2024

Signatures and Consequences of Distributional Reinforcement Learning

Margarida Nunes de Almeida Rodrigues de Sousa

Dissertation presented to obtain the **Ph.D degree in Neurosciences**

International Neurosciences Doctoral Programme

Oeiras, November, 2024

INSTITUTO DE TECNOLOGIA QUÍMICA E BIOLÓGICA ANTÓNIO XAVIER | UNIVERSIDADE NOVA DE LISBOA

Signatures and Consequences of Distributional Reinforcement Learning

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.2.1) [85].

Acknowledgements

I would like to start by thanking Joe. When I joined the lab, I knew little about neuroscience or the dynamics of a systems neuroscience lab and was filled with insecurities. You were always confident we would succeed, always cheerleading, and that support was crucial for me to keep going. Thank you for your endless desire of knowing more and doing better, for asking the right questions, for teaching me, for making me a less telegraphic scientist, for all the great and sometimes crazy ideas. I am really proud of what we ended up doing!

Thank you, Bruno. You played an essential role in my journey, taught me so much about neuroscience, and how to actually make things work in the end. Thanks for the nice dinners and walks, I am glad I met you.

Thank you, Dan—it was exciting to work on extending the distributional learning rules to multiple dimensions. I'm also grateful for the opportunity to join your lab meetings, where I learned so much and gained new insights.

A big thank you to Kenway. Our meetings were very helpful, and I'm grateful that you joined us on this adventure.

Thank you Pawel for performing the challenging dopamine recordings.

Over the years, I've come to see the lab as a kind of extended family, with people coming and going but staying connected.

Margarida, thank you for being such a cheerful, funny person, for always asking the right questions, and for being a good friend from the very start of my PhD—and ever since.

Flipe, thank you for sharing your very clear and at the same time detailed way of thinking about things, you know so much! Thank you for all the help,

for all the discussions, and for the friendship-I'm very lucky to have had you around.

Teresa, we have a kind of telepathic way of communicating. Thanks for always understanding what I am trying to say, and for your encouragement. You've taught me so much, thank you.

Sofia, thank you for your honest advices, your incredible organization and energy, and for your friendship.

Mauro, thank you for the coffee breaks, for the good spirit and for the Suerte!

Sofia CA, you are a really organized and determined person, it was really nice working with you.

Simon, thank you for creating a nice atmosphere around you; I'm glad we ended up as open-lab neighbors and friends.

Renato, thank you for the funny lunches.

Thank you Sofia, Margarida, Flipe, Teresa and Francesco for reading chapters of my thesis and giving important feedback.

Thank you, Pedro, for being there through all the challenging twists and turns along the way.

To my mother, thank you for teaching me that we have the power to achieve things as long as we try and fight for them. Thank you for all your support.

To Isabel, Ana and Fred thank you for always looking out for me and making sure I didn't get completely crazy.

To my father: thank you for always asking when I'd finish, when I'd publish, and whether I'd find a job afterwards.

Thank you to my nephews, Afonso and Miguel, whenever I see you, I always cheer up.

I thank Fundação Champalimaud and Fundação para a Ciência e Tecnologia (FCT) for the funding that made this work possible.

Resumo

A capacidade de prever recompensas é fundamental para o comportamento adaptativo, e os neurónios dopaminérgicos do mesencéfalo (DANs) desempenham um papel essencial neste processo ao sinalizarem erros de previsão de recompensa (RPEs), que informam circuitos receptores sobre as recompensas esperadas no contexto actual [114]. No entanto, a aprendizagem por reforço de diferença temporal (TD), associado aos DANs, aprende apenas a média das recompensas futuras esperadas com desconto temporal, desconsiderando informação sobre as distribuições de magnitude e atraso da recompensa [135]. Neste estudo, introduzimos o TMRL (aprendizagem por reforço de tempo-magnitude), uma variante multidimensional da aprendizagem por reforço distribucional que aprende a distribuição conjunta das recompensas futuras em função do tempo e da magnitude, com um código eficiente adaptado às estatísticas do ambiente.

Observámos, ainda, sinais de cálculos semelhantes ao TMRL na actividade de DANs identificados optogeneticamente em ratos. Identificámos uma diversidade significativa no desconto temporal e na sintonização para a magnitude da recompensa entre DANs, permitindo o cálculo de um mapa probabilístico bidimensional das recompensas futuras a partir de 450 ms de actividade neuronal em resposta a um estímulo preditiva de recompensa. As previsões temporais derivadas deste código correlacionaram-se com o tempo de comportamentos antecipatórios, sugerindo que esta informação orienta decisões sobre quando agir. Finalmente, simulações de comportamento em ambiente de forrageamento evidenciam as vantagens de aceder a uma distribuição conjunta de recompensa em função do tempo e magnitude

quando a recompensa é dinâmica ou os estados internos de necessidade fisiológica são dinâmicos. Estes resultados mostram a riqueza probabilística da informação de recompensa comunicada aos DANs e sugerem uma extensão local-no-tempo dos algoritmos de TD que explica a aquisição e processamento desta informação.

Palavras-chave: Aprendizagem por reforço, gânglios da base, dopamina, tempo, tomada de decisão

Abstract

Learning to predict rewards is fundamental for adaptive behavior. Midbrain dopamine neurons (DANs) play a key role in such learning by signaling reward prediction errors (RPEs) that teach recipient circuits about expected rewards given current circumstances and actions [114]. However, the algorithm that DANs are thought to provide a substrate for, temporal difference (TD) reinforcement learning (RL), learns the mean of temporally discounted expected future rewards, discarding useful information concerning experienced distributions of reward amounts and delays [135]. Here we present time-magnitude RL (TMRL), a multidimensional variant of distributional reinforcement learning that learns the joint distribution of future rewards over time and magnitude using an efficient code that adapts to environmental statistics. In addition, we discovered signatures of TMRL-like computations in the activity of optogenetically identified DANs in mice during behavior. Specifically, we found significant diversity in both temporal discounting and tuning for the magnitude of rewards across DANs, features that allow the computation of a two dimensional, probabilistic map of future rewards from just 450ms of neural activity recorded from a population of DANs in response to a reward-predictive cue. Furthermore, reward time predictions derived from this population code correlated with the timing of anticipatory behavior, suggesting that similar information is used to guide decisions regarding when to act. Finally, by simulating behavior in a foraging environment, we highlight benefits of access to a joint probability distribution of reward over time and magnitude in the face of dynamic reward landscapes and internal physiological need states. These findings demonstrate surprisingly rich

probabilistic reward information that is learned and communicated to DANs, and suggest a simple, local-in-time extension of TD learning algorithms that explains how such information may be acquired and computed.

Keywords: Reinforcement learning, basal ganglia, dopamine, timing, decision-making

Contents

List of Figures

List of Algorithms

1

General introduction

For an agent in an environment, the reinforcement learning (RL) problem is the general challenge of learning how to behave to achieve some goal. The agent, be it natural or artificial, does not have explicit instructions on how to map situations, or states, to actions; instead, it must learn such control policies through trial and error, receiving feedback from the environment only in the form of rewards derived from current objectives. In complex environments, existing dynamical structure can be leveraged if behavioral policies take temporal relationships between states, rewards and actions into account [39]. The underlying policies should be formed when states or actions reduce uncertainty about reward likelihood, magnitude and timing [5]. Animal behavior clearly reflects knowledge of when rewards are expected: rabbits blink when a noxious air puff is expected [104, 148]; rats become more fearful when a shock is expected [76, 25]; thirsty monkeys lick when a water droplet is expected [69]; hungry rats press a lever when it is expected to produce a food morsel [39], to name only a few examples. Furthermore, animals' behavior also reflects knowledge of the distribution of rewards in time. For example, rats adapt nearly immediately to changes in the rate of rewards [40]. Importantly, the speed with which an anticipatory response emerges is proportional to the informativeness of the temporal relation between a reward-predicting stimulus and the reward it predicts—the factor by which the onset of that stimulus shortens the expected interval to the next reward [39]. Additionally, animals also reflect knowledge of expected reward magnitudes: mice increase their licking rate when anticipating larger water droplets [90] and in matching tasks, monkeys allocate their responses

proportionally to the relative rates of reward associated with different options [134], to name only a few examples. Such an account for policy learning seems to require the brain to operate on distributions of rewards in time and magnitude [39].

However, the RL algorithms that have driven startling progress, in the neuroscience of learned behavioral control and in artificial intelligence alike, do not generally learn representations that encode distributions of rewards over time. Here, in Chapter 2 we propose an efficient coding algorithm capable of learning distributions over reward times and magnitudes and test for neural signatures in midbrain dopamine neurons. We then extend the theoretical framework and propose a general multidimensional distributional learning algorithm in Chapter 3.

1.1 Reinforcement Learning

To behave adaptively, animals must learn in complex and dynamic environments to produce actions that are good for them and avoid those that are not. The field of RL provides a normative theoretical framework for adaptive animal behavior [135]. In RL, an agent interacts with the environment by selecting actions, receiving rewards and observing new states, according to the environment transition probabilities.

Classical RL approaches are based on Markov Decision Processes (MDP), where the Markov property is satisfied, i.e. the next state and the reward only depend on the previous state and action, and not on additional information about earlier states or actions. The goal of the agent is to maximize the value function, or cumulative expected future rewards, at each state. Importantly, the agent is not told which actions to take, but instead must discover which actions yield the highest value through trial-and-error. Also, actions may affect not only immediate rewards but also future states and therefore future rewards. A policy is a mapping from states to actions, that defines the probabilities of choosing actions in given states.

There are multiple ways of solving this reward maximization problem: some strategies directly optimize the policy function (policy gradient methods), while others learn the value function and then compute the policy (value-based methods) [71]. Policy gradient methods are advantageous, compared to value-based RL, when the action space is high-dimensional; for example, in the case of an arm reaching a target point in space, where

the action space contains all possible joint positions and velocities. However, policy gradient methods frequently converge to local optimal policies. We will focus on value-based methods throughout this work, since we will study midbrain dopamine neurons (DANs), that are thought to encode a reward prediction error (RPE) consistent with value-based RL and a classical conditions task, where actions are low dimensional, as we will describe in detail later.

An essential aspect of learning to act in an environment is the ability to use past experience to predict the future consequences of an action in a given context. Predictions can be a compressed representation that discard detailed features of environmental structure and are specific to particular, behaviorally relevant events, such as rewards (model-free RL). Alternatively, predictions may represent a model of the world, learning the state-transitions and the reward of each state (model-based RL). The agent then has to plan, simulating future possible trajectories and selecting the optimal one, which can be computationally costly when the decision-tree is wide and/or deep.

Compared to model-based algorithms, model-free algorithms are more sample-efficient and cheaper to learn, at the cost of being less flexible when faced with dynamic environments [30, 26]. For example, when reward is depleted, model-free agents have to relearn the value function through trial-and-error. Model-based agents, on the other hand, can update the reward representation, and re-use the state transition probabilities. In general, early in learning animals are thought to employ a model-based algorithm to estimate value, and later shift towards a model-free algorithm, which is less computationally demanding and can serve well in stable environments [66, 52].

The value function can be learnt through a model-free temporal-difference (TD) algorithm that updates, at each time-step, the current estimate using a RPE, a teaching signal that encodes the inconsistency of the current estimate of the value function. When combined with deep learning, this algorithm outperformed human experts in games such as chess, go, and Atari video games [94].

The successor representation (SR) [27], on the other hand, balances efficiency and flexibility by learning a predictive map of the environment. This map summarizes long-range predictive relationships between states of the environment. The value, in this case, is computed simply as a linear combination of the learnt predictive features, the occupancy matrix, and the

reward function. Therefore, since the rewards and state-transitions are not compressed into a single value, the SR is flexible when the reward signal changes. However, when state-transitions change, they have to be re-learned [97]. Importantly, the SR can also be learnt through TD, where instead of updating it via a RPE, it is updated based on the state occupancy prediction error.

1.2 Reinforcement learning in the brain

In the brain, the basal ganglia (BG) are a group of subcortical nuclei that are highly conserved across vertebrate evolution. The BG receive input from nearly the entire cortical mantle, along with dopaminergic inputs from midbrain DANs, that convey RPEs. The convergence of cortical contextual information with dopamine-driven RPEs makes the BG a candidate substrate for RL, facilitating the mapping of states to actions [54, 29].

The hypothesis is that cortex provides information about both the state of the world and available actions to the input region of the BG, the striatum, which is thought to learn their corresponding value. The striatum influences the activity of the BG output through the direct and indirect pathways. The direct pathway sends inhibitory projections directly to the basal ganglia output nuclei, globus pallidus internal segment and substantia nigra pars reticulata (GPi/SNr). On the other hand, the indirect pathway sends inhibitory projections to the globus pallidus external segment (GPe), which inhibits the subthalamic nucleus (STN), which in turn sends excitatory projections to the GPi/SNr. The output nuclei GPi/SNr are tonically active and exert inhibitory influence over downstream thalamic and brainstem targets, that are thought to inform behavior policies.

A vast majority of neurons in the striatum consist of GABAergic medium spiny projection neurons (MSNs). Action selection is generated from a tight balance between the direct and indirect pathways: activation of the direct pathway medium spiny projection neurons (dMSNs) disinhibits their target structures, consequently promoting specific actions; activation of indirect pathway MSNs (iMSNs) exert a broad inhibitory effects on the targets of basal ganglia output, suppressing competing actions. This loop closes back to cortex via excitatory thalamo-cortical synapses, creating a feedback system that allows for continuous updates to behavior given new feedback from the environment.

Critically, it is thought that DANs in substantia nigra pars compacta and ventral tegmental area (SNc/VTA), that project densely to the striatum, encode a RPE signal [114, 98, 130, 18]. The RPE signal they carry modulates the cortico-striatal synapses [44]. In particular, the overall plasticity of dMSNs increases with an increase in DA (following events or actions that resulted in outcomes better than expected). Conversely, for iMSNs, higher DA values have the opposite effect, reducing the change in plasticity, whereas lower concentrations of DA (for events that were worse than expected) potentiate plasticity [57].

The BG anatomy is convergent: the number of cortical neurons projecting to the striatum was found to be two orders of magnitude greater [68] than the striatal neurons [68], which in turn are two orders of magnitude greater than GPi neurons [106]. This suggests that high-dimensional cortical information is re-coded into a lower-dimensional representation as it passes through the BG. This low-dimensional representation is shaped by RPEs signaled by midbrain DANs, ultimately converging on an efficient representation of the environment that reliably predicts reward. The ability to modulate the cortico-striatal synapses for future action selection, based on previous outcomes, is at the core of adaptive behavior.

1.3 Parallel basal ganglia loops

It has long been proposed that the BG are a set of parallel channels organized topographically on the basis of the input arriving from the cortical hierarchy [55, 36] that vary in degree of abstraction and temporal extendedness [101]. In particular, sensorimotor, associative, and limbic information passes through BG circuits in a largely segregated manner [1], projecting respectively to the dorsolateral, dorsomedial and ventral striatum. Computational models suggest that these different subregions of the striatum may serve distinct functions [13]. Specifically, it is hypothesized that the dorsal striatum functions as an actor, learning policies, while the ventral striatum functions as a critic, evaluating the actor's policy. The actor is believed to update its policy based on RPEs signaled by the critic. Within the dorsal striatum, the dorsomedial region is thought to be learning through model-based RL [120, 4, 157], whereas the dorsolateral region through model-free RL [154, 155].

1.4 Midbrain dopamine neurons encode more than a scalar reward prediction error

In the early stages of electrophysiology, only a handful of DANs were recorded at a time. These initial recordings revealed that the mean population phasic activity was modulated by reward magnitude [114], probability [35], delay [69, 110] and subjective value [77]. This led to the hypothesis that all neurons were encoding the same scalar RPE. Further supporting this idea, optogenetic activating and inhibiting these neurons drives learning that mimics the effects of positive and negative RPE based learning [130, 18].

More recently, many studies have suggested that DANs encode more information than the originally proposed scalar RPE. There is evidence that DANs also respond to new or surprising stimuli [82, 53, 47, 136], suggesting the encoding of a state prediction error, consistent with the estimation of a SR. Other evidence suggests the encoding of an action prediction error in the tail of the striatum [49]. Additionally, DANs in the tail of the striatum have been shown to respond phasically to threatening stimuli, contradicting the classic RPE hypothesis prediction of a negative RPE [93].

These results, together with the parallel architecture of the BG, suggest that a similar computation is being done in each circuit of the BG – a prediction error (PE) – but over different inputs, e.g. reward and state occupancy [79, 43]. Evidence against this model stems from the fact that the majority of DANs seem to respond to rewards; furthermore, the model assumes a labeled line where different PE DANs must project to particular targets, which seems biologically implausible. To address these issues, a feature-specific PE model has recently been proposed [81]. In this model, different circuits use a distributed set of features or representations of the environment in order to estimate value. These features are learnt through linear function approximation, using as teaching signal a feature-specific RPE. Summing each circuit's feature-specific RPE recovers the global RPE. On the other hand, conflicting new data suggest that the activity of DANs can not be explained by TD RL altogether [59, 109, 20].

1.5 Distributional reinforcement learning

Recently, TD algorithms have also been elaborated to express a multiplicity of learned value estimates that differ in their sensitivity to positive and

negative RPEs [24]. These value estimates converge to distinct statistics of the predicted reward distribution, in a theory termed distributional RL [10]. Importantly, having access to these distributed values allows for extracting information about the reward distributions. Evidence of distributional RL-like computations has been reported in midbrain DANs of both mice and primates [23, 100].

Behaviorally, having access to a distributed set of values allows for flexibly changing the attitudes towards risks [84]. A value estimate tuned to low statistics of the reward distribution leads to risk avoiding behaviors, one tuned to the mean, to risk neutral, and one tuned to high statistics of the distribution, to risk seeking behavior. In the field of neuroeconomics, behavioral tasks assess the level of risk of individuals, by estimating the "reward certain" equivalent of a lottery (an option where reward is generated with a given level of uncertainty) [127]. These studies revealed that the risk attitudes are dynamic. For example, monkeys are risk seekers for low rewards and risk avoiders for high rewards [127]. On the other hand, the wealth level modulates risk attitudes: thirstier monkeys, when the stakes are higher, are more risk averse [152]. Distributional RL theory provides an algorithmic understanding of how these risk attitudes can be dynamically modulated.

Distributional RL has also been shown to enhance the performance of deep RL agents on benchmark Atari 2600 video games, a benefit attributed to its effects on representation learning. By estimating a wider range of reward distribution statistics beyond just the expected value, distributional RL serves as an effective auxiliary task [87, 24]. Importantly, as in previously proposed standard TD learning frameworks [94], expected values are used to generate policies in these tasks. Understanding how this richer distributional representation can improve policies remains a critical question. Recently, a risk-sensitive distributional RL approach was introduced [140], along with its generalization — a stochastically dominant distributional RL framework [88] — which may offer new insights for how to use this information to generate behavior policies.

Importantly, the distribution of reward magnitudes was only shown to be present in the population of DANs at the time of reward delivery [24], when the behavior episode has finished. In this work, we showed that this information is also present at the beginning of the episode, when, in principle, it can be used to drive behavior.

1.6 Distributional reinforcement learning in time

Standard TD formulations learn value functions that encode expectations of temporally discounted future reward: delayed rewards are weighted less relative to immediate ones. This produces ambiguity in value representations regarding when future rewards are expected to arrive. To illustrate this ambiguity one may observe the same value corresponding to either a large magnitude but delayed reward, or a smaller magnitude but imminent reward. In principle, knowing in advance the range and likelihood of rewards available, as well as when they are likely to occur, could be useful for planning and for flexible behavior, particularly in the face of non-stationarity environment or dynamic animal internal state (e.g. hunger and satiety) [45].

A population code in which each neuron is tuned to a different timescale of rewards in the environment could enable decoding of information about future reward timing, similar to how distributional reinforcement learning enables the decoding of reward magnitude information. This could be implemented by incorporating a set of neurons with diverse temporal discounting profiles. In particular, since the value of a predictive stimulus reflects the sum of temporally discounted expected future rewards, knowing the population's range of temporal discounts would allow for decoding the dynamics of expected future rewards [142, 138]. Similarly, cells in the entorhinal cortex revealed a spectrum of time constants for encoding instead of future, past times [14].

The value of future rewards should be discounted when there is a risk they may not be obtained in the end. For example, when a squirrel caches an acorn for winter many hazards in the environment could prevent recovery of the nut: the squirrel may forget its location, a competitor may find it, or a fungus could infect it [132]. If the risk is known and equal to a constant hazard rate, reward discounting should follow an exponential decaying function [126]. On the other hand, animals need to perform behaviors over multiple time scales. Therefore, maintaining a bank of temporally discounted values enables dynamic adaptation to changes in the environmental hazard rate and evolving behavioral demands [74].

There are interesting behavior biases that may reflect the multiple temporal discounting reward learning system. For example, in contexts in which a single exponential discounting rate is optimal, humans learn to match that factor [116]. Also interestingly, there is evidence that pigeons, rodents,

monkeys and humans discount hyperbolically, and not exponentially, delayed reward [146]. And hyperbolic temporal discounting of rewards can be a consequence of uncertainty in the underlying hazard rate of the environment, leading to weighting multiple exponential discounted values [126, 34]. Additionally, exponential decreasing curves discriminate better shorter than longer times, which predicts scalar timing [142], i.e., the uncertainty in the the estimation of time is linearly proportional to the magnitude of time, a feature animals exhibit [147].

In this work, we investigated whether midbrain DANs heterogeneously discount delays to reward and whether they contain information about the reward distribution over time. Furthermore, if DANs are encoding a two-dimensional distributional reward code in both magnitude and time, then correcting for tuning diversity along one dimension should uncover the remaining tuning diversity in the other, which might otherwise be masked. Thus, we also examined whether DANs encode information about the joint distribution of reward magnitudes and times.

1.7 Efficient coding of rewards

The brain contains a finite number of neurons, each with limited firing capacity, imposing constraints on its ability to encode information. In the face of such constraints, efficient coding theory suggests that neurons adapt their tuning properties to match the statistical properties of the variables they represent, in a manner that maximizes the information of encoded signals [80, 41]. The classic example being the blowfly compound eye large monopolar cells, that fire proportionally to the distribution of contrast levels measured in natural scenes [80]. Efficient coding has been extensively studied in sensory systems [42, 123, 105], but has received comparatively less attention in the context of reward processing systems [99, 108].

In the context of the population of midbrain DANs, efficient coding predicts that the tuning function with respect to time and magnitude should adapt so as to maximize the encoding of reward time-magnitude information with respect to the current environment. In this work we tested this hypothesis, by estimating the tuning of DANs to different reward time and magnitude probability distributions.

At the behavioral level, results suggest that the distribution of reward timing in an environment may influence how future rewards are temporally

discounted. For example, in humans, temporal discount rates have been shown to decrease with age [48]: as people age and are exposed to longer reward timescales, they tend to discount future rewards less steeply than in earlier life stages. Studies on two closely related primates support a similar adaptation to environmental reward timing. In an intertemporal choice task, marmosets were found to wait significantly longer for food rewards than tamarins [133]. This difference may reflect their natural foraging behaviors: marmosets often consume tree gum, which requires patience as it slowly exudes, while tamarins primarily hunt insects, which demands quick action. From an evolutionary perspective, we hypothesize that tamarins, adapted to shorter food delays, exhibit steeper temporal discounting, while marmosets, accustomed to longer delays, exhibit shallower temporal discounting. Thus, these species-specific temporal preferences may be shaped by their ecological experiences.

Theoretically, we show in Chapter 3 that the distributional reinforcement learning rules converge to an efficient code; a similar result has been derived specifically for the magnitude domain [115]. A general multidimensional efficient coding framework has been proposed [123]; however, no learning rules have been derived. Moreover, the learning rules proposed in the distributional RL framework do not trivially generalize to high-dimensional stimuli, as discussed in Chapter 3. To address this gap, we propose multidimensional distributional learning rules. Recently, concurrent work has introduced similar algorithms for learning distributions of multivariate rewards [149, 159].

2

Dopamine neurons encode a multidimensional probabilistic map of future reward

2.1 Introduction

The field of RL provides a normative theoretical framework for adaptive animal behavior [135]. A core tenet of RL is that behaviors producing maximal expected future reward are the target of learning through interaction with the environment. Relatedly, associative learning, including learning to associate states and actions with future reward as is required by RL, can be viewed through the lens of statistical inference [28]. A major determinant of whether a given observation or action should be associated with future reward is the degree to which its taking place reduces uncertainty about whether, what magnitude of, and when rewards will occur. Such an account for associative learning would seem to require the brain to operate on distributions of events in time [39]. Furthermore, predicting when, and not just whether, behaviorally relevant events such as rewards will occur is often critical for survival. For example, crossing a desert to reach an oasis is only advisable if you can survive the duration of the trip. However, the RL algorithms that have driven startling progress, in the neuroscience of learned behavioral control and in artificial intelligence alike, do not generally learn value representations that encode distributions of rewards over time.

Midbrain dopamine neurons have figured prominently in theories of how RL-like functions may be performed within neural circuits. Specifically,

the phasic activity of midbrain DANs is thought to encode TD RPEs, which serve as teaching signals that are used to update the value of states or actions so as to inform appropriate programs, or policies, for behavioral control [98]. However, standard TD formulations learn value functions that encode expectations of, temporally discounted, future reward: delayed rewards are weighted less relative to immediate ones. This produces ambiguity in value representations regarding when future rewards are expected to arrive. To illustrate this ambiguity one may observe the same value corresponding to either a large magnitude but delayed reward, or a smaller magnitude but imminent reward (Figure 2.1A). In addition, because simple TD learning algorithms learn the average of temporally discounted future reward, they do not learn about the distribution of reward magnitudes. Recently, TD algorithms have been elaborated to express a multiplicity of learned value estimates that differ in their sensitivity to positive and negative RPEs. These value estimates converge to distinct statistics of the predicted cumulative reward distribution, in an innovated theory termed distributional RL [23] (Figure 2.1B). Such innovations have been shown to improve performance of deep RL agents on benchmark tasks due to improved statistical robustness [87, 10], and evidence of distributional RL-like computations has been reported in midbrain DANs of both mice and primates [23, 3, 100], however the direct functional relevance of such distributional mechanisms and representations to behavior is unknown. In the engineering setting, deep RL agents vary in whether and how they make use of knowledge about the distribution over future rewards when selecting actions [9, 88, 140], and decoding of reward distributions from DAN activity has only been demonstrated at the time of reward delivery [23]. In principle, knowing in advance, at the start of an episode, about the range and likelihood of rewards available and when they are likely to occur could be highly useful for planning and flexible behavior, particularly in the face of non-stationarity in either the environment or internal state (e.g. hunger) of the animal [117].

Here, we develop a computational model of efficient multidimensional distributional RL that learns to predict distributions of rewards over both time and magnitude, or distributional time-magnitude RL (TMRL). We then test predictions of the model in vivo by recording from optogenetically identified midbrain DANs in mice during behavior. Specifically, in addition to previously demonstrated variability in the degree to which individual neurons respond to positive and negative RPEs [23] the model predicts

variability in the degree to which individual neurons discount future rewards. We discovered evidence of both types of heterogeneity in DANs. This enabled decoding of future reward times that correlated with the variability in the temporal evolution of behavior, and decoding of reward magnitudes that correlated with behavioral correlates of reward history, suggesting that decoded estimates correspond with animals' expectations about the timing and magnitude of rewards. Furthermore, we show that taking into account the variability in both sensitivity to reward magnitude and delay allows decoding of a two-dimensional distribution, or map, of future reward amount over time from a set of DAN responses to a predictive cue, at the start of an episode. Strikingly, the tuning of individual dopamine neurons for reward time and magnitude was dynamic, adapting to changes in reward statistics, in accordance with the principles of an efficient code that is optimized for encoding information about these two important dimensions of reward. We propose that these data reflect a mechanism by which the brain may use a local-in-time algorithm akin to TD learning to build information-maximising, predictive, and probabilistic models of the environment for use in behavioral control.

Author contributions

Margarida Sousa developed the theory, together with Joe Paton and Daniel McNamee and with input from Kenway Louie. Experiments were designed by Joe Paton, Margarida Sousa and Bruno Cruz. The experimental apparatus was constructed by Bruno Cruz, Pawel Bujalski and Margarida Sousa. All behavioral and electrophysiological experiments that provided data for the study were performed by Pawel Bujalski, and Bruno Cruz performed pilot experiments to establish protocols. Margarida Sousa analysed the neural and behavioral data, with input from Bruno Cruz, Kenway Louie, Daniel McNamee and Joe Paton. Margarida Sousa performed the foraging simulations with input from Daniel McNamee, Kenway Louie and Joe Paton. Joe Paton supervised all aspects of the project.

2.2 Results

2.2.1 Learning and encoding a two-dimensional probabilistic map of future reward

We begin by defining an adaptive distributional code for reward time and magnitude (TMRL). This code takes inspiration from several threads of research on temporal discounting [74, 34, 58, 139], temporal coding in general [117], distributional value codes, and more recent work that extends distributional value codes to the time domain [23, 138, 142]. Classical TD learning produces a global value function that encodes the average of expected future rewards. Temporal discounting of future rewards arises at the computation of the TD RPE $\delta^M(t)$[1] - the temporal difference between the value at the current timestep $V(t)$ and the discounted value, parameterized by a discount factor γ, at the subsequent step $V(t+1)$, plus any incoming reward (Figure 2.1A),

$$\delta^M(t) = r(t) + \gamma V(t+1) - V(t).$$

Though expected reward delays are indeed reflected in the value function because of temporal discounting, delay and magnitude information are compressed into a single scalar value, producing ambiguity between these two dimensions in the value code (Figure 2.1A bottom). Instead of learning a single value function, TD algorithms have recently been elaborated to learn a set of value functions $V_i(t)$ that systematically differ in their sensitivity to positive and negative RPEs (α_i^+ and α_i^-, Figure 2.1B),

$$V_i(t) \leftarrow V_i(t) + \alpha_i^{sgn(\delta_i^M)}\delta_i^M(t),\ \delta_i^M(t) = r(t) + \gamma\tilde{V}(t+1) - V_i(t), i = 1, ..., n_M,$$

where $\tilde{V}(t+1)$ is a random sample from the value distribution at the subsequent step. This causes each value function to converge to a different statistic of the observed distribution of reward magnitudes. Viewed collectively, this set of value functions encodes not just the average magnitude of expected future reward, but the distribution over their magnitudes (distributional TD learning in magnitude, Figure 2.1B). A central insight of our model is that this approach may be generalized to both learn and encode information about the distribution of reward times by assuming that a set

[1] The superscript M refers to the standard RPE.

of value functions $V_j(t)$ are learned that differ in their sensitivity to reward delays, parameterized as the standard temporal discount factor γ_j within the computation of a TD RPE (Figure 2.1C),

$$V_j(t) \leftarrow V_j(t) + \alpha\delta_j^M(t),\ \delta_j^M(t) = r(t) + \gamma_j V_j(t+1) - V_j(t), j = 1, ..., n_T.$$

Multiple timescales for discounting across a set of parallel learning channels alone resolves ambiguity between reward delays and the average reward magnitude that is present in a system with a single temporal discount factor. And crucially, since the RPEs at unpredictable stimuli are expectations of temporally discounted future reward, having knowledge of each channel's temporal discount factor allows for decoding information about the distribution over future reward times (distributional TD learning in time, Figure 2.1C). However, when distributional learning in time is combined with distributional learning of reward magnitude,

$$V_{ij}(t) \leftarrow V_{ij}(t) + \alpha_i^{sgn(\delta_{ij}^M)}\delta_{ij}(t), \delta_{ij}^M(t) = r(t) + \gamma_j\tilde{V}_j(t+1) - V_{ij}(t),$$

the resulting system learns to encode a probabilistic map of future rewards over both dimensions (Figure 2.1D). In such a two-dimensional distributional reward coding system, correcting for tuning diversity across one dimension should reveal the remaining tuning diversity for the other that might otherwise be obscured (Figure 2.1D bottom left).

Critically, because it specifies how temporal and magnitude parameters are learned, the TMRL model adapts to the rewards it experiences. The brain contains a finite number of neurons, and thus faces constraints in its information encoding capacity [6]. In the face of such constraints, efficient coding theory prescribes that the tuning properties of neurons adapt to the statistics of the variable they aim to represent in a manner that maximises overall information content of encoded signals [80, 122]. In the current context, this predicts that the discount factors and value parameters should be adapted so as to maximise the encoding of reward time-magnitude information with respect to the current environment [33]. We return to this aspect of the TMRL model in more detail below.

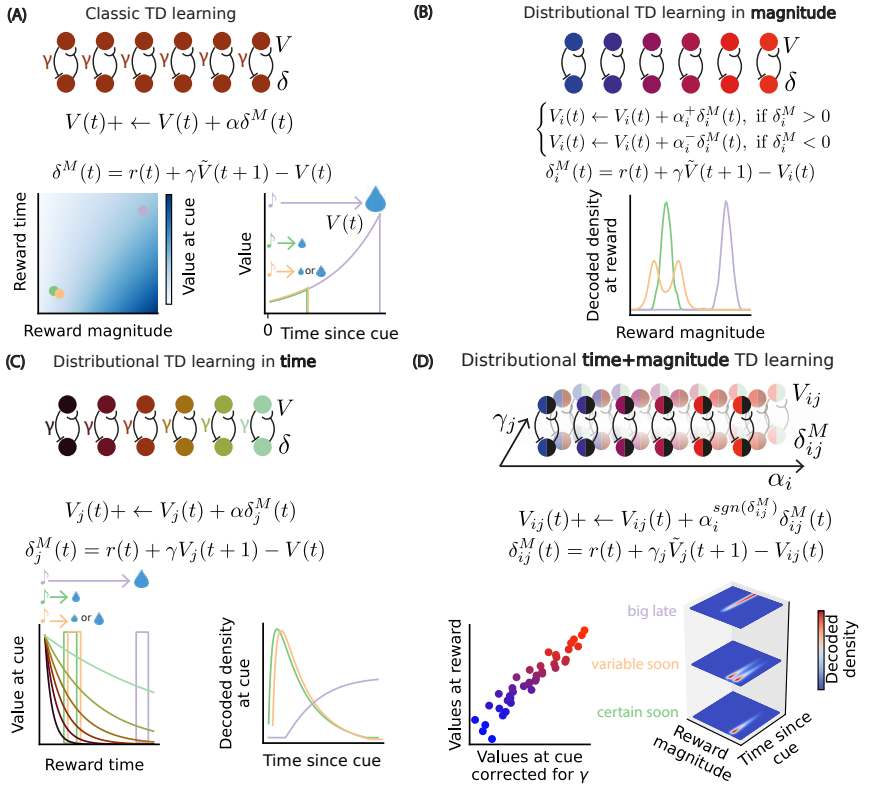

Figure 2.1: Diversity in temporal discounting and relative scaling for positive and negative RPEs facilitates the construction of a distributional map of future reward in time and magnitude. (A) Green cue predicts a certain reward amount after a short delay, orange a variable amount after a short delay and purple a big amount after a long delay. Bottom left: The value at the cue for different reward magnitudes and times. Bottom right: Temporally discounted value for the three different cues as a function of time since cue. (B) In distributional TD learning, units learn a diverse set of values, that allow for decoding the distribution over reward magnitudes represented on the bottom. (C) A population with diversity in temporal discount factors allows for decoding the distribution over future rewards at the cue. Bottom left: temporally discounted values. The orange, green and purple blocks represent the population responses at the three different cues. Bottom right: decoded the future reward distribution over reward time for the three different cues. (D) A population with diverse temporal discount factors and values allows for decoding the map of future reward in time and amount at the cue. Bottom left: Simulated values at reward time as a function of values at the cue corrected for the diversity in temporal discount factor. Bottom right: We use the asymmetries for positive and negative RPEs, the temporal discount factors and the responses at the cue to decode the probability of reward over time and magnitude.

2.2.2 Temporal discount rates vary among dopamine neurons and carry information about the distribution of future reward times

Does the brain use an algorithm similar to TMRL to learn about distributions of rewards along multiple dimensions? We tested for this by recording from midbrain DANs of mice (Figure 2.2) during a simple behavioral task, trace odor conditioning, designed to induce predictions of reward at different delays and magnitudes. Four odor cues (conditioned stimuli, CSs) predicted the same reward amount but with a distinct delay (0, 1.5, 3 or 6 seconds, respectively, Figure 2.3A). A fifth CS predicted, at a delay of 3s, a reward amount sampled from a bimodal probability distribution (Figure 2.3B). Importantly, the mean of the probability distribution of rewards associated with this fifth CS was equal to the fixed reward amount delivered following presentation of the other CSs. We focus our analyses on data from 43 optogenetically identified DANs collected from 6 trained mice. By considering the response of each dopamine neuron to CSs with different delays to reward, we estimated how single neurons discounted rewards over time, parametrizing this function with a temporal discount factor (γ), and a gain parameter (Figure 2.3C). In addition, by examining the responses to different reward amounts, we estimated the value expected by each neuron (reversal point) and the slopes for negative (α^-, represented in blue) and positive (α^+, represented in red) RPEs [24], Figure 2.3D.

We identified significant diversity in temporal discounting across neurons (Figure 2.4A-C). This diversity did not reflect noise in the estimation of γ, as estimates were highly correlated across random partitioning of trials (Figure 2.4B). We note that five cells possessed estimated discount factors that were greater than, and with 99% confidence intervals that were non-inclusive of, one. This may reflect limitations due to the number of reward delays that were probed, since three of these neurons were recorded in the absence of the 0 s reward delay condition. To assess whether this was the case, we substituted the response to the fixed rewards at trial offset for the 0s delayed reward in these cells with the responses to the most unpredictable reward (at the longest delay of 6s), an underestimate of the responses to the missing delay, and calculated the temporal discounts for each neuron. Under this substitution procedure, we obtained temporal discount factors that do not significantly differ from one. For completeness we include these neurons in

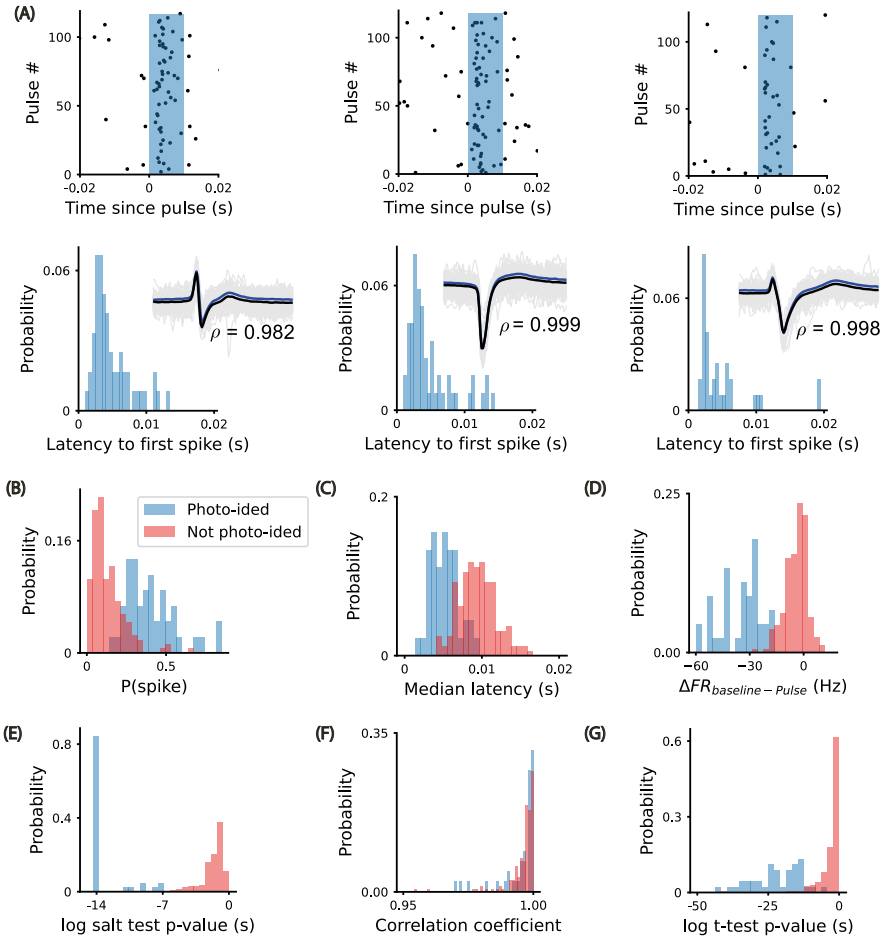

Figure 2.2: Photo-identification of dopamine neurons. (A) Three example photo-ided neurons. Top) Raster plot with single spikes aligned to laser pulse onset (10 ms duration). Bottom) Distribution of latencies to first spike after laser pulse observed in a 1-20ms window Bottom-inset) mean waveform (black) and mean laser-triggered waveform (blue). Distribution of: (B) probability of observing a spike between 1 and 10 ms after laser onset pulse. (C) median latency to first spike in a window between 1ms and 20ms. (D) Differences in firing rate between the baseline and 1-10ms post-pulse window. (E) Log of the p-value of the salt test [75]. (F) Correlation coefficient (ρ) between the mean waveform and the mean laser-triggered waveform. (G) Log of t-test for the difference in firing rates between baseline and pos-pulse firing rate in a window 1-10ms.

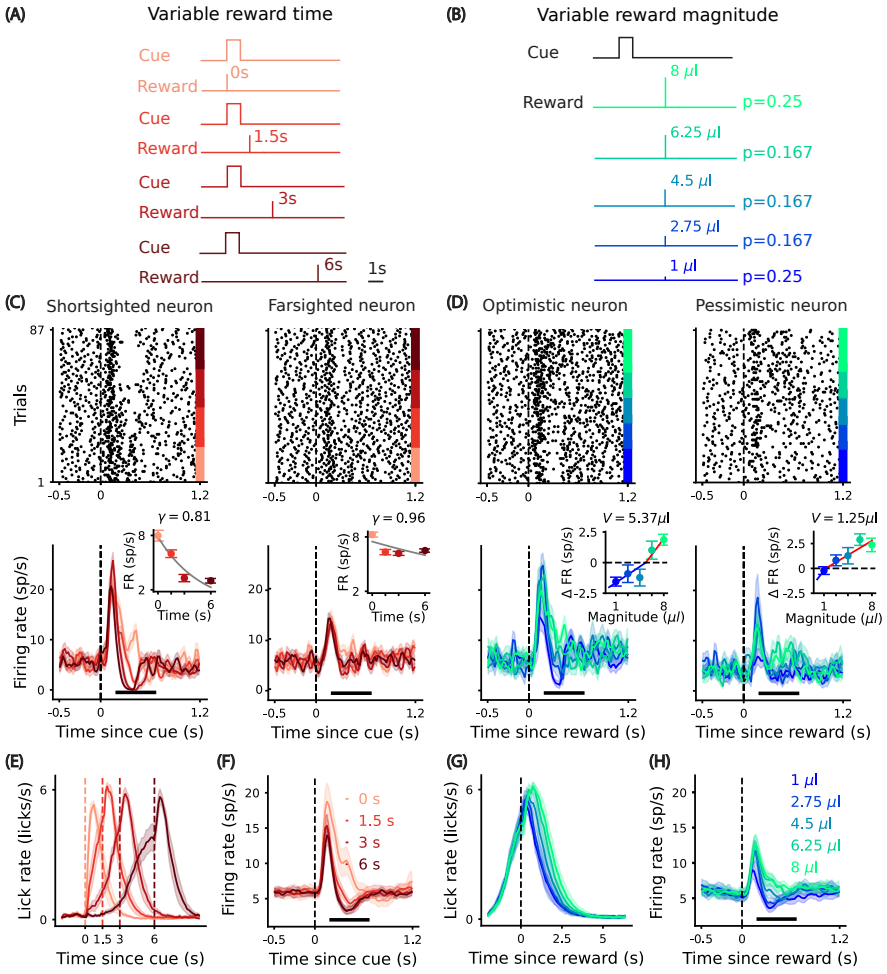

Figure 2.3: Dopamine neurons are modulated by reward magnitude and time. (A) Variable time CSs, odor cues are sampled to produce a uniform distribution of reward times over trials, reward magnitude = 4.5 μl. (B) Variable magnitude CSs. 3s after CS onset, a reward amount sampled from a bimodal distribution is delivered. (C) Raster and mean PSTH aligned to odor onset for two example neurons. Black line: window used to compute responses (200-650ms). Inset: responses to the different delays, gray line: fitted discount function. The error bars are the standard error of the mean over trials. (D) Raster and PSTH aligned to reward delivery for different reward magnitudes. Inset: baseline subtracted responses for the five different reward amounts, blue line: fitted line for the negative responses, red line: fitted line for the positive responses. (E) Mean lick rate averaged over animals aligned to odor onset for different reward times. (F) Mean population PSTH aligned to odor onset. (G) Mean licking rate for all animals aligned to reward delivery for different reward magnitudes. (H) Mean population PSTH aligned to reward delivery.

all analyses.

A key prediction of our distributional code for reward timing is that the probability distribution over future reward time can be decoded from the population responses to a CS (Figure 2.4C). To test this prediction, we focus on the population responses to the CSs that predict a certain reward at fixed delays and assume the system has knowledge of the temporal discount rate of each neuron. Since the population response to a CS reflects value, the sum of temporally discounted future rewards, determining the reward distribution over time presents as a linear regression problem (Figure 2.4D top). The independent variable is a matrix of temporal discounts to the power of the discretized time, the dependent variable is the mean responses to the CS and the regression coefficients are the probabilities of rewards over time (Figure 2.4D, top). Consistent with the model prediction, the resultant densities capture the differences in the timing of rewards for the four CSs (Figure 2.4D, bottom).

We next asked if the decoded estimates correspond with animals' temporal expectations, by comparing trial-by-trial variability in anticipatory licking behavior with the future reward time predicted by the population of DANs. We observed that in trials wherein animals commenced licking earlier or later, the decoded distribution over future rewards exhibited a qualitatively similar shift in time (Figure 2.4E). These data suggest that estimates of future reward times decoded from a dopamine neuron population reflected temporal expectations that animals used to guide behavior.

2.2.3 Dopamine neuron cue responses reflect distributional value information encoded in responses to reward

Next, we sought to combine the distributional time code with the previously proposed distributional code in amount [23], focusing first on the CS predicting a variable reward amount. The distributional RL theory in amount predicts that neurons with asymmetric linear functions for positive and negative RPEs possess reversal points (the reward amount for each neuron that produces zero net change in activity) that correspond to the expectiles of the probability distribution of rewards [23]. Previous work linking distributional codes for reward to the activity of midbrain dopamine neurons has largely focused on reward responses, and not the responses to cues that predict rewards. However, in principle, distributional information

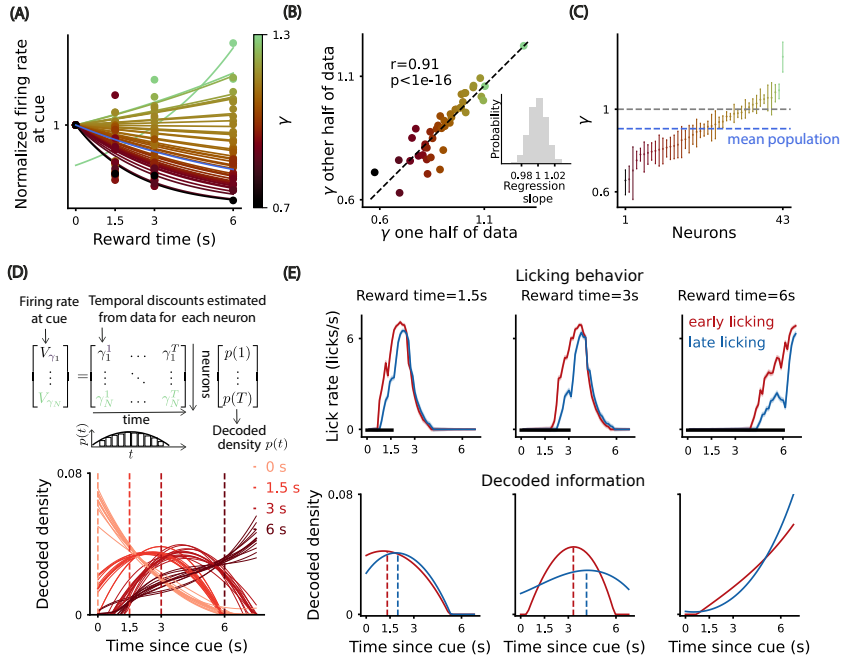

Figure 2.4: Dopamine neurons discount future reward heterogeneously, reflecting information about the timing of future rewards that correlates with licking behavior. (A) The dots are single neuron responses to CSs predicting different reward times, normalized by the responses at delay=0s and the lines are the fitted temporal discount functions, color code: temporal discount factor. The population mean temporal discount function is plotted in blue. (B) Temporal discounts estimated using two random disjoint partitions of the total number of trials. Inset: histogram of the regression slopes for each run is depicted, 95% CI=(0.98,1.02). (C) Cross validation of temporal discount factors of single neurons using 50% of the total number of trials per delay. The error bars: 99% confidence interval, dots: median. Dashed blue line: population mean temporal discount factor, dashed gray line: temporal discount factor equal to one. One-way ANOVA for the difference in population temporal discounts F(42,42957)=20445.18, p-value=1.96e-284. (D) Top: Decode the distribution of future reward corresponds to inverting a linear regression. Middle: A representation of a possible decoded density over a discretized range of times. Bottom: decoded density using the dopamine population responses aligned to odor cue onsets. The light lines represent decoded densities using the responses of 70% of randomly selected trials and the dark lines represent the mean decoded density. (E) Top: Mean licking rate for all mice (n=6), for the trials in which the mice started licking earlier (red) or later (blue). The shaded area depicts the standard error of the mean and the horizontal black line the window used to compute the licking slopes. Bottom: Decoded density using the trials for which the mice started licking earlier (red) or later (blue). The dashed lines depict the maximum of the decoded reward time.

should be propagated backward in time from reward and thus decodable from the responses to reward predictive cues (Figure 2.1D). One reason this may not have been observed in previously reported data is that the diversity in temporal discounting across neurons that we describe in the previous section can occlude distributional reward magnitude information (Figure 2.1A,D). Because we measured diversity in temporal discounting across neurons, we were able to correct for it. Indeed, we found that correcting for the diversity in temporal discounts and gains revealed residual CS responses that are significantly correlated with the reversal points estimated at the time of rewards (Figure 2.5A), indicating stable but mixed selectivity for reward delay and magnitude in single neurons. We then computed probability distributions over reward magnitude from dopamine responses at both the time of CS and the time of reward delivery. The distributions from reward and cue responses were nearly identical (Figure 2.5B), indicating that the consequences of systematic variability in sensitivity to positive and negative errors at reward delivery, previously identified as evidence that the dopaminergic system implements a distributional code for value, are transmitted to the cue response. This establishes the presence, at the time of cue presentation, of information required for the system to foresee impending variability in the magnitude of rewards ahead of time, when it might be used to guide future behavior.

To test whether DAN cue-related activity reflects changes in reward expectation, we leveraged the tendency of subjects to update reward expectation as a function of reward history [78]. Consistent with previous work [112, 16], pupil diameter in mice tracked recent reward history of the cue predicting variable reward amounts at a fixed delay (Figure 2.5C, linear mixed model with a main effect of pupil diameter, and random effect of mouse identity, p-value=0.0 computed by bootstrapping over trials 10,000 times), suggesting a reward history-driven modulation of reward expectation. Given this relationship, we tested if decoded estimates corresponded with animals' reward expectations by comparing trial-by-trial variability in reward history with the future reward magnitude predicted by the population of DANs. We observed that the mean decoded reward magnitude varied as a function of reward history; And yet, there was no difference in the mean decoded reward time as a function of reward history (Figure 2.5D). This provides evidence that the population of DANs is able to reliably encode the time of future rewards, even in the face of changing expectations of average reward

magnitude.

We then tested a key prediction of the distributional TMRL code: that the joint distribution over future reward amounts and times can be decoded, a key prediction of the distributional TMRL code, from the responses of the DAN population at the CS. Conditioned on the CS, the joint probability distribution of reward on each trial can be factorized as the product of the marginal distributions of reward over time, giving a 2D map of future reward magnitude over time (Figure 2.5E) that closely matches the true distributions. Thus, a multi-dimensional probabilistic map of future rewards may be estimated from just 450ms of dopaminergic neural activity at the onset of an episode.

2.2.4 Value and temporal sensitivity efficiently adapt to environment statistics

We have identified diverse temporal discounting and sensitivity to value in midbrain DANs that may be used to estimate the joint probability of future reward time and magnitude. However, does parameter diversity across neurons that enables such a code reflect variability that we as experimenters exploit? Or are the parameters collectively regulated to maximize information about rewards along the hypothesized target dimensions? Evidence of such regulation would not only indicate efficient representational codes for reward within the dopamine system, but by extension, also provide strong evidence that the diversity in parameter tuning is present for the purpose of representing distributional reward information.

We used an efficient population coding framework [41, 84, 115], to derive temporal discount functions that optimally represent the reward times t_r in the environment. Inspired by previous work, we propose that this distribution is efficiently encoded in an expectile code [23]. In particular, we sought to maximize the mutual information between the true expectile reward times predicted at cues s_0, $p(\bar{t_r}|s_0)$, and those encoded by the dopamine neural population $\vec{\delta}$, constrained by the number of neurons N and by the population expected firing rate R. We assume the cue response of each dopamine neuron decays exponentially as a function of reward delay with a time scale τ and a gain parameter a. We parameterize the population with the density of tuning curves d which characterizes the heterogeneous allocation of neurons to reward time scales and a gain g which characterizes the mean

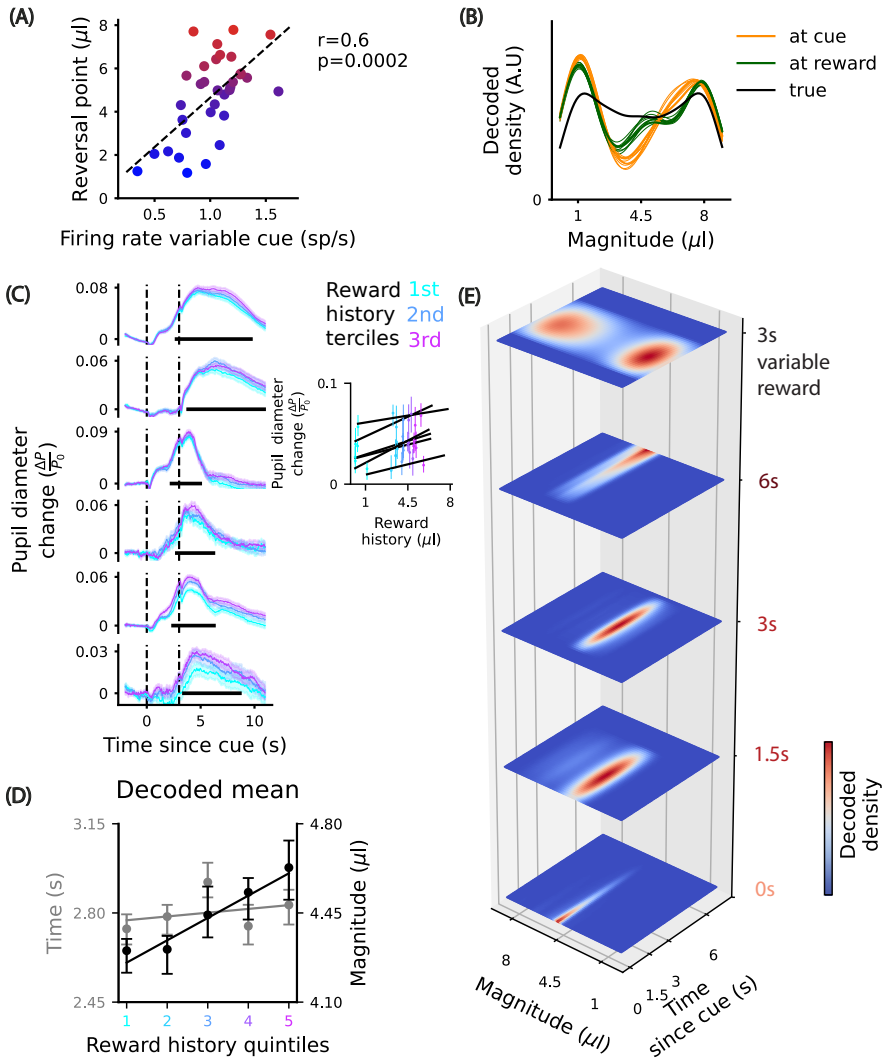

Figure 2.5: Dopamine neurons reflect information about the distribution of future rewards at cue presentation. (A) Reversal points of neurons as a function of the response to the cue associated with variable reward, corrected for the estimated discount function, color code: reversal point. (B) Smoothed distribution over reward amounts decoded from the DAN population response. (C) Mean pupil diameter change for all animals aligned to cue delivery for reward history terciles. Inset: regression of mean pupil peak computed over the full width half maximum window as a function of reward history for individual animals. (D) Conditioning on the reward history, in gray is the mean decoded time and in black the mean decoded magnitude. The error bars correspond to the 50% C.I. over 1000 runs, using the responses of 70% of randomly selected trials. (E) Decoded joint density of reward over magnitude and time.

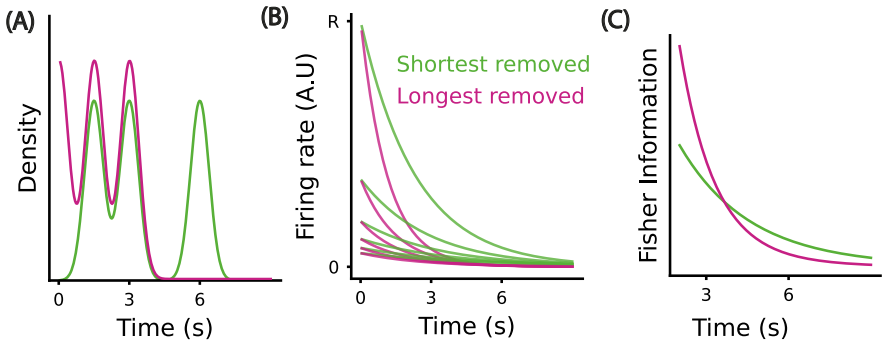

Figure 2.6: Efficient coding predictions and adaptation of dopamine neuron population temporal discount factors with reward occurrence rate. (A) Densities after removing the shortest (green) or longest (pink) reward delay. (B) Optimized tuning functions for the densities of reward time depicted in A. (C) For analytical tractability, we optimize the Fisher information (lower bound on mutual information) [15]. We represent the Fisher Information for the optimized populations.

firing rate across reward times. The solution is that neuron's timescales τ should distribute according to the probability distribution of the current environment expectile reward times $d(\tau) \propto Np(\bar{t_r}|s_0)$ (see Methods and Figure 2.6). Intuitively, if rewards at short delays occur infrequently in a given environment, the system should not waste coding capacity to encode expected future rewards at short timescales, and vice versa if the environment only rarely emits reliable rewards at long delays. Furthermore, by additionally minimizing the mean population firing rate, for each neuron the gain should be inversely proportional to the probability that a randomly chosen reward time will be smaller than it's time scale $P(\tau)$, $g(t_r) = R/(NP(\tau))$ (see Methods). For low reward times, the entire population is active, incurring a large cost for encoding these values. Intuitively, this penalty can be reduced by lowering the gains tuned to late reward time scales, while maintaining optimized coding.

In order to optimize this efficient population code online, we generalize the distributional learning rules (Figure 2.8A) to the time domain, considering multiple channels with different relative scaling for positive and negative reward time prediction errors δ^T[2], that generate a diversity of learnt reward

[2]The superscript T refers to a reward time prediction error.

time scales (Figure 2.8E),

$$\tau_j \leftarrow \tau_j + \alpha_j^{sgn(\delta_j^T)} \delta_j^T.$$

Importantly, these parameters converge to the efficient code that optimally adapts to the statistics of expected reward times in the environment (see Methods).

A critical but untested prediction of the distributional code for value is that the value each neuron expects (as defined by its reversal point) should adapt to changes in the probability distribution of reward magnitudes (Figure 2.8A,B). In addition, the ordering of reversal points in the population should be preserved for different probability distributions. We found that the reversal point ordering is preserved across the two CSs at the reward time (Figure 2.8C) and that the variance of the reversal points for the variable CS is significantly greater than for the certain CS (Figure 2.8D).

The mapping from time scales to temporal discount factors is an exponential mapping that takes into account the fact that steep temporal discounting (i.e. small temporal discount factors) gives rise to very different learned value estimates for rewards that will occur at distinct short delays, but discriminates poorly between rewards occuring at later delays. Conversely, shallow temporal discounting (ie. large temporal discount factors) discriminate between rewards occurring far apart in time, and can thus support encoding of rewards at long time delays. Therefore, the discount factor is a monotonically increasing, exponentially decelerating function of the time scales at which rewards are observed $\gamma = e^{-\frac{1}{\tau}}$ (Figure 2.8E inset).

To test if the dopamine code adapts to the temporal statistics of reward, at the end of each session we modified the temporal reward distributions by removing the CS corresponding to either the shortest or longest reward delay. As predicted by the theory (Figure 2.8F), when removing the longest delay, DANs adapted to improve encoding accuracy on short reward times by decreasing their discount factor (Figure 2.8G). While adaptation when removing the shortest delay was not statistically significant in this dataset (Figure 2.8G), the magnitude of the theoretically predicted adaptation is smaller than that predicted when removing the longest delay, particularly for the larger discount factors reflected in those sessions when the the shortest delay was removed, and the data exhibited a trend in the correct direction, and thus a larger data set may reveal a shift in this condition as well. Alternatively, asymmetry in neural adaptation may be due to a

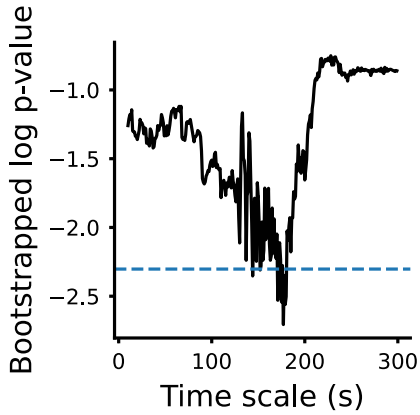

Figure 2.7: Adaptation of dopamine neuron population temporal discount factors with reward occurrence rate. Considering exponential decaying kernels with different time constants to compute the rate of reward occurrence, we compute the p-value bootstrapping 10,000 times, for the null hypothesis that the temporal discounts for higher reward rates are shallower or equal to the lower reward rates. Horizontal blue line: p-value significance equal to 0.1.

rational bias in the neural code towards ensuring that predictions for the very near future are accurately encoded regardless of the temporal reward distribution in the environment, given that passage through intervening delays en route to later rewards is unavoidable. We also tested whether discounting by DANs was updated more continuously by comparing the trial-to-trial adaptation in temporal discount when the rate of reward occurrence is relatively high or low. Indeed, the temporal discounts for low rates were larger than for high rates (Figure 2.7). Regarding the gain, we indeed observe as predicted that individual neuron gains are negatively correlated to temporal discount factors (Figure 2.8H). The lawful, dynamic regulation of temporal discounting in individual DANs that we describe here indicates that principles of efficient coding likely apply to how the brain regulates the time constants over which rewards are predicted. However such lawful adaptation also strengthens our confidence that the dimension over which we as experimenters are decoding expected rewards - time - is indeed an encoding target for the system.

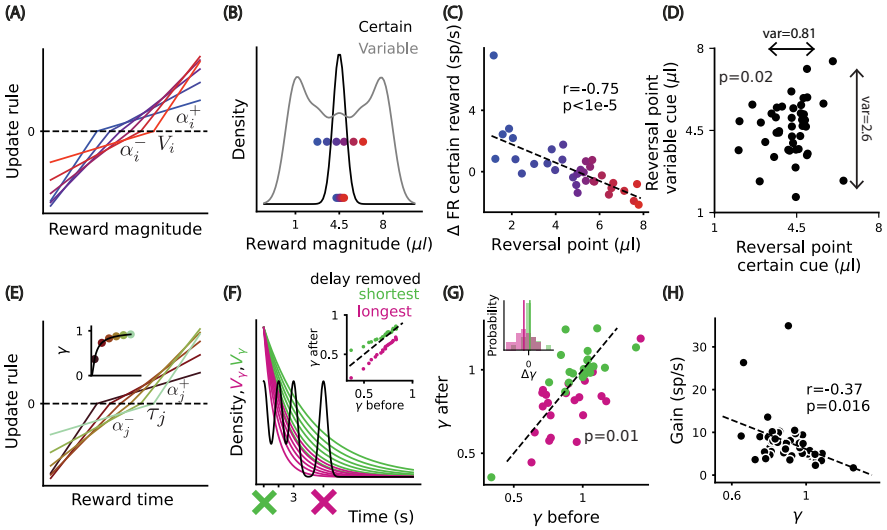

Figure 2.8: Value and temporal sensitivity adapt to changes in reward statistics, in accordance with principles of efficient coding. (A) Distributional value code predicts that units have different asymmetries for positive (α_i^+) and negative (α_i^-) RPEs, generating a set of values V_i, color code: reversal point. (B) Distributional value code predicts that variability in reversal points for the cue predicting a bimodal distribution is greater than for the cue that predicts a certain reward amount. (C) DAN responses to the certain reward amount delivered at a 3s delay, as a function of the reversal point estimated using the responses for the variable reward amounts delivered at a delay of 3s. (D) Reversal points estimated at the cue predicting variable and certain reward magnitudes at the same delay. (E) Considering asymmetric weights for under and over-estimation of reward times generates a diversity of time constants that are mapped to temporal discount factors using the function depicted in the inset, color code: temporal discount factor. (F) Predicted adaptation in temporally discounted values when the reward time distribution is manipulated, by removing the shortest (green curves) or longest delay (magenta curves). The black curve depicts the smoothed distribution of reward times in the first phase of each experimental session. The inset depicts predicted adaptation in temporal discount factors. (G) Experimentally observed adaptation in temporal discount factor estimated from the recorded DANs. Inset: histogram of update in temporal discounts for the two different manipulations. (H) Single neuron gains as a function of temporal discount factors, the dashed line represents the fitted linear regression.

2.2.5 How might information about future reward distribution be used to guide behavior?

The TMRL algorithm is relatively simple and "model-free", meaning that the algorithm itself does not have access to knowledge of how the world transitions between states when learning to assign value to them. However, because the probabilistic map it gives rise to provides rich information about the future, we suspected it might enable behavior similar to that produced by more computationally intensive model-based learning algorithms. A key difference between model-free methods that estimate a single moment vs. the full distribution of values, is that the latter allow for weighting of probabilities [64] depending on context, whereas the former, as a consequence of it having access to only one value estimate, are forced to adapt it through new learning in the face of changes in the environment or the internal state of the agent. We thus reasoned that the very nature of the information TMRL learns can naturally allow for complex computations such as zero-shot adaptation to reward time and magnitude preferences. To examine the potential behavioral benefits of the joint distributional reward representation produced by TMRL in relation to pre-existing algorithms, we performed a series of simulations in which model agents forage for rewards (Figure 2.9). For comparison, we model agents making use of either TMRL, standard TDRL, or the successor representation (SR). Similar to TMRL and TDRL, SR is a predictive model learned using a temporal-difference algorithm. However, SR instead caches temporally discounted expectations of future state occupancy, which can facilitate heuristic planning strategies [27]. Interestingly, the basic mechanism underlying the SR - TD learning of temporally discounted expected future state occupancy - has recently been extended to include a set of SRs that vary in their temporal horizon, a conceptually similar innovation to that reflected in TMRL, except that it focuses on creating multi-scale temporal predictions of future states instead of future rewards [96].

Our simulations took the following form. In a dynamic foraging environment where patches possess different amounts of available reward at different times during the day (Figure 2.9A), simulated mice can adapt policies significantly faster when given access to knowledge of the timing of reward availability (in addition to the average amount of reward each patch contains). The distributional TMRL code generates a representation

that may be used to flexibly adapt policies to preferences towards reward times, by reweighting the map with a reward utility (Figure 2.9B right). Upon detecting that dawn approaches, the amount of time left to forage reduces, and therefore mice should use a steeper temporal discount to read out the value from the distributional map. Using TMRL, mice can adopt a strategy where they instantaneously adapt to changes in reward time-scales from dusk to dawn (Figure 2.9C, F and Figure 2.10A), without having to learn a different set of values associated with a distinct set of states from those used earlier in the night. Indeed, though in principle an agent might solve such problems by learning multiple sets of single value estimates for a larger variety of states, with scale this leads to a severe sample inefficiency of learning [135]. If restricted to using the same set of states, the standard TDRL and SR agents (Figure 2.9F and Figure 2.10A), have to re-learn through experience the value and future occupancy, respectively, for the new time-scale. These simulations provide a simple demonstration of the potential benefit of a model that uses information about the distribution of rewards in time to guide action selection over existing models that do not possess such information, specifically in the context of an environment with temporally delimited reward availability.

The previous example focused on a case where the external world possesses reward dynamics, but biological agents are also subject to internal state dynamics that incentivise using knowledge about reward distributions. For example, in response to changes in wealth or physiological need states, humans and other animals shift their preferences with respect to the variability in the amount of reward, which can be expressed as a utility function [152, 127, 62, 158, 63]. In our example, if the mouse is hungry and needs a significant amount of reward, he might reweight different regions of the TMRL reward map, akin to creating a dynamic utility function that heavily weights large, immediate rewards (Figure 2.9D). This can allow mice to instantaneously employ a policy where they select patch two, predicted to grant the largest reward possibility at a short delay. The SR also allows for a fast adaptation of policies between the sated and the hungry states, by reweighting the rewards (Figure 2.9B Left D, G and Figure 2.10 10B); in contrast, the TDRL algorithm needs to learn through experience to adapt the policy (Figure 2.9G).

Furthermore, the TMRL allows for the reweighting of both the reward magnitude and time by considering the interaction between the temporal

dynamics of the environment and the internal state of the mouse (Figure 2.9E). For example, when it is dawn and the mouse is hungry and has less time to forage, it overweights large immediate rewards (Figure 2.9E), allowing for zero-shot adaptation of the policy (Figure 2.9H, Figure 2.10C), and thus presents a novel computational mechanism by which environmental dynamics interact with internal states to modulate decision-making.

While not intended as exhaustive, the scenarios we simulate here in a foraging environment - temporally delimited and variable magnitude rewards, and internal state dependent modulation of utility - are chosen to illustrate the broad potential importance of the information provided by distributional TMRL for animal behavior. They reflect a small, but illustrative, sample of the rich opportunities for future work to examine how the reward representations generated by a multidimensional distributional algorithm like TMRL can benefit adaptive behavior.

2.3 Discussion

A fundamental facet of intelligence is the ability to use past experience to predict the future [21, 91]. Predictions may incorporate detailed information about how the environment will develop depending on a course of action, constituting models of the world that can be operated on flexibly but laboriously. Alternatively, predictions may take the form of efficient, compressed representations that discard detailed features of environmental structure and are specific to particular, behaviorally relevant events, such as rewards. Within RL, model-based learning algorithms that involve the former, detailed predictions enable more flexible behavior at the cost of computational complexity, while model-free algorithms that target the latter, simpler predictions allow for efficiency at the expense of flexibility. The field of RL provides a growing set of tools for learning simpler and more complex predictions alike in service of adaptive behavior, and there is ample evidence that the brain employs strategies resembling both algorithmic classes depending on, for example, whether behavior is under more explicit, goal-directed or automatic, habitual control [135, 4].

Though often described in categorical terms, recent work has highlighted how predictive representations of intermediate complexity can enable more flexible behavior that tends to characterize model-based algorithms, while using computationally efficient model-free learning algorithms [27, 113].

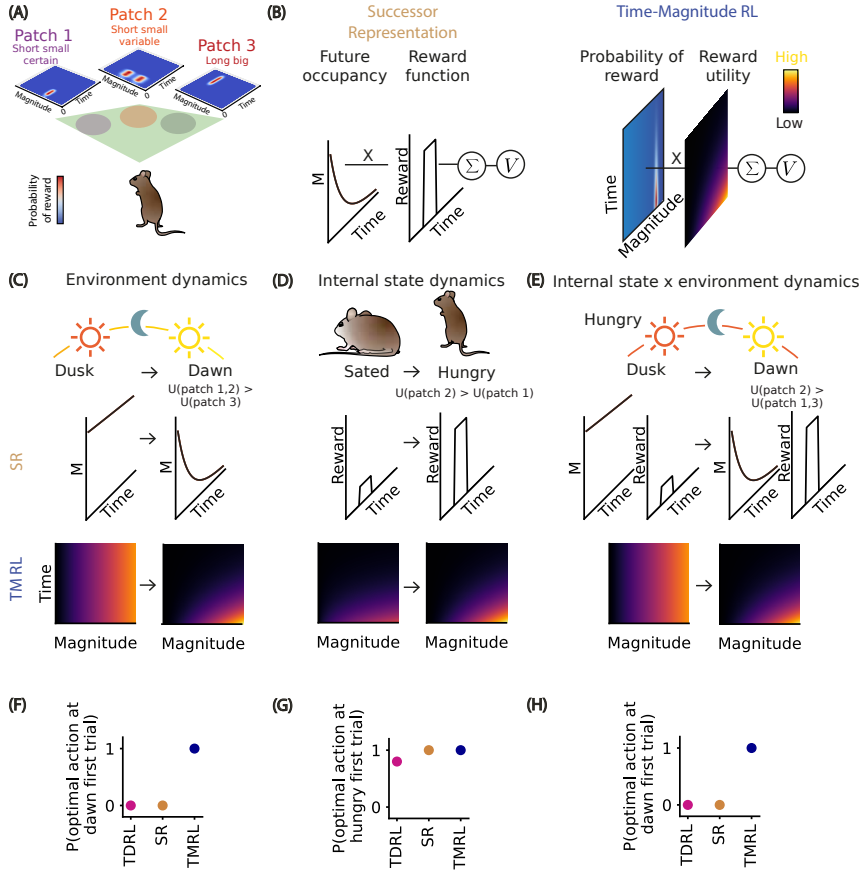

Figure 2.9: Adaptation of value for environment and internal state dynamics. A distributional code allows for flexible adaptation to temporal dynamics and preferences of reward using a model-free RL algorithm. (A) A foraging mouse must decide which patch to choose to maximize cumulative collected rewards in a non-stationary environment. Axes indicate the learned joint probability distribution of reward time and magnitude associated with each patch. (B) In the SR agent, the value of each patch at the start of the day is the product of the temporally discounted future occupancy by the reward at each future time step. In the TMRL agent, the probability distribution over future reward time and magnitude is weighted by a utility function to obtain an estimate that depends on internal state or/and the dynamics of the environment. (C) Adaptation of SR and TMRL agents when the time-scale of the environment changes, from dusk to dawn. (D) Adaptation of the SR and the TMRL when reward is over-valued, that may occur for example when the mouse is sated and becomes hungry. (E) Adaptation of SR and TMRL when the mouse is hungry and has less time to forage at dawn. Probability of choosing the optimal patch at the first trial after dawn (F), after being hungry (G) and after being hungry and after dawn for the three algorithms (H).

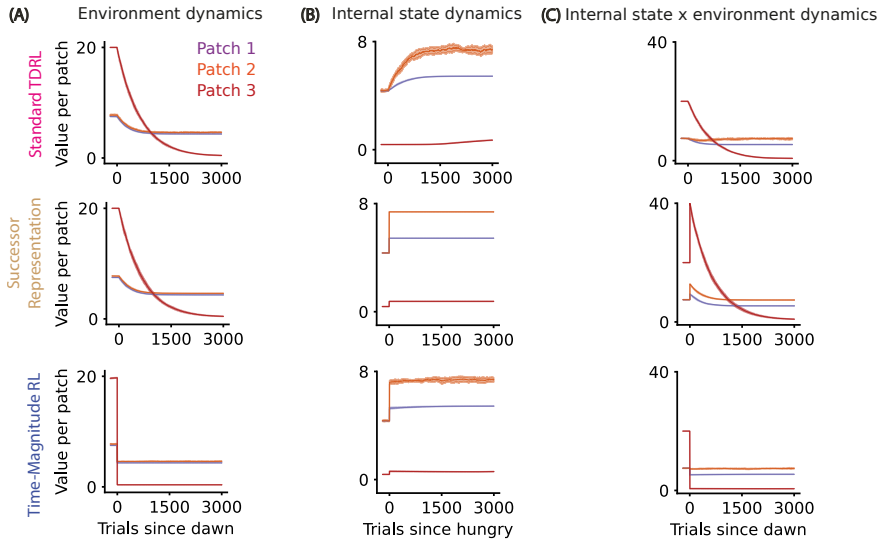

Figure 2.10: Dynamics of value estimates for standard TD, SR and TMRL. (A) Probability of choosing patches one (purple), two (orange) or three (red) as a function of the time since dawn for standard TDRL on the first row, SR on the second row and TMRL on the third row for environment dynamics simulation (Figure 2.9C,F). The error bars represent the standard deviation over 10 runs. (B) Probability of choosing each patch as a function of the time since the animal is hungry (Figure 2.9D,G). The error bars represent the standard deviation over 10 runs. (C) Probability of choosing each patch as a function of the time since dawn when the mouse is hungry for the internal state and environment dynamics simulations (Figure 2.9E,H) . The error bars represent the standard deviation over 10 runs.

Here we present distributional TMRL, a theoretical extension of TD learning that learns an efficient, multi-dimensional, probabilistic map of future reward value over time. This probabilistic map constitutes a kind of 'model' of the world, and yet, the mechanics by which it is learned are less complex than those used to learn the full structure of possible state transitions in an environment. Furthermore, we show how distributional TMRL learns representations that may enable better, i.e. more rewarding, policies for behavioral control when in an environment where exploiting knowledge about the time course of available rewards confers a benefit, or when changes in internal state might favor dynamic attitudes toward risk. Importantly, we present evidence that the brain may make use of TMRL-like computations. Midbrain DANs, a core component of neural systems involved in learning

policies for behavioral control, displayed signatures of a multidimensional distributional code over reward time and value; furthermore, animals appear to use the reward timing information reflected in DAN population activity to guide temporal control of behavior [124]. Lastly, we show that the distributional TMRL code adapts to changes in reward statistics in accordance with information theoretic principles of efficient coding. This provides not only strong evidence that the joint distribution of reward over time and value that we as experimenters decode from neural activity is indeed a functional target for encoding within the system, but potentially opens a fresh perspective on how to interpret individual variability in temporal discounting at a behavioral level. Steep temporal discounting that heavily favors immediate over delayed rewards, leading to apparently maladaptive, impulsive behavior, can be reframed as the optimal solution to a volatile environment. Conversely, shallow discounting, leading to consideration of delayed rewards, is ideal for stable environments that are predictable over long time scales. Thus, changes in environmental volatility may be compensated by adaptive changes in discounting, complementing evidence for volatility-induced changes in learning rates [7, 125]. This suggests that therapeutic attempts to improve impulse control might fruitfully target the environment, by intervening to lengthen the timescale over which predictions are valid, or reorient individuals to the longer timescale structure of rewards that may already exist within an environment.

More than two decades ago, midbrain DANs were first hypothesized to emit a TD RPE to teach recipient circuits accurate reward expectations [114, 98]. This proposal drove the development of a new field of research into whether the brain may be using algorithms like those contained within computational RL models to learn programs for behavioral control, and if so, what form they take. That a compact algorithm like TD learning can estimate an overarching objective around which much of learned behavior can be shaped is not only practically useful in engineering settings, but explains a wide range of neural and behavioral data. However, in its simplest form, TD learning does not capture behaviorally relevant dimensions that characterize rewards because it compresses information regarding reward time, magnitude, and quality into one scalar value that represents, in a common currency, the expected average sum of temporally discounted future rewards. Sensory characteristics, and unambiguous knowledge about magnitude and timing of rewards is lost. Furthermore, experimental data

has shown that DANs can exhibit sensitivity to the sensory characteristics of rewards [119], actions or features of the environment [31], and appear to access internal models regarding environmental structure [118]. It is thus becoming ever more clear that the basic model of DANs emitting a unitary, model free TD RPE requires refinement, at the very least, and some have argued for its replacement altogether [59, 20].

However, the space of possible TD learning algorithms is large and continually growing. Assumptions about the way the state of the environment is encoded, the space of possible actions in TD methods for control, and parameter determination or regulation are just a few of many ways in which TD methods can differ [13]. Given the parallel and modular nature of brain circuits, including those embedding DANs, one class of TDRL extensions that would seem particularly attractive for deriving hypotheses about neural systems posits parallel learning of multiple reward expectations that systematically differ, either quantitatively or qualitatively. For example, multi-system models, where separate model-free and model-based RL mechanisms both contribute to behavioral control have been mapped to distinct, parallel circuitry in the basal ganglia [13, 156], the main target of dopaminergic innervation, and multi-agent and mixture of experts models have been used to explain recent data regarding the functional role of neurons in different regions of the striatum, a major target of dopamine, and spatial heterogeneity in striatal dopamine signaling itself, respectively [50, 22]. In addition, recently described heterogeneity in DAN encoding of task variables has been hypothesized to reflect a vector, as opposed to a scalar, prediction error, still within a TD framework [81], and signatures of multiple separate, qualitatively distinct predictions have been observed in DANs [137]. Here we provide direct evidence that diverse sensitivity to a fundamental dimension along which rewards are distributed, time, can explain another source of variability in response properties across DANs, without abandoning a computational framework for understanding dopaminergic function for which the accumulated experimental support is large. Temporal information is critical for learning systems. Reliable temporal structure in the world forms the basis for identifying predictive relationships that drive associative learning and contribute to causal inference. To extract such structure, the brain must somehow register when certain events occur in relation to each other, creating something akin to a temporal map [5]. We describe the existence of just such a temporal map, specific to reward, that is reflected in a

core component of reward circuitry long known to be critical for behavioral control: midbrain DANs. Importantly, despite our having decoded the joint map of future rewards from populations of DANs, we remain agnostic as to how and where this information is read out by neural circuits in the brain. Indeed our view is that DAN responses derive from upstream representations of value that may be more directly involved in guiding behavior. It will be important to determine in future work whether and in what manner this information is distributed within other brain circuits, and how this information is brought to bear on the problems faced by organisms seeking to thrive within complex and dynamic natural environments.

2.4 Methods

2.4.1 Mice

Young adult (2-7 months old at time of experiments), male DAT-Cre mice expressing Channelrhodopsin-2 (ChR2) in midbrain dopamine cells were used in this study. For this, Ai32(RCL–ChR2(H134R)/EYFP) mice (IMSR$_{JAX}$: 012569) were crossed with DAT-IRES-Cre mice (IMSR$_{JAX}$: 006660). Mice were group-housed (up to 3 mice per cage) until the first of two craniotomies were performed, after which they were single-housed. A temperature (21°C) and humidity-controlled (50%) housing room was maintained with a 12-hour light/dark cycle. Mice were maintained on PicoLab Rodent Diet 20 (5053), and under water deprivation for all behavioural experiments (> 85% body weight from baseline *ad* libitum period before deprivation). All experiments and procedures followed guidelines set and approved by the relevant national and international authorities (Champalimaud Foundation Animal Welfare Body (protocol number: 2017/013), Portuguese Veterinary General Board (Direcção-Geral de Veterinária, project approval 0421/000/000/2018) and European Union Directive 2010/63/EEC).

2.4.2 Surgical Procedures

Each mouse received three surgeries, first one for headpost implantation and then two unilateral craniotomies (performed 1 week apart) above the targeted regions (VTA/SNc). All surgical procedures were carried out under anesthesia with isoflurane (3% for induction; 1–2% for surgery at 0.8 l min

$^{-1}$). Mice were then fixed in a stereotaxic frame and their eyes were protected with a small amount of ophthalmic ointment.

The headpost implantation surgery was performed following the procedure outlined in [2]. Briefly, the hair was shaved down to skin which was then disinfected and incised. After exposing the skull, the soft tissue and periosteum were carefully removed and cleared. Then bregma and lambda were identified, the skull was leveled, and the locations for the craniotomies were marked. Subsequently, the skull surface was prepared for headpost implantation following the procedure suggested in the protocol. The headpost was then placed in a desired location (anterior to bregma and slightly above the skull surface) and cemented in place with dental adhesive (C&B Metabond, Parkell). Following the surgery, mice received carprofen subcutaneously (s.c.) for pain management. Mice were then placed back in their original cages and monitored for 7 days post-surgery to ensure well-being and a full recovery.

Prior to any electrophysical recordings (24-48 hours), a small craniotomy (1.5mm) was performed and sealed with a removable silicone sealant (Kwik-Sil, World Precision Instrument). Two separate craniotomies, one in the left and the other in the right parietal bones, were performed (coordinates: AP: -3.0 mm; ML: ±0.6 mm from the Bregma). Prior to surgery, all mice received carprofen (s.c.), enrofloxacin (s.c.), and dexamethasone (i.m.).

2.4.3 Behaviour & Training

2.4.3.1 Behavioural apparatus

The behavioral setup consisted of an infrared light source, an IR camera, a head-fixation system, a water delivery tube, a custom-built olfactometer [145] with an odor delivery tube, and an air ventilation system that was running during the whole experiment to prevent odor accumulation. The task logic was implemented in a real-time operating system using an Arduino microcontroller (Arduino Mega 2560, Arduino). The behavior of the animal was also monitored via an IR camera (FL3-U3-13S2, FLIR). The videos were acquired with Bonsai [83] at 120fps (640 x 480 pixels) for online licking detection and further offline processing. Briefly, a small area of the image was selected, on each session. For each frame, the sum of all pixels' luminance was computed. The resulting trace was then manually thresholded

and lick events were detected as frames with average luminance above the background.

The odor cues were delivered through the tube of a custom-built olfactometer placed approximately 1cm from the mouse snout, and the odor delivery was controlled via two-way micro-solenoid valves (model LHDA1233115H, Lee Company, CT, USA). Similarly, calibrated water reward was delivered through a lick spout using a two-way micro-solenoid valve.

2.4.3.2 Odor stimuli

During each trial an odor cue was presented for 1 second approximately 1 cm from the snout. Odors were delivered via a computer-controlled olfactometer with a 1,000 ml/minute constant flow. Each odor was dissolved in mineral oil at 1:10 dilution and 15μl of diluted odor solution was applied to the syringe filter (2.7μ m pore, 13mm; Whatman, 6823-1327). Odors were: Cuminaldehyde, (S)-(+)-2-Octanol, (R)-(-)-Carvone, Pentyl acetate and Hexanoic acid.

2.4.3.3 Behavioural task

Mice were water restricted 7 days after head-posting and habituated to head-restrainment for 2-3 days (10-30 min sessions). Within these sessions, mice were allowed to voluntarily lick the spout for a water reward. Following the habituation sessions, mice were trained in an odor-cued classical conditioning task, where an odor cue predicts reward with distinct delay and/or magnitude (*i.e.* volume of water).

The task included five trial types, randomly intermixed. Trial types 1-4 began with a 1-s odor delivery, followed by a delay of 0, 1.5, 3, 6 seconds after odor onset, respectively, and a fixed water reward amount of 4.5 microliters. For one of the animals, the delay of 0s wasn't included in the task. Trial type 5 began with a 1s odor delivery, followed by a 3 seconds delay after odor onset and a reward sampled from a probability distribution with five possible outcomes: 1, 2.75, 4.5, 6.25, 8 microliters with respective probabilities: 0.25, 0.167, 0.167, 0.167, 0.25, such that the mean of this distribution was 4.5 microliters and thus matching the average reward amount delivery across all trial-types. The odor identity associated with each trial type was shuffled for different individual mice. Importantly, at the end of the session, when 200 trials had passed, there was a context switch and either the cue predicting

the shortest or the longest delay was removed. Trial duration was drawn from an exponential distribution (minimum 11s, mean 4.5s, truncated at 21s), resulting in an approximately flat hazard function and an approximately constant reward rate throughout the session.

2.4.4 Electrophysiology

All electrophysiological experiments were conducted while mice were head-restrained. Recordings were performed for up to 6 days following a 24-48 hours recovery period from the craniotomy surgery. Between recording sessions, the craniotomies were covered in the same manner as described above. A two-shank 64-channel silicon probe (ASSY 77-H6, Cambridge NeuroTech) with a tapered optical fibre (Lambda-B fibre 100-µm core NA=0.48, Cambridge NeuroTech) glued to the back, was lowered into the ventral tegmental area (VTA) and substancia nigra pars compacta (SNc). Probe tracks were reconstructed from three different sessions by dipping the probes in DiD, DiO and DiI solutions before insertion. Electrophysiology and laser/LED modulation data were digitized at 30kHz with the Open Ephys acquisition board [121] and recorded with Bonsai [83].

2.4.5 Spike sorting and data processing

In order to remove light artifacts, independent component analysis (ICA) was performed [8]. In particular, we used chuncks of recorded signals from the beginning and end of each session where pulses of light were given, and used fastICA [56] to obtain the independent components and the mixing matrix. Light artifacts were present simultaneously in the majority of the channels, hence we considered as artifacts the two components with highest entropy of the mixing matrix weights, and after visual inspection, removed these components and reconstructed the signals. Data were sorted offline with Kilosort 2.5 spike sorting software
(https://github.com/MouseLand/Kilosort), and manually curated using Phy
(https://github.com/cortex-lab/phy).

2.4.6 Light identification of dopamine neurons

The light evoked photoactivation of midbrain dopaminergic neurons with ChR2 was used to identify neurons as dopaminergic. The protocol consisted of trains of 10 blue (473 nm) light pulses, each 10 ms long, at 1, 5 and 20 Hz (5s inter-train interval) followed by 3 consecutive long pulses, lasting for 1s each. Laser power at the tip of the fiber was on average 20 mW (measured at the tip of the fiber, before each experiment). Optogenetic stimulation was delivered twice at the beginning and twice at the end of the recording session. Units were considered photo-identified using an intersection of different criteria: SALT test [75] p-value <0.001, paired t-test comparing baseline and post-laser onset (1–10ms) firing rate yielding a p-value<0.01, correlation coefficient between laser-triggered waveform, non-evoked waveform >0.9 and probability of eliciting ≥1 spikes within 1-10ms of each pulse >0.1. To be included in the data set, a neuron had to be well-isolated (inter-spike-interval (ISI) violation <0.04 [51]).

2.4.7 Distributional code for reward time model

We describe our model in the context of a class of Markov decision processes (MDPs) which use a deterministic chain of states to model the temporal evolution of a reward function across time. This is known as the complete serial compound representation [86]. Such MDPs $(\mathcal{S}, \mathcal{P}, r)$ are composed of a set of time states $\mathcal{S}$, a reward function $r : \mathcal{S}^N \to \Delta(\mathbb{R})^N$ that for each time states $s_t \in \mathcal{S}$ generates a stochastic reward and a deterministic transition function between time states $\mathcal{P}(s_{t+1}|s_t) = 1$. The value at time state s_t is the expected sum of discounted future rewards over a time range from 0 to L,

$$V(s_t) = \mathbb{E}\left[\sum_{j=0}^{L} \gamma^j r(s_{t+j})\right] = \sum_{j=0}^{L} \gamma^j \bar{r}(s_{t+j}), \tag{2.1}$$

where $\bar{r}(s_t) = \mathbb{E}[r(s_t)]$ and $\gamma \in [0, 1]$ is the temporal discount factor that determines how much the utility of delayed rewards is decreased relative to immediate reward. We highlight that for each time state $s_t \in \mathcal{S}$, there is a corresponding positive real valued time $t \in T \equiv \mathbb{R}^+$. In the next section we will refer to the real valued time.

2.4.7.1 Efficient population coding and the expectiles of reward times

In our experiment, reward time and amount is manipulated across conditions and thus our dopamine population model describes adaptive neural responses to both these features of the environment reward function. Initially, we describe a dopamine neuron's tuning function with respect to reward time $t_r \in T$ since the cue presentation. It is proposed that, across the dopamine neuron population, reward is heterogeneously discounted over time, and reflect this assumption in exponentially decaying neural tuning function parameterized by reward time scale $\tau \in \mathcal{T} \equiv \mathbb{R}^+$, with a gain parameter a

$$\delta_\tau(t_r) = a e^{\frac{-t_r}{\tau}} \quad , \tag{2.2}$$

where a is the response to an immediate reward and τ is the time at which the neural response is reduced to $\frac{1}{e}$ times its initial value. Considering this parametrization, the corresponding temporal discount factors γ (see Eqn. 2.1) arise naturally as a function of the reward time scales τ, specifically $\gamma = e^{-\frac{1}{\tau}}$. These time scales are estimated considering the responses to times for which reward was given. To map the set of neuron's reward time constants $\mathcal{T}$, to the space of possible reward times T we consider a mapping $\iota : \mathcal{T} \rightarrow T$.

In order to predict optimal variability across the reward time scales estimated from dopamine neuron responses, we develop an efficient population coding model[41] for the distribution $p(t_r|s_0) := p(t = t_r|r \neq 0, s_0)$ of reward times t_r after the stimuli presentation at time 0. We propose that this distribution is efficiently encoded in an expectile code, as opposed to the quantile code. Indeed, previous experimental work has shown that the tuning functions of dopamine neurons of reward magnitudes are more consistent with bi-linear functions, that would be predicted for an expectile code, then with heaviside functions, that would be predicted for a quantile code [24]. Thus, for consistency with these previous results, we use an expectile code model based on expectations of reward times. Theoretically, we accomplish this by optimizing our dopamine population model to efficiently encode the distribution of expectiles over reward times. In contrast, the quantile code may be computed in our formalism by efficiently encoding the distribution over reward times [111]. Importantly, though our theoretical formalism is general and integrative over both expectile and quantile codes, our use of the expectile code in particular does not qualitatively change our model predictions.

The expectiles of a distribution generalize the mean statistic analogously to how quantiles generalise the median [111]. Given reward times t_r with probability distribution $p(t_r|s_0)$, the η-expectile $\bar{t_r}$ satisfies [102]:

$$\eta\mathbb{E}[(\bar{t_r} - t_r)^-] = (1-\eta)\mathbb{E}[(\bar{t_r} - t_r)^+] \ .$$

For example, the mean corresponds to the expectile with level $\eta = 0.5$. The expectiles are distributed according to the cumulative distribution,

$$P(\bar{t_r}) \sim \frac{\mathbb{E}[(\bar{t_r} - t_r)^+]}{\mathbb{E}[|\bar{t_r} - t_r|]} \ .$$

An explicit representation of this cumulative distribution has also been derived previously [60].

Given a population activity vector of N dopamine neurons $\vec{\delta}$, we sought to maximize the mutual information $I(\bar{t_r}, \vec{\delta})$ between the true expectile reward times $p(\bar{t_r}|s_0)$ and those encoded by the population $p(\bar{t_r}|\vec{\delta})$, constraining on the number of neurons and on the population expected firing rate R. We consider a population with homogeneous derivatives,

$$\delta_\tau(t_r) = -\int_{t_r}^{+\infty} \delta'_\tau(s)ds \ , \tag{2.3}$$

that approximately tiles the range of possible reward times from 0 to L, have constant Fisher information and where each neuron is subject to independent Poisson noise. A population that linearly tiles the exponential decay half-lives approximately satisfies these conditions, however this construction has edge effects at 0 which does not affect the capability of this model to encode future (non-zero) reward times. Since computing the mutual information is analytically intractable, a lower bound is optimized instead, namely the Fisher information [15]. We parameterize the optimized dopamine tuning curves $\vec{\delta}_{(d,g)}$ with an invertible neural *density* function[3] $d : \mathcal{T} \to \mathcal{T}$ which characterizes the heterogeneous allocation of neurons to reward time scales $\tau \in \mathcal{T}$ and a *gain* function $g : T \to T$ which characterizes the mean firing rate across reward times $t \in T$,

$$\begin{aligned} \text{Population mapping: } \vec{\delta}(t_r) &\to \vec{\delta}_{(d,g)}(t_r) \\ \text{Neuron mapping: } \delta_\tau(t_r) &\to \delta_{(d(\tau),g(t_r))}(t_r) \ , \end{aligned} \tag{2.4}$$

[3]Note that this is a distinct concept to a probability density function. It is necessarily invertible and, when maximally efficient from a neural coding perspective, a one-dimensional neural density function corresponds to the corresponding cumulative probability distribution function of the encoded stimulus [41].

where $d(\tau)$ is the density of tuning curves at reward time scale τ in the optimized population and $g(t_r)$ is the population mean firing rate for reward time t_r. The Fisher information of the optimized population is given by [41],

$$I_f^d(t_r) \propto g d^2 \ .$$

After constraining on the number of neurons N, the solution for the density function d is given by,

$$d(\tau) \propto N p(\iota^{-1}(\bar{t_r})|s_0) \ , \tag{2.5}$$

i.e., the population reward time scales should distribute proportionally to the expectile rewards times in the environment. On the other hand, additionally constraining on the mean population firing rate R, considering the population defined in Eqn. 2.3, the population gain function g should satisfy,

$$\int_{-\infty}^{+\infty} p(t_r|s_0) \int_{t_r}^{+\infty} -d(u)g(u)dudt_r = R \ . \tag{2.6}$$

Integrating by parts to obtain the mean population firing rate as,

$$\int_{-\infty}^{+\infty} P(t_r|s_0)d(t_r)g(t_r)dt_r = R \ , \tag{2.7}$$

where $P(t_r|s_0)$ is the cumulative distribution of $p(t_r|s_0)$. Therefore the solution for the population gain is given by,

$$g(t_r) = \frac{R}{NP(\iota(\tau)|s_0)} \ , \tag{2.8}$$

i.e., for each neuron the gain should be inversely proportional to the probability that a randomly chosen reward time will be smaller than it's time scale. For small reward times, the entire population is active, incurring a large metabolic cost for encoding these values. Intuitively, this metabolic penalty can be reduced by lowering the gains tuned to long reward times and therefore large cumulative probabilities $P(\iota_t(\tau)|s_0)$ in Eqn. 2.8.

2.4.7.2 Distributional learning mechanism

In our analysis thus far, we have identified a particular density profile of neural tuning functions that optimally encode the distribution of reward times following a stimulus presentation. In this section, we describe how

such a neural population may, in principle, be optimized via an online learning process.

We pursue an algorithmic strategy inspired by the distributional learning of reward magnitudes as proposed previously [23]. In this approach, multiple channels with different relative scaling for over- or under-estimation of reward magnitude (α_i^+, α_i^-) leads to a diversity of learned values,

$$\begin{cases} V_i \leftarrow V_i + \alpha_i^+ \delta_i^M, \text{ if } \delta_i^M > 0 \ , \\ V_i \leftarrow V_i + \alpha_i^- \delta_i^M, \text{ if } \delta_i^M < 0 \ . \end{cases}$$

which collectively encode the entire reward magnitude distribution.

With respect to the distribution of reward times $p(t_r|s_0)$, we consider multiple channels with different relative scaling for over- and under-estimation of reward times thus generating a heterogeneous set of time scales. If t_r is the time of reward on a given trial, the corresponding prediction error is given by $\delta_j^T = t_r - \tau_j$ and the update rules are

$$\begin{cases} \tau_j \leftarrow \tau_j + \alpha_j^+ \delta_j^T, \text{ if } \delta_j^T > 0 \ , \\ \tau_j \leftarrow \tau_j + \alpha_j^- \delta_j^T, \text{ if } \delta_j^T < 0 \ . \end{cases}$$

These learning rules converge to the *quantiles* of the probability distribution over expectile reward times $p(\bar{t_r}|s_0)$ [23]. This is because these quantiles are uniformly distributed in the cumulative probability space (i.e. the domain of the cumulative distribution function $P(\bar{t_r}|s_0)$) over expectile reward times, and satisfy the optimal information-theoretic condition defined above (Eqn. 2.5).

2.4.7.3 Multi-dimensional integration over reward magnitude and time

Importantly, while in the distributional code for magnitude, the slope asymmetry $\kappa = \frac{\alpha^+}{\alpha^+ + \alpha^-}$ controls the level of *optimism* in individual units [23], in the temporal coding model developed here it controls the reward time scale and therefore the temporal discount factor, also known as *impatience* [46]. Integrating the distributional models in reward magnitude and time, leading to the time magnitude reinforcement learning (TMRL), each neuron is characterized by an optimism level κ and an impatience level η, corresponding to the temporal discount factor γ_j (which is induced from a corresponding reward time scale τ_j). Therefore the temporal-difference vector RPEs take

the form,

$$\delta_{ij}^{M}(s_t) = r + \gamma_j \tilde{V}_j(s_{t+1}) - V_{ij}(s_t) \ , \tag{2.9}$$

where $\tilde{V}_j(s_{t+1})$ is a random sample from the value distribution at the next state s_{t+1} temporally discounted by γ_j. This is denoted by the *imputation* step, that implies a non-local update rule, as shown in previous work [111, 24]. Understanding if, in the population of dopamine neurons, this manifests as a non-local update rule or may be implemented locally remains an open question. Finally, the multi-dimensional distributional value update rule is given by,

$$V_{ij} \leftarrow V_{ij}(s_t) + \alpha_i^{sgn(\delta_{ij}^{M})} \delta_{ij}^{M}(s_t) \ . \tag{2.10}$$

Algorithm 1 Time-Magnitude RL (TMRL)

Inputs: Set of optimism levels over reward magnitudes indexed by i, $\kappa_i = \alpha_i^+/(\alpha_i^+ + \alpha_i^-)$, and over reward times indexed by j, $\eta_j = \alpha_j^+/(\alpha_j^+ + \alpha_j^-)$.
Initialize $V_{i,j}$ for all pairs i, j and states $s_t \in \mathcal{S}$.
Loop for each episode:
  Initialize state s_0.
  For each time-step t and state s_t:
    Observe r and s_{t+1}
    For all i,j:
      $\delta_{ij}^{M}(s_t) = r + \gamma_j \tilde{V}_j(s_{t+1}) - V_{ij}(s_t)$
      $V_{ij} \leftarrow V_{ij} + \alpha_i^{sgn(\delta_{ij}^{M})} \delta_{ij}^{M}(s_t)$
    If $r > 0$:
      For all j:
        $\delta_j^T = t - \tau_j$
        $\tau_j \leftarrow \tau_j + \alpha_j^{sgn(\delta_j^T)} \delta_j^T$
        $\gamma_j = e^{-1/\tau_j}$
  Until terminal state.

2.4.8 Data analysis

Spike counts were binned in 2-ms windows and smoothed by convolving with a gamma probability distribution kernel (shape parameter k =2 and scale parameter θ =25ms).

Responses to reward were defined as the average activity from 200 to 650 ms after cue and reward onset, baseline subtracted by the mean activity over trials from -1000 ms to 0 ms relative to cue onset. Responses to cue

were defined as the average activity from 200 to 650 ms after cue onset. This window was selected in order to exclude the initial response to the solenoid valve opening, that was in the majority of the neurons not selective to reward amount or delay, as also shown in previous literature [129]. For each dopamine neuron j, we assumed the tuning function of reward time (t_r) was given by an exponential decaying function [69],

$$\delta_j(t_r) = a_j \gamma_j^{t_r} \quad , \tag{2.11}$$

where γ_j is the temporal discount factor and a_j is a gain parameter. The parameters were estimated minimizing the mean squared error between $\delta_j(t_r)$ and the mean responses at the cue for each reward time.

To test whether the estimated temporal discount factors did not reflect noise we divided the trials in two random partitions, such that each partition contained the same number of trials for each reward delay. Then we estimated the temporal discounts for each partition and measured the linear regression slope, correlation coefficient and two-tailed p-value for these sets of estimates. We repeated 10,000 times this procedure, and computed the mean correlation coefficient, the geometric mean of the obtained p-values and the 95% confidence interval of the regression slopes.

In order to test if the single neuron's estimated temporal discounts are significantly different from the population mean temporal discount, we took randomly selected 50% of trials per delay, estimated the temporal discounts and repeated 1000 times, to obtain confidence intervals.

To test if changes in the reward time statistics would lead to an adaptation of the population temporal discounts, we removed either the longest or shortest delay and estimated the temporal discount before and after this context switch. The responses to the same delays were used to estimate the temporal discounts before and after the manipulation. After the context switch, we did not consider the first 5 trials. To assess the degree of relative adaptation of the population activity when removing the shortest or longest delay at the end of the session, we bootstrapped considering 10,000 resamples and computed the two-tailed p-value and 95% confidence interval (for the null hypothesis that there is no adaptation in the mean population temporal discounts). We also performed a test for adaptation of the population temporal discounts separately, for each type of context switch, bootstrapping using 10,000 resamples and computing the one-tailed p-value and 95% confidence interval.

To test whether discounting by dopamine neurons was updated more continuously and determine the time-scale of adaptation, we computed the rate of reward occurrence, by convolving the occurrence of rewards with exponential discounting kernels with different time scales. We then measured the adaptation in temporal discounts for a given time scale by comparing the update in temporal discounts for relatively low and high rates (1st and 3rd terciles respectively), bootstrapping the one-way p-value for the null hypothesis that the temporal discounts for higher rates is shallower than for lower rates (considering 10,000 resamples) (Figure 2.7).

On the other hand, for each dopamine neuron i, we assumed the tuning function of reward magnitude (r) was given by a bilinear function,

$$\begin{cases} \delta(r) = \alpha_i^+(r - V_{\kappa_i}), \text{ if } r > V_{\kappa_i} \ , \\ \delta(r) = \alpha_i^-(r - V_{\kappa_i}), \text{ if } r < V_{\kappa_i} \ , \end{cases}$$

where α_i^+ if the slope for the positive RPEs, α_i^- the slope for negative RPEs, $\kappa_i = \frac{\alpha_i^+}{\alpha_i^+ + \alpha_i^-}$ is the asymmetry in slopes for positive and negative RPEs and V_{κ_i} is the reversal point. As in previous work [23], the reversal point was defined as the magnitude M that maximized the number of positive responses to rewards greater than M plus the number of negative responses to rewards less than M. After measuring reversal points, we fit linear functions separately to the positive (α_i^+) and negative (α_i^-) domains of each cell.

We measure the correlation between the firing rate at the cue corrected for the temporal discount and gain (Equation 2.11) and report the correlation coefficient, two-tailed p-value and 95% confidence interval.

To test if the variance of estimated reversal points for the variable cue was significantly greater than for the certain one, we bootstrapped considering 10,000 resamples and computed the one-tailed p-value and 95% confidence interval.

To classify an event as a lick it had to have a minimum duration of 0.015s. Licks were binned in 0.13s windows and smoothed by convolving with a exponential decaying function with a time scale of 0.77s to obtain lick rates. The anticipatory licking slope, described in Figure 2.4G, was defined as the linear regression slope in licking rate from t =0.01s to t = reward time.

Pupil diameter was estimated from video frames. Frames were first cropped to regions-of-interest around the eye and then analyzed using

DeepLabCut (DLC) [89] . For training the DLC network, 8 points around the pupil were labelled in a subset of 20 frames from the training sessions videos one week prior to recordings and all recording sessions. Subsequently the network was trained and the points for the remaining frames were predicted. The session was included in the analysis, if the median DLC confidence was higher than 0.85. The pupil diameter was estimated as the average over the distance between pairs of points opposite diametrically and weighted by the mean confidence of these pairs. Frames with a DLC confidence smaller than the median session confidence minus 0.1 were removed. Additionally, frames where the absolute difference between the estimated diameter were bigger than 10 were removed. The pupil diameter for these frames were then interpolated. To remove fast fluctuations that do not reflect pupil diameter dynamics, we used a low-pass butter filter, with a cutoff of 10Hz. The pupil diameter was normalized by the 0.1-quantile of pupil diameter over the session. We estimated the time-scale each animal used to integrate rewards by searching for the time-scale that maximized the mean difference between reward history terciles (normalized by the maximum) over the full width half maximum window. For each animal, the reward history was then computed by convolving the trial-to-trial reward magnitudes with an exponential decay kernel with the previously defined time-scale.

2.4.8.1 Statistics

Statistical analyses were performed in Python. No statistical methods were used to predetermine sample sizes. Performed statistical tests, sample size, relevant statistics and P values are reported throughout the text or in the respective figure caption.

2.4.9 Future reward distribution decoding

2.4.9.1 Decoding future distribution of reward times

As defined before in Equation 2.1, the value considering a time horizon of L is given by,

$$V_{\gamma_j} = \sum_{t}^{L} \gamma_j^t \bar{r}(s_t) \ . \tag{2.12}$$

The problem of determining the expected future rewards over time $\bar{\mathbf{r}}(t) = (\bar{r}(s_1), \ldots, \bar{r}(s_L))$ can be seen as a linear regression problem. The data points

correspond to the population temporal discounts $\{\gamma_1, \ldots, \gamma_N\}$, the basis functions are given by $\phi(\gamma) = (\gamma, \gamma^2, \ldots, \gamma^T)$ and the targets are $\mathbf{V}_\gamma = \{V_{\gamma_1}, \ldots V_{\gamma_N}\}$, which are the single neuron's responses at the cue. Assuming a uniform prior and a Gaussian likelihood functions with inverse variance β_j the log of the posterior is given by,

$$\ln p(\bar{\mathbf{r}}(s_t)|\mathbf{V}_\gamma) = \sum_{j=1}^{N} -\frac{\beta_j}{2}(V_{\gamma_j} - \bar{\mathbf{r}}(s_t)^T \phi(\gamma_j))^2 + \ln\left(\frac{\mathbf{1}}{L}\right) \ .$$

When the reward is either 0 or 1 the expected rewards at a given time state s_t corresponds to the probability of rewards at that time, therefore the log posterior becomes,

$$\ln p(t_r|\mathbf{V}_\gamma) = \sum_{j=1}^{N} -\frac{\beta_j}{2}(V_{\gamma_j} - \bar{\mathbf{r}}(t)^T \phi(\gamma_j))^2 + \ln\left(\frac{\mathbf{1}}{L}\right) \ . \tag{2.13}$$

In practice, for each neuron we considered V_j the responses of each neuron normalized by the estimated gain and β_j was the estimated inverse of the variance of V_{γ_j}. The solution that maximized (2.13) was determined analytically using singular value decomposition (SVD), as proposed in [151, 138]. To obtain a probability distribution, the solution was normalized. A qualitative search was done on the smoothing parameter, to maximize the similarity between the estimated and the true probability distribution.

For Figure 2.4E, we computed the mean decoded densities using the population responses in trials with anticipatory licking slope greater or smaller than the median. For Figure 2.5E, we computed the mean decoded reward times and magnitudes using the population responses in trials with reward magnitude history equal to the quintiles.

2.4.9.2 Decoding future distribution of rewards amounts

The population firing rate at the cue corrected for the diversity in gain and temporal discount factor ($\vec{\delta}_{\text{cue}}$) is significantly correlated with the reversal points estimated at the reward delivery. We therefore assume an additive Gaussian noise model,

$$p(\mathbf{V}_\kappa|\vec{\delta}_{\text{cue}}) \sim \mathcal{N}(\mathbf{V}_\kappa|v\vec{\delta}_{\text{cue}}, \beta^{-1}) \ . \tag{2.14}$$

In practice, for each neuron, β_i was the inverse of the estimated variance of V_{κ_i}. Each dopamine neuron i has a different level of optimism κ_i, by

weighting asymmetrically positive (α_i^+) and negative (α_i^-) RPEs, with $\kappa_i = \frac{\alpha_i^+}{\alpha_i^+ + \alpha_i^-}$, predicting different values, V_{κ_i}. We measure the responses of each neuron at only five different reward amounts in a small range of rewards, from 1μl-8μl. We use Monte Carlo simulations to estimate the bias and variance we are inducing in the estimation of the κ parameter. For these simulations we assume the noise in the responses is Gaussian and use the variance of each neuron's responses for each reward amount to generate responses. We observe that there is a systematic relationship in the induced bias and variance: the variance increases quadratically with the distance to the mean of the reward magnitudes and the absolute bias increases linearly with the distance to the mean of the reward amounts, and it is positive or negative for reversal points smaller or greater than the mean reward amounts, respectively. We therefore subtract the induced bias to the estimated κ_i's. We use a isotonic regression to model the relationship between the reversals V_{κ_i}'s and κ_i's [17], which only assumes this is an increasing function. To estimate the piecewise linear functions in isotonic regression we take in account the variance in the estimation of each κ_i. We assume the population of N neurons is minimizing the loss function [111]

$$\sum_{i=1}^{N} \mathbb{E}_{r\sim R}\left[\rho_{\hat{\kappa}_i}(r - V_{\hat{\kappa}_i})\right] = \mathbb{E}_{r\sim R}\left[\sum_{i=1}^{N} \rho_{\hat{\kappa}_i}(r - V_{\hat{\kappa}_i})\right] \ ,$$

where

$$\rho_\kappa(u) = \begin{cases} \kappa u^2, \text{ if } u > 0 \ , \\ (1-\kappa)u^2, \text{ else.} \end{cases} \tag{2.15}$$

Considering asymmetric Gaussian likelihood functions [107] with inverse variance β_i, we obtain the posterior over reward amounts,

$$p(r|\mathbf{V}_\kappa, \vec{\delta}_{\text{cue}}) \propto \sum_{i=1}^{N} \rho_{\kappa_i}(r - V_{\kappa_i})\beta_i \approx \sum_{i=1}^{N} \rho_{\hat{\kappa}_i}(r - v\delta_{\text{cue}})\beta_i. \tag{2.16}$$

2.4.9.3 Decoding future rewards over time and amount

By integrating constant reward magnitudes, the TMRL model (equation 2.10) converges to a diverse set of temporally discounted values at the cue that we represented by $\mathbf{V}_\gamma = \{V_{\gamma_1}, \dots V_{\gamma_N}\}$ in section 2.4.9.1. And on the other hand, the model converges at the time of reward delivery to a diverse set of values for variable rewards, that we represented by $\mathbf{V}_\kappa = \{V_{\kappa_1}, \dots V_{\kappa_N}\}$ in

section 2.4.9.2. Finally, in the case of variable reward magnitudes, the values at the cue are influenced both by the variability in temporal discounting and magnitude optimism level and the general posterior distribution over reward magnitude and time is given by,

$$p(r,t|\mathbf{V}_\kappa,\mathbf{V}_\gamma) \propto p(\mathbf{V}_\kappa,\mathbf{V}_\gamma|r,t)p(r,t) \quad . \tag{2.17}$$

There is evidence in our data set that the diversity in dopamine tuning functions to reward time is independent of the diversity in reward magnitude, hence we can factorize the above equation,

$$\begin{aligned} p(r,t|\mathbf{V}_\kappa,\mathbf{V}_\gamma) &\propto p(\mathbf{V}_\kappa|r,t)p(\mathbf{V}_\kappa|r,t) \\ &\propto p(r,t|\mathbf{V}_\kappa)p(r,t|\mathbf{V}_\gamma)p(\mathbf{V}_\kappa)p(\mathbf{V}_\gamma). \end{aligned}$$

Conditioned on the cues, the amounts and times of reward are independent in our behavioral task, hence we assume the joint distribution over reward amounts and times can be factorized,

$$p(r,t|\mathbf{V}_\kappa,\mathbf{V}_\gamma) \propto p(r|\mathbf{V}_\kappa)p(t|\mathbf{V}_\gamma)p(\mathbf{V}_\kappa)p(\mathbf{V}_\gamma). \tag{2.18}$$

Thus we derive a probabilistic decoding framework for independently decoding reward time and magnitude. This, despite the limited numbers of neurons in our dataset, allows for the decoding of the reward distribution over these two dimensions in practice. In principle, with a sufficiently large number of neurons this assumption can be relaxed and equation 2.17 can be used to decode the joint probability distribution.

2.4.10 Foraging simulations

In the foraging simulations in Figure 2.6 we implemented a temporal-difference learning algorithm for estimating the value of each patch (TDRL). Simultaneously, we implemented a temporal-difference algorithm for estimating the occupancy matrix, that was multiplied by the reward, to obtain the successor representation (SR) estimation of value for each patch. For the TMRL we considered a set of values with different temporal discounts and optimism levels, and with these, we decoded the probability distribution over future reward time and magnitude for each patch. Initially the values for all algorithms were set to zero. The learning rate was set to 0.02. The environment was non-stationary: patch one retrieved a reward of reward of

2.5 between time-step 0 and 3, patch two a reward of 1 or 4 between time-step 0 and 3 and patch three a reward of 4 between time-step 6 and 11.

We model the utility function in time as a decaying function in time parametrised by the temporal discount factor (γ^t) and the utility in magnitude as a non-decreasing function in magnitude parametrised by a power (r^p). We consider the utility matrix U defined over a discretized time range, $\{1, \ldots, j, \ldots, T\}$, and magnitude range, $\{r_1, \ldots, r_i, \ldots, r_N\}$,

$$U = \begin{bmatrix} u_{11} & \ldots & u_N u_{1N} \\ & \ddots & \\ u_{N1} & & \ldots u_{NN} \end{bmatrix} = \begin{bmatrix} r_1^p \gamma^1 & \ldots & r_N^p \gamma^1 \\ & \ddots & \\ r_N^p \gamma^T & & \ldots r_N^p \gamma^T \end{bmatrix}. \tag{2.19}$$

The change in internal state from hungry to sated was modelled as a change in the intrinsic value or utility of reward magnitude. When the mice was sated the temporal discount was set to $\gamma = 0.6$ and the utility function in reward magnitude was considered to be linear ($\{r_1, \ldots, r_N\}$ in 2.19) [62]. When the mice became hungry, the utility function in reward magnitudes was considered to be convex with a power of $p = 2$ ($\{r_1^2, \ldots, r_N^2\}$ in equation 2.19) and the temporal discount factor was kept the same. On the other hand, the change in time-scale of the environment, when the mouse had more or less time to forage, from dusk to dawn, was modelled as a change in the temporal discount factor from $\gamma = 1$ to $\gamma = 0.6$. The interaction between both dynamics in internal state and the environment, when the animal was hungry and there was an adaptation in the time-scale of the environment from dusk to dawn was modelled as a change in the temporal discount from $\gamma = 1$ to $\gamma = 0.6$ and a change in the reward magnitude utility from linear to convex with a power of $p = 2$.

Since the TMRL allows for the decoding of a probability distribution over future reward time and magnitude, it can immediately recompute the value of each patch for the new state. In particular, for each time-step in the future, the probability distribution over reward magnitudes p_{ij} is normalized, such that for all j,

$$\sum_i p_{ij} = 1.$$

Then, each entry of this matrix is multiplied by the respective element in the utility matrix and summed to obtain the value for each patch,

$$V = \sum_{i,j} p_{ij} u_{ij}. \tag{2.20}$$

The SR agent can also recompute the value function for the hungry state by flexibly updating the reward. However, for the adaptation in the temporal discount factor from dusk to dawn, the agent needs to re-learn through experience to update the temporally discounted future occupancy. On the other hand, the standard value TDRL agent needs to learn through experience to compute the value of each patch for all types of state changes. In Figures 2.9F-H the fraction of times the optimal action was selected after a state change, considering a greedy policy, is reported.

3

The unified framework for multi-dimensional distributional neural learning

3.1 Introduction

Efficient neural coding, a long-established predictive framework for computational neuroscience [6, 8], has been re-energized by novel theoretical developments and high-dimensional neural population recordings [131]. In particular, recent work has generalized classical analyses for single units sensitive to one-dimensional stimulus features, to populations of neurons encoding multi-dimensional stimulus spaces [153, 41]. However, these theories only describe the offline optimization of efficient population codes and lack a model of learning through online interactions with an environment. In a distinct area of computational neuroscience overlapping with machine learning, many advances in state-of-the-art RL are variants of the novel distributional RL suite of algorithms for which there is evidence of its implementation in the brain [23]. Distibutional RL facilitates the online learning of distributional population codes (each "unit" corresponding to a quantile of the reward distribution) however only for one-dimensional feature spaces (i.e. reward magnitude). We present a unifying framework that integrates efficient coding theory and distributional RL as special cases and combines the best of both of these theoretical paradigms. Algorithmically, our novel framework results in online learning equations for dynamically updating multi-dimensional distributional population codes, stimulus sample

by stimulus sample. We show that these population codes converge on the globally optimal (efficient) codes asymptotically and analytically characterize the non-asymptotic finite-sample suboptimality associated with learning variability. As an example case, we model the transformations of place cell tuning due to the animal experiencing local environment deformations [73], which could not be previously modeled in an efficient coding framework since it requires learning a two-dimension probability distribution. More generally, we suggest that our unified theory provides a complete framework for learning optimal neural distributional population codes in the brain.

Author contributions

Margarida Sousa and Daniel McNamee developed the theory, with input from Joe Paton. Margarida Sousa performed the simulations supervised by Daniel McNamee.

3.2 Results

3.2.1 One-dimension: efficient coding and distributional learning

We begin by introducing an efficient population coding framework for one-dimension stimulus [41], followed by a description of how the distributional learning rules converge to this code.

The fundamental theoretical hypothesis of efficient coding is that neural tuning functions are organized to maximize the amount of information about an input stimulus in the neural responses. This is formalized as the *mutual information* $I(\vec{r}, s)$ between the neural population response $\vec{r}$ and the stimulus s. For one-dimensional stimuli s, Ganguli and Simoncelli [41] parametrize the tuning functions of a population of N neurons by a neural *density* function $d(s)$ which characterizes the heterogeneous allocation of neurons to stimulus level $s \in \mathcal{S}$ and a *gain* function $g : \mathcal{S} \to \mathbb{R}$ which characterizes the mean firing rate across stimuli. The mutual information $I(\vec{r}, s)$ can then be maximized analytically with respect to $d(s)$ and $g(s)$ using calculus of variations. The optimal density function is given by $d(s) = Np(s)$, i.e., neurons should distribute proportionally to the distribution of stimulus, as represented in Figure 3.1A. The optimal gain $g(s)$ should be constant and equal to the mean population firing rate. However, no learning rules that converges to the efficient code were proposed in this work.

More recently, distributional learning algorithms, which learn an approximate representation of a value distribution, have been introduced in the value-based reinforcement learning context [24]. These algorithms learn an efficient code but now for a value distribution rather than a stimulus distribution as described in the previous paragraph. To learn the τ-quantile θ_τ given a sample s, the update rule is (Figure 3.1 B):

$$\begin{cases} \theta_\tau \leftarrow \theta_\tau + \tau, \text{ if } s > \theta_\tau \\ \theta_\tau \leftarrow \theta_\tau - (1-\tau), \text{ otherwise.} \end{cases} \tag{3.1}$$

In summary, while efficient coding theory can optimize neural populations to encode possibly multi-dimensional distributions, distributional reinforcement learning can learn a one-dimensional value distribution online. We ask whether these two foundational theories in computational neuroscience may be unified. Our starting point for our theory is the theoretical observation that both of these algorithm classes are optimizing distributional encoding according to the so-called *Wasserstein metric* [65] which we explore in the next paragraph.

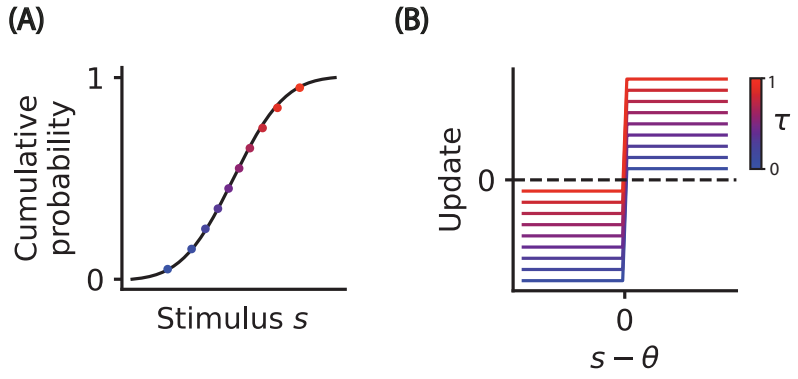

Figure 3.1: Efficient coding for 1-dimensional stimuli. (A) The optimal *neural density* derived in [41], that corresponds to the cumulative probability function, represented in black. The dots are the quantiles color coded by the τ level. (B) Quantile regression update rules, each curve is color coded by the τ level.

3.2.2 Towards online learning of multi-dimensional distributions: the Wasserstein metric

We will now introduce the Wasserstein metric, which will form the basis of our theory regarding how to learn to efficiently encode multidimensional probability distributions. Let μ and ν be an initial and target distributions defined over $\mathbb{R}^d$. Let $\mathcal{T} : U \to V$ denote all transport maps from μ to ν. The p-Wasserstein corresponds to the minimal cost of transporting μ to ν,

$$W_p(U, V) = \left(\inf_{T \in \mathcal{T}} \int ||u - T(u)||^p du \right)^{\frac{1}{p}}. \tag{3.2}$$

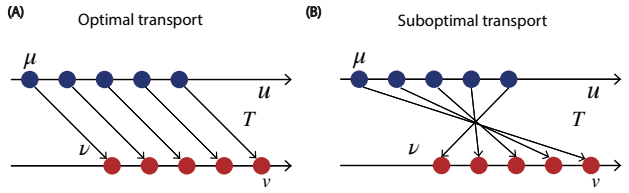

Figure 3.2: (A) Optimal transport from initial distribution μ to target distribution ν is represented. (B) A suboptimal transport from μ to ν is represented, since each point in μ is not mapped to the closest point in ν.

The Wasserstein distance has an intuitive interpretation: imagine there is a pile of dirt, represented by the distribution μ, that needs to be transported into a hole, represented by the distribution ν. The cost of transporting a unit of dirt from a point u to $v = T(u)$ is given by the distance $||u - T(u)||^p$. The Wasserstein distance represents the cost of transporting all the dirt from μ to ν in the most efficient way possible. Thus, it is sometimes referred to as the *earth movers distance* for discrete distributions.

For the 1-dimensional case, the optimal transport has a closed form solution that is given by the mapping from the quantiles of μ to the quantiles of ν,

$$W_p(U, V) = \left(\int_0^1 |F_U^{-1}(\tau) - F_V^{-1}(\tau)|^p d\tau \right)^{\frac{1}{p}}, \tag{3.3}$$

where F denotes the cumulative distribution. In particular, this corresponds to the transport map from the distribution of stimulus to the distribution of neural population responses proposed in Ganguli and Simoncelli's work [41]. However, this transport map can not be trivially generalized to higher dimensions, because unlike in the one-dimensional case, where quantiles

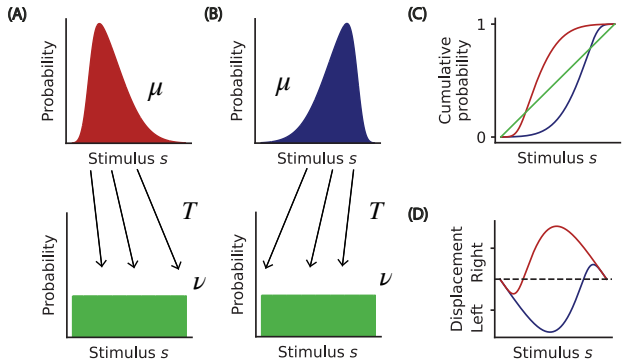

Figure 3.3: (A,B) Optimal transport map from a right (red) and left (blue) skewed continuous distribution to a uniform distribution. (C) The cumulative for all distributions represented in A and B. (D) Distortion induced by each of the transports described in A and B. Since the distribution represented in red has more mass on the left, in order to distribute uniformly the mass, it is transported to the right, and vice-versa for the distribution represented in blue.

are uniquely defined, in higher dimensions, there are multiple ways of partitioning probability distributions into equal-mass parts (Figure 3.4).

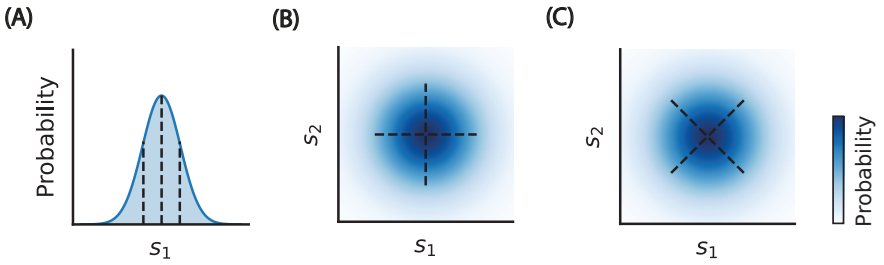

Figure 3.4: (A) In one-dimension the probability distribution can be uniquely divided into equal mass parts. (B,C) However, in 2-dimensions there are infinite ways of dividing it into equal mass parts.

3.2.3 Learning a multidimensional distributional code

In order to overcome this degeneracy problem for high dimensions, we add the Wasserstein metric as an additional cost for the efficiency of the transport map from the initial to the target distribution, the stimulus distribution. In practice, to solve this problem, we resort to a classic result in fluid

dynamics that states that solving the Fokker-Planck equation (FPE) is the gradient flow of a certain function (e.g KL-divergence) in the 2-Wasserstein space of probability measures [61]. We implement this gradient flow non-parametrically using particle approximation techniques [19] by considering the family of delta diracs $\mathcal{P} = \{p_{\boldsymbol{\theta}} : p_{\boldsymbol{\theta}} = \sum_{i=1}^{N} \frac{1}{N} \delta_{\theta_i}\}$ as an approximation of the target distribution

$$p_{\boldsymbol{\theta}_{t+1}} = \underset{p \in \mathcal{P}}{\operatorname{argmin}} \{KL(p||p(\boldsymbol{s}))) + W_2^2(p, p_{\boldsymbol{\theta}_t})\} \tag{3.4}$$

where $p(\mathbf{s})$ can also be discretized and represented by samples from the target distribution.

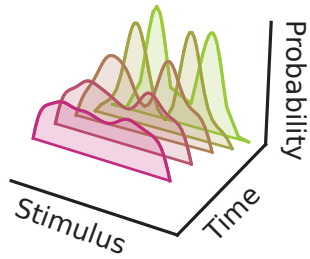

Figure 3.5: An example trajectory of particle learning rules defined in equation 3.4 from an initial uniform distribution to a target bimodal distribution.

By considering a discrete approximation of Equation 3.4 [19] we get

$$p_{\theta_{t+1}} = \underset{p \in \mathcal{P}}{\operatorname{argmin}} \{\underbrace{-\mathbb{E}_p[\log p(\boldsymbol{\theta}|\mathbf{s})]}_{F_1} + \underbrace{\mathbb{E}_p[\log p] + \frac{1}{2h} W_2^2(p, p_{\boldsymbol{\theta}_t})}_{F_2}\} \tag{3.5}$$

with global convergence guarantees in the large sample limit. The F_1 term attracts the particles towards the stimulus samples and the F_2 term regularizes the particle iteration. The discrete gradient flow rules can be derived for each particle and we obtain the distributional neural learning (DNL) update rules

$$\theta_i^t \leftarrow \theta_i^{t-1} - \alpha \left(\frac{\partial F_1}{\partial \theta_i} + \frac{\partial F_2}{\partial \theta_i} \right), \tag{3.6}$$

where

$$\frac{\partial F_1}{\partial \theta_i} = -\nabla_{\theta_i} \log p(\theta_i|\mathbf{s}),$$

and

$$\frac{\partial F_2}{\partial \theta_i} = \sum_j 2c\left(\frac{d_{ij}}{\lambda} - 1\right) e^{\frac{-d_{ij}}{\lambda}} (\theta_i - \theta_j^{t-1}), \tag{3.7}$$

where $d_{ij} = \|\theta_i - \theta_j\|^2$, c is the weight given to regularizing the iterations between particles, and λ defines how close particles should be (Figure 3.6). If $\frac{d_{ij}}{\lambda} > 1$, then θ_i is pulled towards θ_j, with force proportional to $(\frac{d_{ij}}{\lambda} - 1)e^{\frac{-d_{ij}}{\lambda}}$. If $\frac{d_{ij}}{\lambda} < 1$, then θ_i is pushed away.

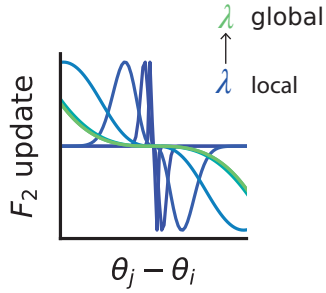

Figure 3.6: The F_2 update for different values of λ: as λ increases, the update rule becomes more global in its effect.

When considering a joint distribution over two dependent random variables that cannot be factorized, DNL is capable of accurately representing the full joint distribution (Figure 3.7 A), when comparing with the factorized 1-dimensional quantile regression learning (Figure 3.7 B).

3.2.4 Learning efficiency for the multidimensional distributional case

In the non-asymptotic finite-sample regime, our theory enables us to characterize the degree to which DNL results in neural coding trajectories which deviates from the optimal transport maps. Essentially, DNL generates globally optimal learning trajectories when the Jacobian of the updates (Equation 3.6) remains symmetric. This is guaranteed for convex target density functions (see a radially symmetric density example in Figure 3.8 and full proof in [67]). Otherwise, DNL generates locally optimal maps which integrate to globally *twisted* transport maps (see rotation and shearing example in Figure 3.8).

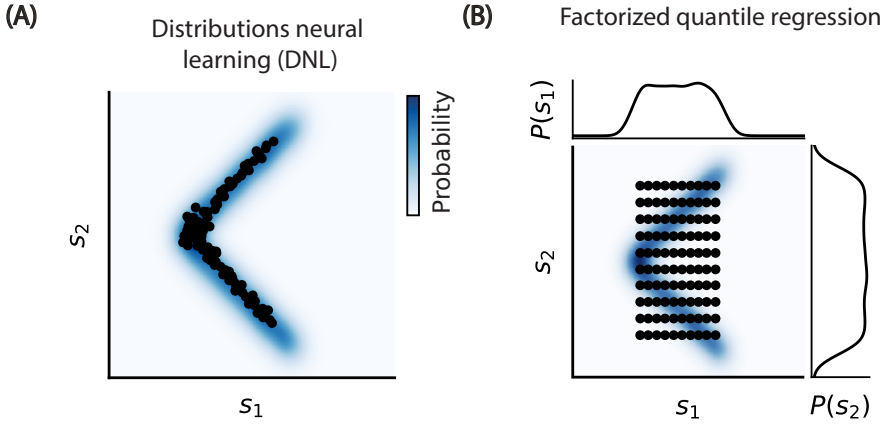

Figure 3.7: (A) The joint distribution over s_1 and s_2 is shown as a blue color map, the learnt particles are represented by black data points. (B) The quantiles of the factorized joint distribution are represented by black data points.

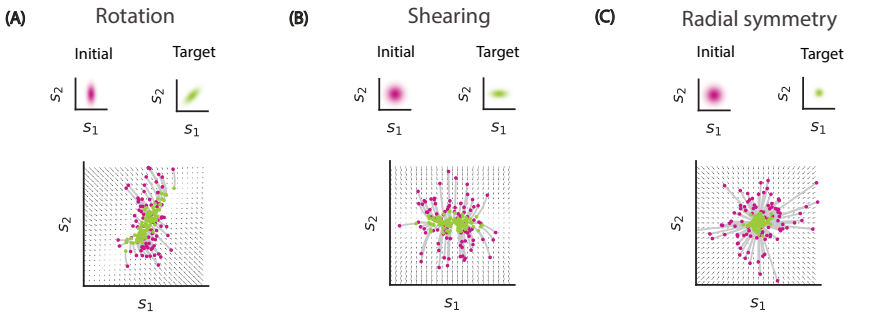

Figure 3.8: In (A) and (B) we simulate cases where the Jacobian of the particle update rules is not symmetric and therefore generates globally *twisted* transport maps. In (C) we simulate the case where the initial and target distributions are radial symmetric, and therefore the Jacobian of the update is symmetric and the transport map is optimal.

3.2.5 Preliminary results: distributional neural learning improves generalization

We simulate a deep version of DNL and compare its generalization performance with the previously proposed Deep Quantile Regression (DQR) algorithm. Both networks receive input gratings with varying orientations, and the joint distribution over orientations and rewards is shown in Figure

3.9B. A subset of orientations (indicated by the dashed vertical line in Figure 3.9B) was withheld during training. To assess each algorithm's ability to capture the full target distribution, we measured the KL divergence between the target and decoded distributions. The interaction term among particles in the DNL loss function introduces an inductive bias for smooth densities, enabling DNL to converge to connected solutions, which improves its generalization over DQR.

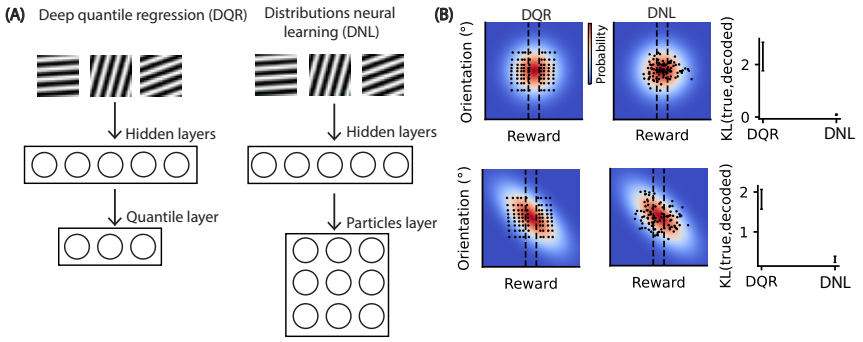

Figure 3.9: (A) We train a DQR to predict $N = 100$ reward quantiles and a DNL to predict the $N = 100$ particles over rewards and magnitudes. The DQR and DNL networks have two hidden layers with 512 and 256 units. The Adam optimizer was used, with an initial learning rate of 0,0001. (B) The dots correspond to a quantile for a given orientation in the DQR case and a particle in the DNL case. The target distributions are represented in the background. On the right, the 95% C.I. of KL divergence between the quantiles or particles and the target distribution is shown for 10 runs.

3.2.6 Preliminary results: multidimensional distributional learning rules model adaptation of place cell

We next applied our model to transformations of place cell tuning due to local environment deformations [73], which could not be previously modeled in an efficient coding framework since it requires learning a two-dimension random vector, where the x and y space positions are correlated.

Krupic et al [73] allowed rodents to experience a locally transformed environment (solid black line in Figure 3.10 left) after exploring the complete environment (dashed line in Figure 3.10 left). Place cells in the hippocampus adapted their tuning to the new environment. We modeled each place cell tuning in terms of each neuron's preferred spatial position $\{\theta_i\}_{i=1}^{N}$, where

N is the number of place cells. The space distribution p_θ for which this population is optimized was interpolated from $\{\theta_i\}_{i=1}^N$ conceptualized as approximate Dirac delta functions in probability space. Initially we gave samples of x and y positions that covered a rectangle until the particles converged. Then, only samples from the trapezoid represented with a full line in Figure 3.10 were given. We compare the mean displacement and the vector fields with the adaptation observed in the place cells (Figure 3.10).

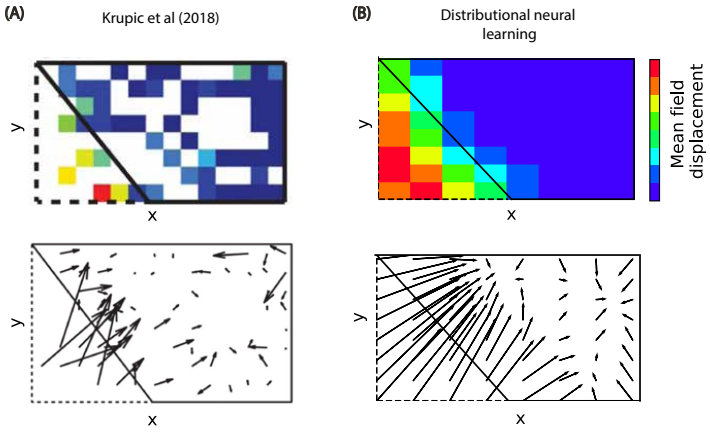

Figure 3.10: (A) Adapted from [73]. The complete environment is represented in a full line and the transformed environment in a dashed line. The mean field distortion (top) and vector field (bottom) of place cells are represented after the lower triangle was inaccessible. (B) Distributional neural learning modeling results. $N = 100$ particles were considered. The parameters for the update rules were set to: $c = 0.03$ and $\lambda = 0.07$.

3.3 Discussion

A general multidimensional efficient coding framework has been proposed previously [123]; however, no learning rules were derived. Moreover, the learning rules within the distributional RL framework do not generalize to stimuli of dimensionality greater than one, as discussed in Chapter 3.2.2. To bridge this gap, we propose multidimensional distributional learning rules. We define sufficient conditions for the neural coding trajectories to deviate from global optimal transport maps, in the non-asymptotic finite-sample case. An important future step will be to determine the learning rule parameters that bias learning trajectories towards the optimal transport. Our

preliminary results suggest that DNL can enhance generalization, relative to the factorized DQR. Understanding if DNL also improves transfer learning is an important future step.

As an empirical demonstration, we also model transformations in place cell tuning in response to local environmental changes. Extending this approach to model changes in grid cell tuning in the entorhinal cortex when the environment shape changes would be an interesting direction for future work [72] as is modeling how grid cell fields are attracted towards reward locations [12].

Referring back to the first chapter, the time-magnitide RL (TMRL) model we proposed does not converge to the efficient code when reward magnitude and time are correlated, in contrast to our more general DNL theory. Therefore, a specific signature of DNL in midbrain dopamine neurons could be identified in an experiment with a condition whereby a cue predicts a reward distribution with correlated reward timing and magnitude.

4

General discussion

In standard value-based RL, delay and magnitude information are compressed into a single scalar value, producing ambiguity between these two dimensions. Here we present time-magnitude RL (TMRL), a multidimensional variant of distributional reinforcement learning that learns the joint distribution of future rewards over time and magnitude using an efficient code that adapts to environmental statistics. We discovered signatures of TMRL-like computations in the activity of optogenetically identified DANs in mice during behavior. Specifically, we found significant diversity in both temporal discounting and tuning for the magnitude of rewards across DANs, features that allow the computation of a two-dimensional, probabilistic map of future rewards from a brief snapshot of neural activity recorded from a population of DANs in response to a reward-predictive cue. Furthermore, the decoding of future reward times correlated with the variability in the temporal evolution of behavior, and the decoding of reward magnitudes correlated with behavioral correlates of reward history, suggesting that decoded estimates correspond with animals' expectations about the timing and magnitude of rewards.

Additionally, by simulating behavior in a foraging environment, we highlight benefits of access to a joint probability distribution of reward over time and magnitude in the face of dynamic reward landscapes and internal physiological need states. These findings demonstrate surprisingly rich probabilistic reward information that is learned and communicated to DANs, and suggest a simple, local-in-time extension of TD learning algorithms that explains how such information may be acquired and computed.

However, TMRL does not converge to the efficient code when times and magnitudes of rewards are correlated. In order to overcome this problem, we propose a framework for learning general multidimensional distributions. We demonstrate its validity, and show preliminary results that suggest it leads to better generalization when compared with the factorized distributional reinforcement learning algorithm.

We will now outline a set of key observations and questions that I believe our work raises.

4.1 Time-magnitude RL (TMRL) is a model-free algorithm that learns a 'model' of the environment

In RL there is a trade-off between statistically efficient use of experience and computational tractability [26]. Model-based RL, while initially slower to learn, offers flexibility in adapting to changes in the rewards in the environment. In contrast, model-free RL achieves faster initial learning, but at the expense of adapting slower when the rewards in the environment change. Our proposed model, TMRL, combines the best of both worlds by enabling the decoding of future reward distributions across both time and magnitude. This allows TMRL to learn via a simple model-free mechanism while retaining the flexibility to adapt when rewards in the environment shift that one typically associates to model-based approaches.

Having access to a distribution over reward magnitudes at future times also allows for planning what actions to choose at each future time-step. In contrast to TMRL, the standard TD learning (considering time-steps as states), does not allow for planning, as the value of a future time-step is only accessible when the agent has reached it. Another advantage of TMRL, described in Chapter 2, is that it allows for zero-shot adaptation to dynamics in risk preferences and temporal discounting, or their interaction—a feature beneficial in dynamic environments or when internal states fluctuate.

4.2 How does temporal discounting arise within neural circuits?

When modeling behavior and neural data using reinforcement learning frameworks, we must make assumptions about how to represent states, actions, rewards, as well as how to set parameters such as learning rate and temporal discount factor. In this work, we modeled the diverse reward-timing tuning functions of DANs as exponentially decaying functions. However, our data is also compatible with the hypothesis that distinct circuits within the basal ganglia have access to representations of different time-scales (or state features), presumably inherited from the cortical hierarchical organization of time-scales [101]. As a consequence of receiving input representations over distinct time-ranges, the resulting value estimates and RPEs will reflect these multiple time-scales. Supporting this view, dopamine transients in the dorsolateral, dorsomedial or ventral striatal regions are known to signal RPEs based on distinct time-scales [95].

4.3 Why is there adaptation in dopamine neurons' time-scales if the system has at its disposal estimates of future rewards at different time-scales?

In our dataset, we observed that when possible reward times in the task are shortened (by removing the cue that predicts the longest interval), the temporal discount factors of DANs adapt to optimally encode the updated reward times (Chapter 2, Figure 2.5). Similar adaptation has been described in striatal neurons; e.g. MSNs rescale their responses to track the latency to next reward as it changes in a serial fixed interval task [92]. At first sight, such rescaling might seem redundant, given that the system already has access to reward estimates across different time-scales [95].

One reason for this redundancy may be that, depending on both the environment's state and the animal's internal state, representations with varying degrees of abstraction and temporal extension are needed to drive appropriate behaviors. In cortex, these dimensions appear to be correlated: more abstract regions, such as the prefrontal cortex, represent information over longer time-scales, while sensorimotor areas encode less abstract information over shorter time-scales [101]. As a result, these features cannot

be dissociated, which may explain why representations of values in the striatum, and of RPEs in midbrain DANs, span multiple time-scales.

On the other hand, to flexibly coordinate behavior across a wide range of time-scales—from milliseconds to days—it is beneficial to generate reward predictions over multiple time horizons. However, the basal ganglia have a limited number of circuits, so each loop must adapt to a specific range of reward time-scales. For instance, a child, with limited experience of extended time periods, may rely on shorter time ranges for reward predictions, whereas an adult can integrate expectations over longer spans. The adaptation of each circuit's time-scale enables an efficient representation of rewards. Furthermore, having access to multiple reward time-scales ensures rapid adaptation when behavior demands shift, allowing a new circuit to take over control of behavior.

4.4 How is the distributional map read out?

We have demonstrated that DANs contain information about the joint reward distribution in time and magnitude at the beginning of an episode (Chapter 2, Figure 2.5). An open question is how and where this information is read out by neural circuits in the brain. One possibility is that this information is distributed across different circuits, each characterized by a reward expectation with a specific level of optimism and temporal discounting, and downstream systems do not explicitly have access to the full distributional reward map.

Importantly, if the distribution were to be readout, knowledge of the tuning function of each DAN towards magnitude and time would be necessary. Notably, the dorsal-ventral axis of the striatum has been shown to encode a gradient of expectations of reward over distinct time-scales and reward optimisms [144, 95]. The projections from the striatum to the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) are topographically organized: ventral striatal regions project to more medial areas, while dorsal striatal regions project to more lateral areas of the VTA/SNc [11]. This suggests that DANs may also be topographically organized according to their tuning to reward magnitudes and times. Such an organization could provide a mechanism for downstream systems to infer reward tuning properties based on spatial arrangement, and successfully decode the distributional map of rewards.

4.5 Are animals using the distributions over reward magnitude and time to guide behavior?

Having access to information is not the same as using it to guide behavior. In our work, we demonstrated that DANs have access to distributions over both reward magnitudes and timings at the beginning of an episode, when it might be useful to drive behavior. We found that the mean decoded reward time correlated with the timing of anticipatory licking, while the mean decoded reward magnitude correlates with pupil diameter. However, these behavioral readouts do not confirm that the full reward distribution is being utilized. Investigating whether animals actually use distributional reward information is a gap in the studies on distributional reinforcement learning [23, 100].

Confirming that animals use distributional information is challenging. For example, risk sensitivity may arise from estimating both the mean and variance of rewards, making it difficult to isolate the specific role of the full distributional information. While *in silico* simulations suggest that access to reward magnitude distributions can improve generalization and allow for transfer learning [23], designing tasks to test these advantages and directly link them to DANs seems unfeasible.

4.6 First phase of dopamine neurons' responses reflects prior distribution of rewards in the environment?

Historically, the response dynamics of DANs has been described as having two distinct phases. The initial phase, often described as a salience or sensory signal, is unaffected by reward features. In contrast, the second phase is clearly modulated by reward features, such as magnitude and time [128, 32]. However, our findings suggest that the initial phase of DAN responses may encode a prior distribution over expected rewards in the task.

In our task, we use solenoid valves to deliver the various odors. Solenoid valves, when switched on or off, produce a characteristic sound. We hypothesized that neural activity elicited by this sound, common across all cues, would arrive at dopamine neurons prior to any information about cue identity. The reason for this is twofold: firstly, the solenoid opening precedes odor delivery. Secondly, due to differences in processing time required to extract simple auditory information as compared to complex olfactory

information. Indeed, our analysis revealed that the first-phase response of dopamine neurons was much more similar across cues than the second-phase responses (see Figure 4.1B). Similarly, the decoded distribution of future reward times (see Figure 4.1D) was again analogous across cues when using the first phase responses but not when using the second phase (cf. with Figure 2.4D). Moreover, the decoded distribution during the first phase was broader than the distributions derived from second-phase responses (cf. Figure 2.3D). Finally, we decoded the joint distribution over both reward times and magnitudes (Figure 4.1C) and found it to be closer to the prior distribution of rewards than would be expected by chance (Figure 4.1E). Collectively, these results suggest that DANs may have access to the full distribution of possible reward magnitudes and times in the environment early on, followed by a cue specific reward time-magnitude distribution.

This preliminary finding can be reconciled with previous studies on the response dynamics of dopamine neurons [69, 70, 127], where odors were not preceded by auditory cues. One possible explanation is that low-level features of the stimuli (e.g., contrast in the case of visual stimuli) may be accessible to midbrain dopamine neurons earlier in time [141, 103].

4.7 Future directions

The behavioral tasks used to probe DANs' distributional reward representations are relatively simple [23]. In contrast, risk preference tasks reveal how animals weigh reward uncertainty, while intertemporal choice tasks provide behavioral readouts on how animals discount delays to rewards. DANs have been recorded during both tasks independently [69, 127]. Combining these two tasks while recording DAN activity could offer valuable insights into whether the time-magnitude distributional map decoded from DANs can predict choices across varying reward delays and magnitudes. Investigating whether biases in choice behavior align with biases in DANs' distributional reward maps is an important future research direction, that helps connecting distributional reward representations with behavioral policies.

With acute recordings, we captured DAN responses only after the animal had been extensively trained. We revealed that after training, DANs have access to a distribution of rewards over magnitudes and times. An exciting future direction would be to investigate how the reward distributional map continuously emerges during training time.

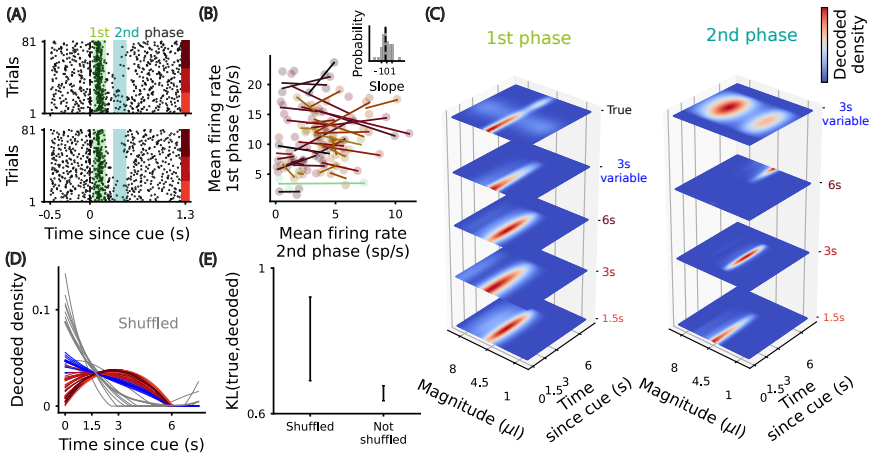

Figure 4.1: First phase of midbrain dopamine neurons encodes a distribution that is similar across cues and closer to prior distribution of the rewards in the task than what would be expected by chance. (A) Raster aligned to odor valve opening for two example neurons. The green shaded area depicts the window used to compute the first phase responses (50-200ms) and the blue to the second phase (300-450ms). (B) Second-phase responses vary more with the upcoming delay than first-phase responses. First phase mean responses across the 1.5s, 3s and 6s cues for different neurons as a function of the second phase mean responses. The slopes are the fitted linear regression models for each neuron. Points and slopes are color coded by the estimated temporal discount factor. Inset: distribution of slopes across neurons. The variation in first phase responses is smaller than in second phase responses over neurons, since the mean absolute slope (vertical) is smaller than one ($p=0.001$, 95% C.I.=(0.0099,0.71), bootstrapping 10,000 times). (C) Decoded joint density of reward over magnitude and time, using the first (left) and second (right) phase population responses aligned to the different cues. (D) Decoded density over reward time using the first phase dopamine population responses for all cues. The gray lines depict the decoded density when the population temporal discount factors are shuffled. The light lines represent decoded densities using the responses of 70% of randomly selected trials and the thicker lines represent the mean decoded densities. (E) 90% C.I. of the mean Kullback-Leibler (KL) divergence between the true prior distribution over reward times and magnitudes in the task and the decoded from the dopamine neurons when the population tuning with respect to reward time and magnitude is or not shuffled considering 100 runs of the decoder.

In our task, we have observed that DANs adapt to reward time statistics. When we remove the cue predicting the longest delay, temporal discounts become steeper, compared to when removing the shortest delay. Evaluating whether temporal discounting behaviorally reflects the reward timing structure in the environment is a possible future direction. This finding could offer valuable insights for therapeutic strategies aimed at improving impulse control. This could involve lengthening the timescale over which predictions remain valid or reorient individuals to the longer timescale structure of rewards that may already exist within an environment.

In the third chapter of this thesis, we proposed learning rules for general probability distributions across multiple features, that we named distributional neural learning (DNL). The TMRL model does not converge to the efficient code when reward magnitude and time are correlated, in contrast to our more general DNL theory. Therefore, a specific signature of DNL in midbrain DANs could be identified in an experiment with a condition whereby a cue predicts a reward distribution with correlated reward timing and magnitude.

Furthermore, the DNL framework introduced offers a promising foundation for developing a decision-making theory based on the distribution of reward magnitudes and times. An interesting path is to derive policies not resorting to dynamic programming and Bellman updates, inspired by the Todorov's Linear Markov Decision Processes framework [143]. On the other hand, including preferences towards risks in magnitude and temporal discounting is another possible future direction. For example, Yaari's theory of decision-making is based on distorting the cumulative distribution of reward magnitudes, that can be seen as weights applied to the quantiles of the reward distribution [150]. Extending distortion risk measures to encompass both the time and magnitude dimensions would be an interesting direction for further research [38]. Additionally, generalizing policies based on the conditional value at risk or stochastic dominance, to the time-magnitude domain, is another exciting direction [37, 140, 88].

Bibliography

[1] G. E. Alexander and M. D. Crutcher. "Functional architecture of basal ganglia circuits: neural substrates of parallel processing". In: *Trends in neurosciences* 13.7 (1990), pp. 266–271 (cit. on p. 5).

[2] "Appendix 1: IBL protocol for headbar implant surgery in mice (2020)". In: () (cit. on p. 37).

[3] R. Avvisati et al. "Distributional coding of associative learning within projection-defined populations of midbrain dopamine neurons". In: *bioRxiv* (2022), pp. 2022–07 (cit. on p. 12).

[4] B. W. Balleine and J. P. O'doherty. "Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action". In: *Neuropsychopharmacology* 35.1 (2010), pp. 48–69 (cit. on pp. 5, 31).

[5] P. D. Balsam and C. R. Gallistel. "Temporal maps and informativeness in associative learning". In: *Trends in neurosciences* 32.2 (2009), pp. 73–78 (cit. on pp. 1, 35).

[6] H. B. Barlow et al. "Possible principles underlying the transformation of sensory messages". In: *Sensory communication* 1.01 (1961), pp. 217–233 (cit. on pp. 15, 55).

[7] T. E. Behrens et al. "Learning the value of information in an uncertain world". In: *Nature neuroscience* 10.9 (2007), pp. 1214–1221 (cit. on p. 34).

[8] A. J. Bell and T. J. Sejnowski. "An information-maximization approach to blind separation and blind deconvolution". In: *Neural computation* 7.6 (1995), pp. 1129–1159 (cit. on pp. 39, 55).

[9] M. G. Bellemare, W. Dabney, and R. Munos. "A distributional perspective on reinforcement learning". In: *International conference on machine learning*. PMLR. 2017, pp. 449–458 (cit. on p. 12).

[10] M. G. Bellemare, W. Dabney, and M. Rowland. *Distributional reinforcement learning*. MIT Press, 2023 (cit. on pp. 7, 12).

[11] A. Björklund and S. B. Dunnett. "Dopamine neuron systems in the brain: an update". In: *Trends in neurosciences* 30.5 (2007), pp. 194–202 (cit. on p. 70).

[12] C. N. Boccara et al. "The entorhinal cognitive map is attracted to goals". In: *Science* 363.6434 (2019), pp. 1443–1447 (cit. on p. 65).

[13] A. M. Bornstein and N. D. Daw. "Multiplicity of control in the basal ganglia: computational roles of striatal subregions". In: *Current opinion in neurobiology* 21.3 (2011), pp. 374–380 (cit. on pp. 5, 35).

[14] I. M. Bright et al. "A temporal record of the past with a spectrum of time constants in the monkey entorhinal cortex". In: *Proceedings of the National Academy of Sciences* 117.33 (2020), pp. 20274–20283 (cit. on p. 8).

[15] N. Brunel and J.-P. Nadal. "Mutual information, Fisher information, and population coding". In: *Neural computation* 10.7 (1998), pp. 1731–1757 (cit. on pp. 25, 42).

[16] T. Cash-Padgett et al. "Opposing pupil responses to offered and anticipated reward values". In: *Animal cognition* 21 (2018), pp. 671–684 (cit. on p. 22).

[17] N. Chakravarti. "Isotonic median regression: a linear programming approach". In: *Mathematics of operations research* 14.2 (1989), pp. 303–308 (cit. on p. 50).

[18] C. Y. Chang et al. "Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors". In: *Nature neuroscience* 19.1 (2016), pp. 111–116 (cit. on pp. 5, 6).

[19] C. Chen et al. "A unified particle-optimization framework for scalable Bayesian sampling". In: *arXiv preprint arXiv:1805.11659* (2018) (cit. on p. 60).

[20] L. T. Coddington and J. T. Dudman. "The timing of action determines reward prediction signals in identified midbrain dopamine neurons". In: *Nature neuroscience* 21.11 (2018), pp. 1563–1573 (cit. on pp. 6, 35).

[21] K. J. W. Craik. *The nature of explanation*. Vol. 445. CUP Archive, 1967 (cit. on p. 31).

[22] B. F. Cruz et al. "Action suppression reveals opponent parallel control via striatal circuits". In: *Nature* 607 (2022), pp. 521–526. URL: https://api.semanticscholar.org/CorpusID:250337715 (cit. on p. 35).

[23] W. Dabney et al. "A distributional code for value in dopamine-based reinforcement learning". In: *Nature* 577.7792 (2020), pp. 671–675 (cit. on pp. 7, 12, 14, 20, 23, 44, 47, 55, 71, 72).

[24] W. Dabney et al. "Distributional reinforcement learning with quantile regression". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018 (cit. on pp. 7, 17, 41, 45, 57).

[25] M. Davis, L. S. Schlesinger, and C. Sorenson. "Temporal specificity of fear conditioning: Effects of different conditioned stimulus–unconditioned stimulus intervals on the fear-potentiated startle effect." In: *Journal of Experimental Psychology: Animal Behavior Processes* 15.4 (1989), p. 295 (cit. on p. 1).

[26] N. D. Daw, Y. Niv, and P. Dayan. "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nature neuroscience* 8.12 (2005), pp. 1704–1711 (cit. on pp. 3, 68).

[27] P. Dayan. "Improving generalization for temporal difference learning: The successor representation". In: *Neural computation* 5.4 (1993), pp. 613–624 (cit. on pp. 3, 29, 31).

[28] P. Dayan and T. Long. "Statistical models of conditioning". In: *Advances in neural information processing systems* 10 (1997) (cit. on p. 11).

[29] K. Doya. "Complementary roles of basal ganglia and cerebellum in learning and motor control". In: *Current opinion in neurobiology* 10.6 (2000), pp. 732–739 (cit. on p. 4).

[30] N. Drummond and Y. Niv. "Model-based decision making and model-free learning". In: *Current Biology* 30.15 (2020), R860–R865 (cit. on p. 3).

[31] B. Engelhard et al. "Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons". In: *Nature* 570.7762 (2019), pp. 509–513 (cit. on p. 35).

[32] N. Eshel et al. "Dopamine neurons share common response function for reward prediction error". In: *Nature neuroscience* 19.3 (2016), pp. 479–486 (cit. on p. 71).

[33] A. L. Fairhall et al. "Efficiency and ambiguity in an adaptive neural code". In: *Nature* 412.6849 (2001), pp. 787–792 (cit. on p. 15).

[34] W. Fedus et al. "Hyperbolic discounting and learning over multiple horizons". In: *arXiv preprint arXiv:1902.06865* (2019) (cit. on pp. 9, 14).

[35] C. D. Fiorillo, P. N. Tobler, and W. Schultz. "Discrete coding of reward probability and uncertainty by dopamine neurons". In: *Science* 299.5614 (2003), pp. 1898–1902 (cit. on p. 6).

[36] N. N. Foster et al. "The mouse cortico–basal ganglia–thalamic network". In: *Nature* 598.7879 (2021), pp. 188–194 (cit. on p. 5).

[37] C. Gagne and P. Dayan. "Peril, prudence and planning as risk, avoidance and worry". In: *Journal of Mathematical Psychology* 106 (2022), p. 102617 (cit. on p. 74).

[38] A. Galichon and M. Henry. "Dual theory of choice with multivariate risks". In: *Journal of Economic Theory* 147.4 (2012), pp. 1501–1516 (cit. on p. 74).

[39] C. R. Gallistel and J. Gibbon. "Time, rate, and conditioning." In: *Psychological review* 107.2 (2000), p. 289 (cit. on pp. 1, 2, 11).

[40] C. Gallistel et al. "The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect." In: *Journal of experimental psychology: Animal behavior processes* 27.4 (2001), p. 354 (cit. on p. 1).

[41] D. Ganguli and E. P. Simoncelli. "Efficient sensory encoding and Bayesian inference with heterogeneous neural populations". In: *Neural computation* 26.10 (2014), pp. 2103–2134 (cit. on pp. 9, 23, 41–43, 55–58).

[42] D. Ganguli and E. P. Simoncelli. "Neural and perceptual signatures of efficient sensory coding". In: *arXiv preprint arXiv:1603.00058* (2016) (cit. on p. 9).

[43] S. J. Gershman et al. "Explaining dopamine through prediction errors and beyond". In: *Nature Neuroscience* (2024), pp. 1–11 (cit. on p. 6).

[44] S. Ghosh and A. M. Zador. "Corticostriatal plasticity established by initial learning persists after behavioral reversal". In: *eneuro* 8.2 (2021) (cit. on p. 5).

[45] J. Gibbon. "Scalar expectancy theory and Weber's law in animal timing." In: *Psychological review* 84.3 (1977), p. 279 (cit. on p. 8).

[46] P. W. Glimcher and E. Fehr. *Neuroeconomics: Decision making and the brain*. Academic Press, 2013 (cit. on p. 44).

[47] L. S. Gonzalez et al. "Ventral striatum dopamine release encodes unique properties of visual stimuli in mice". In: *Elife* 12 (2023), e85064 (cit. on p. 6).

[48] L. Green, J. Myerson, and P. Ostaszewski. "Discounting of delayed rewards across the life span: age differences in individual discounting functions". In: *Behavioural processes* 46.1 (1999), pp. 89–96 (cit. on p. 10).

[49] F. Greenstreet et al. "Action prediction error: a value-free dopaminergic teaching signal that drives stable learning". In: *BiorXiv* (2022), pp. 2022–09 (cit. on p. 6).

[50] A. A. Hamid, M. J. Frank, and C. I. Moore. "Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment". In: *Cell* 184.10 (2021), pp. 2733–2749 (cit. on p. 35).

[51] D. N. Hill, S. B. Mehta, and D. Kleinfeld. "Quality metrics to accompany spike sorting of extracellular signals". In: *Journal of Neuroscience* 31.24 (2011), pp. 8699–8705 (cit. on p. 40).

[52] P. C. Holland. "Relations between Pavlovian-instrumental transfer and reinforcer devaluation." In: *Journal of Experimental Psychology: Animal Behavior Processes* 30.2 (2004), p. 104 (cit. on p. 3).

[53] J. C. Horvitz. "Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events". In: *Neuroscience* 96.4 (2000), pp. 651–656 (cit. on p. 6).

[54] J. C. Houk, J. L. Adams, and A. G. Barto. "A model of how the basal ganglia generate and use neural signals that predict reinforcement". In: (1994) (cit. on p. 4).

[55] B. J. Hunnicutt et al. "A comprehensive excitatory input map of the striatum reveals novel functional organization". In: *elife* 5 (2016), e19103 (cit. on p. 5).

[56] A. Hyvarinen. "Fast and robust fixed-point algorithms for independent component analysis". In: *IEEE transactions on Neural Networks* 10.3 (1999), pp. 626–634 (cit. on p. 39).

[57] Y. Iino et al. "Dopamine D2 receptors in discrimination learning and spine enlargement". In: *Nature* 579.7800 (2020), pp. 555–560 (cit. on p. 5).

[58] M. Janner, I. Mordatch, and S. Levine. "Generative temporal difference learning for infinite-horizon prediction". In: *arXiv preprint arXiv:2010.14496* (2020) (cit. on p. 14).

[59] H. Jeong et al. "Mesolimbic dopamine release conveys causal associations". In: *Science* 378.6626 (2022), eabq6740 (cit. on pp. 6, 35).

[60] M. C. Jones. "Expectiles and M-quantiles are quantiles". In: *Statistics & Probability Letters* 20.2 (1994), pp. 149–153 (cit. on p. 42).

[61] R. Jordan, D. Kinderlehrer, and F. Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM journal on mathematical analysis* 29.1 (1998), pp. 1–17 (cit. on p. 60).

[62] A. Kacelnik and M. Bateson. "Risky theories—the effects of variance on foraging decisions". In: *American zoologist* 36.4 (1996), pp. 402–434 (cit. on pp. 30, 52).

[63] J. H. Kagel, L. Green, and T. Caraco. "When foragers discount the future: Constraint or adaptation?" In: *Animal Behaviour* 34 (1986), pp. 271–283 (cit. on p. 30).

[64] D. Kahneman and A. Tversky. "Prospect theory: An analysis of decision under risk". In: *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013, pp. 99–127 (cit. on p. 29).

[65] L. V. Kantorovich. "Mathematical methods of organizing and planning production". In: *Management science* 6.4 (1960), pp. 366–422 (cit. on p. 57).

[66] S. Killcross and E. Coutureau. "Coordination of actions and habits in the medial prefrontal cortex of rats". In: *Cerebral cortex* 13.4 (2003), pp. 400–408 (cit. on p. 3).

[67] Y.-H. Kim and E. Milman. "A generalization of Caffarelli's contraction theorem via (reverse) heat flow". In: *Mathematische Annalen* 354.3 (2012), pp. 827–862 (cit. on p. 61).

[68] A. E. Kincaid, T. Zheng, and C. J. Wilson. "Connectivity and convergence of single corticostriatal axons". In: *Journal of Neuroscience* 18.12 (1998), pp. 4722–4731 (cit. on p. 5).

[69] S. Kobayashi and W. Schultz. "Influence of reward delays on responses of dopamine neurons". In: *Journal of neuroscience* 28.31 (2008), pp. 7837–7846 (cit. on pp. 1, 6, 46, 72).

[70] S. Kobayashi and W. Schultz. "Reward contexts extend dopamine signals to unrewarded stimuli". In: *Current Biology* 24.1 (2014), pp. 56–62 (cit. on p. 72).

[71] J. Kober, J. A. Bagnell, and J. Peters. "Reinforcement learning in robotics: A survey". In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1238–1274 (cit. on p. 2).

[72] J. Krupic et al. "Grid cell symmetry is shaped by environmental geometry." In: *Nature* 518.7538 (2015), pp. 232–235 (cit. on p. 65).

[73] J. Krupic et al. "Local transformations of the hippocampal cognitive map." In: *Science* 359.6380 (2018), pp. 1143–1146 (cit. on pp. 56, 63, 64).

[74] Z. Kurth-Nelson and A. D. Redish. "Temporal-difference reinforcement learning with distributed representations". In: *PLoS One* 4.10 (2009), e7362 (cit. on pp. 8, 14).

[75] D. Kvitsiani et al. "Distinct behavioural and network correlates of two interneuron types in prefrontal cortex". In: *Nature* 498.7454 (2013), pp. 363–366 (cit. on pp. 18, 40).

[76] J. LaBarbera and R. M. Church. "Magnitude of fear as a function of expected time to all aversive event". In: *Animal Learning & Behavior* 2.3 (1974), pp. 199–202 (cit. on p. 1).

[77] A. Lak, W. R. Stauffer, and W. Schultz. "Dopamine prediction error responses integrate subjective value from different reward dimensions". In: *Proceedings of the National Academy of Sciences* 111.6 (2014), pp. 2343–2348 (cit. on p. 6).

[78] B. Lau and P. W. Glimcher. "Dynamic response-by-response models of matching behavior in rhesus monkeys". In: *Journal of the experimental analysis of behavior* 84.3 (2005), pp. 555–579 (cit. on p. 22).

[79] B. Lau, T. Monteiro, and J. J. Paton. "The many worlds hypothesis of dopamine prediction error: implications of a parallel circuit architecture in the basal ganglia". In: *Current Opinion in Neurobiology* 46 (2017), pp. 241–247 (cit. on p. 6).

[80] S. Laughlin. "A simple coding procedure enhances a neuron's information capacity". In: *Zeitschrift für Naturforschung c* 36.9-10 (1981), pp. 910–912 (cit. on pp. 9, 15).

[81] R. S. Lee et al. "A feature-specific prediction error model explains dopaminergic heterogeneity". In: *Nature neuroscience* (2024), pp. 1–13 (cit. on pp. 6, 35).

[82] T. Ljungberg, P. Apicella, and W. Schultz. "Responses of monkey dopamine neurons during learning of behavioral reactions". In: *Journal of neurophysiology* 67.1 (1992), pp. 145–163 (cit. on p. 6).

[83] G. Lopes et al. "Bonsai: an event-based framework for processing and controlling data streams". In: *Frontiers in neuroinformatics* 9 (2015), p. 7 (cit. on pp. 37, 39).

[84] K. Louie. "Asymmetric and adaptive reward coding via normalized reinforcement learning". In: *PLoS computational biology* 18.7 (2022), e1010350 (cit. on pp. 7, 23).

[85] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: `https://github.com/joaomlourenco/novathesis/raw/main/template.pdf` (cit. on p. i).

[86] E. A. Ludvig, R. S. Sutton, and E. J. Kehoe. "Stimulus representation and the timing of reward-prediction errors in models of the dopamine system". In: *Neural computation* 20.12 (2008), pp. 3034–3054 (cit. on p. 40).

[87] C. Lyle, M. G. Bellemare, and P. S. Castro. "A comparative analysis of expected and distributional reinforcement learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 4504–4511 (cit. on pp. 7, 12).

[88] J. Martin et al. "Stochastically dominant distributional reinforcement learning". In: *International conference on machine learning*. PMLR. 2020, pp. 6745–6754 (cit. on pp. 7, 12, 74).

[89] A. Mathis, P. Mamidanna, and K. M. Cury. "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning". In: *Nature Neuroscience* 21 (2018), pp. 1281–1289 (cit. on p. 48).

[90] S. Matias et al. "Activity patterns of serotonin neurons underlying cognitive flexibility". In: *Elife* 6 (2017), e20552 (cit. on p. 1).

[91] D. McNamee and D. M. Wolpert. "Internal models in biological control". In: *Annual review of control, robotics, and autonomous systems* 2.1 (2019), pp. 339–364 (cit. on p. 31).

[92] G. B. Mello, S. Soares, and J. J. Paton. "A scalable population code for time in the striatum". In: *Current Biology* 25.9 (2015), pp. 1113–1122 (cit. on p. 69).

[93] W. Menegas et al. "Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli". In: *Nature neuroscience* 21.10 (2018), pp. 1421–1430 (cit. on p. 6).

[94] V. Mnih et al. "Human-level control through deep reinforcement learning". In: *nature* 518.7540 (2015), pp. 529–533 (cit. on pp. 3, 7).

[95] A. Mohebi et al. "Dopamine transients follow a striatal gradient of reward time horizons". In: *Nature Neuroscience* 27.4 (2024), pp. 737–746 (cit. on pp. 69, 70).

[96] I. Momennejad and M. W. Howard. "Predicting the future with multi-scale successor representations". In: *BioRxiv* (2018), p. 449470 (cit. on p. 29).

[97] I. Momennejad et al. "The successor representation in human reinforcement learning". In: *Nature human behaviour* 1.9 (2017), pp. 680–692 (cit. on p. 4).

[98] P. R. Montague, P. Dayan, and T. J. Sejnowski. "A framework for mesencephalic dopamine systems based on predictive Hebbian learning". In: *Journal of neuroscience* 16.5 (1996), pp. 1936–1947 (cit. on pp. 5, 12, 34).

[99] A. Motiwala et al. "Efficient coding of cognitive variables underlies dopamine response and choice behavior". In: *Nature Neuroscience* 25.6 (2022), pp. 738–748 (cit. on p. 9).

[100] T. H. Muller et al. "Distributional reinforcement learning in prefrontal cortex". In: *Nature Neuroscience* 27.3 (2024), pp. 403–408 (cit. on pp. 7, 12, 71).

[101] J. D. Murray et al. "A hierarchy of intrinsic timescales across primate cortex". In: *Nature neuroscience* 17.12 (2014), pp. 1661–1663 (cit. on pp. 5, 69).

[102] W. K. Newey and J. L. Powell. "Asymmetric least squares estimation and testing". In: *Econometrica: Journal of the Econometric Society* (1987), pp. 819–847 (cit. on p. 42).

[103] S. K. Ogawa et al. "Organization of monosynaptic inputs to the serotonin and dopamine neuromodulatory systems". In: *Cell reports* 8.4 (2014), pp. 1105–1118 (cit. on p. 72).

[104] T. Ohyama and M. D. Mauk. "Latent acquisition of timed responses in cerebellar cortex". In: *Journal of Neuroscience* 21.2 (2001), pp. 682–690 (cit. on p. 1).

[105] B. A. Olshausen and D. J. Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381.6583 (1996), pp. 607–609 (cit. on p. 9).

[106] G. Percheron et al. "The primate motor thalamus analysed with reference to subcortical afferent territories". In: *Stereotactic and functional neurosurgery* 60.1-3 (1993), pp. 32–41 (cit. on p. 5).

[107] V. Picheny et al. "Bayesian quantile and expectile optimisation". In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 1623–1633 (cit. on p. 50).

[108] R. Polanía, M. Woodford, and C. C. Ruff. "Efficient coding of subjective value". In: *Nature neuroscience* 22.1 (2019), pp. 134–142 (cit. on p. 9).

[109] L. Qian et al. "The role of prospective contingency in the control of behavior and dopamine signals during associative learning". In: *bioRxiv* () (cit. on p. 6).

[110] M. R. Roesch, D. J. Calu, and G. Schoenbaum. "Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards". In: *Nature neuroscience* 10.12 (2007), pp. 1615–1624 (cit. on p. 6).

[111] M. Rowland et al. "Statistics and samples in distributional reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5528–5536 (cit. on pp. 41, 42, 45, 50).

[112] P. H. Rudebeck et al. "A role for primate subgenual cingulate cortex in sustaining autonomic arousal". In: *Proceedings of the National Academy of Sciences* 111.14 (2014), pp. 5391–5396 (cit. on p. 22).

[113] E. M. Russek et al. "Predictive representations can link model-based reinforcement learning to model-free mechanisms". In: *PLoS computational biology* 13.9 (2017), e1005768 (cit. on p. 31).

[114] W. Schultz, P. Dayan, and P. R. Montague. "A neural substrate of prediction and reward". In: *Science* 275.5306 (1997), pp. 1593–1599 (cit. on pp. v, vii, 5, 6, 34).

[115] H. H. Schütt, D. Kim, and W. J. Ma. "Reward prediction error neurons implement an efficient code for reward". In: *Nature Neuroscience* (2024), pp. 1–7 (cit. on pp. 10, 23).

[116] N. Schweighofer et al. "Low-serotonin levels increase delayed reward discounting in humans". In: *Journal of Neuroscience* 28.17 (2008), pp. 4528–4532 (cit. on p. 8).

[117] K. H. Shankar and M. W. Howard. "A scale-invariant internal representation of time". In: *Neural Computation* 24.1 (2012), pp. 134–193 (cit. on pp. 12, 14).

[118] M. J. Sharpe et al. "Dopamine transients are sufficient and necessary for acquisition of model-based associations". In: *Nature neuroscience* 20.5 (2017), pp. 735–742 (cit. on p. 35).

[119] M. J. Sharpe et al. "Dopamine transients do not act as model-free prediction errors during associative learning". In: *Nature communications* 11.1 (2020), p. 106 (cit. on p. 35).

[120] M. W. Shiflett, R. A. Brown, and B. W. Balleine. "Acquisition and performance of goal-directed instrumental actions depends on ERK signaling in distinct regions of dorsal striatum in rats". In: *Journal of Neuroscience* 30.8 (2010), pp. 2951–2959 (cit. on p. 5).

[121] J. H. Siegle et al. "Open Ephys: an open-source, plugin-based platform for multichannel electrophysiology". In: *Journal of neural engineering* 14.4 (2017), p. 045003 (cit. on p. 39).

[122] E. P. Simoncelli and B. A. Olshausen. "Natural image statistics and neural representation". In: *Annual review of neuroscience* 24.1 (2001), pp. 1193–1216 (cit. on p. 15).

[123] E. C. Smith and M. S. Lewicki. "Efficient auditory coding". In: *Nature* 439.7079 (2006), pp. 978–982 (cit. on pp. 9, 10, 64).

[124] S. Soares, B. V. Atallah, and J. J. Paton. "Midbrain dopamine neurons control judgment of time". In: *Science* 354.6317 (2016), pp. 1273–1277 (cit. on p. 34).

[125] A. Soltani and A. Izquierdo. "Adaptive learning under expected and unexpected uncertainty". In: *Nature Reviews Neuroscience* 20.10 (2019), pp. 635–644 (cit. on p. 34).

[126] P. D. Sozou. "On hyperbolic discounting and uncertain hazard rates". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265.1409 (1998), pp. 2015–2020 (cit. on pp. 8, 9).

[127] W. R. Stauffer, A. Lak, and W. Schultz. "Dopamine reward prediction error responses reflect marginal utility". In: *Current biology* 24.21 (2014), pp. 2491–2500 (cit. on pp. 7, 30, 72).

[128] W. R. Stauffer et al. "Components and characteristics of the dopamine reward utility signal". In: *Journal of Comparative Neurology* 524.8 (2016), pp. 1699–1711 (cit. on p. 71).

[129] W. R. Stauffer et al. "Components and characteristics of the dopamine reward utility signal". In: *Journal of Comparative Neurology* 524.8 (2016), pp. 1699–1711. DOI: https://doi.org/10.1002/cne.23880. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.238

80. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.23880 (cit. on p. 46).

[130] E. E. Steinberg et al. "A causal link between prediction errors, dopamine neurons and learning". In: *Nature neuroscience* 16.7 (2013), pp. 966–973 (cit. on pp. 5, 6).

[131] N. A. Steinmetz et al. "Challenges and opportunities for large-scale electrophysiology with Neuropixels probes". In: *Current opinion in neurobiology* 50 (2018), pp. 92–100 (cit. on p. 55).

[132] J. R. Stevens and D. W. Stephens. "Patience". In: *Current Biology* 18.1 (2008), R11–R12 (cit. on p. 8).

[133] M. Stevens et al. "Phenotype–environment matching in sand fleas". In: *Biology Letters* 11.8 (2015), p. 20150494 (cit. on p. 10).

[134] L. P. Sugrue, G. S. Corrado, and W. T. Newsome. "Matching behavior and the representation of value in the parietal cortex". In: *science* 304.5678 (2004), pp. 1782–1787 (cit. on p. 2).

[135] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on pp. v, vii, 2, 11, 30, 31).

[136] Y. K. Takahashi et al. "Dopamine neurons respond to errors in the prediction of sensory features of expected rewards". In: *Neuron* 95.6 (2017), pp. 1395–1405 (cit. on p. 6).

[137] Y. K. Takahashi et al. "Dopaminergic prediction errors in the ventral tegmental area reflect a multithreaded predictive model". In: *Nature neuroscience* 26.5 (2023), pp. 830–839 (cit. on p. 35).

[138] P. Tano, P. Dayan, and A. Pouget. "A local temporal difference code for distributional reinforcement learning". In: *Advances in neural information processing systems* 33 (2020), pp. 13662–13673 (cit. on pp. 8, 14, 49).

[139] S. Thakoor et al. "Generalised policy improvement with geometric policy composition". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 21272–21307 (cit. on p. 14).

[140] T. Théate and D. Ernst. "Risk-sensitive policy with distributional reinforcement learning". In: *Algorithms* 16.7 (2023), p. 325 (cit. on pp. 7, 12, 74).

[141] J. Tian et al. "Distributed and mixed information in monosynaptic inputs to dopamine neurons". In: *Neuron* 91.6 (2016), pp. 1374–1389 (cit. on p. 72).

[142] Z. Tiganj et al. "Estimating scale-invariant future in continuous time". In: *Neural Computation* 31.4 (2019), pp. 681–709 (cit. on pp. 8, 9, 14).

[143] E. Todorov. "Linearly-solvable Markov decision problems". In: *Advances in neural information processing systems* 19 (2006) (cit. on p. 74).

[144] I. Tsutsui-Kimura et al. "Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task". In: *Elife* 9 (2020), e62390 (cit. on p. 70).

[145] N. Uchida and Z. F. Mainen. "Speed and accuracy of olfactory discrimination in the rat". In: *Nature neuroscience* 6.11 (2003), pp. 1224–1229 (cit. on p. 37).

[146] A. Vanderveldt, L. Oliveira, and L. Green. "Delay discounting: Pigeon, rat, human—does it matter?" In: *Journal of Experimental Psychology: Animal learning and cognition* 42.2 (2016), p. 141 (cit. on p. 9).

[147] E. Weber. "De pulsu, resorptione, auditu et tactu: annotationes anatomicae et physiologicae, auctore. prostat apud CF Koehler". In: *ProQuest Number: INFORMATION TO ALL USERS* (1834) (cit. on p. 9).

[148] N. E. White et al. "Coefficients of variation in timing of the classically conditioned eyeblink in rabbits". In: *Psychobiology* 28.4 (2000), pp. 520–524 (cit. on p. 1).

[149] H. Wiltzer et al. "Foundations of multivariate distributional reinforcement learning". In: *arXiv preprint arXiv:2409.00328* (2024) (cit. on p. 10).

[150] M. E. Yaari. "The dual theory of choice under risk". In: *Econometrica: Journal of the Econometric Society* (1987), pp. 95–115 (cit. on p. 74).

[151] A. E. Yagle. "Regularized matrix computations". In: *matrix* 500 (2005), p. 10 (cit. on p. 49).

[152] H. Yamada et al. "Thirst-dependent risk preferences in monkeys identify a primitive form of wealth". In: *Proceedings of the National Academy of Sciences* 110.39 (2013), pp. 15788–15793 (cit. on pp. 7, 30).

[153] T. E. Yerxa et al. "Efficient sensory coding of multidimensional stimuli". In: *PLoS computational biology* 16.9 (2020), e1008146 (cit. on p. 55).

[154] H. H. Yin, B. J. Knowlton, and B. W. Balleine. "Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning". In: *Behavioural brain research* 166.2 (2006), pp. 189–196 (cit. on p. 5).

[155] H. H. Yin, B. J. Knowlton, and B. W. Balleine. "Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning". In: *European journal of neuroscience* 19.1 (2004), pp. 181–189 (cit. on p. 5).

[156] H. H. Yin et al. "Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill". In: *Nature neuroscience* 12.3 (2009), pp. 333–341 (cit. on p. 35).

[157] H. H. Yin et al. "The role of the dorsomedial striatum in instrumental conditioning". In: *European Journal of Neuroscience* 22.2 (2005), pp. 513–523 (cit. on p. 5).

[158] J. Yoshimura et al. "Dynamic decision-making in uncertain environments I. The principle of dynamic utility". In: *Journal of ethology* 31 (2013), pp. 101–105 (cit. on p. 30).

[159] P. Zhang et al. "Distributional reinforcement learning for multi-dimensional reward functions". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1519–1529 (cit. on p. 10).

ITQB NOVA, Avenida da República, 2780-157 Oeiras, Portugal | Tel +351 214 469 100 | itqb.unl.pt