.

While this ISI-only CSC state representation is successful in explaining many properties of the dopamine response to conditioned stimuli, it fails to predict the result of our experiments. As the ISI period is identical between conditions and there is no representation of the ITI period, the TD error for Odor A is unchanged between conditions (Fig. 5f).

An extension of this ISI-only model, is to model the ISI and ITI using CSC, resetting with each odor. We attempted this model (CSC with ITI states model), which while it predicts a decrease for Degradation, it also predicts a decrease in the Cued Reward condition (Fig. 5f), conflicting with our result. Consecutive states locked to odors make it difficult to learn as the particular state immediately before Odor A will be different in almost every trial, given the randomness introduced by the variable ITI length and the random presentation of odors.

Rather than represent the ITI with many consecutive states, it is possible to represent it a single state. This model, which we term Cue-Context model, is functionally similar to the previously developed cue competition model^{35,54}. Our Cue-Context model extends the original CSC model with a state that is constantly on (the ‘context’) during both the ISI and ITI (Fig. 5b). This model successfully predicts the pattern of experimental results we observed, with a decrease in the Odor A response during Degradation and a smaller decrease during Cued Reward (Fig. 5f). This can be understood as the context acquiring value, lesser in the Cued Reward condition because the increased value (more rewards) is attributed to both the context and the new odor compared the Degradation condition, where the increased value is attributed fully to the context. By increasing the context and thus value during the ITI, the TD error at Odor A is diminished. While this produces the qualitatively correct pattern of result, only with a temporal discount factor that is well-below reported values^{130–133} does it produce the quantitatively correct pattern (Fig. 6).

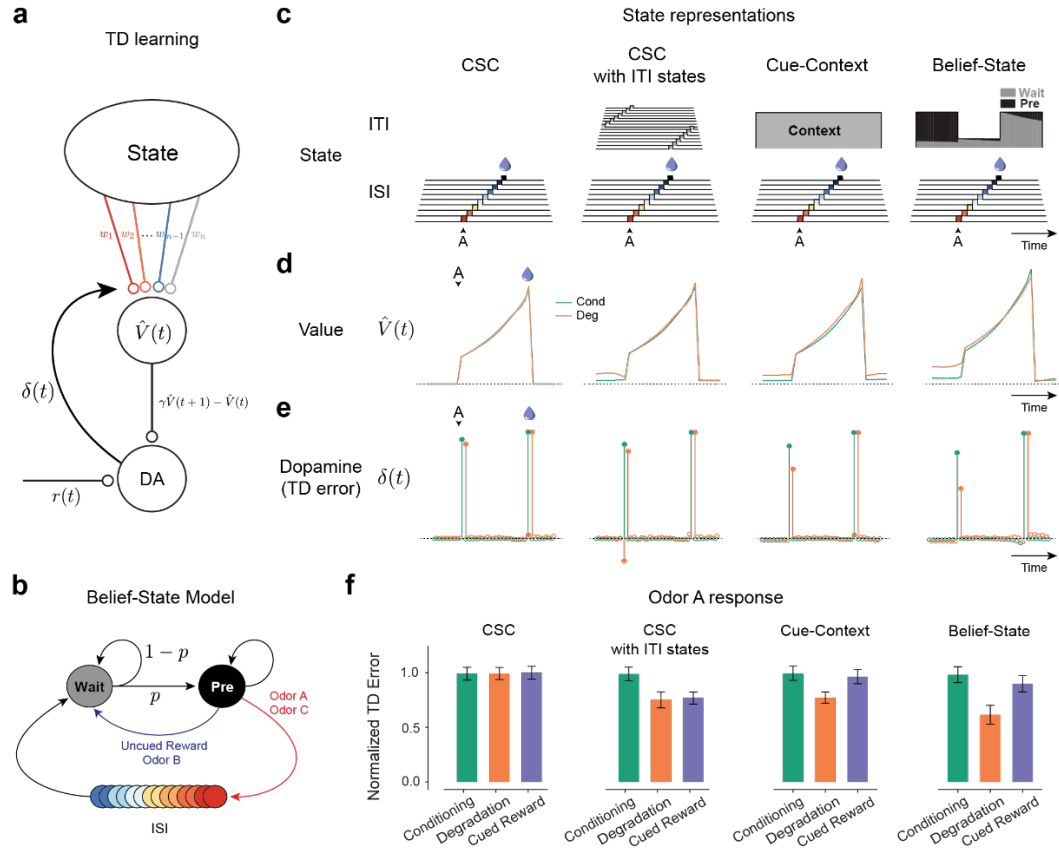


Figure 5 | TD learning models can explain dopamine responses in contingency degradation with appropriate ITI representation

- Temporal Difference Zero, TD(0), model – The state representation determines value. The difference in value between the current and gamma-discounted future state plus the reward determines the reward prediction error or dopamine. This error drives updates in the weights.
- Belief-State Model: After the ISI, the animal is in the Wait state, transitioning to the pre-transition (‘Pre’) state with fixed probability p . Animals only leave Pre state following the observation of odor or reward.
- State representations: from the left, Complete Serial Compound (CSC) with no ITI representation, CSC with ITI states, Cue-Context model, and the Belief-State model.
- Value in Odor A trials of each state representation using TD(0) for Conditioning and Degradation conditions
- TD error is the difference in value plus the reward.
- Mean normalized TD error of Odor A response from 25 simulated experiments. Error bars are SD.

We therefore considered whether further information about the experimental design could be used to refine the state representation. Inspired by previous work showing that dopamine neurons use hidden state inference in a task with stochastically timed rewards^{134,135}, we considered the ITI as being represented by the belief between two states (Belief-State TD model; Fig. 5b). These states we term, the ‘Wait’ state, which reflects early ITI, a minimum fixed amount of waiting period in which there is no chance of an

uncued reward or odor being delivered, and the ‘pre-transition’ (Pre) state, in which there is an imminent chance of reward or odor being delivered. The transition and observation matrix, which are used to compute the belief between each state, were derived from the experimental settings and assume a fixed probability of transition from the Wait to Pre state, modeling a growing anticipation of the next trial beginning. Using this state representation improved the quantitative accuracy of the model for a given γ versus the Cue-Context and accurately predicting the experimental data at our chosen γ (Fig 5f., Fig. 6), a γ which is consistent with previously reported results^{130–133}.

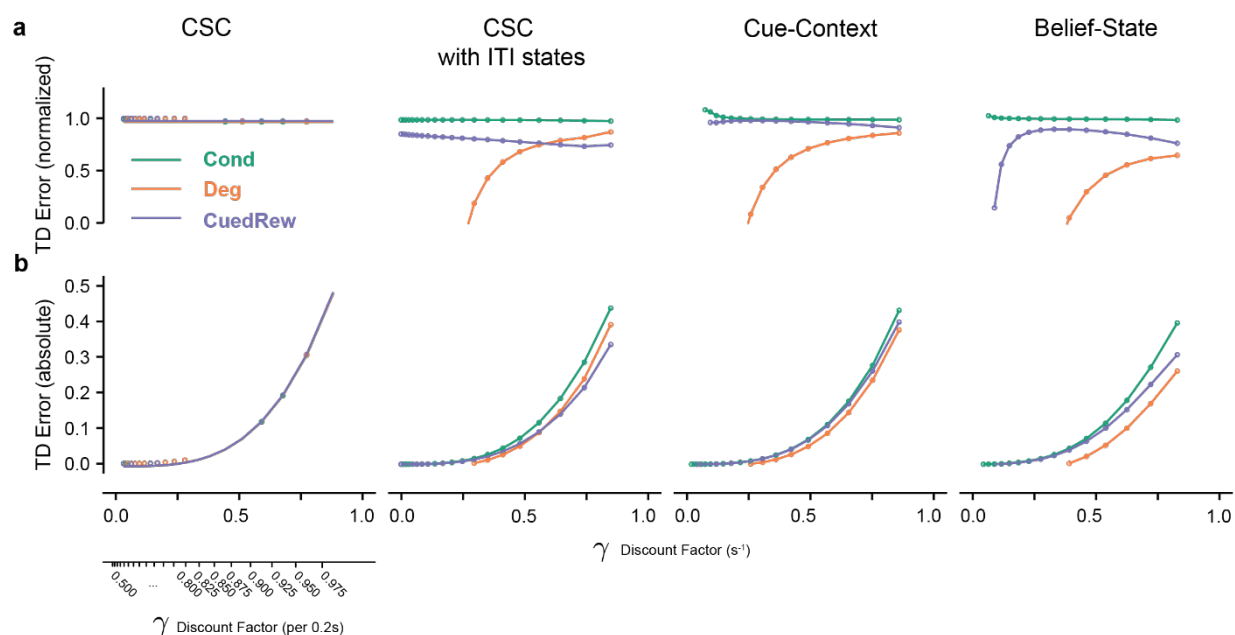


Figure 6 | Discount factor determines modeled contingency degradation effect size

Influence of discount factor (γ) on relative predicted odor A response relative to Conditioning (a) or absolute (b), where reward size = 1 for four models presented in Figure 5. The bottom right scale shows the discount factor converted to step size (0.2s); other axes use per second discount. The tested range is 0.5-0.975 discount per 0.2s in 0.025 steps.

To test which of these two models - Cue-Context and Belief-State - best describes the state representation driving dopamine responses and behavior, we focused our analysis on the ITI period: whereas the Cue-Context representation models the ITI as a single, homogenous state (the context), the

Belief-State model captures temporal heterogeneity by modeling it as the gradual transition between two states – capturing a growing anticipation of the next trial or reward (Fig. 7a).

In Pavlovian settings, anticipatory licking (as opposed to consummatory licking) has been used as a measure of current value – for example, animals will lick more to cues that predict greater rewards¹³⁶. Odor B provides an opportune moment to examine whether the ITI is a heterogeneous interval. This is because Odor B predicts no reward within the current trial (and thus no consummatory licking) but also provides further information that no odor or uncued reward will be delivered for the length of one trial. In this way, Odor B provides 100% certainty to the animal they are in the ‘Wait’ state. Consistent with this understanding of the task structure, the delivery of Odor B during the Degradation condition prompted animals to stop licking. Both the Cue-Context and Belief-State models capture this effect. The crucial difference is how the lick rate recovers. In the Cue-Context model, ITI value is related to a single state, which without reward, decreases at the rate of α (learning rate). In our Belief-State model, value continually increases (Fig. 7c) across the entire ITI, as the increased belief that the next trial is imminent increases continuously. We find the lick pattern following Odor B matches the pattern of the belief state model and not the Cue-Context model, with a sudden decrease in licking followed by a gradual increase in the Degradation condition that is unrelated to the ISI length (Fig. 7c, summarized in Fig. 7d).

This pattern of licking behavior also suggests that the animals do not develop more complex models of timing. Odor B predicts approximately ten seconds with no reward. An ideal agent would not lick during this time, waiting until the transition to an uncued reward is possible. The mice instead begin resume licking within 2-3 seconds of Odor B delivery, with the lick rate increasing over several seconds.

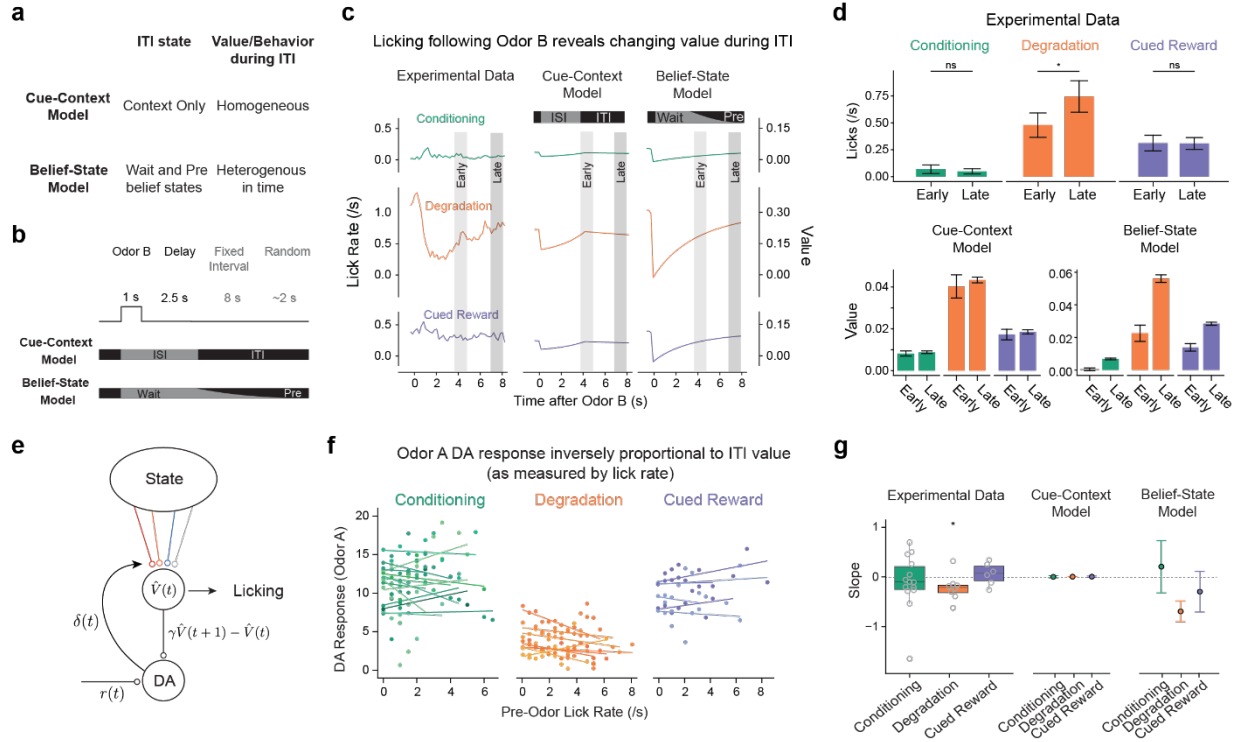


Figure 7 | Belief-State model, but not Cue-Context model, explains variance in behavior and dopamine responses

- Cue-Context model and Belief-State model differ in their representation of the ITI.
- Odor B predicts no reward and at least 10 s before the start of the next trial.
- Odor B induces a reduction in licking, particularly in the Degradation condition, which matches the pattern of value in the Belief-State model better than the Cue-Context model.
- Quantified licks (top) from experimental data in early (3.5-5s) and late (7-8s) post cue period. Error bars are SEM, *, $P < 0.05$, paired t -test. Value from Cue-Context and Belief-State model for the same time period, error bars are SD.
- If licking is taken as a readout of value, then ITI licking should be inversely correlated with dopamine.
- Per animal linear regression of Odor A dopamine response (z-score axonal calcium) on lick rate in 2s before cue delivery.
- Summarized slope coefficients from experimental data (left) and models (right). Boxplot shows median and IQR, one sample t -test.

While the value from the model explains the time course of licking following Odor B, this account does not, by itself, explain the decrease in anticipatory licking in response to Odor A (Fig. 1d). This decreased responding is a consistent feature of contingency degradation⁶. It however can be explained if considered a problem of optimizing a fixed amount of effort over a session. If each lick carries a small, fixed effort cost, a rational agent will lick proportional to the total amount of rewards. Indeed, the number of anticipatory licks (all licks excluding the 2 seconds immediately after rewards) are roughly proportional

to the number of rewards, being approximately double in the CuedRew and Deg conditions (Fig. 8a). A reasonable way to allocate these fixed licks is using the relative value at a given time point. Thus, the increased ITI value and fixed ISI value can explain the reallocation of licking from the ISI to the ITI in the Degradation condition and thus the decreased responding (Fig. 8b-d).

Having shown the lick rate is explained by the changing value in the Belief-State model, we wished to test whether this could be used to explain trial-by-trial variance in the dopamine response. Continuing with the assumption that licking is a moment-by-moment measure of value, our Belief-State model predicts there should be an inverse correlation between the pre-odor lick rate and the Odor A dopamine response. To test this, we correlated the number of licks in the two seconds before the cue to the Odor A response on a trial-by-trial basis, regressing a linear model independently for each mouse pooling the last two sessions of each condition under the assumption that the task was well-learned in these sessions.

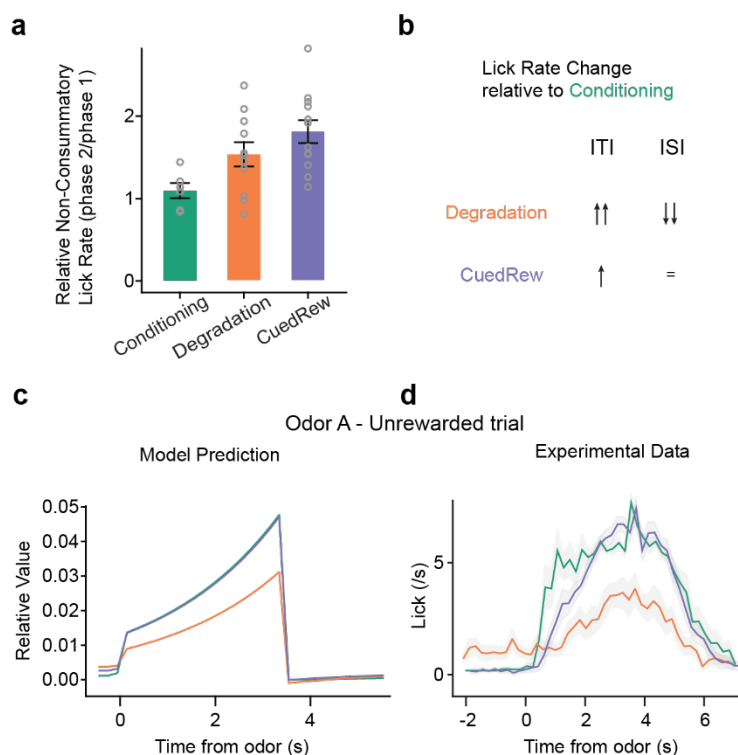


Figure 8 | Relative value explains decreased anticipatory licking during ISI during contingency degradation

(a) Mean non-consummatory lick rate normalized to the Conditioning phase, suggesting that the Degradation and Cued Reward conditions elicit approximately twice the lick rate of the Conditioning condition, and thus proportional to the total reward quantity. Consummatory licks were considered any licks occurring in the 2 seconds following reward delivery.

(b) Summary of lick rate changes relative to the Conditioning phase during the inter-trial interval (ITI) and the inter-stimulus interval (ISI).

(c) Average relative value (current value/session total value, scaled by total reward) during odor A trial.

(d) Experimental data showing the actual lick rates recorded during Odor A

unrewarded trials, compared over time, which aligns with the assumptions and predictions made in a, b, and c.

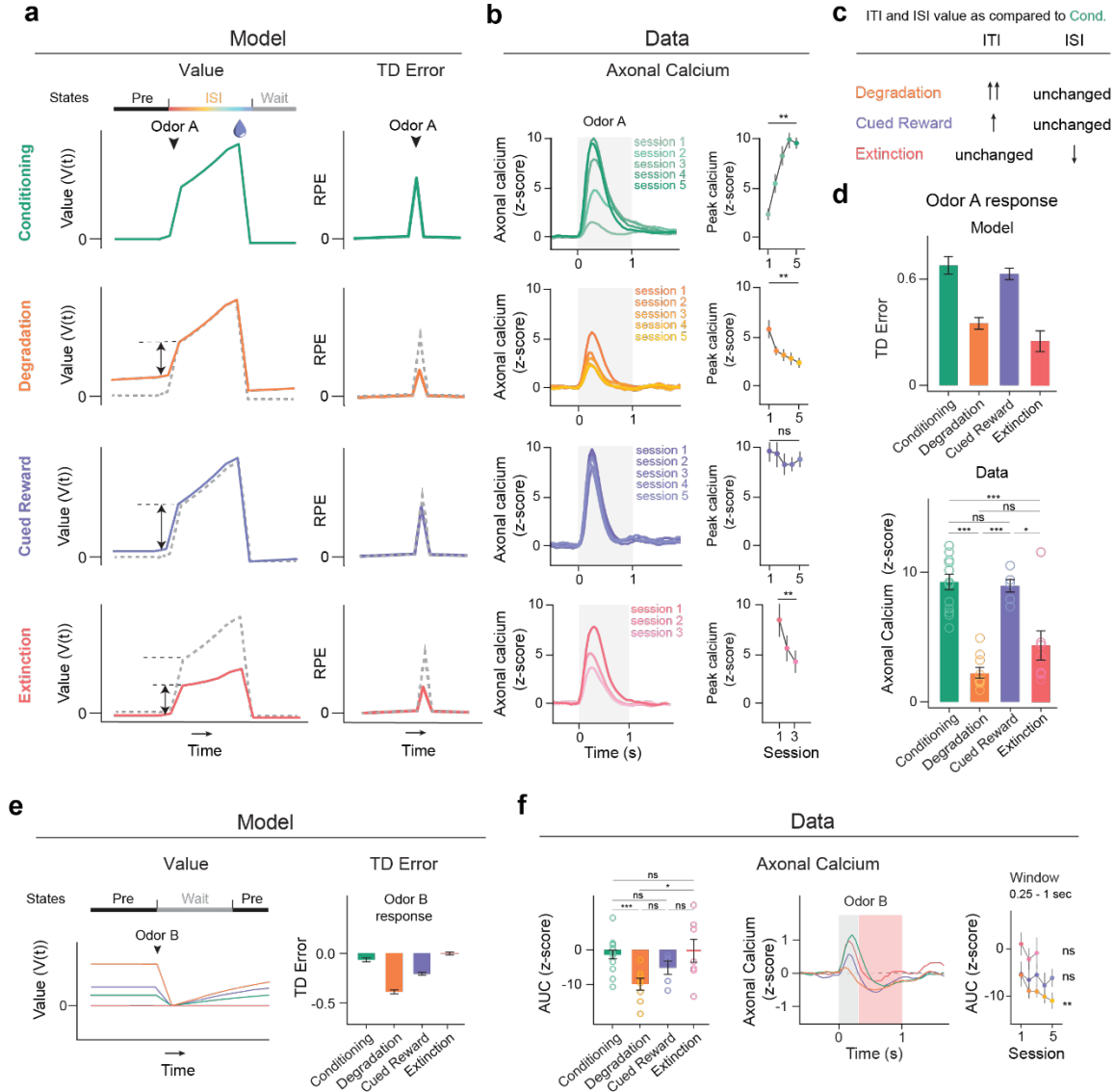
Only in the Degradation condition was there a significant negative correlation between the lick-rate and the Odor A response for the population (Fig. 7f, g). This can be explained by the Belief-State model, as ITI value varies depending on the length of the ITI but cannot be explained by the Cue-Context model, in which ITI value is fixed (Fig. 7g). The modeling suggests that the lack of a significant trend in the remaining two conditions is due to the lower variance in value in the pre-cue period, with an average of 0.28 ± 0.87 and 0.46 ± 1.11 licks (mean $\pm$ s.d.) in the 2 second pre cue period in Cond and CuedRew respectively (versus 1.51 ± 1.52 in Deg), leading to underpowered analysis.

In summary, the ITI state representation is essential to explaining the relative effects of contingency degradation and additional cued rewards on the Odor A response. Complex ITI representations, such as CSC, are inefficient, whereas modeling it as a single state (Cue-Context), does not capture the heterogeneity of the ITI. Our Belief-State model, representing the ITI using two states, is sufficient to explain the experimental results. Intuitively, by capturing more information about the task, this model allows for more efficient learning than the considered alternatives.

Other aspects of dopamine responses and model predictions

Having identified a sufficient model, we next examined how well this model matched the other experimental results. Figure 9a visualizes the value as predicted by the Belief-State model across four conditions tested (Conditioning, including Recovery; Degradation, Cued Reward and Extinction). In the Odor A rewarded trial, the value during the ISI remained unchanged in the first three conditions, significantly decreased in Extinction, closely mirroring the (prospective) reward probability ($P(R|A)$). For the reasons discussed above, the pre-ISI period, reflecting the pre-transition state ('Pre'), showed a modest increase in the Cued Reward case and a significant rise in the Degradation condition. The TD errors upon Odor A presentation, reflective of the difference in value between Pre state and the first ISI substate, diminished in both Degradation and Extinction. In both these conditions, contingency is reduced by increasing $P(R|A-)$ and decreasing $P(R|A)$, respectively¹³⁷. Notably, our model suggested two distinct mechanisms underlying these two processes: an increase in Pre state value in Degradation and a decrease

in ISI value in Extinction (Fig. 9c). Our Belief-State TD learning model well-matched the experimental results (Fig. 9b, d), including the previously Extinction, which we did not examine when constructing the model.



- Plots averaged from one representative simulation of Odor A rewarded trial ($n = 4,000$ simulated trials) for four distinct conditions using the Belief-State model. Graphs are for the corresponding value function (left) and TD error (right) of cue response for Odor A rewarded trials.
- Signals from dopamine axons (mean) across multiple sessions of each condition (left). Mean peak dopamine axonal calcium signal (z-scored) for the first to last session in Phase 2 for four contingency conditions (right). Error bars represent SEM. ns, $P > 0.05$; **, $P < 0.01$, Student's paired t -test. The Belief-State model captures the modulation of Odor A dopamine response in all conditions.
- Degradation, Cued Reward and Extinction conditions differ in how their ITI and ISI values change compared to the Conditioning phase.

- d) Mean peak TD error by Belief-State model and dopamine axonal signal (z-scored) to Odor A for four distinct conditions. Error bars represent SEM. ns, $P > 0.05$; *, $P < 0.05$; ***, $P < 0.001$, Welch's t -test. The model's prediction captured the pattern in the dopamine data well.
- e) Averaged traces from a representative simulation of the Odor B trial ($n = 4,000$ simulated trials) across four distinct conditions using the Belief-State model. Graphs are for the value function and TD errors of cue response for Odor B trials.
- f) Z-scored dopamine axonal signals to Odor B quantified from the red-shaded area to quantify the later response only. The bar graph (left) shows the mean z-scored Odor B AUC from 0.25s-1s response from the last session of each condition. Error bars are SEM. * $P < 0.05$; ***, $P < 0.001$, Welch's t -test. The line graph (right) shows the mean z-scored AUC over multiple sessions for each condition. Statistical analysis was performed on data from the first and last sessions of these conditions. Error bars are SEM.

We next examined how well this model matched all other experimental results. As the model predicted, the other distinct difference between degradation and extinction will be that degradation affects TD error for all cues due to changes in the Pre state value, while extinction impacts only the specific cue undergoing extinction. According to this prediction, we examined the Odor B trials. In the Belief-State model, Odor B is a transition from the Pre to Wait state and thus the TD error is the difference between these two state values. Therefore, we anticipated the most negative response in the Deg group, owing to a higher Pre state value, and relatively unchanged 'Wait' value. We also expected an unchanged response in Extinction in comparison to Conditioning. Experimentally, the response to Odor B was biphasic, featuring an initial positive response followed by a later negative response. Such a biphasic response has been previously noted, with general agreement that the second phase is correlated with value¹³⁸. By quantifying just the later response (250ms-1s window), there was a close match between the model and the data for Odor B response (Fig. 9e, f).

The Belief-State model shows that TD errors at reward omission are based on the difference between the final ISI substate and Wait state values. The Wait state value, generally lower than the Pre state value, has minor changes across conditions. This results in consistent TD errors at reward omission across Conditioning, Degradation, and Cued Reward conditions due to similar ISI values, but a significant reduction in Extinction due to a lower ISI value, which highly aligned with experimental results (Fig. 10). TD errors at predicted rewards, reflecting the difference between actual reward and ISI values, exhibit minimal changes across Conditioning, Degradation and Cued Reward conditions, which is also consistent

with data. We show that the model effectively predicts the patterns in reward omission and anticipated rewards.

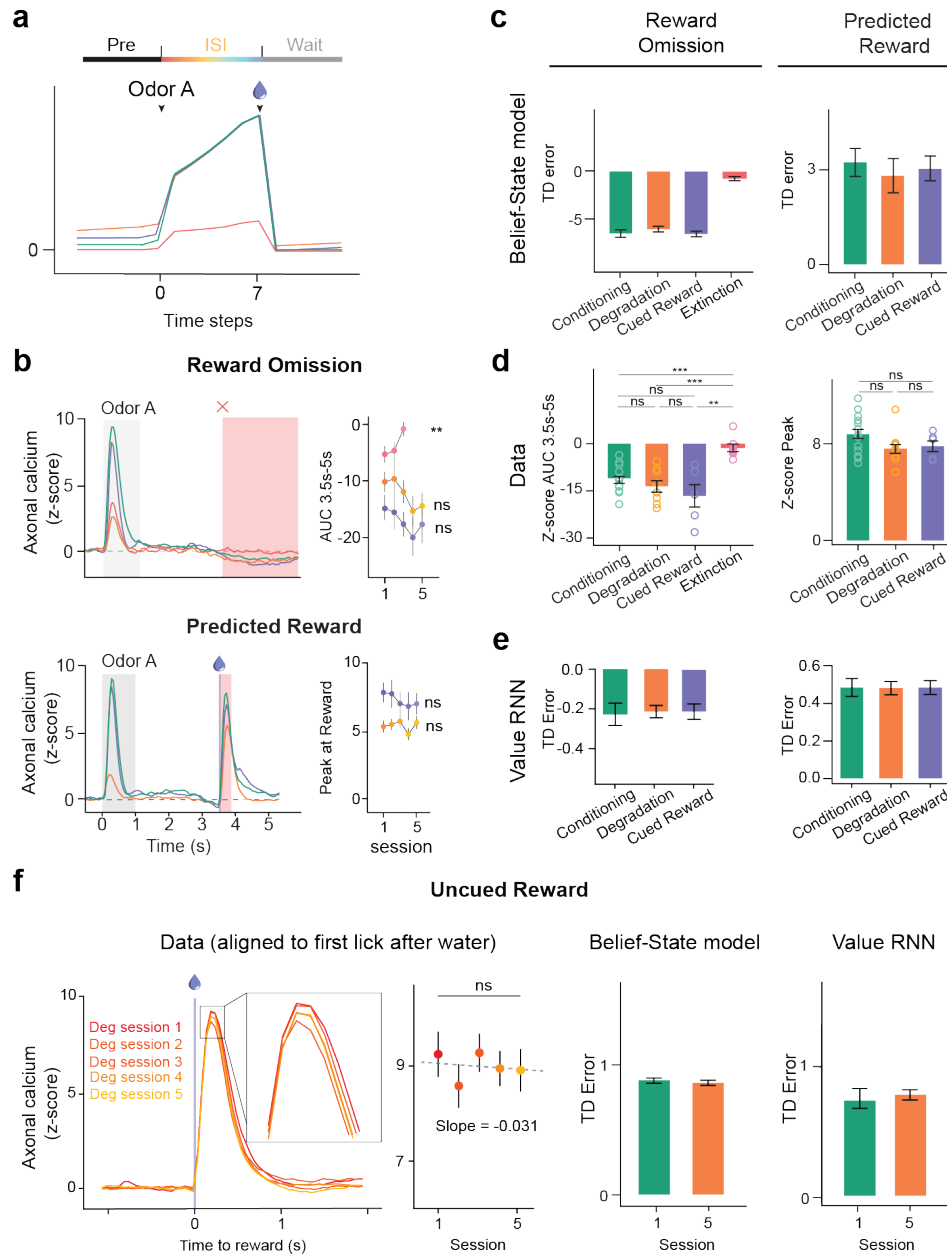


Figure 10 | Comparison of reward and omission responses between experimental data, Belief-State model and value RNN predictions

- Plots averaged from one representative simulation of Odor A rewarded trial ($n = 4,000$ simulated trials) for four distinct conditions using the Belief-State model. Graphs are for the corresponding value function of Odor A rewarded trials, with Pre state, ISI state and Wait state annotated.
- Z-scored DA axonal signals to reward omission and predicted reward following Odor A quantified from the red shaded area. Line graphs (right) show the mean z-scored response over multiple sessions for each

condition. Statistical analysis was performed on data from the first and last sessions of these conditions. Error bars are SEM. ns, $P > 0.05$; **, $P < 0.01$, paired t -test.

- (c) The predictions of the Belief-State model for reward omission and predicted reward (mean, error bars: SD).
- (d) The experimental data for reward omission and predicted reward (mean, error bars: SEM). ns, $P > 0.05$; **, $P < 0.01$; ***, $P < 0.001$, Welch's t -test.
- (e) The predictions of the Value RNN models for reward omission and predicted reward (mean, error bars: SD).
- (f) The experimental data, TD error prediction by Belief-State model and Value RNN model for uncued reward response in Degradation. While the Belief-State model captured the downward trend in response magnitude, none of the three statistical tests showed significant changes.

In total, the above results indicate that the TD model with proper task states can effectively recapitulate nearly all aspects of phasic dopamine responses across various trial types and task events.

Recurrent neural networks that learn to predict values through TD learning can explain dopamine responses

The models discussed above, while effective, are ‘hand-crafted’ and tuned to the particular task setting. While there is evidence that dopamine neurons rely on belief-state inference in computing TD error^{135,139,140}, the question of how the brain learns such a state-space is less well understood. Previous work has shown that RNNs, trained to estimate value directly from observations (‘value RNNs’), develop belief-like representations despite not being explicitly trained to do so. This approach substitutes the ‘hand-crafted’ states for an RNN that is trained using the TD errors (Fig. 11a).

Here, we applied the same value RNN to our contingency manipulation experiments. We generated training sets, consisting only of the odor and reward timings, that matched the three conditions. The RNNs were first trained on the Phase 1 Conditioning and then either on the Phase 2 contingency Degradation or Cued Reward conditions (Fig. 11b). To simulate animal’s internal timing uncertainty, the reward timing was jittered 0.5 seconds on a random selection of trials. Testing sets were then generated for each RNN without the jitter, with the value and TD-error readout of these test sets being presented here. RNNs were trained with a different number of hidden units, from 5 to 50.

Following training, RNNs closely matched the experimental results (Fig. 11c). Like the TD models used in the above section, the decrease in Odor A response is explained by an increase in the value during the ITI period, not a shift in the value during the ISI (example Fig. 11d).

We were interested in understanding the inferred state spaces used by the RNN models. To visualize this, we applied canonical correlation analysis (CCA)^{141,142} to the activity of the hidden units. Briefly, in each condition, we applied principal component analysis (PCA) to identify the principal components (PCs) that explained 80% of the variance (mean number of components = 4.26) then used CCA to project the PCAs into a single space. CCA finds linear combinations of each of the PCs that maximally correlate – allowing us to identify hidden units encoding the same information in the different RNNs. We then used the combination of PCA and CCA to create a map from hidden unit activity to a common state. A decoder trained only on this three-dimensional space explained 99.7% of the variance in value.

In all conditions, without any stimuli, the activity will end up at the fixed point. This can be understood as the ITI state. In all conditions, the Odor A trajectory is similar, indicating a shared representation of the ISI period (Fig. 11e). Furthermore, in the Cued Reward condition, the Odor C trajectory is nearly identical to that of Odor A, potentially reflecting generalization. In the Degradation condition, delivering Odor B causes a trajectory that is significantly longer than the other two conditions, potentially corresponding to the Wait state.

We were interested in comparing the state space to our Belief-State model. To do this, we calculated the beliefs at each given time point in the simulated experiment and used a linear decoder to measure how accurately these beliefs could be decoded from the hidden unit activity. As previously noted⁴¹, the unit activity became more belief-like with increasing hidden size (Fig. 11f). Notably the decoding performance was higher for the Degradation condition at each hidden size. This is explained by the better decoding of the Wait state (Fig. 11f, right panel). As evident in the visualized activity in the state spaces, the RNNs trained on the Degradation condition developed distinct trajectories in the ITI compared to the other two conditions (Fig. 11g), taking a longer period of time to return to the fixed ITI point and following a similar pattern regardless of the particular trial type. In all RNNs that predicted Degradation (Odor A Deg < Odor A Cond), the Wait state was decodable with a minimum performance of $R^2=0.57$.

This suggests that it is the delivery of rewards during the ITI that reshapes the state space to be heterogeneous, while in the other conditions this is not necessary and thus the ITI has a relatively fixed

state space representation. That the RNN can learn a belief-like representation from limited information, using only the TD error as feedback, suggests a generalized method by which the brain can construct state spaces using TD algorithms.

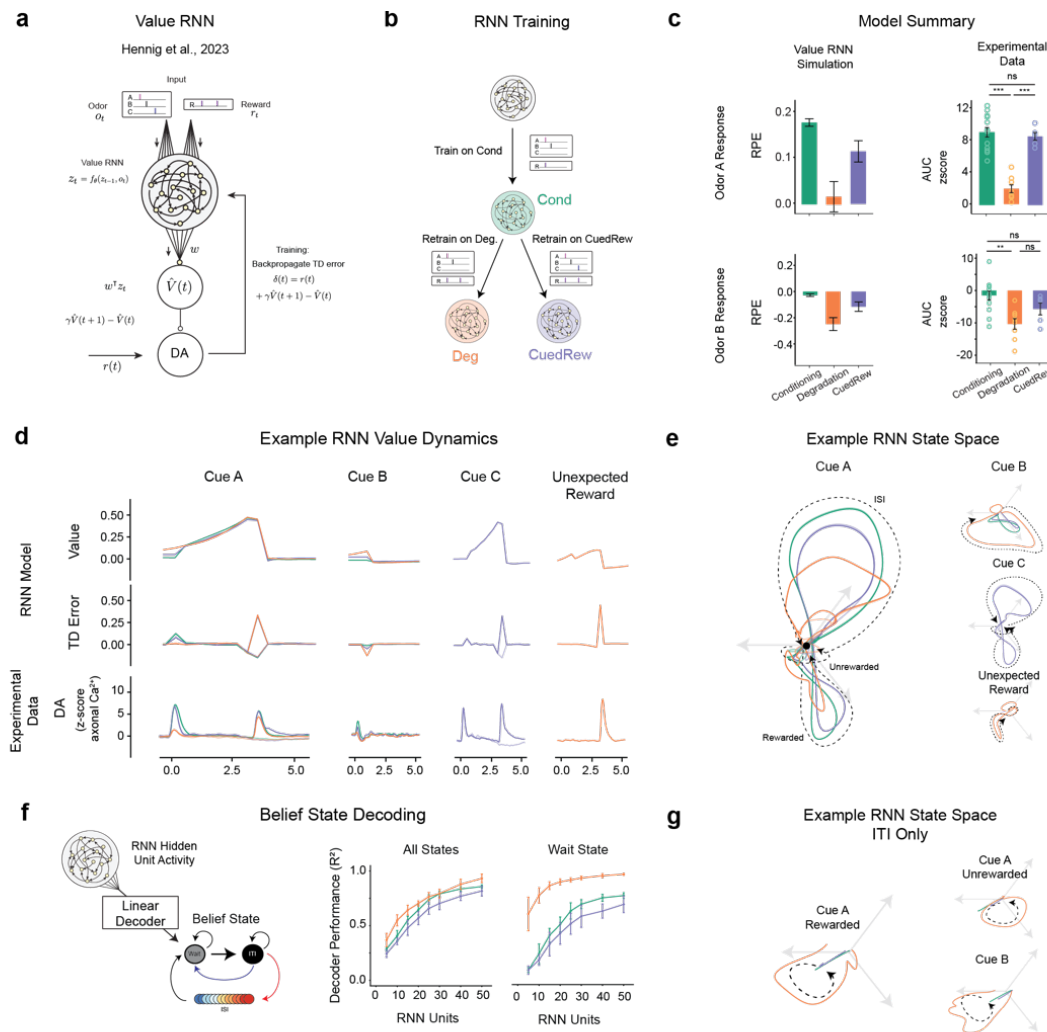


Figure 11 | Value RNNs recapitulate experimental results using state-spaces akin to hand-crafted belief state model

- The Value RNN replaces the hand-crafted state space representation with an RNN that is trained only on the observations of cues and rewards. The TD error is used to train the network.
- RNNs were initially trained on simulated Conditioning experiments, before being retrained on either Degradation or Cued Reward conditions.
- The predictions of the RNN models (mean, error bars: SD) closely match the experimental results
- Example value, TD error and corresponding average experimental data from a single RNN simulation. Notably, decreased degradation Odor A response is explained by increased value in the pre-cue period.
- Hidden neuron activity projected into 3D space using CCA from the same RNNs used in (d). The Odor A ISI representation is similar in each of the three conditions, and similar to the Odor C representation. Odor B representation is significantly changed in the Degradation condition.

- (f) Correspondence between RNN state space and Belief-State model. A linear decoder was trained to predict beliefs using RNN hidden unit activity. With increasing hidden units, the RNN becoming increasingly belief-like. The improved performance of the decoder for the Degradation condition is explained by better decoding of the Wait state. Better Wait state decoding is explained by altered ITI representation:
- (g) Same RNNs as in (d) and (e), hidden unit activity projected into state-space as (e) for the ITI period only reveals ITI representation is significantly different in the Degradation case.

A retrospective learning model, ANCCR, cannot explain the dopamine responses

While our analysis using the TD learning models with explicit state representations and the value-RNNs suggest that TD learning models are sufficient to explain our experimental results, we have not yet considered whether alternative definitions of contingency would also provide an account of our results. The ANCCR (adjusted net contingency for causal relations) is a recently described new model, proposed as an alternative account of the TD explanation of dopamine activity (Fig. 12a). The authors have previously shown that this model can account for contingency degradation⁸¹ and suggested TD accounts could not. ANCCR builds upon the authors' previous observation that the retrospective information ('which cues precede reward?') can be used to explain animal behavior previously unexplained by prospective accounts¹²⁰. Accordingly, the ANCCR model begins with the calculation of the retrospective contingency, using eligibility traces as a principled method to compute contingency in continuous time, rather from trial-by-trial probabilities. At the time of a reward (or a 'meaningful event'), the difference between the eligibility of cues at the time of reward and the average cue eligibility is computed. This generalizes the trial-based definition of $\Delta P_{\text{retro}}(A)$ to continuous time. From this retrospective contingency and the average event rates, the model proposes the prospective contingencies are inferred using a Bayes-type computation. From the prospective and retrospective contingencies, a weighted sum ('net') contingency is calculated, by learning all pairwise event contingencies, a retrospective map is established. This map can be used to calculate the change in expectation of reward for a given event, considering other explanations. It is this 'adjusted net contingency' that ANCCR proposes is represented in the dopamine signal.

To test the ANCCR model, we used the authors' published code to model the same 25 simulated experiments used in our TD modeling. For this experiment, ANCCR has 12 parameters, at least 6 of which have a significant impact on the modeled response. We first tried using the parameters published in Garr et

al. (2023) and Jeong et al. (2022), with the Garr parameters being presented, being closer to the experimental results. While the ANCCR model accurately predicted a decreased response for Odor A during contingency degradation, it predicted a similar response in the Cued Reward condition, conflicting with the experimental results (Fig. 12b). We varied the parameter (w) which controls the relative amount of retrospective and prospective information used to calculate the contingency. This parameter controlled the relative size of the decrease, while sensitive to parameter choice governing the eligibility decay rate and learning rates, in general the halving of $P(A|R)$ produces a large decrease in the retrospective contingency ($P(A|R)-P(A)$), whereas the increase in $P(R)$ slightly decreases the prospective contingency ($P(R|A)-P(R)$).

We next considered whether this was a problem of parameter selection and therefore simulated the first 5 virtual experiments for the parameter search space used in Garr et al. (2023), trying a total of 21,000 parameter combinations, including those in the two previous studies^{81,119} (indicated as 1, 2 and 3). Fig. 12c plots the Odor A dopamine response in the Degradation and Cued Reward case for each of these combinations, normalized by the response during Conditioning. No parameter combination predicted the correct pattern of experimental results, quantitatively or qualitatively (Fig. 12c).

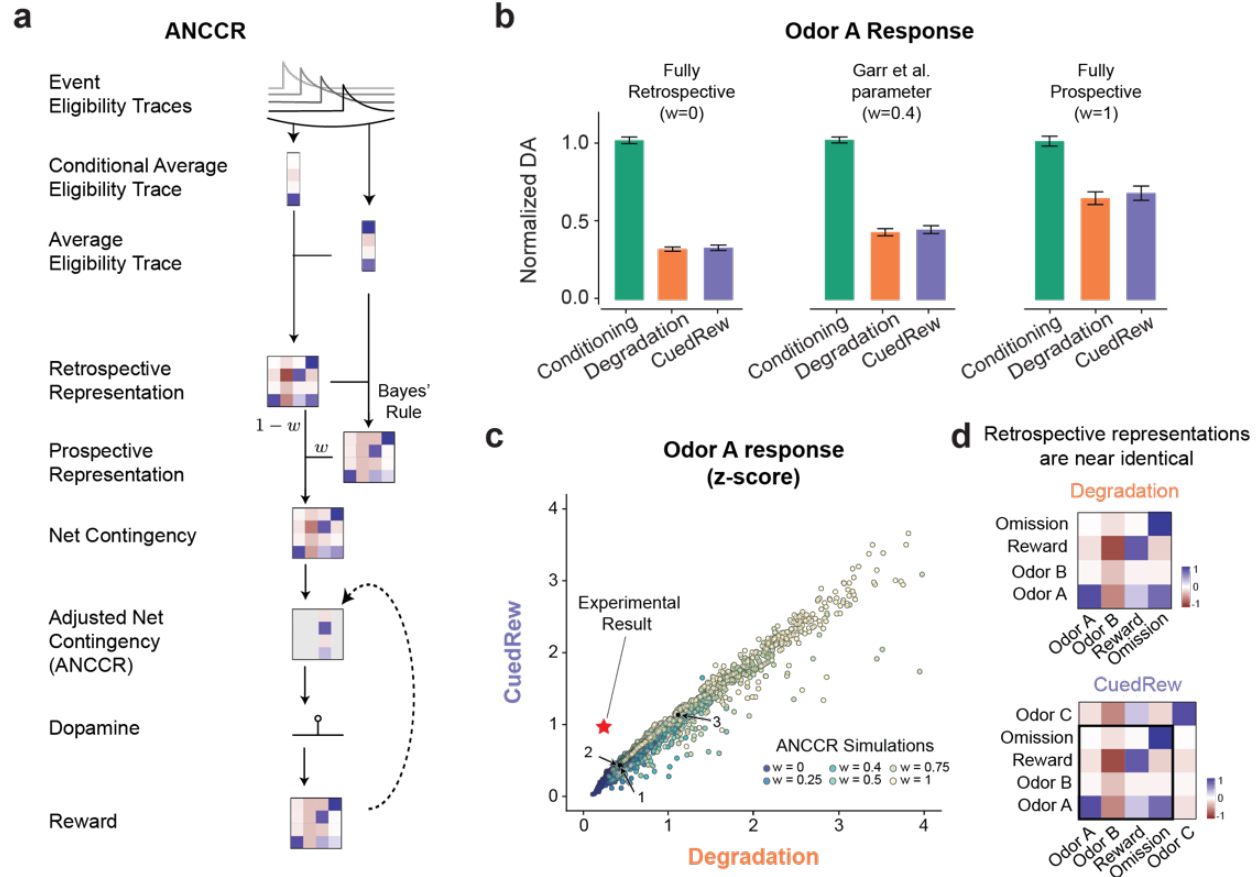


Figure 12 | ANCCR does not explain the experimental results

- Simplified representation of ANCCR model. Notably, the first step is to estimate retrospective contingency using eligibility traces.
- Simulations of the same virtual experiments used in Figure 3 using ANCCR, using the parameters in Garr et al., 2023, varying the prospective-retrospective weighting parameter (w). Error bars are SD. In all cases, the predicted odor A response is similar in the Degradation and Cued Reward conditions
- No parameter combination explains the experimental result. Searching 21,000 parameter combinations across six parameters (T ratio = 0.2-2, α = 0.01-0.3, k = 0.01-1 or $1/(\text{mean inter-reward interval})$, w = 0-1, threshold = 0.1-0.7, α_R = 0.1-0.3). Experimental result plotted as a star. Previously used parameters (Garr et al., 2023 as 1, Jeong et al., 2022 as 2 and 3) indicated. Dots are colored by the prospective-retrospective weighting parameter (w), which has a strong effect on the magnitude of Phase 2 response relative to Phase 1.
- As the contingency is calculated as the first step, and the contingencies are similar in Degradation and Cued Reward conditions, there is little difference in the retrospective contingency representation between the two conditions, explaining why, regardless of parameter choice, ANCCR predicts similar responses.

Discussion

Here we examined behaviors and VS dopamine signals in a Pavlovian contingency degradation paradigm, including a pivotal control. Our results show that dopamine cue responses, similar to behavioral conditioned responses, were attenuated when stimulus-outcome contingency was degraded by the uncued delivery of additional rewards. Crucially, neither dopamine signals nor conditioned responses were affected in a control condition in which the delivery of additional rewards were cued by a different stimulus despite a similar number of rewards being administered. This result not only demonstrated that the above results were not due to satiety, but also provided key insights into possible mechanisms underlying contingency degradation.

Contrary to the claim in previous studies, our modeling showed that aspects of dopamine responses can be comprehensively explained by TD^{81,119} learning models, if the model is equipped with proper state representations reflecting the task structure. These TD learning models also readily explained dopamine cue responses in the control condition with cued rewards – the results which strongly violated the predictions of a contingency-based retrospective causal learning model (ANCCR) or the ΔP definition of contingency. The results indicate that dopamine signals, as well as conditioned responses, primarily reflected the prospective, but not retrospective, stimulus-outcome relations. Rather than discarding the notion of contingency altogether, we propose that these results point toward a novel definition of contingency grounded in the TD learning framework. These results bear significant implications for the theory of associative learning and the nature of dopamine signals, which help resolve some of previously unresolved controversies.

TD learning model as a model of associative learning

Historically, Pavlovian contingency degradation paradigms have played a pivotal role in the development of animal learning theories^{107,143}, yet the exact underlying mechanisms remain to be determined^{65,111}. Here we show that the effect of contingency manipulations, both on behavior and

dopamine responses, can be explained by TD learning models. As our systematic investigation revealed, the failure of previous efforts to explaining contingency degradation with TD learning models is due to the use of inappropriate state representations, either not considering the ITI period or a simplistic way the ITI states were implemented. We show two types of TD learning models that explain the basic behavioral and dopamine results. The first model (Cue-Context model) uses a contextual stimulus as one of the states that continuously exists throughout the task period, which is equivalent to the cue-competition model traditionally considered in the animal learning theory literature. The second model (Belief-State model) explicitly models the task structure as transitions across the ISI and the two ITI states – “wait” and “pre-transition” states, and TD learning operates on beliefs (posterior probabilities) over these discrete states. In both models, the reduction in dopamine cue responses occur due to an increase in the value preceding a cue presentation, which decreases^{65,111,143} the change in value (reward expectation) induced by the cue, rather than due to a decrease in the absolute level of the value induced by the cue. This raises the question of why cue-induced anticipatory licking is reduced during contingency degradation. We provide a potential mechanism: the animal distributes anticipatory behavior depending on the relative values across different states.

Our results favor the Belief-State model over the Cue-Context model; both the dopamine and behavioral data were better explained by the Belief-State model. One could argue that it is unclear whether the animal can learn “sophisticated” state representations such as those used in our Belief-State model. In support of a belief state TD learning model, our analysis of anticipatory licking indicated that the reward expectation was modulated in a manner intricately linked to different task states: the ISI, waiting, and pre-transition states. Furthermore, we show that recurrent neural networks, which were trained to predict values (value-RNN), acquired the activity patterns that can be seen as representing beliefs, merely from observations without explicitly instructed to develop such representations, similar to our previous work using different behavioral tasks. Critically, when trained on contingency degradation sessions the RNNs developed more heterogenous representations of the ITI, capturing the same phenomenon as the Belief-State model.

It has been shown that TD learning models can explain a wealth of phenomena studied in the animal learning theory literature^{41,144}. The present study adds to this list Pavlovian contingency degradation – a classic phenomenon long studied in psychology and now in neurobiology. These results indicate that TD learning models provide a foundation with which to understand associative learning while the RNN-based approach provides a principled way to apply TD learning with little assumptions about state representations.

State representations as population activity dynamics

In RL, the “state” is a critical component which represents the set of observable and inferred variables necessary to compute value and policy. The artifice of the state representations used in RL modeling and in neurobiology and various assumptions have been criticized¹⁴⁵. For instance, it is implausible to have separate sets of neurons activated sequentially (i.e. CSCs) for separate cues¹⁴⁵. Furthermore, states are often defined within each “trial”; how can states be defined when there are no obvious trial structures¹⁴⁵? The success of value-RNNs in replicating aspects of dopamine signals and the acquisition of belief-like representations provide crucial insights into how biological circuits may represent states.

The critiques on the state representations used in TD are at odd in light of recent successes in RL which have been applied to fairly complex situations, containing many stimuli and often lacking obvious trial structures¹¹⁴. A key to this question, in part, lies in the use of neural networks, which take complex and arbitrary inputs and generalize them based on similarity. Our results in value-RNNs can provide further insights. As shown in our previous and the present work, in the value-RNNs, there is a stable fixed point (attractor), which corresponds to the ITI state. The ITI state, thus, exists by default. Furthermore, different stimuli induce stimulus-specific trajectories in the population activity state space. We can find close correspondence between population dynamics and distinct states assumed in belief state TD learning models. These results indicate that the population activity patterns in a network, including attractors and stimulus-specific trajectories, represent distinct states such as those assumed in our belief state TD learning models. Although the activity representing different trajectories likely involves overlapping sets of neurons,

as long as the activity patterns are discriminable, they can be trained to compute values properly by adjusting downstream synaptic weights.

Overall, while TD learning models with hand-crafted state representations help develop conceptual understanding, the RNN-based approach can provide crucial insights into how hand-crafted state representations could be implemented in neural networks. In the future, it is of great interest to examine whether neural activity in the brain exhibits patterns of activity predicted by the value-RNN models. The prefrontal cortex is a strong candidate area, receiving dopaminergic innervation from the VTA, with disruptions to this signal blunting changes to contingency degradation although this study used instrumental, not Pavlovian, conditioning. However other areas, such as the hippocampus also contribute task-relevant information during degradation to the prefrontal cortex¹⁴⁶ and neural-network modeling approaches that reflect the brain's functional organization¹⁴⁷ might provide more insight than our model which treats the state-machinery of the brain as a single recurrent neural network.

Limitations of the ANCCR model as a model of associative learning and dopamine

The present study unveiled limitations of the recently proposed causal learning model, ANCCR^{81,119}. The Degradation and Cued Reward conditions are minimally different and thus provide a strong test of the algorithm design. Our results indicate that the ANCCR model fails to explain the observed results despite our extensive examination of its parameter space. Crucially, the ANCCR model suffers from the same flaw as the ΔP definition of contingency. By calculating contingency by subtracting the background event rate, the model loses information that allows it to attribute increased value in the same manner as the TD models. Given the similarity in event rate between the conditions, the retrospective representations (and average eligibility trace) remain similar, with the computed retrospective contingencies in the Degradation condition being a subset of the Cued Reward contingencies. This explains why ANCCR predictions are similar for the two conditions independent of the parameter choice, as the rest of the model depends on this computed retrospective contingency as input. Thus, the failure of the ANCCR to explain the Cued Reward condition reflects the fundamental construction of the ANCCR model.

The failure of the ANCCR model doesn't not exclude some of the interesting ideas integrated into ANCCR, including how it uses retrospective information to learn state space. Rather it is its reliance on contingency that constitutes its core deficit. A recent report¹¹⁹ demonstrated that ANCCR is able to explain the dopamine response in outcome-selective contingency degradation. This is a result of the multidimensional tracking of cue-outcome contingencies in ANCCR. We show that both the belief state model and the value RNN, if trained on each reward separately and total value is taken to be the absolute difference of the two separate values, successfully predicts the experimental results of Garr et al. (2023) (Figure 13).

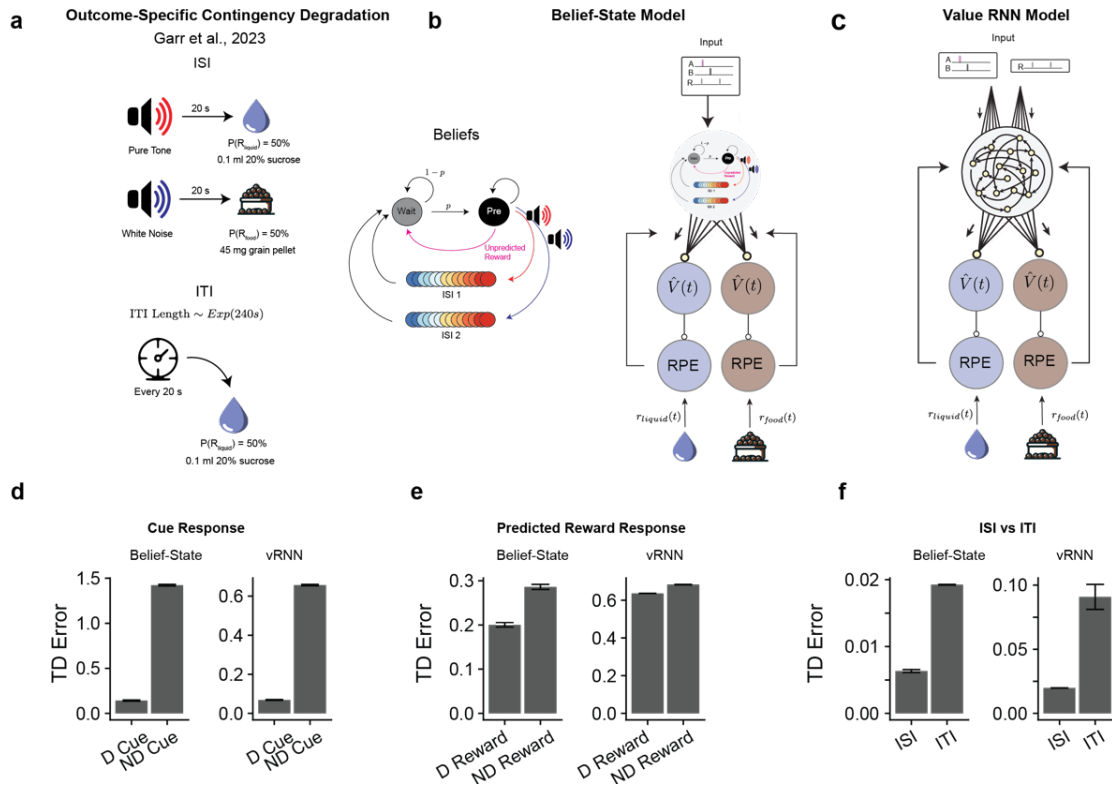


Figure 13 | Outcome-specific contingency degradation explained by belief state model and value RNN model.

(a) Experimental design of Garr et al., two cues predicted either a liquid or food reward. During degradation, every 20 s the liquid reward was delivered with 50% probability. The ITI length was drawn from an exponential distribution with mean of 4 minutes.

(b) Belief-State model design. The Belief-State model was extended to include a second series of ISI substates to reflect the two types of rewarded trials. The model was then independently trained on the liquid reward and food reward.

(c) The value RNN model design – as (b) but replacing the belief state model with the value RNN, using a vector-valued RPE as feedback, with each channel reflecting one of the reward types.

(d-f) Summary of predicted RPE responses from Belief-State Model and Value RNN (vRNN). The RPE was calculated as the absolute difference between the liquid RPE and food RPE. Other readout functions (e.g. weighted sum) produce similar results. Both model predictions match experimental results with degraded (D) cue (panel d) and degraded reward (e) having a reduced dopamine response versus non-degraded (ND). Furthermore, average RPE during ISI (3 seconds after cue on) and ITI (3 seconds before ITI) capture measured experimental trend. Error bars are SEM.

A similar approach using “multi-threaded predictive models” was used to successfully explain dopamine data in other situation¹²⁸. While this leaves open questions about how this is implemented biologically (the same being true for ANCCR), it does demonstrate that more complex contingency manipulation can still be explained by TD models. In fact, recent studies have provided evidence for heterogeneous responses to different types of rewards in dopamine neurons. While further evidence is required to solidify this understanding, this provisional assumption shows the potential that the TD models for multiple outcomes can be achieved in neural circuitry by concurrently running parallel circuits.

Methods

Animals

A total of 31 mice were used. 18 wildtype mice (8 males and 10 females) at 3-6 months of age were used to collect only behavioral data. For fiber photometry experiments, 13 double transgenic mice resulting from

the crossing of DAT-Cre (Slc6a3tm1.1(cre)Bkmn; Jackson Laboratory, 006660)⁴¹ with Ai148D (B6.Cg-Igs7tm148.1(tetO-GCaMP6f,CAG-tTA2)Hze/J; Jackson Laboratory, 030328)⁴² (DAT::cre x Ai148, 7 males and 6 females) at 3–6 months of age were used. Mice were housed on a 12 hr /12 hr dark/light cycle. Ambient temperature was kept at 75 ± 5 °F and humidity below 50%. All procedures were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Animal Care and Use Committee.

Surgery

Mice used for fiber photometry recordings underwent a single surgery to implant a multifiber cannula and head fixation plate implantation 2-3 weeks prior to the beginning of the behavioral experiment. All surgeries were performed under aseptic conditions. Briefly, mice were anesthetized with an intraperitoneal injection of a mixture of xylazine (10 mg/kg) and ketamine (80 mg/kg) and placed in a stereotaxic apparatus in a flat skull position. During surgery, the bone above the Ventral Striatum area was removed using a high-speed drill. A custom multifiber cannula (6 fibers, 200 μ m core diameter, 0.37 NA, Doric Lenses) was lowered over the course of 10 min to target 6 subregions in the Ventral Striatum. The regions are co-ordinates relative to Bregma (in mm) were: Lateral nucleus accumbens (lNAc, AP:1.42, ML:1.5, DV:-4.5); Medial NAc (mNAc, AP:1.42, ML:1, DV:-4.5); anterior lateral olfactory tubercle (alOT, AP:1.62, ML:1.3, DV:-4.8); posterior lateral OT (plOT, AP:1.00, ML:1.3, DV:-5.0); anterior medial OT (amOT, AP:1.62, ML:0.8, DV:-4.8); posterior medial OT (pmOT, AP:1.00, ML:0.8, DV:-5.0). Dental cement (MetaBond, Parkell) was then used to secure the implant and custom headplate and to cover the skull. Mice were singly housed after surgery and post-operative analgesia was administered for 3 days (buprenorphine ER-LAB 0.5 mg/ml). Mice used for behavioral training only underwent a similar surgical process, with a single surgery to only implant a head fixation plate.

Behavioral training

After the recovery from headplate-implantation surgery, animals were given ad libitum access to food and water for 1 week. Before experiments and throughout the duration of the experiments, mice were water restricted to reach 85–90% of their initial body weight and provided approximately 1–1.5 mL water per day in order to maintain the desired weight and were handled every day. Mice were habituated to head fixation and drinking from a waterspout 2-3 days prior to the first training session. All tasks were run on a custom-designed head-fixed behavior set-up, with software written in MATLAB and hardware control achieved using a BPod state machine (1027, Sanworks). A mouse lickometer (1020, Sanworks) was used to measure licking as infra-red beam breaks. The water valve (LHDA1233115H, The Lee Company) was calibrated, and a custom-made olfactometer was used for odor delivery. The odor valves (LHDA1221111H, The Lee Company) were controlled by a valve driver module (1015, Sanworks) and a valve mount manifold (LFMX0510528B, The Lee Company). All components were controlled through the Bpod state machine. Odors (1-hexanol, d-limonene, and ethyl butyrate, Sigma-Aldrich) were diluted in mineral oil (Sigma-Aldrich) 1:10, and 30 μ L of each diluted odor was placed on a syringe filter (2.7- μ m pore size, 6823-1327, GE Healthcare). Odorized air was further diluted with filtered air by 1:8 to produce a 1 liter/min total flow rate. The identity of the rewarded and non-rewarded odors were randomized for each animal.

In Conditioning sessions, there are three types of trials: (1) trials of Odor A (40% of all trials) associated with a 75% chance of water delivery after a fixed delay (2.5 s), (2) trials of unrewarded Odor B (20% of all trials) as control to ensure that the animals learned the task, and (3) background trials (40% of all trials) without odor presentation. Rewarded odor A trials consists of 2s pre-cue period, 1s Odor A presentation, 2.5s fixed delay prior to a 9 μ L water reward and 8s post-reward period. Unrewarded Odor B trials consist of a 2s pre-cue period, 1s Odor B presentation, and 10.5s post-odor period. Background trials in the Conditioning phase span a 13.5s eventless period. Trial type was drawn pseudo-randomly from a scrambled array of trial types maintaining a constant trial type proportion. Inter-trial-intervals (ITI) following the post-reward period were drawn from an exponential distribution (mean: 2s).

Learning was assessed principally by anticipatory licking detected at the waterspout for each trial type, with mice performing 100-160 trials per session until they reach an asymptotic task performance, typically after 5 sessions.

After the Conditioning phase, the mice were divided into three groups to undergo different conditions: Degradation (Deg group), Cued Reward (CuedRew group), and Conditioning (Cond group). The Deg group experienced contingency decrease during the Degradation phase. In the Degradation phase, Odor A still delivers water reward with 75% probability, and Odor B remains unrewarded. The difference was the introduction of uncued rewards (9 μ L water) in 75% of background trials to diminish the contingency. Animals underwent 5 sessions, each with 100-160 trials, to adapt their conditioned and neural responses to the new contingency. Degradation changed the cue value relative to the background trial but did not impact the reward identity, reward magnitude, or delay to/probability of expected reward.

The CuedRew group was included to account for potential satiety effects due to the extra rewards the Deg group mice received in the background trials. Unlike the Deg group, the CuedRew group's background trials were substituted with rewarded Odor C trials, where mice received additional rewards signaled by a distinct odor (Odor C). Rewarded odor C trials have the same trial structure as the rewarded odor A trials and animals were given 5 sessions, with 100-160 trials each, to adapt their conditioned response and neural responses to this manipulation.

The Cond group proceeded with an additional five Conditioning sessions, keeping the trial structure and parameters unchanged as in the Conditioning phase.

Post-degradation: eight mice were randomly chosen from the Deg group for the reinstatement phase, replicating the initial Conditioning conditions. After three reinstatement sessions, once the animals' performance rebounded to pre-degradation levels, we initiated the extinction process. This involved the delivery of both odors A and B without rewards, effectively extinguishing the cue-reward pairing. To mitigate the likelihood of animals generating a new state to account for the sudden reward absence, a shorter reinstatement session was conducted prior to the Extinction session on the Extinction day. Extinction was conducted over three days, each day featuring 100-160 trials. After Extinction, a second reinstatement

session was implemented, re-introducing the 75% reward contingency for odor A. All eight animals resumed anticipatory licking within ten trials during this reinstatement.

Fiber photometry

Fiber photometry allows for recording of the activity of genetically defined neural populations in mice by expressing a genetically encoded calcium indicator and chronically implanting optic fiber(s). The fiber photometry experiment was performed using a bundle-imaging fiber photometry setup (BFMC6_LED(410-420)_LED(460-490)_CAM(500-550)_LED(555-570)_CAM(580-680)_FC, Doric Lenses) that collected the fluorescence from a flexible optic fiber bundle (HDP(19)_200/245/LWMJ-0.37_2.0m_FCM-HDC(19), Doric Lenses) connected to a custom multifiber cannula containing 6 fibers with 200- μ m core diameter implanted during surgery. This system allowed chronic, stable, minimally disruptive access to deep brain regions by imaging the top of the patch cord fiber bundle that was attached to the implant. Interleaved delivery 473 nm excitation light and 405 nm isosbestic light (using LEDs from Doric Lenses) allows for independent collection of calcium-bound and calcium-free GCaMP fluorescence emission in two CMOS cameras. The effective acquisition rate for GCaMP and isosbestic emissions was 20Hz. The signal was recorded during each session when the animals were performing the task. Recording sites which had weak or no viral expression or signal were excluded from analysis.

The global change of signals within a session was corrected by a linear fitting of dopamine signals (473nm channel) using signals in the isosbestic channel during ITI and subtracting the fitted line from dopamine signals in the whole session. The baseline activity for each trial (F_0 each) was calculated by averaging activity in the pre-stimulus period between -2 to 0 seconds before an odor onset for odor trials or water onset for uncued reward trials. Z-score was calculated as $(F - F_0 \text{ each}) / \text{STD_ITI}$ with STD_ITI is the standard deviation of the signal during the ITI.

To quantify Odor A responses, we looked for ‘peak responses’ by finding the point with the maximum absolute value during the 1 s window following the stimulus onset in each trial. To quantify Odor B responses, we measure area under curve by summing the value during the 250 ms to 1s window following

the stimulus onset in each trial. This is to separate out the initial activation in the no-go response that we consistently observed, and which may carry salience or surprise information independent of value. To quantify reward responses, we looked for ‘peak responses’ by finding the point with the maximum absolute value during the 1-s window following the reward onset in each trial. To quantify reward responses, we looked for area under curve by summing the value during the 1.5s window following the reward omission in each trial.

Histology

To verify the optical fiber placement and GCaMP expression, mice were deeply anesthetized with an overdose of ketamine-medetomidine, and perfused transcardially with 0.9% saline followed by 4% paraformaldehyde (PFA) in PBS at the end of all experiments. Brains were removed from the skull and stored in PFA overnight followed by 0.9% saline for 48 hours. Coronal sections were cut using a vibratome (Leica VT1000S). Brain sections were imaged using fluorescent microscopy (AxioScan slide scanner, Zeiss) to confirm GCaMP expression and the location of fiber tips. Brain section images were matched and overlaid with the Paxinos and Franklin Mouse Brain Atlas cross-sections to identify imaging location.

Computational Modeling

Simulated Experiments

To compare these models, we generated 25 simulated experiments of each group, matching trial statistics to the experimental settings, but increasing the number of trials to 4,000 in each phase to allow to test for steady-state response in both these TD simulations and the ANCCR simulations. We then calculated the state representation of the simulated experiments for each of four state representations and ran the TD learning algorithm with no eligibility trace, called TD(0), using these state representations (Fig. 5a). While TD(0) has a learning rate parameter (α), it did not influence the steady-state results, which are presented, and thus the only parameter which influenced the result was γ , the temporal discount factor, set to 0.925 for all simulations using a timestep of $\Delta t = 0.2s$ (Fig. 6 shows the γ parameter search space).

CSC-TD model with and without ITI states

We initially simulated the Conditioning, Degradation, and Cued Reward experimental conditions using the CSC-TD model, adapted from Schultz et al. (1997)⁷¹. The cue length was fixed at 1 unit of time, with time unit size set to 0.2 s, and the ISI was matched to experimental parameters at 3.5 s. Simulated cue and reward frequencies were matched to experimental parameters, separately simulating Conditioning then Degradation and Conditioning then Cued Reward. In complete serial compound, also known as tapped-delay, each cue results in a cascade of discrete substates that completely tile the ISI. TD error and value were then modelled using a standard TD(0) implementation, using $\alpha = 0.1$, $\gamma = 0.925$. Reported values are the average of the last 200 instances averaged for 25 simulations. The model was run with states tiling the ISI only (CSC) or tiling the ISI and ITI until the next cue presentation (CSC with ITI states).

Context-TD model

The Context-TD model, which is an extension of the CSC-TD model, includes context as an additional cue, with the parameters from the simulation experiment design otherwise retained. For each phase (Conditioning, Degradation, Cued Reward) a separate context state was active for the entire phase, including the ISI and ITI, and off otherwise. This corresponds to the additive Cue-Context model previously described^{65,111,143}. TD errors reported are the average of the last 200 instances averaged for 25 simulations, except for Extinction which corresponded to third day of training.

Belief-State model

We simulated the TD error signaling in all four conditions (Conditioning, Degradation, CuedRew, Extinction) using a previously described belief-state TD model¹³⁹. In each simulation, the observation and transition matrix were calculated to represent that state space represented in Figure 3b, matching the experimental parameters. The difference from the original description was the splitting of the ITI period into two recurrent states: ‘Wait’ and pre-transition(‘Pre’), with a probabilistic transition from to the pre-

transition state. The observation and transition matrices were recalculated for each phase. How this state space is learnt on a trial-by-trial basis precise is beyond the scope of this study and the steady-state TD errors are not dependent on the implementation. As for the other models, TD errors reported are the average of the last 200 instances averaged for 25 simulations, except for Extinction which corresponded to third day of training.

RNN Modeling

We implemented value RNNs, as described previously, to model the responses in the three conditions (Cond., Deg., CuedRew). Briefly, simulated tasks were generated to match experimental parameters. We then trained recurrent network models, in PyTorch, to estimate value. Each value RNN consisted of between 5 and 50 GRU cells, followed by a linear readout of value. The timestep size was 0.5 s and gamma was 0.83 to match 0.925 for 0.2 s timesteps used in the TD simulations. To replicate the actual training process, we initially trained the RNNs on the Cond simulations, then on either the Deg or CuedRew conditions.

To visualize the state space used, we used a two-step canonical correlation analysis process. First, the principal components of the hidden-unit activity of each RNN was computed. Then we used the Python package rcca, to perform a three-way canonical correlation analysis to map the principal components from each RNN into a single 3-dimensional space.

We measured belief R^2 as previously described. For each simulation, we calculated the beliefs from the observations of cues and rewards. We then used multivariate linear regression to decode these beliefs from hidden unit activity.

ANCCR model

The ANCCR model is a recent alternative explanation of dopamine function⁸¹. While two previous studies have tested contingency degradation with ANCCR, they did not include the cued-reward controls. We implemented the ANCCR model using the code provided on the repository site (<https://github.com/namboodirilab/ANCCR>) and matching the simulation parameters to the experiment.

We used the set of parameter values used in the previous study, trying both Jeong et al., (2022) and Garr et al. (2023). The total parameter space searched was: T ratio = 0.2-2, α = 0.01-0.3, k = 0.01-1 or 1/(mean inter-reward interval), w = 0-1, threshold = 0.1-0.7, α_R = 0.1-0.3. The presented results use the parameters from Garr et al. (2023), as they were a better fit (T ratio = 1, α = 0.2, k = 0.01, w = 0.4, threshold = 0.7, α_R = 0.1). Additionally, we varied the weight of prospective and retrospective processes (w) to examine whether the data can be explained better by choosing a specific weight. Data presented are the last 200 instances averaged for the same 25 simulations used in the TD simulations.

Statistical analysis

Data analysis was performed using third party packages (e.g. Scipy, Statsmodel, etc.) in Python. All code used for analysis is available on request. Our behavioral data and dopamine response data have passed the normality test. For statistical comparisons of the mean, we used Student's t-test with a significance threshold of 0.05, adjusted according to Bonferroni correction. We used Welch's test for dopamine response to various events due to unequal variance between groups. Paired t-tests were conducted when the same mouse's performance was being compared across two different sessions. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications. The assumptions of the t-test were tested using the Shapiro-Wilk test to check for normality and Levene's test to check for equal variance.

Chapter II. Odor Valence Representation in Olfactory Tubercle during contingency learning

Introduction

One of the primary functions of the nervous system is to assign value to sensory stimuli that animals have encountered in the environment and adjust behavior accordingly. The olfactory system, as the earliest direct link of the nervous system to the external environment^{87,148,149}, plays an important role in learning stimulus value and guide behavior in the environment. The olfactory tubercle (OT), a key component of the olfactory pathway, is recognized for its integral role in olfactory processing^{15,23,100}. It receives a distinct stream of olfactory input, primarily from tufted cells in the olfactory bulb¹⁰², and substantial dopaminergic

input from the ventral tegmental area (VTA). This unique positioning enables the OT to integrate these inputs, thus playing a crucial role in assigning valence to olfactory stimuli.

Furthermore, the OT is important to motivation and reward processing, with studies showing its enhanced role compared to other regions like the nucleus accumbens in these functions^{18,99,150–153}. Specifically, the OT's response to cocaine administration¹⁹ and its activation leading to odor preferences underscore its significance in the reward system^{17,21,104}. Notably, the OT's higher concentration of dopaminergic inputs, compared to the posterior piriform cortex, points to its distinct role in olfactory processing and odor-reward contingency representation^{21,22}. In fact, colleagues in the Murthy lab have identified that odor responses in the OT are modulated by the prediction of reward or punishment regardless of odor identity^{22,23}, and also, many literatures reported that neurons in OT encode odor valence^{16,28,105,106}.

Anatomically, the OT is part of the ventral striatum, predominantly composed of spiny projection neurons (SPNs) expressing D1- and D2-type dopamine receptors^{24,26}. These two groups of SPNs have been found to differ in many aspects, including their spatial distributions within the OT and input/output projection patterns^{154,155}. Furthermore, D1- and D2-type neurons have different affinities to dopamine and coupled to distinct intracellular signaling pathways. These differences may result in different plasticity rules for D1- and D2-SPNs. These observations raising questions about their specific roles in shaping odor valence representation and contingency learning. To explore the role of these neuronal subtypes in the OT, I have employed the longitudinal optical imaging technique, specifically two-photon microscopy, to record and track the activities of these neuronal types in behaving mice. This will allow an examination of the evolution and stability of OT neural representation and the differential roles of these neuronal types in the learning process.

In the real world, predictive relationships between particular cues and outcomes can change. Contingency degradation models erosion of predictive relationships: once predictive cue becomes no longer predictive of an outcome. Behavioral studies have identified **contingency** as a fundamental factor modulating associative learning as measured based on conditioned responses. Studies have shown that the odor responses in the OT influences reward behaviors and odor-preferences¹⁷, and are highly plastic¹⁰⁵,

allowing the valence representation to be rapidly updated in different circumstances. However, it's unclear how neural representations in the OT change as odor-outcome contingency is degraded. It is possible that neuronal representations in the OT simply reflect current contingencies, but also possible that they maintain a long-term memory of the association at both single-neuron and population levels, even after alterations in the contingency. By imaging populations of D1- and D2-type neurons in the OT during contingency degradation procedures, my research aims to investigate the stability and evolution of neural responses to rewarded odors and their correlation with behavioral changes. This approach will test whether OT representations are static or continuously updated with new learning, and whether there is a subset of neurons that maintain stability amidst these changes. It is also possible that the two populations of neurons have different evolution of representation during/following contingency degradation. In total, I aimed at characterizing the stability of neural representations in the OT under changing contingencies.

Results

In this study, I investigated the development and stability of population representations in D1 and D2 receptor-expressing neurons in the olfactory tubercle (OT), as well as their modifications under different conditions. To achieve this, I conducted two-photon calcium imaging of these two neuronal populations in the OT of mice trained in an odor-reward Pavlovian conditioning task. Specifically, I used an adeno-associated virus for conditional expression of the calcium indicator GCaMP7s in the presence of Cre recombinase. This was unilaterally injected into the right-hemisphere OT of either *Tac1-Cre* or *Adora2A-Cre* mice, followed by the implantation of a 1.1 mm custom-made cannula targeting the OT, as shown in Fig. 14a.

After recovery, a Gradient-Index (GRIN) lens was inserted into the cannula for imaging. The mice were then subjected to water restriction, accustomed to head fixation, and trained in an odor-reward Pavlovian conditioning task. All mice ($n = 15$) were initially trained on two odors: Odor A, predicting a

water reward with 75% probability, and Odor B, indicating no reward. In Phase 1, Odor A accounted for 40% of trials, Odor B for 20%, and the remaining 40% were blank trials. Results indicated that mice developed anticipatory licking to Odor A but not Odor B within five sessions (Fig. 14e).

In Phase 2, the mice were divided into two groups (Fig. 14b). The first group ('Cond', $n = 7$) continued the same task as in Phase 1, showing no significant behavioral changes over five additional sessions (Fig. 14e). The second group ('Deg', $n = 8$) underwent contingency degradation, where blank trials were replaced with 'background water' trials delivering a reward 75% of the time. This group exhibited a decrease in anticipatory licking to Odor A ($t_8 = -0.36$, $P = 0.00074$, paired t-test, session 5 vs session 10) and an increase in non-licking trials before water delivery in Odor A trials (Fig. 14e). This suggests that the mice were sensitive to contingency and adjust their behavior accordingly. We confirmed that anticipatory licking behavior was consistent between mice with and without a GRIN lens implant.

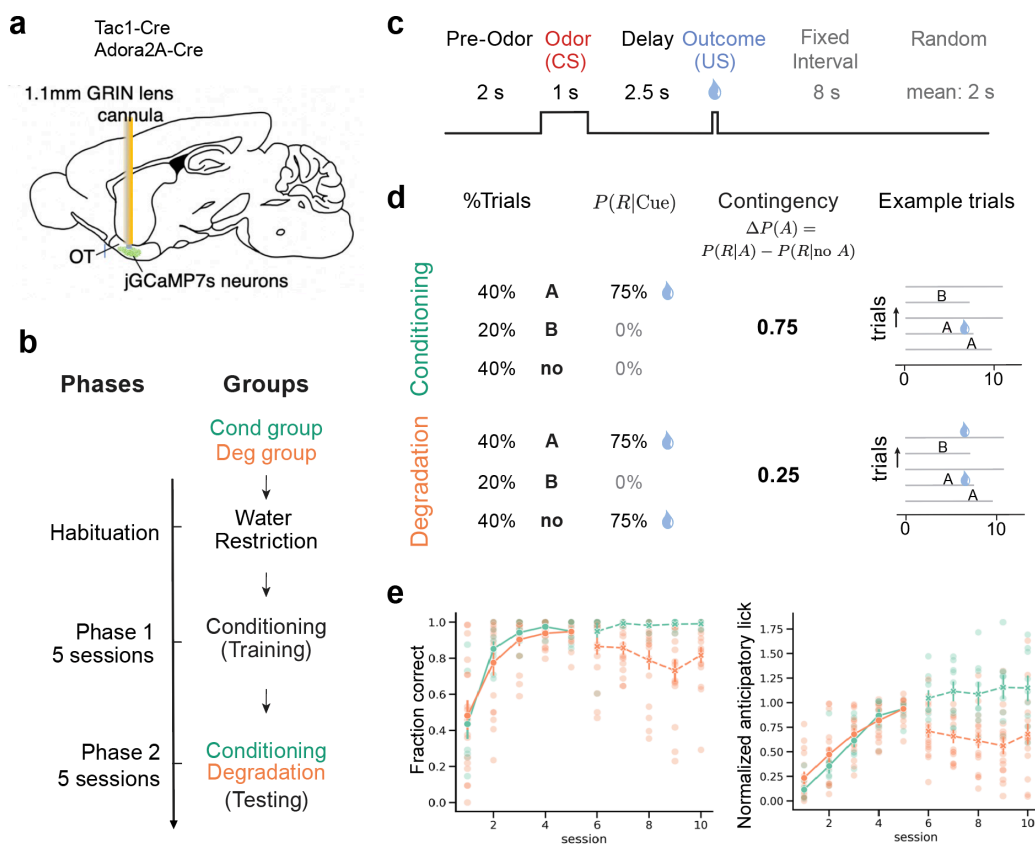


Figure 14 | Task design and behavioral responses

- (a) Configuration of two-photon imaging. A sagittal view from either Tac1-Cre or Adora2A-Cre mouse shows targeting in the OT.
- (b) Experimental design. Two groups of mice were subjected to two conditions of contingency learning. All animals underwent Phases 1 and 2.
- (c) Trial timing.
- (d) Trial parameters per condition. In Conditioning and Degradation, Odor A predicts a 75% chance of reward (9 μ L water) delivery, and Odor B indicates no reward. In Degradation, blank trials were replaced with uncued rewards (75% reward probability).
- (e) Fraction correct (percentage of >1 lick in the delay window in Odor A trials) and average lick rate in 3s post-cue (Odor A) by session. Error bars represent SEM.

OT neuronal activity robustly and bidirectionally reflects odor-outcome contingency

Our imaging results showed that both D1 neurons from Tac1-Cre mice ($n = 6$ mice) and D2 neurons from Adora2A-Cre mice ($n = 6$ mice) developed a range of excitatory and inhibitory responses to odors over several days of training (Figure 15a, b). During the conditioning phase, both D1 and D2 neurons quickly developed both excitatory and inhibitory responses to Odor A on the first day of training. As training progressed, there was a clear increase in the number of responsive neurons and in the magnitude of their responses, which paralleled the increased value of the contingency associated with the odor (Figure 15b, d). However, during the degradation phase, both the magnitude of responses and the number of significantly responsive D1 and D2 neurons diminished considerably. This reduction exemplifies the effect of contingency degradation, a phenomenon that was not observed at the population level in the neurons of animals that continued the cue-reward association training ('Cond group'). The classic go/no-go model would predict opposite activity patterns in the two pathways during movement, however, previous studies have shown that in vivo recordings from the two types of SPNS have revealed very similar activity patterns. Our results further support this by demonstrating that both D1 and D2 neurons showed quite similar changes in response to changes in contingency at the population level.

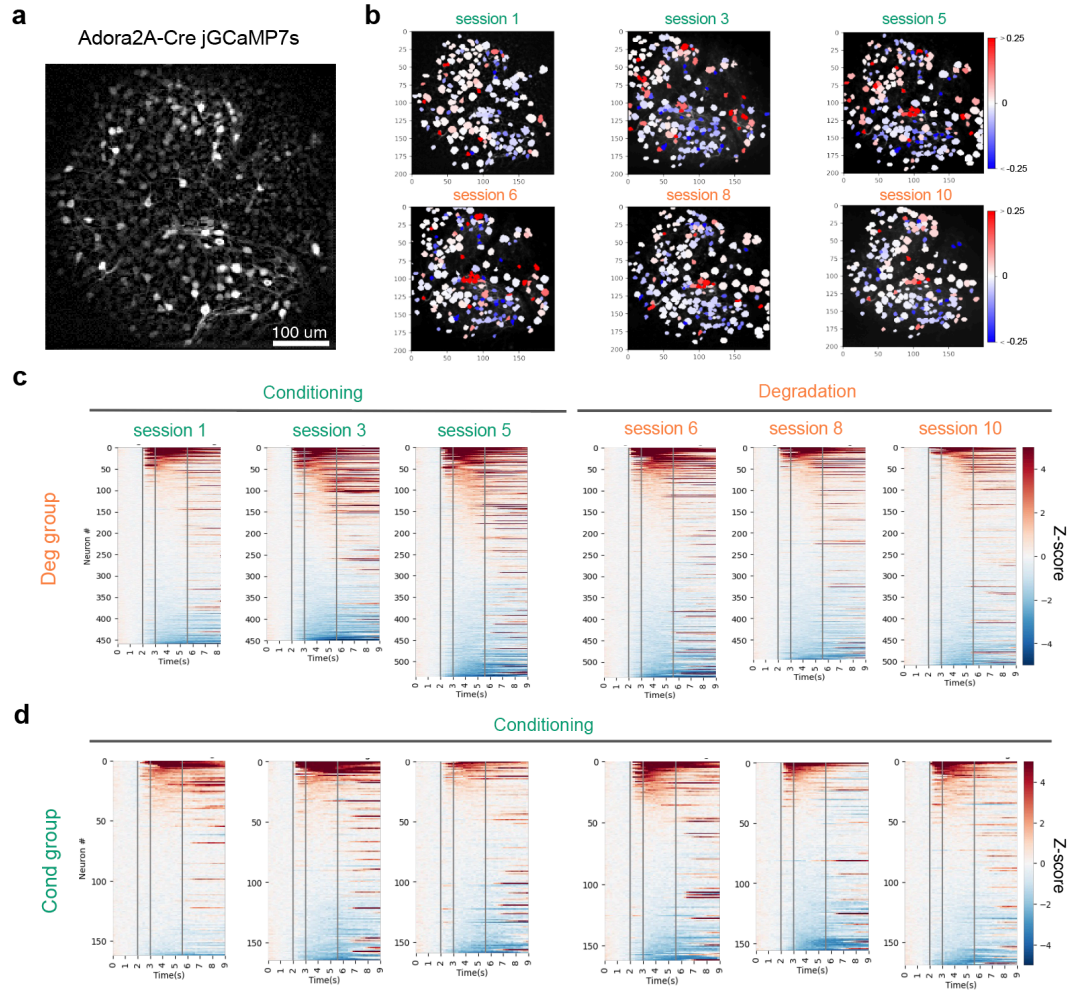
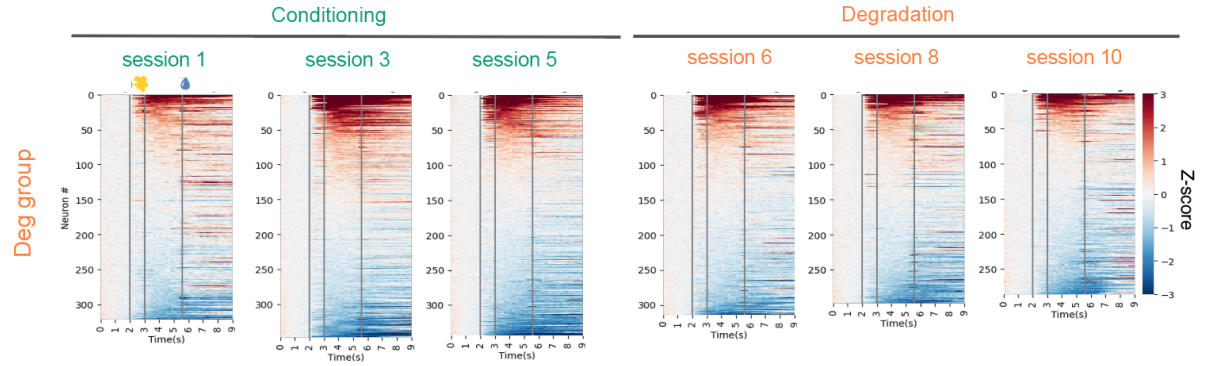


Figure 15 | Two- photon calcium imaging of D1- and D2- type neurons in the olfactory tubercle (OT) during odor- outcome association learning.

- (a) Field of view of GCaMP7s expressing neurons in a Adora2A-Cre mouse.
- (b) Neuronal activity in field of view of a Adora2A-Cre mouse in Odor A trials across days of training. Colored overlay indicate significantly responsive neurons. Session 1-5 are from Conditioning phase and session 6-10 are collected from Degradation phase. Peak activity (df/f) is shown in the 1 s period after Odor A delivery prior to reward.
- (c,d) Heatmap showing mean Odor A response time courses of all recorded D2 (rows) during Conditioning and Degradation sessions, sorted by response magnitude (excitatory to inhibitory). Vertical dashed lines are for Odor A onsets, offsets, and water delivery timepoint.

a Tac1-Cre jGCaMP7s



b

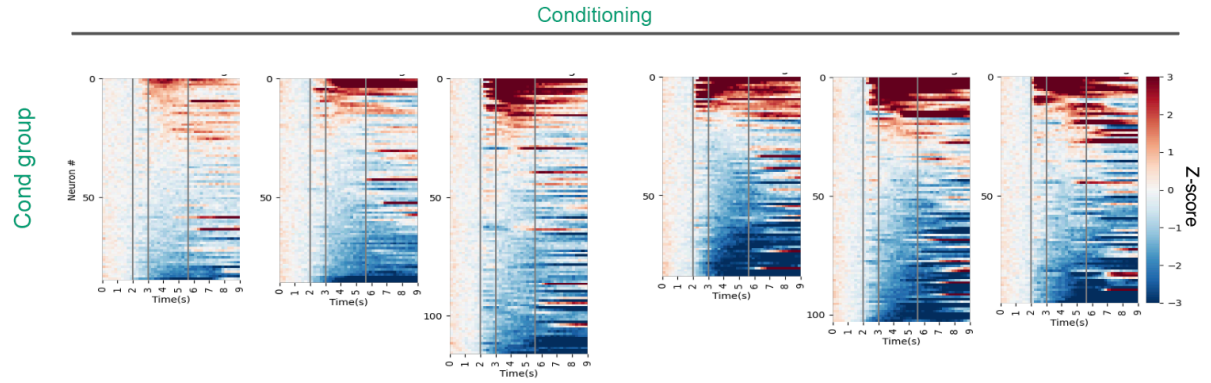


Figure 16 | D1- type neurons in the olfactory tubercle

(a,b) Heatmap showing mean Odor A response time courses of all recorded D1 (rows) during Conditioning and Degradation sessions, sorted by response magnitude (excitatory to inhibitory). Vertical dashed lines are for Odor A onsets, offsets, and water delivery timepoint.

a

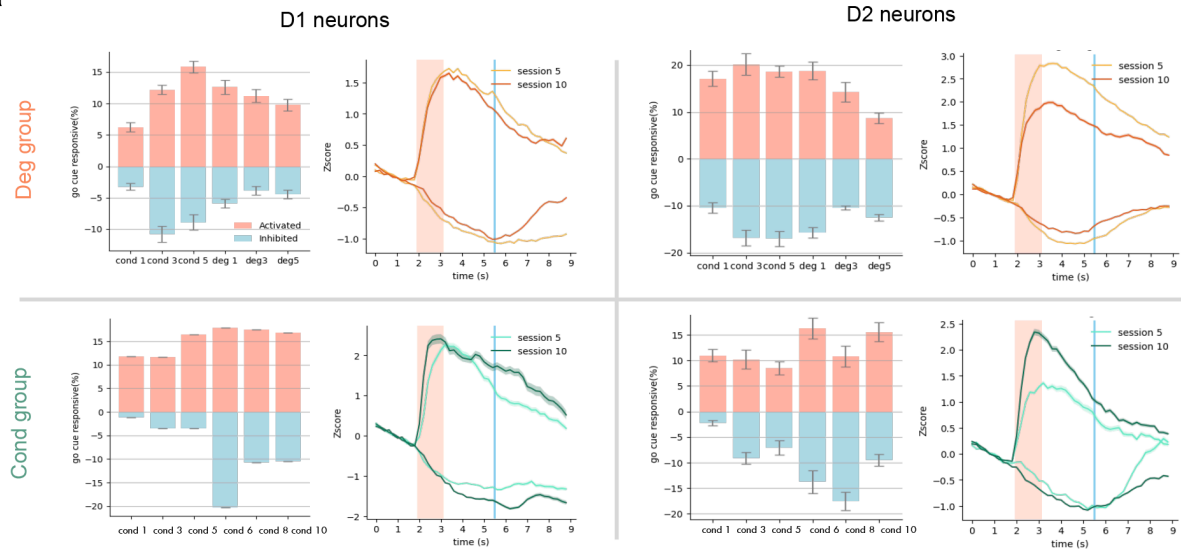


Figure 17 | Population level quantification of D1/D2 neural responses

- (a) The bar graphs delineate the percentage of neurons with notable responses, while the adjacent line graphs offer a detailed view of the response time courses for neurons that were either significantly activated or suppressed by the conditioned odor.

To dissect the distinct behaviors of D1 and D2 neurons in response to contingency degradation, I quantified their responses at the population level first (Figure 17a). Initially, a significant proportion of both neuron types exhibited either activation or suppression in response to the conditioned odor cue. This behavior marks the associative learning phase, where connections between the odor cue and reward are established. However, noticeable changes were observed compared to the conditioned (Cond) group as the task progressed into the degradation sessions, particularly in the D2 neuron group. The introduction of contingency degradation resulted in a pronounced population-level reduction in the number of significantly responsive neurons ($p < 0.001$, Student t-test, cond 5 session vs deg 5 session) and average response magnitude of these neurons among D2 neurons. On the other hand, D1 neurons showed a less pronounced decrease in the number of activated neurons ($p < 0.01$, Student t-test, cond 5 session vs deg 3 session) and maintained similar levels of average response magnitude after degradation, implying that D1 neurons may encode the cue-reward association more resiliently to changes in contingency, potentially representing a different facet of associative learning or preserving a memory of the odor cue.

These results reveal a nuanced divergence in how D1 and D2 neurons in the OT process contingency information. The data indicate that D2 neurons adjust their activity more dynamically in accordance with reward prediction changes—an essential aspect of adaptive learning. In contrast, the sustained activity of D1 neurons, which remains relatively unaffected by contingency degradation, poses interesting questions regarding their specific contributions to olfactory learning and odor information encoding. This differential responsiveness to contingency degradation between D1 and D2 neurons enriches our understanding of the neural substrates of associative learning in the olfactory tubercle.

Longitudinal analysis of individual neurons during contingency learning

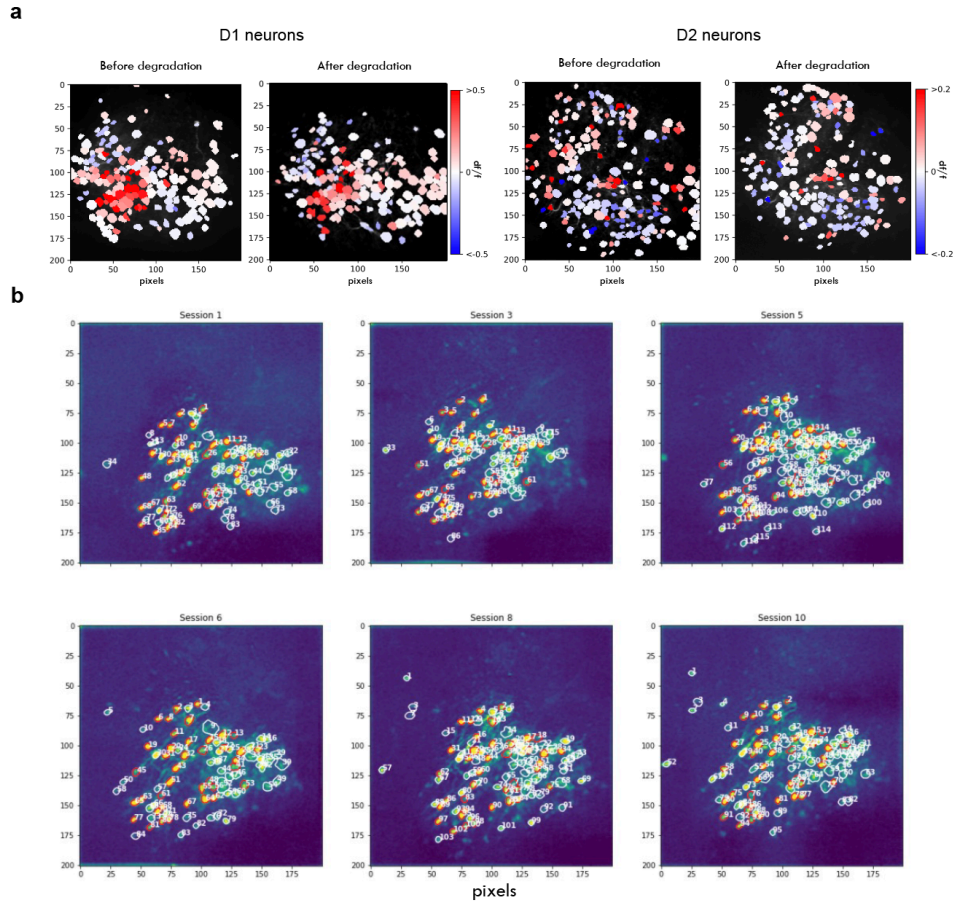


Figure 18 | Neural response dynamics in D1 and D2 Neurons during contingency degradation and longitudinal tracking across Training Sessions

(a) Activity maps of D1 and D2 neurons in the olfactory tubercle before and after contingency degradation. Each point represents a single neuron with color coding indicative of response magnitude: red for excitatory and blue for inhibitory responses. The left panels show D1 neuron responses, and the right panels show D2 neuron responses, illustrating the change in neural activity patterns due to the degradation process.

(b) Longitudinal tracking of an example FOV across multiple training sessions (1, 3, 5, 6, 8, 10). The white contours indicate all the identified neurons from the FOV, with red contours representing neurons that could be tracked over all 6 sessions. This longitudinal tracking could provide visualization and analysis on the progression and plasticity of neuronal responses as the mice learn the task odors and then experience the contingency degradation.

We next performed a longitudinal analysis of individual neurons during contingency learning to discern how alterations in contingency influence the odor valence representation at the single neuron level. In addition to the population level analysis presented above, it is important to test more detailed patterns within each population. For instance, an intriguing possibility is that most individual D1 or D2 neurons are subject to modulation reflective of the changing contingency, while there exist some neurons exhibiting

stable representation throughout over days. During each imaging session, I carefully adjusted the depth of the imaging plane to closely align with that of the previous days. This was achieved by comparing the field of view (FOV) with images of the FOV from previous sessions. Such adjustments were crucial for ensuring the capture of the same or highly overlapping neuronal populations across different days of experiment. To ensure accurate tracking of same neurons across sessions, semi-automated multi-session registration techniques were employed. Utilizing the multi-session registration algorithm in CaImAn facilitated the initial identification of consistent neurons, although this method's performance can be variable¹⁵⁶. To mitigate this remaining variability, I performed an automated grid search to ascertain the optimal adjustments required for aligning the spatial footprints. For further precision in neuron identification, spatial footprints of cells were plotted alongside an index, followed by meticulous manual registration and correction. This dual strategy of automated alignment complemented by manual refinement ensures the fidelity of neuron tracking throughout the entire period of our experiments.

The methods chosen for this analysis aim to provide a window into the complex dynamics of neural representations within the OT during associative learning. This analysis is still ongoing. By examining how stable individual neurons represent the valence of the sensory stimuli and how plastic they adapt their representation in response to altered contingency at both the population- and single cell-level, we gain a deeper understanding of the neuronal underpinnings that support odor-reward association learning.

Discussion

The olfactory system in rodents, particularly mice and rats, provides an excellent experimental system for studying examining associative learning. These animals demonstrate a remarkable ability to rapidly master tasks involving odor cues, establishing links between odors, motor responses, and rewards. In our study, we have chosen mice as our model organism to explore the acquisition and representation of valence signals in the OT, a structure integral to both the odor processing and reward learning. Utilizing head-restrained mice for high-resolution, longitudinal two-photon calcium imaging permits in-depth

exploration of neural representational development and stability in the OT during continual learning. As mice adapted their behavior in response to altered contingencies, I monitored the evolving neural representations in two distinct populations within the OT. Notably, D2 neurons showed dynamic bidirectional changes in response to contingency shifts, while D1 neurons maintained relatively more stable activity patterns despite degradation.

Longitudinal tracking allowed for a single-cell level analysis. From the continual learning group ('Cond' group), in which at the population level, the activities remained largely stable during training, I discovered instability in individual D1 and D2 neurons (data analysis in progress). This finding has sparked our interest in understanding the stability of OT valence encoding in response to external meaningful cues and whether representational drift is a factor in the OT. Ultimately, this project aims to dissect the activity contributions across neural sub-populations. We seek to distinguish between drift-related activities (i.e., those independent of changes in task parameters) and activities directly related to alterations in task parameters and statistics. This distinction is crucial for a deeper understanding of how the OT encodes and processes associative learning and the stability of these neural representations.

Representational stability and drift

Research on the hippocampus and neocortex have documented that neural activity patterns can remain stable over several days, yet may undergo substantial changes over more extended periods, which have been characterized as the representation drift^{157–164}. The concept of the 'neural representation' refers to the specific spatiotemporal activity patterns in a neuron population that correlate with sensory experiences or actions¹⁶⁵, and the term 'representation drift' describes the process wherein neurons that initially responded to certain stimulus or experience become less responsive over time, while other neurons become active. In the hippocampus, it has been found that the subset of place cells encoding spatial positions can vary, even in unchanging environments. While it has long been assumed implicitly that neural representation of external environment or experience was stably encoded in a specific group of neurons, the phenomenon of representation drift has been found and supported by longitudinal calcium imaging,

electrophysiology, and fMRI studies in various brain regions, including the hippocampus^{157,163,164}, posterior parietal cortex, anterior piriform cortex¹⁶¹, and primary visual cortex¹⁶⁰.

In our research, I developed a behavioral task that enabled mice to learn and flexibly update cue-outcome associations over approximately two weeks. This project presents two conceptual advances: Firstly, it is known that reinforcement learning leads to changes in neural activities, especially in ventral striatal areas, including the OT. However, the stability of representations formed through reinforcement learning in striatal regions has yet to be explored compared to sensory and motor cortical areas, or hippocampal structures. My project offers an opportunity to use rodents' rapid and robust olfactory learning to investigate the stability of population representations in the OT under various conditions. Secondly, the OT's local networks are predominantly inhibitory, contrasting with other regions where representational stability is often discussed in the context of largely excitatory recurrent neural networks. Our studies are among the first to apply advanced computational algorithms and conceptual frameworks to understand the stability of population representations in the OT.

The concept of representational drift remains a vigorously debated and unresolved topic in neuroscience. Our research will contribute crucial biological data to enhance the understanding of neural representations in the brain.

Methods

Animals

Adult male and female (Tac1)-cre (B6;129S-Tac1^{tm1.1}(cre)Hze/J, Jackson Laboratory; RRID:IMSR JAX:021877) heterozygous mice and (Adora2A)-cre (B6.FVB(Cg)-Tg(Adora2a-cre)KG139Gsat/Mmucd, GENSAT; MGI:4361654) were 3–8 months of age at the start of the experiments. I used Tac1-Cre mice to image D1-type neurons and Adora2A- Cre mice to image D2- type neurons. Animals were housed on a 12-hour dark/12 hour light cycle (dark from 07:00 to 19:00) and performed a task at the same time each day. Animals were singly housed after surgery, and were water restricted during training and imaging. Mice received water every day by experimenters and the body weights were kept >80% of their weights with freely available water. All procedures were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Animal Care and Use Committee.

Cannula assembly

In our lab, I assembled custom-designed cannulas according to my colleague's method described previously. I used biocompatible polyimide tubing from MicroLumen, 6.2 mm in length and 1.1 mm in diameter with an ultra-thin wall. This material has been shown to induce minimal inflammatory response in the brain. For the base of the cannula, a 150 μ m thick quartz coverslips (sourced from Electron Microscopy Sciences) was cut into 1 mm diameter disks using a laser cutter with fine-tune later parameters. Under a surgical microscope, these quartz disks were held in place by a vacuum, attached to the polyimide tube using Norland Optical Adhesive NOA 68 from Edmund Optics, and then cured with a UV light source from ThorLabs. Prior to surgical implantation, each cannula was meticulously inspected and disinfected using the UV light source to ensure sterility and integrity.

Surgery

Prior to any behavioral training, naïve mice underwent a series of surgical procedures including virus injection, cannula implantation, and head fixation plate implantation. The mice were first anesthetized using an intraperitoneal injection of a xylazine (10 mg/kg) and ketamine (80 mg/kg) mixture. The mouse was then secured in a stereotaxic apparatus and a craniotomy of 1.8 mm was performed at coordinates 1.5 mm anterior-posterior (AP) and 1.5 mm medial-lateral (ML) at the right hemisphere. A 20 G needle facilitated the removal of tissue to a depth of 2.5 mm below the brain surface before the virus injection. For the virus injection, a glass micropipette was used to attached to a nanoinjector (MO-10, Narishige) to introduce 300 nL of pGP-AAV9-syn-FLEX-jGCaMP7s-WPRE virus (Addgene) at a depth of 4.90 mm dorsoventral (DV), injecting at a rate of 100 nL/min. Following a 20-minute waiting period after the injection, the micropipette was carefully retracted.

Subsequently, the sterilized cannula was maneuvered using a 20G blunt needle attached to a vacuum. It was gradually lowered to a depth of 4.9 mm DV over a 40-minute duration. The cannula was then firmly attached to the skull with cyanoacrylate glue, and a head fixation plate was similarly secured in place behind

the cannula. We used Dental cement (MetaBond) to encase the skull and headplate and sealed the cannula opening with a silicone sealant (KwikSil). Post-surgery, the mice were housed individually.

After allowing a 4-week recovery period for virus expression and to mitigate any inflammatory response from the cannula insertion, a 1 mm diameter, 3.4 mm long 0.5NA GRIN lens (Thorlabs) was inserted into the cannula. Following this, the behavioral training and imaging sessions began.

Behavioral training

Following headplate-implantation surgery, animals were allowed a recovery period of one month with ad libitum access to both food and water, ensuring complete recuperation before the commencement of experimental procedures. Subsequent to the recovery period and prior to the initiation of experimental tasks, as well as for their duration, the mice were subjected to water restriction protocols. These were designed to reduce their body weight to 85–90% of their pre-restriction weight, a common practice in behavioral neuroscience to enhance motivation for water-based rewards. Mice received 1–1.5 mL of water daily to maintain this desired weight range and were handled daily to promote acclimation to human interaction.

All experimental tasks were conducted on a specialized head-fixed behavioral set-up, custom-designed for this study to facilitate a controlled environment for head-restrained neuroimaging and behavioral tracking. Odor delivery was executed using a custom-built olfactometer system, as described previously^{166,167}. This system delivered odors at a consistent flow rate of 1.5 L/min and at a concentration of 20% saturation for a duration of 1 seconds per presentation.

Odorants, specifically 1-hexanol and d-limonene sourced from Sigma-Aldrich, were diluted in mineral oil (also from Sigma-Aldrich) at a 1:10 ratio. For each experiment, 1.1 ml of the diluted odor solution was placed in individual tubes connected to the olfactometer. The assignment of which odor served as the rewarded or non-rewarded stimulus was randomized across subjects to preclude any bias.

Licking behavior—a key indicator of task engagement and learning—was monitored via a capacitance sensing Arduino circuit, which provided a sensitive and real-time measurement of interactions with the water delivery spout. Control and recording of behavioral events were performed using a Python-based

software, with custom adaptations from the ScopeFoundry platform (<http://www.scopefoundry.org/>). Data acquisition was enhanced by employing National Instruments DAQ hardware for high-fidelity recording of the behavioral data.

In Conditioning sessions, there are three types of trials: (1) trials of Odor A (40% of all trials) associated with a 75% chance of water delivery after a fixed delay (2.5 s), (2) trials of unrewarded Odor B (20% of all trials) as control to ensure that the animals learned the task, and (3) background trials (40% of all trials) without odor presentation. Rewarded odor A trials consists of 2s pre-cue period, 1s Odor A presentation, 2.5s fixed delay prior to a 9 μ L water reward and 8s post-reward period. Unrewarded Odor B trials consist of a 2s pre-cue period, 1s Odor B presentation, and 10.5s post-odor period. Background trials in the Conditioning phase span a 13.5s eventless period. Trial type was drawn pseudo-randomly from a scrambled array of trial types maintaining a constant trial type proportion. Inter-trial-intervals (ITI) following the post-reward period were drawn from an exponential distribution (mean: 2s).

Learning was assessed principally by anticipatory licking detected at the waterspout for each trial type, with mice performing 100-160 trials per session until they reach an asymptotic task performance, typically after 5 sessions.

After the Conditioning phase, the mice were divided into two groups to undergo different conditions: Degradation (Deg group) and Conditioning (Cond group). The Deg group experienced contingency decrease during the Degradation phase. In the Degradation phase, Odor A still delivers water reward with 75% probability, and Odor B remains unrewarded. The difference was the introduction of uncued rewards (9 μ L water) in 75% of background trials to diminish the contingency. Animals underwent 5 sessions, each with 100-160 trials, to adapt their conditioned and neural responses to the new contingency. Degradation changed the cue value relative to the background trial but did not impact the reward identity, reward magnitude, or delay to/probability of expected reward.

The Cond group proceeded with an additional five Conditioning sessions, keeping the trial structure and parameters unchanged as in the Conditioning phase.

Two-photon imaging of calcium activity

In vivo imaging was performed using a custom-built microscope, as previously described in relevant literature^{168,169}. The imaging process was conducted at a rate of 5 Hz using an air objective (10×, Leica) and a wavelength of 930 nm, facilitated by a Ti:sapphire laser (Chameleon Ultra, Coherent). The management of image acquisition, scanning, and stimulus delivery was handled by custom-written LabVIEW (National Instruments) software²³ available on GitHub.

Before the two-photon imaging, the position of the GRIN lens and the approximate neuronal imaging plane were initially determined using a camera. The head fixation plate was then mounted on an adjustable pitch and roll platform (Thorlabs), allowing for manual adjustments to align the lens parallel to the objective.

Calcium imaging data analysis

The imaging data underwent motion correction using the Non-Rigid Motion Correction (NoRMCorre) technique¹⁷⁰. Following this, the activity of individual neurons was isolated and background noise was reduced using CaImAn¹⁵⁶, supplemented by manual refinement for accuracy. The quantification of putative calcium transient events in each neuron was based on activity levels that were three standard deviations above a temporally proximal baseline, lasting longer than five frames. Neurons with fewer than three recorded transients were excluded from the analysis.

Given the inherent baseline fluorescence fluctuations in individual neurons, the activity of each neuron during individual trials was normalized to a 2-second pre-trial baseline.

Histological confirmation of imaging site

Upon completing the imaging experiments, the mice were subjected to transcardial perfusion, after which their brains were carefully extracted from the skull. Coronal floating sections of these brains were then prepared using a vibratome (Leica VT1000S). To ascertain the location of GCaMP expression and the trajectory of the cannula, these brain sections were imaged using the Zeiss Axio Scan slide scanner at the Harvard Center for Biological Imaging.

The images of the brain sections were meticulously matched and overlaid with corresponding cross-sections from the Paxinos and Franklin Mouse Brain Atlas. This process was crucial for accurately identifying the specific imaging locations within the brain. In the cohort of Tac1-Cre mice, seven out of eight showed confirmed imaging locations in the olfactory tubercle (OT). However, in the remaining cases, the cannula tip was found in either the ventral pallidum or the nucleus accumbens. Among the Adora2A-Cre mice, eight out of nine had confirmed OT imaging locations.

For the subsequent analysis, only those mice were considered which had both confirmed OT imaging locations and successful imaging outcomes, ensuring the validity and accuracy of the study's results.

Conclusion

The ability to learn to associate particular stimuli with a specific outcome is crucial in life, which plays essential roles in various activities such as hunting, eating and escaping. While previous psychological experiments have shown that the strength of the association is modulated critically by the **contingency**, this concept is relatively understudied in neuroscience. In this dissertation, I sought to reveal the neural underpinnings of contingency, focusing on the phenomenon known as contingency degradation, where the introduction of an unpredicted reward alongside a predicted one diminishes the previously established association with a conditioned stimulus^{42,62,171}.

The study began with establishing a behavioral paradigm for analyzing contingency degradation in a Pavlovian odor-reward association task in mice. The observed behavioral responses, specifically anticipatory licking to the reward-predicting odor cue, were significantly affected by contingency degradation. This finding is in line with previous research in the field^{42,62,171}.

The midbrain dopamine neurons are traditionally believed to signal reward prediction errors - the discrepancy between the predicted and actual rewards - a concept paralleling the 'TD error' in TD learning algorithms^{8,9,115,172,173}. However, recent theories have proposed alternative interpretations of dopamine signals^{81,119}, leading to debates about the nature of dopamine signals. Additionally, more evidence shows that dopamine signals are more heterogeneous than previously thought^{9,124,174,175}. For example, dopamine

signals may support different forms of learning depending on their projection to different striatal target areas: previous data in the Uchida lab suggest that VS projecting dopamine signals are most consistent with traditional RPE signal, whereas dopamine neurons projecting to the tail of the striatum (TS) are activated by novelty and threats^{124,175}. Considering these findings, here I asked how contingency degradation affects dopamine signals in VS, and whether observed signals are consistent across the VS. Utilizing multifiber photometry in the VS, we revealed that dopamine responses in all subregions were consistently attenuated when stimulus-outcome contingency was degraded by the uncued delivery of additional rewards. Our modeling showed that aspects of dopamine responses can be comprehensively explained by TD learning models, if the model is equipped with proper state representations reflecting the task structure.

Within the VS, the reinforcement learning framework posits that RPEs are used to update values, or predictions about future rewards, which can then guide behavior towards reward acquisition^{8,9}. As a component of the VS, OT neurons have been reported to encode odor valence^{22,23,105}, and the odor responses are highly plastic^{20,21,31,176}, which are particularly relevant to the task of odor-reward association degradation. However, the specific neural representations in the OT during contingency degradation remained elusive. Leveraging advancements in imaging techniques, particularly two-photon calcium imaging, has facilitated longitudinal recordings of neural populations at a single-cell resolution. My research has revealed that both D1- and D2-type OT neurons exhibit a sensitivity to contingency. These neurons demonstrate a diminished response to the previously rewarded odor cue following contingency degradation. This suggests that the OT's population activity dynamically reflects changes in contingency, aligning closely with the conditioned behaviors observed in Chapter I of this dissertation. In this conclusion, I want to further discuss the following three ideas:

TD learning and contingency

Learning predictive or causal relationships requires properly assigning credits to those events that are responsible for the outcomes observed. A key to this process is considering counterfactuals⁵⁰ – would a particular outcome occur had I not seen that cue, or had I not taken that action? In the present study, we show that TD learning models with ITI representations learn and predict value in the time before cue

presentation. The cue-associated TD error is then calculated as the difference in value in the presence and absence of that cue. Consequently, computation of TD errors effectively subtracts the prediction of value in the absence of the cue – i.e. the counterfactual prediction. More generally, the computation of TD error or its variants can be seen as subtracting out counterfactuals. In a class of RL algorithms commonly used in artificial intelligence applications (advantage actor-critic algorithms), the actor decides which action to take for a given state and the critic evaluates the action by computing the advantage function, defined as:

$$A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$$

where $Q(s_t, a_t)$ is the state-action value function¹⁷⁷. If this is taken to be the immediate reward of the action plus the expected return of the new state, $Q(s_t, a_t) = r_t + \gamma V(s_{t+1})$ then the advantage function can be approximated by the TD error $A(s_t, a_t) = Q(s_t, a_t) - V(s_t) = (r_t + \gamma V(s_{t+1})) - V(s_t) = E[\delta_t]$ (ref.^{178,179}).

As discussed in recent work¹⁷⁹, in fully observable environments without confounds, the advantage function is exactly equivalent to the Neyman-Rubin definition of causal effect of an action: the difference in outcomes given an action versus outcomes otherwise. In this context, the definitions of causality, contingency and TD error align – all emphasizing the consideration of counterfactual prediction: that is, the difference between potential future outcomes (following action) and the alternative when the action is not taken. TD error can therefore be both a measure of contingency and useful in establishing causal relationships, without invoking retrospective computations.

TD errors improve over the ANCCR and ΔP definitions because the comparison to the reward probability of US given CS is not simply the reward probability given absence of CS, but to $V(s)$, which is the γ -discounted sum of all future rewards given the current state, with the state encapsulating all environmental information. As demonstrated by our modeling, the heterogeneous state representation during the absence of events (the ITI) is critical to the accuracy of our models to match the experimental data.

While these relationships between TD error and contingency hold in fully observable environment, our value-RNN approach may extend these results to more complex/realistic environments. Veitch et al. (2019) has demonstrated network embeddings, like our value-RNN, can reduce the problem of inferring causality to a problem of predicting outcomes¹⁸⁰. These networks do not require full knowledge of the environment to succeed but rather learn to extract sufficient information to establish causality. Ultimately, TD error could provide pivotal signals for contingency – the essential quantity for causal inference.

Dopamine signaling as a candidate for modulating odor contingency-dependent representation in OT

What is the mechanism by which neural representations in the OT are modulated during contingency degradation? The prevailing view in the literature suggests that various regions within the mammalian brain reflect contingency^{3,4,6,57,60,181}, with some regions being essential for its detection⁵. This potentially leads to the hypothesis that the representations in the OT may be relayed from these areas through efferent connections. Nonetheless, here I propose an alternative hypothesis: the modulation within the OT is driven by a simple tripartite plasticity rule involving dopamine as well as pre- and post-synaptic neurons. As many studies have reported, the dopamine input from the VTA and SNc is critical to normal functioning of the striatum: the striatum responds to fluctuations in DA signaling which changes striatal function^{155,182–185}. It has been specifically noted that VTA-OT projecting neurons are important in forming odor preferences (REF). Therefore, to investigate the role of dopamine in modulating OT neural representation further, optogenetic tools could be utilized to modulate dopaminergic input to the OT, examining its influence on contingency-dependent activity. By replacing the unpredicted background reward with optogenetic stimulation of dopamine fibers that innervate the OT, this manipulation will enable examining whether pure dopamine could elicit similar degradation effect in terms of subsequent behavioral adjustments and neural responses to rewarded odor. Furthermore, to test the necessity of dopaminergic modulation in the representation of contingency, dopaminergic activity within the OT could be inhibited concurrently with the delivery of the background reward during contingency degradation phases. This will provide insight into the role of dopamine in contingency learning.

An interesting hypothesis to test is that contingency learning depends on temporal relationship between reward-evoked dopamine and input from the olfactory bulb through a mechanism of spike timing-dependent plasticity (STDP) like bidirectional modulation on synaptic connection. Studies have posited that dopamine may act as a gate for synaptic plasticity^{182,184,186}, providing a biological mechanism for updating contingency based on error signals. The role of dopamine in the striatum may not be simply defined as reinforcing or depressing, as some in vitro studies suggested that the induction of long-term potentiation or long-term depression will be biased by the physiological states ('up' and 'down' states) of the striatal neurons¹⁸⁴. Previous study in *Drosophila* has shown that dopamine's action has bidirectional synaptic modulation in the mushroom body based on different temporal relation between dopamine stimulation and Kenyon cell presynaptic activation¹⁸⁷. It is interesting to investigate dopamine's temporal sensitivity during contingency learning: 1) whether there is evidence for STDP-like, dopamine-dependent bidirectional modulation of synaptic strength at the behavioral timescale, and 2) how the temporal intervals between dopamine stimulation and pre- and post-synaptic activities in OT D1/D2 type neurons influence hetero-synaptic connectivity. These questions underscore the need for a nuanced understanding of the interplay between temporal dynamics and synaptic plasticity in neural circuits underlying contingency learning.

Potential neural populations that encode ISI value prediction during contingency degradation

In light of the observations that OT neurons modify their response when contingencies are altered, a critical question arises: which brain areas sustain the prediction of odor valence during the inter-stimulus interval (ISI) as postulated by our TD learning model in Chapter I? Regions implicated in higher-order cognitive functions, such as the orbitofrontal cortex (OFC)^{57,137}, anterior cingulate⁵⁷, basolateral amygdala⁶, and medial prefrontal cortex⁶⁰, are known to reflect contingency and adapt their neural activities accordingly. Among these, the OFC is particularly thought to mediate contingency learning by transferring cue-reward information to the VTA¹³⁷. Notably, a recent study has shown that certain OFC subpopulations retain a long-term memory of the original cue-reward association at both the individual neuronal and

collective population levels, even after the contingency changed, whereas other subpopulations adjust their activities to mirror the new contingencies¹³⁷.

Our preliminary data, albeit not presented in this dissertation, also indicate a similar dichotomy at the single-cell level in the OT: some D1- and D2-type neurons adjust their responses to the altered contingency, while others maintain their reaction to the rewarded cue post-degradation. This observation opens up one entry point for future research to dissect the neural substrates encoding the stable ISI value. Moreover, it highlights the importance of employing longitudinal neuronal tracking techniques, such as in vivo multi-photon recording, for two primary reasons: firstly, to analyze the evolution of responses during learning and the preservation of these learned responses, necessitating longitudinal monitoring of the same neuron sets across multiple days of behavioral learning; secondly, it allows the sampling of heterogeneity and understanding of spatial distribution by imaging a large amount of neurons which are cell-type specific or projection-defined. These insights open avenues to explore the complexity of OT neural representation and the predictive value that may originate from distinct neuronal subpopulations.

Together, these results provide new insights into the neural circuits involved in the representation of contingency in the mammalian brain. Further neuroscientific investigation will be needed to decipher the specific causal role of dopamine in detecting contingencies and determine the brain regions encoding these representations that influence behavior. Investigating diverse circuits within the neocortical, striatal, and limbic systems will enhance our comprehension of the neural mechanisms underpinning contingency, a vital element of an animal's dynamic strategy in response to everchanging world for survival.

References

1. Everitt, B. J. & Robbins, T. W. Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. *Annu Rev Psychol* **67**, 23–50 (2016).
2. Allan, L. G., Hannah, S. D., Crump, M. J. C. & Siegel, S. The Psychophysics of Contingency Assessment. *J Exp Psychol Gen* **137**, 226–243 (2008).
3. Chudasama, Y. *et al.* The role of the anterior cingulate cortex in choices based on reward value and reward contingency. *Cereb Cortex* **23**, 2884–2898 (2013).
4. Balleine, B. W. & Dickinson, A. Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407–419 (1998).
5. Lex, B. & Hauber, W. Disconnection of the Entorhinal Cortex and Dorsomedial Striatum Impairs the Sensitivity to Instrumental Contingency Degradation. *Neuropsychopharmacology* **35**, 1788–1796 (2010).
6. Bermudez, M. A. & Schultz, W. Responses of amygdala neurons to positive reward-predicting stimuli depend on background reward (contingency) rather than stimulus-reward pairing (contiguity). *J Neurophysiol* **103**, 1158–1170 (2010).
7. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science (1979)* **275**, 1593–1599 (1997).
8. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural Circuitry of Reward Prediction Error. <https://doi.org/10.1146/annurev-neuro-072116-031109> **40**, 373–394 (2017).

9. Watabe-Uchida, M. & Uchida, N. Multiple Dopamine Systems: Weal and Woe of Dopamine. *Cold Spring Harb Symp Quant Biol* (2019) doi:10.1101/sqb.2018.83.037648.
10. Daw, N. D. *Advanced Reinforcement Learning. Neuroeconomics: Decision Making and the Brain: Second Edition* (Elsevier Inc., 2013). doi:10.1016/B978-0-12-416008-8.00016-4.
11. Daw, N. D. & Doya, K. The computational neurobiology of learning and reward. *Curr Opin Neurobiol* **16**, 199–204 (2006).
12. Sutton, Richard, S. Learning to Predict by the Method of Temporal Differences. *Mach Learn* **3**, 9–44 (1988).
13. Sutton, R. & Barto, A. *Introduction to Reinforcement Learning*. (1998).
14. de Jong, J. W. *et al.* A Neural Circuit Mechanism for Encoding Aversive Stimuli in the Mesolimbic Dopamine System. *Neuron* **101**, 133-151.e7 (2019).
15. Wesson, D. W. & Wilson, D. A. Sniffing out the contributions of the olfactory tubercle to the sense of smell: hedonics, sensory integration, and more? *Neurosci Biobehav Rev* **35**, 655–668 (2011).
16. Zhang, Z. *et al.* Activation of the dopaminergic pathway from VTA to the medial olfactory tubercle generates odor-preference and reward. *Elife* **6**, (2017).
17. FitzGerald, B. J., Richardson, K. & Wesson, D. W. Olfactory tubercle stimulation alters odor preference behavior and recruits forebrain reward and motivational centers. *Front Behav Neurosci* **8**, (2014).
18. Ikemoto, S. Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens-olfactory tubercle complex. *Brain Res Rev* **56**, 27–78 (2007).
19. Ikemoto, S. Involvement of the olfactory tubercle in cocaine reward: intracranial self-administration studies. *J Neurosci* **23**, 9305–9311 (2003).

20. Gadziola, M. A. & Wesson, D. W. The neural representation of goal-directed actions and outcomes in the ventral striatum's olfactory tubercle. *Journal of Neuroscience* **36**, 548–560 (2016).
21. Gadziola, M. A., Tylicki, K. A., Christian, D. L. & Wesson, D. W. The olfactory tubercle encodes odor valence in behaving mice. *J Neurosci* **35**, 4515–4527 (2015).
22. Millman, D. J. & Murthy, V. N. Rapid Learning of Odor-Value Association in the Olfactory Striatum. *J Neurosci* **40**, 4335–4347 (2020).
23. Martiros, N., Kapoor, V., Kim, S. E. & Murthy, V. N. Distinct representation of cue-outcome association by D1 and D2 neurons in the ventral striatum's olfactory tubercle. *Elife* **11**, (2022).
24. Cansler, H. L., Wright, K. N., Stetzik, L. A. & Wesson, D. W. Neurochemical organization of the ventral striatum's olfactory tubercle. *J Neurochem* **152**, 425–448 (2020).
25. Xiong, A. & Wesson, D. W. Illustrated review of the ventral striatum's olfactory tubercle. *Chem Senses* **41**, 549–555 (2016).
26. Ashwell, K. W. Cyto- and chemoarchitecture of the monotreme olfactory tubercle. *Brain Behav Evol* **67**, 85–102 (2006).
27. RICHARDNEWMAN, S. An Experimental Study of the Ventral Striatum of the Golden Hamster. 1. Neuronal Connections of the Nucleus Accumbens. *J Comp Neurol* 191~167-192 (1980).
28. Ikemoto, S. Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens-olfactory tubercle complex. *Brain Res Rev* **56**, 27–78 (2007).
29. Wesson, D. W. & Wilson, D. A. Smelling sounds: Olfactory-auditory sensory convergence in the olfactory tubercle. *Journal of Neuroscience* **30**, 3013–3021 (2010).

30. Murata, K., Kanno, M., Ieki, N., Mori, K. & Yamaguchi, M. Mapping of Learned Odor-Induced Motivated Behaviors in the Mouse Olfactory Tubercle. *J Neurosci* **35**, 10581–10599 (2015).
31. Gadziola, M. A. & Wesson, D. W. The neural representation of goal-directed actions and outcomes in the ventral striatum's olfactory tubercle. *Journal of Neuroscience* **36**, 548–560 (2016).
32. Zhang, Z. *et al.* Whole-Brain Mapping of the Inputs and Outputs of the Medial Part of the Olfactory Tubercle. *Front Neural Circuits* **11**, 52 (2017).
33. Zhang, Z. *et al.* Activation of the dopaminergic pathway from VTA to the medial olfactory tubercle generates odor-preference and reward. *Elife* **6**, (2017).
34. Hall, G. Pavlovian Conditioning. *Animal Learning and Cognition* 15–43 (1994) doi:10.1016/b978-0-08-057169-0.50008-5.
35. Rescorla, R. A. Behavioral studies of Pavlovian conditioning. *Ann. Rev. Neurosci* (1988).
36. Thorndike, E. L. Animal intelligence; experimental studies. *Animal intelligence; experimental studies* (2012) doi:10.5962/BHL.TITLE.55072.
37. Wagner, R. A. R. & A. R. A theory of classical conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory* **21**, 64–99 (1972).
38. Wasserman, E. A. , and R. R. M. What's elementary about associative learning? *Annual review of psychology* **48**, 573–607 (1997).
39. Miller, R. R. , R. C. B. and N. J. G. Assessment of the Rescorla-Wagner model. *Psychol Bull* **117**, 363 (1995).
40. Pearce, J. M. , and M. E. B. Theories of associative learning in animals. *Annual review of psychology* **52**, 111–139 (2001).

41. Gershman, S. J. A Unifying Probabilistic View of Associative Learning. *PLoS Comput Biol* **11**, e1004567 (2015).
42. Rescorla, R. A. PROBABILITY OF SHOCK IN THE PRESENCE AND ABSENCE OF CS IN FEAR CONDITIONING. *J Comp Physiol Psychol* **66**, 1–5 (1968).
43. Gibbon, J., Berryman, R. & Thompson, R. L. Contingency Spaces and Measures in Classical and Instrumental Conditioning 1. *J Exp Anal Behav* **21**, 585–605 (1974).
44. Hallam, S. C., Grahame, N. J. & Miller, R. R. *Exploring the Edges of Pavlovian Contingency Space: An Assessment of Contingency Theory and Its Various Metrics. LEARNING AND MOTIVATION* vol. 23 (1992).
45. Cheng, P. W. From covariation to causation: A causal power theory. *Psychol Rev* **104**, 367–405 (1997).
46. Gallistel, C. R., Craig, A. R. & Shahan, T. A. Contingency, Contiguity, and Causality in Conditioning: Applying Information Theory and Weber’s Law to the Assignment of Credit Problem. *Psychol Rev* (2019) doi:10.1037/rev0000163.
47. Jenkins, H. M. & Ward, W. C. *Psychological Monographs: General and Applied JUDGMENT OF CONTINGENCY BETWEEN RESPONSES AND OUTCOMES I*. vol. 79 (1965).
48. Allan, L. G. Human contingency judgments: Rule based or associative? *Psychol Bull* **114**, 435–448 (1993).
49. Griffiths, T. L. & Tenenbaum, J. B. Structure and strength in causal induction☆. *Cogn Psychol* **51**, 334–384 (2005).
50. Pearl, J. Causality: Models, reasoning, and inference, second edition. *Causality: Models, Reasoning, and Inference, Second Edition* 1–464 (2011) doi:10.1017/CBO9780511803161.

51. Balsam, P. D. & Gallistel, C. R. Temporal maps and informativeness in associative learning. *Trends Neurosci* **32**, 73–78 (2009).
52. Papini, M. R. & Bitterman, M. E. *The Role of Contingency in Classical. Rychological Rminu* vol. 97 (1990).
53. RESCORLA, R. A. PAVLOVIAN CONDITIONING AND ITS PROPER CONTROL PROCEDURES. *Psychol Rev* **74**, 71–80 (1967).
54. A., R. R. A theory of Pavlovian conditioning : Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory* **2**, 64–69 (1972).
55. RESCORLA, R. A. PROBABILITY OF SHOCK IN THE PRESENCE AND ABSENCE OF CS IN FEAR CONDITIONING. *J Comp Physiol Psychol* **66**, 1–5 (1968).
56. Lex, B. & Hauber, W. Disconnection of the Entorhinal Cortex and Dorsomedial Striatum Impairs the Sensitivity to Instrumental Contingency Degradation. *Neuropsychopharmacology* **35**, 1788–1796 (2010).
57. Jackson, S. A., Horst, N. K., Pears, A., Robbins, T. W. & Roberts, A. C. Role of the Perigenual Anterior Cingulate and Orbitofrontal Cortex in Contingency Learning in the Marmoset. *Cereb Cortex* **26**, 3273–3284 (2016).
58. Vaghi, M. M. *et al.* Action-Outcome Knowledge Dissociates From Behavior in Obsessive-Compulsive Disorder Following Contingency Degradation. *Biol Psychiatry Cogn Neurosci Neuroimaging* **4**, 200–209 (2019).
59. Balleine, B. W. & Dickinson, A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407–419 (1998).

60. Coutureau, E., Esclassan, F., Di Scala, G. & Marchand, A. R. The role of the rat medial prefrontal cortex in adapting to changes in instrumental contingency. *PLoS One* **7**, (2012).
61. Dickinson, A., Balleine, B. W. & Sydney, U. *Causal Cognition and Goal-Directed Action*. <https://www.researchgate.net/publication/232603232> (2012).
62. Hammond, L. J. The effect of contingency upon the appetitive conditioning of free-operant behavior. *J Exp Anal Behav* **34**, 297–304 (1980).
63. Bush, R. R. , and F. M. Stochastic models for learning. (1955).
64. Gourley, S. L. *et al.* The orbitofrontal cortex regulates outcome-based decision-making via the lateral striatum. *European Journal of Neuroscience* **38**, 2382–2388 (2013).
65. Madarasz, T. J. *et al.* Evaluation of ambiguous associations in the amygdala by learning the structure of the environment. *Nat Neurosci* **19**, 965–972 (2016).
66. Wasserman, E. A., Chatlosh, D. L. & Neunaber, D. J. Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learn Motiv* **14**, 406–432 (1983).
67. Chatlosh, D. L., Neunaber, D. J. & Wasserman, E. A. Response-outcome contingency: Behavioral and judgmental effects of appetitive and aversive outcomes with college students. *Learn Motiv* **16**, 1–34 (1985).
68. Shanks, D. R. & Dickinson, A. Associative Accounts of Causality Judgment. *Psychology of Learning and Motivation - Advances in Research and Theory* **21**, 229–261 (1988).
69. Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning 1988 3:1* **3**, 9–44 (1988).
70. Sutton, R. S. & Barto, A. G. *Time-Derivative Models of Pavlovian Reinforcement*. (1990).

71. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science (1979)* **275**, 1593–1599 (1997).
72. Daw, N. D., Tourtezky, D. S., Daw, N., Courville, A. & Touretzky, D. Representation and Timing in Theories of the Dopamine System. *Neural Comput* **18**, 1637–1677 (2006).
73. Mirenowicz, J. & Schultz, W. Importance of unpredictability for reward responses in primate dopamine neurons. *J. Neurophysiol.* **72**, 1024–1027 (1994).
74. Schultz, W. Predictive reward signal of dopamine neurons. *J Neurophysiol* **80**, 1–27 (1998).
75. Schultz, W. & Dickinson, A. Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* **23**, 473–500 (2000).
76. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science (1979)* **275**, 1593–1599 (1997).
77. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
78. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science (1979)* **275**, 1593–1599 (1997).
79. Amo, R. *et al.* A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nature Neuroscience* 2022 25:8 **25**, 1082–1092 (2022).
80. Kim, H. R. *et al.* A Unified Framework for Dopamine Signals across Timescales. *Cell* **183**, 1600-1616.e25 (2020).
81. Jeong, H. *et al.* Mesolimbic dopamine release conveys causal associations. *Science (1979)* **378**, (2022).
82. Parent, A. *et al.* Organization of the basal ganglia: The importance of axonal collateralization. *Trends Neurosci* **23**, (2000).

83. Prensa, L. & Parent, A. The nigrostriatal pathway in the rat: A single-axon study of the relationship between dorsal and ventral tier nigral neurons and the striosome/matrix striatal compartments. *Journal of Neuroscience* **21**, 7247–7260 (2001).
84. Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. *Trends Neurosci* **12**, 366–375 (1989).
85. Medina, L. & Reiner, A. Neurotransmitter Organization and Connectivity of the Basal Ganglia in Vertebrates: Implications for the Evolution of Basal Ganglia (Part 2 of 2). *Brain Behav Evol* **46**, 247–258 (1995).
86. Stephenson-Jones, M. Pallidal circuits for aversive motivation and learning. *Curr Opin Behav Sci* **26**, 82–89 (2019).
87. Schneider, G. Brain structure and its origins. *Angewandte Chemie International Edition*, 6(11), 951–952. 1–20 (2013).
88. Graybiel, A. M. Habits, Rituals, and the Evaluative Brain. <https://doi.org/10.1146/annurev.neuro.29.051605.112851> **31**, 359–387 (2008).
89. Redgrave, P., Vautrelle, N. & Reynolds, J. N. J. Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. *Neuroscience* **198**, 138–151 (2011).
90. Smith, Y., James, D., Redgrave, P. & Kimura, M. Symposium Thalamic Contributions to Basal Ganglia-Related Behavioral Switching and Reinforcement. (2011) doi:10.1523/JNEUROSCI.4634-11.2011.
91. Van der Meer, M. A. A. & Redish, A. D. Ventral striatum: A critical look at models of learning and evaluation. *Curr Opin Neurobiol* **21**, 387–392 (2011).
92. Brog, J. S., Salyapongse, A., Deutch, A. Y. & Zahm, D. S. The patterns of afferent innervation of the core and shell in the “Accumbens” part of the rat ventral striatum: Immunohistochemical detection of retrogradely transported fluoro-gold. *Journal of Comparative Neurology* **338**, 255–278 (1993).

93. Stuber, G. D., Hnasko, T. S., Britt, J. P., Edwards, R. H. & Bonci, A. Dopaminergic terminals in the nucleus accumbens but not the dorsal striatum corelease glutamate. *Journal of Neuroscience* **30**, 8229–8233 (2010).
94. Saunders, B. T. & Robinson, T. E. The role of dopamine in the accumbens core in the expression of pavlovian-conditioned responses. *European Journal of Neuroscience* **36**, 2521–2532 (2012).
95. Ikemoto, S. & Panksepp, J. The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. *Brain Res. Brain Res. Rev.* **31**, 6–41 (1999).
96. Wightman, R. M. *et al.* Dopamine release is heterogeneous within microenvironments of the rat nucleus accumbens. *Eur J Neurosci* **26**, 2046–2054 (2007).
97. Collins, A. L. & Saunders, B. T. Heterogeneity in striatal dopamine circuits: Form and function in dynamic reward seeking. *J Neurosci Res* **1**, 1–24 (2020).
98. Aragona, B. J. *et al.* Regional specificity in the real-time development of phasic dopamine transmission patterns during acquisition of a cue-cocaine association in rats. *Eur J Neurosci* **30**, 1889–1899 (2009).
99. Wright, K. N. & Wesson, D. W. The tubular striatum and nucleus accumbens distinctly represent reward-taking and reward-seeking. *J Neurophysiol* **125**, 166–183 (2021).
100. Yamaguchi, M. Functional Sub-Circuits of the Olfactory System Viewed from the Olfactory Bulb and the Olfactory Tubercle. *Front Neuroanat* **11**, 33 (2017).
101. Purves D, A. G., Fitzpatrick D, et al. Neuroscience: Central Projections of the Olfactory Bulb. 2nd edn, (Sinauer Associates, 2001). - Google Search.
102. Igarashi, K. M. *et al.* Parallel Mitral and Tufted Cell Pathways Route Distinct Odor Information to Different Targets in the Olfactory Cortex. *Journal of Neuroscience* **32**, 7970–7985 (2012).

103. Zhou, L., Furuta, T. & Kaneko, T. Chemical organization of projection neurons in the rat accumbens nucleus and olfactory tubercle. *Neuroscience* **120**, 783–798 (2003).
104. Zhang, Z. *et al.* Whole-brain mapping of the inputs and outputs of the medial part of the olfactory tubercle. *Front Neural Circuits* **11**, 1–13 (2017).
105. Gadziola, M. A., Tylicki, K. A., Christian, D. L. & Wesson, D. W. The olfactory tubercle encodes odor valence in behaving mice. *Journal of Neuroscience* **35**, 4515–4527 (2015).
106. Wesson, D. W. & Wilson, D. A. Sniffing out the contributions of the olfactory tubercle to the sense of smell: Hedonics, sensory integration, and more? *Neurosci Biobehav Rev* **35**, 655–668 (2011).
107. Pavlovian conditioning: It's not what you think it is. <https://psycnet.apa.org/record/1988-19216-001>.
108. Gershman, S. J. & Ullman, T. D. Causal implicatures from correlational statements. *PLoS One* **18**, (2023).
109. Balsam, P. D. & Gallistel, C. R. Temporal maps and informativeness in associative learning. *Trends Neurosci* **32**, 73–78 (2009).
110. Gallistel, C. R., Fairhurst, S. & Balsam, P. The learning curve: Implications of a quantitative analysis. *Proc Natl Acad Sci U S A* **101**, 13124–13131 (2004).
111. Pearce, J. M. & Bouton, M. E. Theories of Associative Learning in Animals. <https://doi.org/10.1146/annurev.psych.52.1.111> **52**, 111–139 (2003).
112. Learning and Behavior - Mark E. Bouton - Oxford University Press. <https://global.oup.com/academic/product/learning-and-behavior-9780878933853?cc=us&lang=en&>.
113. Gershman, S. J. Context-dependent learning and causal structure. *Psychon Bull Rev* **24**, 557–565 (2017).

114. Mnih, V. *et al.* Human-level control through deep reinforcement learning. (2015)
doi:10.1038/nature14236.
115. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward.
Science (1979) **275**, 1593–1599 (1997).
116. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific
signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88
(2012).
117. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural Circuitry of Reward Prediction
Error. *Annu Rev Neurosci* **40**, 373–394 (2017).
118. Niv, Y. Reinforcement learning in the brain. *J Math Psychol* **53**, 139–154 (2009).
119. Garr, E. *et al.* Mesostriatal dopamine is sensitive to specific cue-reward contingencies.
doi:10.1101/2023.06.05.543690.
120. K Namboodiri, V. M. & Stuber, G. D. The learning of prospective and retrospective
cognitive maps within neural circuits. *Neuron* vol. 109 3552–3575 Preprint at
<https://doi.org/10.1016/j.neuron.2021.09.034> (2021).
121. Kim, C. K. *et al.* simultaneous fast measurement of circuit dynamics at multiple sites
across the mammalian brain. **13**, 325 (2016).
122. Bäckman, C. M. *et al.* Characterization of a mouse strain expressing Cre recombinase
from the 3' untranslated region of the dopamine transporter locus. *Genesis* **44**, 383–
390 (2006).
123. Daigle, T. L. *et al.* A suite of transgenic driver and reporter mouse lines with enhanced
brain cell type targeting and functionality. *Cell* **174**, 465 (2018).
124. 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**, (2017).

125. Akam, T., Costa, R. & Dayan, P. Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. *PLoS Comput Biol* **11**, e1004648 (2015).
126. Takahashi, Y. K. *et al.* Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat Neurosci* **14**, 1590–1597 (2011).
127. Starkweather, C. K. & Uchida, N. Dopamine signals as temporal difference errors: recent advances. *Curr Opin Neurobiol* **67**, 95–105 (2021).
128. Takahashi, Y. K. *et al.* Dopaminergic prediction errors in the ventral tegmental area reflect a multithreaded predictive model. *Nat Neurosci* **26**, (2023).
129. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science (1979)* **275**, 1593–1599 (1997).
130. Kobayashi, S. & Schultz, W. Influence of reward delays on responses of dopamine neurons. *J Neurosci* **28**, 7837–7846 (2008).
131. Fiorillo, C. D., Newsome, W. T. & Schultz, W. The temporal precision of reward prediction in dopamine neurons. **11**, (2008).
132. Enomoto, K. *et al.* Dopamine neurons learn to encode the long-term value of multiple future rewards. doi:10.1073/pnas.1014457108.
133. Masset, P., Tano, P., Kim, H. R., Malik, A. N. & Uchida, N. Multi-timescale reinforcement learning in the brain. doi:10.1101/2023.11.12.566754.
134. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat Neurosci* **20**, 581–589 (2017).
135. Starkweather, C. K., Gershman, S. J. & Uchida, N. The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. *Neuron* **98**, 616–629.e6 (2018).

136. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* 2012 482:7383 **482**, 85–88 (2012).
137. Namboodiri, V. M. K. *et al.* Single-cell activity tracking reveals that orbitofrontal neurons acquire and maintain a long-term memory to guide behavioral adaptation. *Nat Neurosci* **22**, 1110–1121 (2019).
138. Schultz, W. Dopamine reward prediction-error signalling: a two-component response. (2016) doi:10.1038/nrn.2015.26.
139. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat Neurosci* **20**, 581–589 (2017).
140. Lak, A., Nomoto, K., Keramati, M., Sakagami, M. & Kepecs, A. Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. *Curr Biol* **27**, 821–832 (2017).
141. Bach, F. R. & Jordan, M. I. Kernel Independent Component Analysis. *Journal of Machine Learning Research* **3**, 1–48 (2002).
142. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat Neurosci* **18**, 1025–1033 (2015).
143. Wagner, A. R. *A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement.* <https://www.researchgate.net/publication/233820243> (2014).
144. Niv, Y. Reinforcement learning in the brain. *J Math Psychol* **53**, 139–154 (2008).
145. Namboodiri, V. M. K. How do real animals account for the passage of time during associative learning? *Behavioral neuroscience* **136**, 383 (2022).

146. Piquet, R., Faugère, A. & Parkes, S. L. A hippocampo-cortical pathway detects changes in the validity of an action as a predictor of reward. *Curr Biol* **34**, 24-35.e4 (2023).
147. Delamater, A. R., Siegel, D. B. & Tu, N. C. Learning about reward identities and time. *Behavioural processes* **207**, (2023).
148. Ache, B. W. & Young, J. M. Olfaction: Diverse species, conserved principles. *Neuron* **48**, 417–430 (2005).
149. Kaas, J. H. The evolution of the complex sensory and motor systems of the human brain. *Brain Res Bull* **75**, 384–390 (2008).
150. Hagamen, T. C., Greeley, H. P., Hagamen, W. D. & Reeves, A. G. Behavioral asymmetries following olfactory tubercle lesions in cats. *Brain Behav Evol* **14**, 241–250 (1977).
151. Koob, G. F., Riley, S. J., Smith, S. C. & Robbins, T. W. Effects of 6-hydroxydopamine lesions of the nucleus accumbens septi and olfactory tubercle on feeding, locomotor activity, and amphetamine anorexia in the rat. *J Comp Physiol Psychol* **92**, 917–927 (1978).
152. Koob, G. F. & Volkow, N. D. Neurocircuitry of addiction. *Neuropsychopharmacology* **35**, 217–238 (2010).
153. Wesson, D. W. & Wilson, D. A. Smelling sounds: olfactory-auditory sensory convergence in the olfactory tubercle. *J Neurosci* **30**, 3013–3021 (2010).
154. Bolam, Hanley, Booth & Bevan. <Bolam_et al _JAnat_2000_Synaptic organisation of the basal ganglia.pdf>. **44**, 527–542 (2000).
155. Gerfen, C. R. & Surmeier, D. J. Modulation of striatal projection systems by dopamine. *Annu Rev Neurosci* **34**, 441–466 (2011).
156. Giovannucci, A. *et al.* CalmAn an open source tool for scalable calcium imaging data analysis. *Elife* **8**, (2019).

157. Rubin, A., Geva, N., Sheintuch, L. & Ziv, Y. Hippocampal ensemble dynamics timestamp events in long-term memory. *Elife* **4**, (2015).
158. Rotenberg, A., Abel, T., Hawkins, R. D., Kandel, E. R. & Muller, R. U. Parallel Instabilities of Long-Term Potentiation, Place Cells, and Learning Caused by Decreased Protein Kinase A Activity. (2000).
159. Driscoll, L. N., Duncker, L. & Harvey, C. D. Representational drift: Emerging theories for continual learning and experimental future directions. *Curr Opin Neurobiol* **76**, (2022).
160. Marks, T. D. & Goard, M. J. Stimulus-dependent representational drift in primary visual cortex. *Nature Communications* 2021 12:1 **12**, 1–16 (2021).
161. Schoonover, C. E., Ohashi, S. N., Axel, R. & Fink, A. J. P. Representational drift in primary olfactory cortex. *Nature* **594**, 541–546 (2021).
162. Geva, N., Deitch, D., Rubin, A. & Ziv, Y. Time and experience differentially affect distinct aspects of hippocampal representational drift. *Neuron* **111**, 2357-2366.e5 (2023).
163. Kentros, C. G., Agnihotri, N. T., Streater, S., Hawkins, R. D. & Kandel, E. R. Increased attention to spatial context increases both place field stability and spatial memory. *Neuron* **42**, 283–295 (2004).
164. Ziv, Y. *et al.* Long-term dynamics of CA1 hippocampal place codes. *Nat Neurosci* **16**, 264–266 (2013).
165. Rule, M., O’Leary, T., neurobiology, C. H.-C. opinion in & 2019, undefined. Causes and consequences of representational drift. *Elsevier*.
166. Soucy, E. R., Albeanu, D. F., Fantana, A. L., Murthy, V. N. & Meister, M. Precision and diversity in an odor map on the olfactory bulb. *Nat Neurosci* **12**, 210–220 (2009).
167. Uchida, N., Cury, K. & Soucy, E. Olfactometer.

168. Gire, D. H., Kapoor, V., Arrighi-Allisan, A., Seminara, A. & Murthy, V. N. Mice develop efficient strategies for foraging and navigation using complex natural stimuli. *Current Biology* **26**, 1261–1273 (2016).
169. Petzold, G. C., Hagiwara, A. & Murthy, V. N. Serotonergic modulation of odor input to the mammalian olfactory bulb. *Nat Neurosci* **12**, 784–791 (2009).
170. Pnevmatikakis, E. A. & Giovannucci, A. NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. *J Neurosci Methods* **291**, 83–94 (2017).
171. Dickinson, A. Goal-directed behavior and future planning in animals. *Animal Thinking: Contemporary Issues in Comparative Cognition* **8**, 79–91 (2011).
172. Schultz, W. Dopamine reward prediction error coding. *Dialogues Clin Neurosci* **18**, 23–32 (2016).
173. Sutton, R. S. & Barto, A. G. Reinforcement learning: An Introduction Second edition. *Learning* **3**, 322 (2012).
174. Engelhard, B. *et al.* Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* **570**, 509–513 (2019).
175. Menegas, W. *et al.* Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *Elife* **4**, e10032 (2015).
176. Oetl, L. L. *et al.* Phasic dopamine reinforces distinct striatal stimulus encoding in the olfactory tubercle driving dopaminergic reward prediction. *Nat Commun* **11**, (2020).
177. Baird, L. Residual Algorithms: Reinforcement Learning with Function Approximation. *Machine Learning Proceedings 1995* 30–37 (1995) doi:10.1016/b978-1-55860-377-6.50013-x.
178. Schulman, J., Moritz, P., Levine, S., Jordan, M. & Abbeel, P. High-Dimensional Continuous Control Using Generalized Advantage Estimation. (2015).

179. Pan, H.-R., Gürtler, N., Neitz, A. & Schölkopf, B. Direct Advantage Estimation. (2021).
180. Veitch, V., Wang, Y. & Blei, D. M. Using Embeddings to Correct for Unobserved Confounding in Networks. (2019).
181. Namboodiri, V. M. K. *et al.* Single-cell activity tracking reveals that orbitofrontal neurons acquire and maintain a long-term memory to guide behavioral adaptation. *Nat Neurosci* **22**, 1110–1121 (2019).
182. Reynolds, J. N. & Wickens, J. R. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw* **15**, 507–521 (2002).
183. Centonze, D., Picconi, B., Gubellini, P., Bernardi, G. & Calabresi, P. Dopaminergic control of synaptic plasticity in the dorsal striatum. *European Journal of Neuroscience* **13**, 1071–1077 (2001).
184. Calabresi, P., Picconi, B., Tozzi, A. & Di Filippo, M. Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci* **30**, 211–219 (2007).
185. Calabresi, P., Maj, R., Mercuri, N. B. & Bernardi, G. Coactivation of D1 and D2 dopamine receptors is required for long-term synaptic depression in the striatum. *Neurosci Lett* **142**, 95–99 (1992).
186. Keiflin, R. & Janak, P. H. Dopamine Prediction Errors in Reward Learning and Addiction: From Theory to Neural Circuitry. *Neuron* **88**, 247–263 (2015).
187. Handler, A. *et al.* Distinct Dopamine Receptor Pathways Underlie the Temporal Sensitivity of Associative Learning. *Cell* (2019) doi:10.1016/j.cell.2019.05.040.