Expectile-based Distributional Reinforcement Learning and Dopamine-associated Mental Disorders

M. Farrugia-Roberts, N. Kruck, T. Premrudeepreechacharn, P. Santhanakrishnan, S. Yang

Spring Semester, 2020

Abstract

Recent evidence suggests a link between expectile-based distributional reinforcement learning and the phasic activity of midbrain dopaminergic neurons. We review the long-neglected concept of expectile statistics, the recent trend of distributional reinforcement learning, and traditional models used to understand the brain's dopaminergic reward system and its links to several mental disorders. We propose a new architectural model of the brain's reward system as an expectile-based distributional extension of the traditional neural actor-critic architecture. We also derive an expectile imputation strategy that is more efficient and biologically plausible than existing optimisation-based strategies. Finally, we develop a detailed hypothesis regarding a uniquely distributional mechanism of learning distortions characteristic of several dopamine-associated mental disorders, and conduct exploratory simulation experiments to establish the conceptual feasibility of our hypothesis.

1 Introduction

We investigate links between *reinforcement learning (RL)*, the subfield of machine learning concerning the design of *computational agents* that learn how to behave through interaction with their environment [1], and *reinforcement learning*, the subfield of neuroscience concerning the neural systems with which animals learn how to behave through interaction with their environment [2]. We begin in sections 1.1 and 1.2 with a review of relevant ideas from each of these field.

In section 1.3 we integrate recent advancements from each field [3, 4] into a new model for distributional RL in the brain, we outline a detailed hypothesis regarding its links to dopamine-associated mental disorders, and we take steps towards deriving a simple, efficient, and biologically plausible expectile imputation strategy to accompany our model.

The remainder of the report is organised as follows: In section 2 we outline two exploratory simulation experiments relevant to our hypotheses. In sections 3 and 4 we present and discuss our findings from these experiments. In extensive appendices, we provide a detailed description of selected mathematical aspects of our model, related algorithms, and some alternative models.

1.1 Review of computational reinforcement learning

Here we give a semi-formal overview of the foundational concepts of computational reinforcement learning (RL). For a comprehensive introduction, see [1]. We formalise the interaction between a computational *agent* and its *environment* as a *Markov decision process*: In each of a sequence of time *steps*, the agent perceives a *stimulus* x describing[1] the environment's *state* s, and selects and performs an *action* a; then the environment delivers a numerical *reward* r, and transitions to a new state s'. RL algorithms aim to create agents with decision-making *policies* that achieve high *return* in expectation. We define return as a geometrically discounted sum of the rewards from each step. The expectation is taken over randomness in the agent's actions and the environment's reward and transition dynamics.

Temporal difference (TD) learning is a well-studied framework for RL algorithms [1] wherein the algorithm directly estimates the *value* of each stimulus (defined as the conditional expected return given the stimulus) from experience, and the agent uses these estimates to select actions. These values (and their estimates) are summarised by functions $v(x)$ (and $\hat{v}(x)$). TD algorithms center around the precise method of *updating* $\hat{v}(x)$ based on observed interactions with the environment.

In simple environments, where the stimulus identifies the state[1], we may use *tabular* TD algorithms. These algorithms estimate $\hat{v}$ as a *table* of estimates (with one estimate per state) and update the table according to a *Bellman update equation* such as[2,3]

$$\hat{v}(x) \overset{+}{\leftarrow} \alpha(r + \gamma\hat{v}(x') - \hat{v}(x)) \tag{1}$$

where x, r, x' is a sequence of states and rewards drawn from experience in the environment, $\alpha > 0$ is a *learning rate* used to modulate the speed and stability of the update process, and $\gamma \in [0, 1]$ is a discount factor used in defining the return. Note that (1) simply adjusts $\hat{v}(x)$ towards $r + \gamma\hat{v}(x')$ in proportion to the difference between these two quantities. This latter quantity is a subsequent estimate of the true $v(x)$ based on additional experience r, x' available at a later time, and accordingly this difference is the so-called *temporal difference*.

In environments with more possible stimuli such a table would be intractably large or infinite, but TD RL is still possible. One may instead estimate a finite parametrisation of v using function approximation techniques such as artificial neural networks [5], leading to so-called *deep reinforcement learning* (successfully applied in e.g. [6, 7, 8]). Update equation (1) generalises to to update the parameters of the function approximator based on the gradient of a mean squared error loss function—the loss function defining expected values (see section 1.1.1). In practice, these updates tend to converge to an accurate estimate of v as an instance of *stochastic semi-gradient descent* [1].

[1] The stimulus is the information available to the agent. It may be the state's identity ($x = s$), or an incomplete set of state 'features'.

[2] Many variations on this equation exist. In this review we focus on *1-step TD for state values*. For our experiments, we switch to using *state-action values*—we maintain a value estimate for each state-action combination. The updates retain the same general structure [1].

[3] Here, $a \overset{+}{\leftarrow} b$ denotes additive assignment, as in $a \leftarrow a + b$.

Finally, we link value estimation and action selection. Some algorithms (e.g. *Q-learning*, *SARSA* [1], *deep Q-learning* [6]) estimate the value of each action alongside each state, and select using these estimates. Alternatively, *actor-critic* methods [1, 9, 7] learn their *decision policy* directly alongside their value function, mutually optimising both: Interactions controlled by the policy (termed the 'actor') determine value function updates, and the predictions of the value function (termed the 'critic') inform policy updates. Actor-critic methods are of special interest for their use as architectural models of neural reward systems (see section 1.2.2).

1.1.1 Expectile statistics and generalised expectations

The expected value μ_X of a random variable X minimises the expected squared distance from the random variable. That is,

$$\mu_X = \arg\min_{\mu} \mathbb{E}\left[(X-\mu)^2\right]. \tag{2}$$

Expectiles are a class of summary statistics generalising the expected value [10, 11]. Given an *asymmetry parameter* $\tau \in (0,1)$, the τ-expectile of X, $\epsilon_X(\tau)$, minimises an *asymmetric* version of this expected squared distance, weighting squared positive distances by τ and squared negative distances by $1-\tau$:[4]

$$\epsilon_X(\tau) = \arg\min_{\epsilon} \mathbb{E}\left[[\![X > \epsilon]\!]^{\tau}_{1-\tau}(X-\epsilon)^2\right] \tag{3}$$

The expected value μ_X is recovered as the 0.5-expectile. Expectiles with $\tau > 0.5$ are more sensitive to 'higher-than-expected' outcomes than they are to 'lower-than-expected' outcomes, and so they lie above the expected value of X. We can thus view these expectiles as *optimistic expectations* (they 'expect' the value of X to be higher than its expected value). Likewise, expectiles with $\tau < 0.5$ lie below μ_X and represent *pessimistic expectations*. In this sense, the asymmetry parameter $\tau \in (0,1)$ captures the degree of *positive outlook* of an expectile.

Since their introduction, expectiles have been neglected [12], perhaps because sample expectiles lack the simple formula and interpretation of the expected value or of *quantiles*. Quantiles and expectiles are deeply related [13, 14], and expectiles may be preferable in some contexts [15, 16, 17]. Moreover, viewing expectiles as *generalised expectations* at varied *degrees of positive outlook* may improve interpretability, and appendix A.1 gives an efficient procedure for computing sample expectiles.

1.1.2 Distributional reinforcement learning

While traditional TD algorithms approach the goal of maximising expected return by directly estimating expected return, recent *distributional RL* algorithms estimate richer characterisations of conditional return distributions than just their expected values. For example, one might model conditional return distributions with a parametric family of distributions and estimate their parameters (one set of parameters for each stimulus) [18], one might learn a finite non-parametric estimate of each conditional distribution or its cumulative distribution functions [19, 20], or, more generally, one might learn an arbitrary set of statistics of each distribution such as their *quantiles* [21, 22] or *expectiles* [3]. Moreover, many of these algorithms have tabular and function-approximation variants. Notably, [23] extends the *actor-critic* framework to incorporate a *distributional critic*.

Each distributional RL method brings its own *distributional Bellman update equation*, an analog of (1) extended to maintain a *suite of estimates* $\hat{s}_1(x), \hat{s}_2(x), \ldots, \hat{s}_K(x)$. The precise form of the new update equations must depend on the statistics being estimated, but [3] provides a unified framework for distributional updates of arbitrary statistics. The following is an example of an algorithm based on their general framework:

Algorithm 1 (update framework)**.** Given experience x, r, x':

1. *Impute* the estimates $\hat{s}_1(x'), \ldots, \hat{s}_K(x')$ into a *sample* of points $z'_1, \ldots, z'_N$ with the return distribution of x'.
2. Transform each z'_i into $z_i = r + \gamma z'_i$, yielding a sample $z_1, \ldots, z_N$ with the return distribution of x.
3. With these z_i, update each $\hat{s}_k(x)$ towards the desired statistics, e.g. by following the gradient of a *loss function*.

In pursuit of an *expectile-based* distributional RL algorithm, [3] proposes to estimate each conditional return distribution with expectiles at a fixed set of asymmetry parameters $\tau_1, \ldots, \tau_K$. We denote the estimates $\hat{\epsilon}_x(\tau_1), \ldots, \hat{\epsilon}_x(\tau_K)$ for stimulus x.

Their algorithm, *Expectile Distributional RL (EDRL)*, follows the framework above. Step 3 uses the optimisation target of equation (3) as a loss function, moving each $\hat{\epsilon}_x(\tau_k)$ towards the τ_k-expectile of $z_1, \ldots, z_N$, directed by its gradient:[3,4]

$$\hat{\epsilon}_x(\tau_k) \overset{+}{\leftarrow} \frac{2\alpha}{N}\sum_{i=0}^{N}[\![z_i > \hat{\epsilon}_x(\tau_k)]\!]^{\tau_k}_{1-\tau_k}(z_i - \hat{\epsilon}_x(\tau_k)) \tag{4}$$

This update equation resembles equation (1): $\hat{\epsilon}_x(\tau_k)$ moves towards each z_i in 'proportion' to the difference $z_i - \hat{\epsilon}_x(\tau_k)$—except with separate proportionality constants for positive and negative differences. We return to this insight in section 1.3.

Furthermore, step 1 requires an *expectile imputation strategy*—a method for converting a set of expectiles $\epsilon(\tau_1), \ldots, \epsilon(\tau_K)$ into a sample $z_1, \ldots, z_N$ with those expectiles. Some statistics, such as quantiles, permit a closed-form imputation strategy, but [3] finds only an *optimisation-based* strategy for expectiles. They compute an imputed sample *numerically* as the root of the vector function $f(z_1, \ldots, z_{N=K}) = (f_1, \ldots, f_K)$ with

$$f_k = -\frac{2}{N}\sum_{i=0}^{N}[\![z_i > \epsilon(\tau_k)]\!]^{\tau_k}_{1-\tau_k}(z_i - \epsilon(\tau_k)). \tag{5}$$

Each f_k is the gradient of the optimisation target of (3) defining the τ_k-expectile of the sample, evaluated at $\epsilon(\tau_k)$, and thus a root of f corresponds to a sample with the required expectiles[5].

Distributional RL shows promise over traditional value-based RL. Its algorithms [20, 21, 22, 23, 3] outperform state-of-the-art traditional RL approaches in popular benchmarks [24]. [20, 4] offer some explanations for this improvement. Besides performance, [18] advocates a second class of benefits of distributional RL: Learning about return distributions enables making decisions according to alternative *risk preferences*. For example, one may control exposure to large negative returns[6] by selecting actions based on risk-sensitive measures such as *value at risk* or *conditional value at risk* (a.k.a. *expected shortfall*). These measures may be computed from distributional information such as quantiles [18] or expectiles [16].

[4] Here, $[\![P]\!]^a_b$ denotes a *generalised Iverson bracket*, evaluating to a if P is a true proposition, or to b otherwise.

[5] Steps 1 and 3 both involve the same gradient expression, but they differ importantly in use: Step 3 varies expectile estimates based on a fixed sample; Step 1 varies sample points to match fixed expectiles.

[6] Pursuing high expected return may permit rare ruinous outcomes.

1.2 Review of reinforcement learning in the brain

In this section, we provide a review of animal reinforcement learning and associated mental disorders. We focus on dopaminergic neurons—long implicated in reward-based learning since [25] observed that their electrical stimulation positively reinforced behaviour in rats. While other neurotransmitters such as serotonin have significant effects on learning behaviour and cognition [26], they are beyond our review's scope.

Dopamine is a neurotransmitter with significant effects on cognition and behaviour including motor control, motivation and decision making. A large majority of dopaminergic neurons are found in the ventral part of the mesencephalon (midbrain). The mesencephalic dopaminergic system consists of three major nuclei—the substantia nigra pars compacta (SNpc), the ventral tegmental area (VTA) and the retrorubtal field [27], and comprises of several dopaminergic networks with separate projection pathways such as the nigrostriatal pathway and the mesolimbic pathway.

1.2.1 The reward prediction error hypothesis

The *reward prediction error hypothesis* attempts to explain the phasic activity of mesencephalic dopaminergic neurons. Under this hypothesis, the brain uses a temporal difference (TD) algorithm to learn to predict and maximise subjectively rewarding experiences [1, 28, 29, 30]. *Reward predictions* made by the brain correspond to the *value estimates* of a TD algorithm, and *reward prediction errors* (RPEs) correspond to the *temporal difference* from equation (1). Moreover, the phasic activity of dopaminergic neurons is a neurophysiological signal encoding the magnitude of these RPEs. This link between computational and biological RL was made after [31, 32] observed that when a subject is presented with an unexpected reward, dopaminergic neurons respond strongly, but once the subject is trained and an association is built between a predictive stimulus and the reward, the response of the dopaminergic neurons reduces.

In the literature, RPE is defined as the quantitative discrepancy between the result expected when the cue was presented, and the result that was actually experienced [33]. When a cue is first encountered, followed by an unexpected reward, there is a large discrepancy between the expectations and actual occurrence. This leads to the production of a large RPE. However, once the individual learns that the cue reliably predicts a particular event, there is little RPE as the discrepancy between what is expected and what actually occurs is reduced. Therefore, RPE facilitates learning about reward-predictive cues and allows more accurate predictions about future rewards [34].

Dopaminergic neurons not only respond to the reward itself but also the predictive stimulus [29]. When the reward itself is presented, there is no response if the reward had already been predicted with the stimulus. When the reward received is greater than predicted, dopaminergic neurons respond strongly, reflecting a positive RPE. On the other hand, rewards which are smaller than predicted elicit a negative response (depressive activity), reflecting a negative RPE. This activity of dopaminergic neurons largely occurs in the SNpc, which is part of a larger cortico-basal ganglia-thalamo-cortical (CBGTC) circuit that is critical to cognitive action selection.

There are two core pathways in the CBGTC, namely the direct Go pathway which responds to positive RPE and the indirect No-Go pathway which responds to negative RPE [35]. The striatum, which is the major input region of the CBGTC, contains two major type of neurons corresponding to the direct Go pathway and the indirect No-Go pathway. The Go pathway neurons predominantly contain D1 receptors on which dopamine has an excitatory effect. By contrast, No-Go pathway neurons contain D2 receptors via which, dopamine has an inhibitory effect [36]. Therefore, when there is fast and high-amplitude phasic dopaminergic firing, encoding positive RPE, the direct Go pathway is activated via D1 receptors, driving reinforcement of actions that lead to positive rewards. When there is a dip in dopamine firing, encoding negative RPE, the indirect pathway neurons with D2 receptors are disinhibited, promoting avoidance learning.

1.2.2 The neural actor-critic model

The traditional actor-critic model as briefly described in section 1.1 consists of a critic sub-network computing the weighted sum of future rewards and an actor sub-network which utilizes these predictions to choose actions that maximize the weighted sum. Multiple studies reviewed by [37] and [38] assigned the actor and critic functions to specific brain regions; suggestions have been made that dopaminergic signals from the VTA to the ventral striatum correspond to the critic, whereas signals from the SNpc to the dorsal striatum corresponds to the actor as it learns the action-selection policy, in contrast to earlier works that assigned the actor-critic system to two chemically different compartments within the striatum, the matrix- and striosomal neurons [39].

The model is illustrated in figure 1. Information on states of the internal or external environment is represented by the input from the frontal cortex to the ventral striatum. Outputs of the ventral striatum (the critic), predictions of state values, are used to calculate the RPE in the VTA and the SNpc. The dopaminergic signal resulting from the RPE computed in the VTA/SNpc then projects back to the dorsal and ventral striatum (the actor and critic, respectively) [38], inducing the 'policy updates' mentioned above.

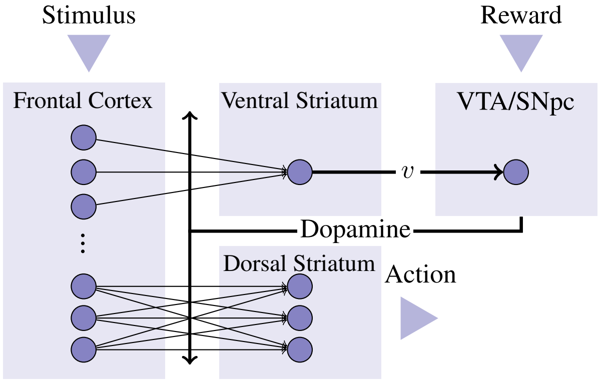

Figure 1: Traditional actor-critic architecture in the biological system, adapted from [38].

Disturbances in this actor-critic system, such as the impairment of the critic in response to cocaine exposure, can have detrimental effects on action selection processes that [38] describes as "the coach gone awry and an actor running loose", and lead to aberrant behavior observed in mental disorders, which will be discussed in the following section.

1.2.3 Dopamine-associated mental disorders

Dysfunctions in dopaminergic neurons have been implicated in multiple neuropsychiatric disorders such as schizophrenia [40], addiction [41], and Parkinson's disease [42].

Insights on mechanisms and complications of mental disorders associated with reward processing systems might be explained by a distorted reinforcement learning system. The list of mental disorders discussed in this work is not exhaustive and leaves out clinically relevant disorders such as depression, which is characterized by an overly pessimistic view on the model of the world [43].

Drug Addiction A pattern often seen in drug addiction patients is that despite of some kind of awareness that persisting drug intake will result in serious harm, compulsive drug-seeking behavior will be continued. This indicates that the conscious assessment of consequences in the higher levels of the brain may still be intact but due to an overvaluation of drugs, the action selection process in the lower levels is distorted and choices that lead to drug receipt will be preferred over other, less harmful options [44, 45].

The main aspect that sets drug addiction apart from other dopamine-associated mental disorders is the direct or indirect neuropharmacologically induced increase of dopamine levels in response to cocaine, nicotine, heroin and other substances of abuse [46]. In TDRL, an effective learning process requires a shift of the RPE signal to zero, in which case no dopamine signal production occurs (phasic dopamine is a learning signal encoding RPE, as mentioned in 1.2.1) and learning stops [47]. However, the DA increase caused by drugs always produces a positive RPE signal, increasing the value of states leading to drug receipt, which consequently increases the likelihood for the agent to select an action that leads to these states [46]. Repetitive selection of actions in favor of drug consumption will ultimately 'overshadow' the values of other reinforcing rewards [45].

The dysregulation of motivational circuits that is found in drug addiction has been described by a three-step cycle mediated by three major neurocircuits; the basal ganglia, extended amygdala, and prefrontal cortex, as proposed by [48]. The first stage, defined as the *binge/intoxication stage*, involves a drug-mediated dopamine release (among other neurotransmitters) into the ventral striatum, causing the individual to feel 'high' by the emulation of dopamine increase frequencies associated with rewards. New associations with drug availability are established for previously neutral stimuli, increasing incentive salience. The subsequent *withdrawal/negative affect stage* is characterized by stress responses such as chronic irritability, dysphoria or decreased motivation for natural rewards, which are the result of dysregulated components of the reward system including the decreased dopaminergic transmission in the nucleus accumbens. The third stage, termed the *preoccupation/anticipation stage*, is responsible for relapses in humans and involves impaired inhibition of maladaptive behavior [48]. The No-Go-system would normally inhibit the Go-system responsible for the drive of craving and engagement of habits; this inhibitory control is missing in addiction, preventing individuals from successful cognitive inhibition of craving [48].

The shift from the binge/intoxication stage to the withdrawal stage that drives the drug-taking behavior likely represents a shift from impulsivity to compulsivity, or a shift from positive to negative reinforcement. Initial drug use for the feeling of reward is followed by drug use to avoid negative states [49, 48]. Considering the fact that only certain substance users go through this transition [48] and that heritability exists in addiction [50], genetic differences between individuals and epigenetic changes also need to be kept in mind; however, this is beyond the scope of our review.

Gambling addiction The link between gambling disorder and dopamine is suggested by the fact that uncertainty is both the core feature of gambling games and the main cause of dopamine activation. As well as yielding reward prediction error, the variability of rewards boosts the incentive salience of its stimuli [51]. Other dopamine response to reward proximity [52, 53] or stress [54] may also reinforce the addiction. In addition, overactive dopamine firings encourage the causal binding of unrelated events [55], which leads to a sense of control and distorted reasoning seen in the case of 'near misses' or 'losses disguised as wins' [56, 57].

Other studies proposed models where dopamine is not mainly responsible for the disorder, yet plays significant role in the complete process. [58] attributed gambling disorder to the impairment of model-based learning and decision making, which results in insensitivity to the changes in the environment. [59] explained the addiction with broken state representation and categorization formed by a sequence of rewards with a few big strikes at the start and many small losses following. [60] modeled the gambling disorder as a lack of directed exploration.

Parkinson's Disease Parkinson's Disease (PD) is one of the most common age-related neurodegenerative diseases. PD mainly affects the motor control system, and is caused by loss of dopaminergic neurons, with also genes and environmental factors contributing to define the mechanism and cause of disease [61, 62, 63]. This midbrain dopaminergic neuron loss in nigrostriatal pathway results in a lower tonic dopamine level and lower phasic dopamine bursts, equivalently increasing recruitment of the indirect No-Go pathway and decreasing that of Go pathway, facilitating punishment learning and impaired reward learning, respectively. Punishment learning is the ability of a subject to learning to avoid negative stimuli, and reward learning is learning to choose positive stimuli [64]. The medication for PD patients, such as dopamine D2 receptor agonists (DA) or dopamine replacement therapy (L-dopa), reverses this No-Go bias. However, some patients with PD, up to 20%, tend to have manifested at least one impulse control disorder (ICD) and related behaviors, such as compulsive buying and pathological gambling, at some point [65].

The overmedication hypothesis is proposed to explain this phenomenon as a possible side effect of overdosing the ventral striatum, which is less affected in PD patients compared to the dorsal striatum [66]. Coupled with DA medication, PD, affecting the actor, patients with ICD, affecting the critic, tend to learn and perform better in trials with positive reward than in negative rewards, suggesting suboptimal learning from negative RPE in the critic [67]. Interestingly, [68] finds that in addition to verifying impaired loss learning in PD patients when applied DA medication with fMRI, on the fact that DA induce greater ventral striatal activity to positive RPE in PD patients with ICD, resulting in a 'better than expected' outcome.

1.3 Distributional reinforcement learning in the brain

Though the RPE hypothesis of the phasic signalling of midbrain dopaminergic neurons has been acknowledged as one of the most successful and elegant achievements of computational neuroscience [69, 70], recent work suggests that traditional temporal difference RL, if it indeed tells the story of RL in brains, may not tell the whole story. Analysing reward response data from an optogenetic study of midbrain dopaminergic neurons in mice [71], [4] found that while aggregate phasic activity followed the predictions of traditional RL, responses differed significantly across individual neurons. Earlier work (e.g. [72]) reports similar diversity amongst neurons, but [4] offers an explanation in terms of the new *distributional RL* paradigm.

In particular, [4] analysed the responses of 40 dopaminergic neurons to various reward magnitudes and found the following:

1. The neurons switched from decreased to increased activity (compared to baseline) at different reward magnitudes.
2. Each neuron responded linearly to reward magnitude on either side of its individual reversal point, but the slopes on either side (and also between neurons) differed.

These findings suggest the neurons did not signal a uniform RPE according the same prediction and subjective reward signals as if by equation (1)—otherwise, their reversal points would coincide, and they would each behave linearly.

Instead, [4] equates each neuron's piece-wise linear response with the characteristic asymmetry of expectile statistics, and suggests that these responses signal the *asymmetrically-scaled* prediction errors of equation (4) as if implementing *expectile-based distributional RL*. In this view, the two slopes of each neuron determine an individual asymmetry parameter τ, and the reversal point of the neuron corresponds to the predicted τ-expectile of the reward distribution. [4] supports this *distributional hypothesis* of midbrain phasic dopaminergic signalling with an additional finding:

3. The slopes and reversal points of the neurons, if interpreted as a set of τ-expectiles, qualitatively resemble the expectiles of the underlying reward distribution.

This *distributional hypothesis* is appealing. Though this analysis is based on a small number of neurons, existing evidence seems to support the general idea that dopaminergic neurons could operate in a complex heterogeny [73, 74, 75]. Moreover, the advantages of distributional over traditional RL in terms of learning performance and flexibile risk preferences (see section 1.1.2) may justify its natural existence. This hypothesis may represent another event in the history of mutually-beneficial discovery between the fields of artificial intelligence and neuroscience [1, 4]. Finally, a model of learning in the brain based on distributional RL affords a number of new ways for things to 'go awry', with new potential to explain the mechanisms of learning-associated mental disorders [4]. We build directly on this hypothesis in our work.

In a similar vein, it has recently been proposed to model the distorted learning of patients with certain dopamine-associated mental disorders using a generalised RL framework with parallel treatment of positive and negative rewards [76, 77]. Their proposed model, *split Q-learning*, splits positive and negative reward into two separate streams and maintains separate estimates of the return derived drawn from each stream. Weighting these streams asymmetrically affords great flexibility in modelling the effect that mental disorders have on the reward processing systems [78]. This asymmetry is reminiscent of expectile-based distributional RL, however we note that even distributional RL only considers a single stream delivering the sum of positive and negative rewards. We take some inspiration from [77] in our experimental design (see section 2.2).

1.3.1 A distributional neural actor-critic model

Building on the distributional hypothesis of [4], we propose that an extended neural actor-critic architecture implements expectile-based distributional RL in the brain. We continue to associate the *ventral striatum* with the *critic* and the *dorsal striatum* with the *actor*, and their input from the *frontal cortex* still represents the environmental stimulus. We propose that the outputs of the ventral striatum represent a suite of predictions, each representing an *expectile* of the future reward *distribution* at some degree of asymmetry. Furthermore, we suppose that the VTA and SNpc, or neighbouring areas, continuously perform some kind of *imputation* so as to compute asymmetrically-scaled RPEs as if by equation (4)[7]. The *dopaminergic neurons* of the VTA/SNpc are responsible for communicating these asymmetrically-scaled RPEs to the remainder of the system to drive the learning process, as per the distributional hypothesis.

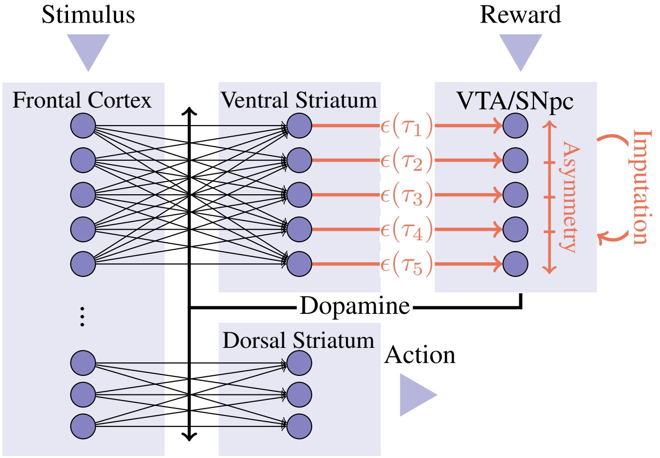

Figure 2: *Distributional* neural actor-critic architecture, inspired by expectile-based distributional RL and the imputation framework.

The model assumes significant complexity between the ventral striatum and VTA/SNpc. Expectile predictions must reach dopaminergic neurons with corresponding asymmetries. Distributional learning may also require the dopaminergic neurons' outputs to selectively project back to the source of their respective inputs [4][8]. Correlating the asymmetry parameters of dopaminergic neurons with their spatial location in the VTA/SNpc may simplify the implementation. [73] reported spatial diversity in the response characteristics of midbrain dopaminergic neurons (though not the same kind of response diversity explored in [4]). Furthermore, neuromodulatory systems are known for exhibiting complex heterogeneity [74]. In particular, [75] discusses molecular and genetic diversity amongst dopaminergic neurons which, we propose, could help establish the kind of spatial organisation discussed above.

[7] We discuss this link in greater detail in appendix A.2.

[8] The discussion in [4] (supplemental information) on this point appears to neglect the importance of imputation [3]. Properly considering imputation may turn out to simplify the required architecture.

1.3.2 An alternative expectile imputation strategy

Imputation is a bottleneck in current expectile-based distributional RL algorithms and in their biological plausibility. [3] mitigated EDRL training time with additional computational resources to perform their optimisation-based imputation strategy. If the brain performs imputation, we suppose it is not through this computationally demanding optimisation-based strategy.

An efficient replacement strategy could streamline expectile-based distributional RL and provide a more biologically plausible account of the brain's reward processing according to the distributional hypothesis[9]. We derive an efficient alternative strategy—with, in places, the parallel flavour of biologically plausible neural computation—in algorithm 2. The strategy is based on an insight from one of the original papers introducing expectiles [11] relating the expectiles of a random variable to the variable's cumulative distribution function (CDF):

Theorem 1. *Let X be a random variable with mean μ_X and CDF F_X. The solutions to equation (3) form a strictly monotonically increasing function of τ, $\epsilon_X(\tau) : (0,1) \to \mathbb{R}$, with range R_X and inverse function $\tau_X(\epsilon) : R_X \to (0,1)$. Moreover, if F_X is continuously differentiable, then τ_X has continuous derivative τ'_X, and for $x \neq \mu_X$ (that is, for $\tau_X(x) \neq \frac{1}{2}$),*

$$F_X(x) = \frac{\tau'_X(x)(\mu_X - x) - \tau_X(x)(1 - 2\tau_X(x))}{(1 - 2\tau_X(x))^2}, \qquad (6)$$

where this equation holds in the limit for $x = \mu_X$.

Proof. In their first theorem, [11] proves the existence, uniqueness, and strict monotonicity of the function ϵ_X, and a similar relationship to equation (6). We merely simplify this relationship, noting $\epsilon'_X(\tau_X(x)) = 1/\tau'_X(x)$ since $\epsilon_X = \tau_X^{-1}$. □

Algorithm 2 (alternative imputation strategy)**.** Given a set of asymmetry parameters $\tau_1, \dots, \tau_K$ and their corresponding expectiles $\epsilon_1, \dots, \epsilon_K$ for a random variable X with CDF F_X:

1. Interpret $(\epsilon_1, \tau_1), \dots, (\epsilon_K, \tau_K)$ as coordinates of the function τ_X (with $\tau_X(\epsilon_k) = \tau_k$).
2. Approximate τ'_X with coordinates $(\epsilon_1, \tau'_1), \dots, (\epsilon_K, \tau'_K)$ (with $\tau'_k \approx \tau'_X(\epsilon_k)$) based on neighbouring points of τ_X.
3. Approximate F_X with coordinates $(\epsilon_1, F_1), \dots, (\epsilon_1, F_K)$ (with $F_k \approx F_X(\epsilon_k)$) using equation (6). Exclude any k for which $\tau_k = \frac{1}{2}$, noting that for such k, $\epsilon_k = \mu_X$.
4. Use *inversion sampling* to produce a sample from F_X. That is, return the inverse image under F_X of a uniformly random sample over $[0,1]$. A large sample will share this CDF. Note: To accurately invert F_X, we must interpolate between and extrapolate beyond our coordinates.

This strategy represents a hopeful step towards better expectile imputation strategies for both computers and brains. In our implementation, we approximate τ'_X by fitting a quadratic polynomial to the immediately neighbouring points along the inverse of τ_X, and we interpolate between our coordinates with radial basis functions and extrapolate beyond them with an exponential decay. We note that other detailed implementations may be preferable, but are yet to be explored. See appendix A.4 for further discussion.

1.3.3 Imputation fidelity and imputation distortion

An expectile imputation strategy should faithfully recover a distribution similar to the one from which its input expectiles were estimated. We refer to this concept as the strategy's *fidelity*.

The distribution of asymmetry parameters[10] may underpin the fidelity of an imputation strategy. A faithful and efficient imputation strategy may rely on its inputs representing a sufficiently *full* spectrum of asymmetries (that is, with some optimistic, some balanced, and some pessimistic expectiles), and its fidelity may plausibly suffer given a different distribution of asymmetry parameters (for example, with only optimistic or only pessimistic expectiles[11]). Such inputs might lead to the strategy producing a *distorted* imputed distribution, and the distorted distribution may be systematically biased relative to the original distribution, following the mismatch between the assumed asymmetry distribution of the strategy and the real asymmetry distribution of the input expectiles.

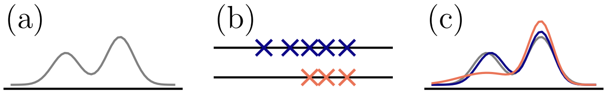

Figure 3: Imputation distortion. **(a):** An example reward distribution. **(b):** Its expectiles at various asymmetries (orange: missing pessimistic expectiles). **(c):** Hypothetical (orange: distorted) imputed distribution.

Such an *imputation distortion* may lead to a distortion in the learned distributions in a distributional RL algorithm, and thereby to distorted behaviour. According to the imputation-based update framework, imputation forms a crucial part of every parameter update during learning. If a systematic bias alters distributional information as it is 'backed up' from the estimates at one state to the estimates at the previous state, then this bias could prevent accurate learning of future rewards' distributions. Unable to learn about future consequences despite experience, an agent may in turn show distorted behaviour.

This leads to our core hypothesis regarding a uniquely distributional mechanism for distorted learning in mental disorder patients. We stress that the distributional perspective is integral to our hypothesis: In the traditional, expected-value-based RL paradigm, there is no distribution of asymmetry parameters determining a suite of estimates maintained by the learner, and no concept of imputation through which updates can be corrupted.

Hypothesis. The brain implements expectile-based distributional RL according to our distributional neural actor-critic model. In healthy individuals, predictions are maintained across a wide distribution of asymmetry parameters. In some individuals, the effective distribution of asymmetry parameters is altered (for example, by selectively disabling neurons involved in processing pessimistic expectiles). This impairs the fidelity of the imputation calculation, leading to a systematic distortion of distributional information during distributional updates. This in turn affects the distributional predictions learned by the reward system, and, finally, manifests as a behavioural distortion characteristic of a dopamine-associated mental disorder.

[9] It may be possible to achieving distributional RL in the brain outside the imputation framework. Such a method would be consistent with the distributional hypothesis, but would invalidate our model.

[10] We distinguish *two distributions*: The *distribution of future rewards* which a distributional RL algorithm seeks to estimate, and the *distribution of asymmetry parameters* determining the suite of statistics with which the algorithm accomplishes this estimation.

[11] While each expectile contains information about an entire distribution, for efficiency reasons a strategy may not utilise all information in each expectile when expecting a comprehensive set of expectiles.

2 Methods

In this section we detail our exploratory simulation experiments investigating the fidelity of different imputation strategies and its susceptibility to a shift in the asymmetry distribution, and measuring the effect of a shifted asymmetry distribution on distributional learning dynamics.

2.1 Imputation fidelity experiments

We designed a simulation experiment to explore the fidelity of each imputation strategy—both the optimisation-based strategy from [3] as summarised in section 1.1.2, and our alternative strategy as derived in section 1.3.2.

The simulation begins with a large sample randomly drawn from a test distribution, such as a multi-modal Gaussian mixture model. From this sample, we calculate the τ_k-expectile for the set of asymmetry parameters $\tau_k = \frac{2k-1}{2K}$ ($k = 1, \ldots, K$). To test an imputation strategy, we use the strategy to impute these expectiles into an *imputed* sample. Then, we compute the expectiles of the imputed sample.

Comparing the original sample to the imputed sample, and comparing the original sample's expectiles to those of the imputed sample, reveals the level of fidelity with which the imputation strategy achieves its goal. Any divergence between the two distributions may signify an undesired imputation distortion. To amplify and detect even small distortions, we continue our simulation by iterating the cycle described above: Use the previous imputed sample's expectiles to impute another sample, take expectiles, and so on.

To explore the effect of the asymmetry distribution on this fidelity, we repeated the above simulation for each imputation strategy with the pessimistic expectiles withheld from the imputation step. That is, we use an *optimistic* asymmetry distribution, spanning from $\tau_{\lfloor \frac{k}{2} \rfloor} = \frac{1}{2}$ to $\tau_K = \frac{2K-1}{2K}$.

In our experiments, the various simulation parameters were as follows: We used a Gaussian mixture model with two equally-weighted components with unit variance and means ± 5 as the test distribution for all simulations. We set K and thereby the asymmetry distributions as outlined in table 1. We continued each simulation for at least 50 iterations. We used the method from appendix A.1 to efficiently calculate sample expectiles.

Strategy	Asymmetries	K	τ_k used
Optimisation	Full	11	$\frac{1}{22}, \frac{3}{22}, \ldots, \frac{21}{22}$
Optimisation	Optimistic	21	$\frac{1}{2}, \frac{23}{42}, \ldots, \frac{41}{42}$
Alternative	Full	101	$\frac{1}{202}, \frac{3}{202}, \ldots, \frac{201}{202}$
Alternative	Optimistic	101	$\frac{1}{2}, \frac{103}{202}, \ldots, \frac{201}{202}$

Table 1: Parameters for fidelity experiments. The optimisation-based imputation strategy uses smaller K because of its computational complexity (for both asymmetry distributions, we impute 11 expectiles).

2.2 Iowa gambling task experiments

The Iowa gambling task (IGT) [79, 80] is a popular paradigm for studying decision-making in humans and its links to reinforcement learning [81]. In the task, participants must repeatedly choose between four decks of cards, A, B, C, and D, to draw from. The cards from each deck are accompanied by positive rewards, and sometimes also by negative penalties. Various reward schemes are used in the literature, as summarised by [81] (supplemental information), but in all cases decks A and B yield high positive rewards but a negative expectation and decks C and D yield moderate rewards with a positive expectation.

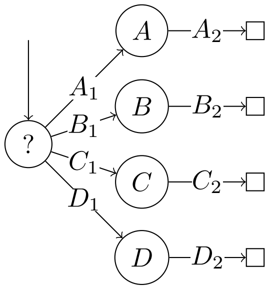

Figure 4: Our 'two-step' Markov decision process implementing an Iowa gambling task. In each episode, the agent begins in state ? and may choose action A_1, B_1, C_1, or D_1, for no reward. Actions A_2, B_2, C_2, and D_2 are available from the respective resulting states. These actions are rewarded according to the reward scheme outlined in table 2, and lead to a terminal state. Thus, the rewards are delivered with a one-step *delay* after the agent chooses a 'deck'.

Table 2: Reward scheme for our Iowa gambling task (IGT) environment. The agent always gets the positive reward. The negative reward is added to the positive reward with the given probability. We also repeated our experiments with a modified reward scheme, to highlight the potential for the asymmetry distribution to influence behaviour in distributional RL. In our modified task, deck A' replaced deck A.

Deck	Positive reward	Negative reward	Prob. of neg. reward	Expected total reward
A	100	-250	50%	-25
B	100	-1250	10%	-25
C	50	-50	50%	25
D	50	-250	10%	25
A'	135	-250	50%	10

In order to explore the impact of the asymmetry distribution on expectile-based distributional RL, we implemented the IGT as a simple Markov decision process (figure 4, reward scheme in table 2). Crucially, our model of the task separates the action of selecting a deck from the action receiving a reward. Learning this task requires updates to 'back up' information learned about each reward distribution to the state from which decisions are made. If this update involves an imputation step distorting the backed-up distribution (e.g. due to a skewed asymmetry distribution), then this distortion may affect the learned distributions in the decision-making state and the agent's behaviour. A successful agent will learn to prefer C_1 and D_1 over A_1 and B_1 based on experience with A_2, B_2, C_2, and D_2. To emphasise the potential for imputation distortion to influence behaviour, we also explored a modified reward scheme with A_2's positive reward increased (with deck A still inferior in expectation).

In our experiments, we compared traditional Q-learning [1] and EDRL ([3], section 1.1.2). For EDRL, we used [3]'s optimisation-based imputation strategy (with $K = 11$), or our alternative imputation strategy (with $K = 101$). In particular, we used EDRL-style *updates* in a Q-learning-style algorithm, learning a suite of expectile estimates for each of the eight state-action combinations. We used the 0.5-expectile to select greedy actions. As in section 2.1, we used uniformly spaced asymmetry distributions and also optimistically-shifted asymmetry distributions. We trained each algorithm for 100,000–200,000 episodes with exploration probability ϵ=0.2, learning rate α=0.001, and discount factor γ=1.

3 Results

We present the results of the simulations outlined in section 2. We summarise our imputation fidelity findings in figure 5, and our IGT findings in figures 6 and 7.

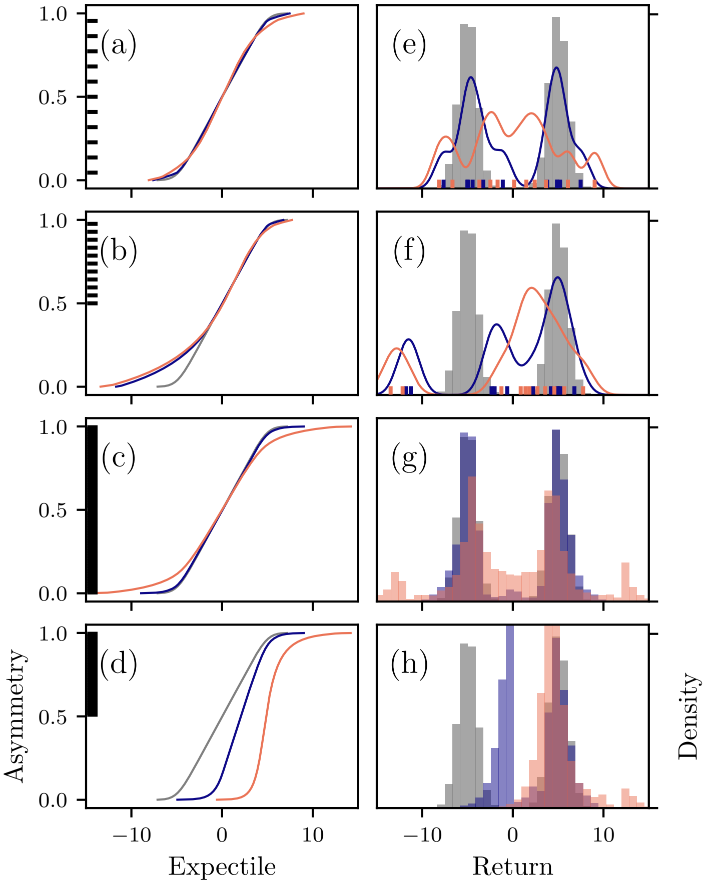

Figure 5: Imputation Fidelity. **Strategy key:** Each row contains the results for one strategy and asymmetry distribution, in the same order as the rows of table 1. **Colour key:** Gray: original distribution; purple: first imputed distribution; orange: distribution after 50 imputations. **(a)–(d):** Expectiles plotted against the corresponding asymmetry parameters. Black marks on the left denote the asymmetry parameters used in imputation (the curves always show a full set of expectiles). **(e), (f):** Optimisation strategy yields samples of $K = 11$ points, scattered along the horizontal axis against a histogram of the original sample. Curves show kernel density estimate (bandwidth parameter 0.18). **(g), (h):** Histograms for original *and* imputed samples.

The optimisation-based strategy showed excellent expectile fidelity and a certain robustness to a skewed asymmetry distribution. This method sustained expectiles *within* the asymmetry distribution in both cases (figure 5a, b). However, perhaps due to the limited output sample size, the samples appear to have failed to capture the true shape of the original distribution (figure 5e, f), despite matching well at certain expectiles.

The alternative strategy showed little divergence with a full asymmetry distribution (figure 5c, g), but a dramatic shift under a skewed asymmetry distribution (figure 5d, h). Notably, 5d shows that this shift impacts not only the ‘missing’ expectiles, but also those within the optimistic asymmetry distribution, including the 0.5-expectile. In further experiments (not shown) we observed a reversed effect with a *pessimistic* asymmetry distribution, and a significant divergence over a larger number of iterations, even with the full asymmetry distribution.

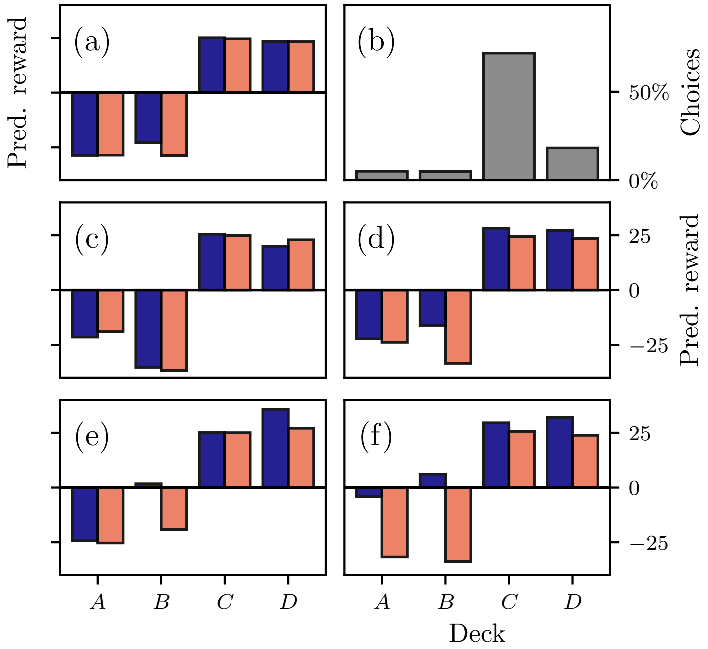

Figure 6: Learning the IGT. **(a):** Q-learning’s expected value estimates for 8 actions $A_1, A_2, B_1, B_2, C_1, C_2, D_1, D_2$, respectively (i.e. left/purple bars: estimates ‘backed up’ to decision-making state, right/orange bars: estimates learned from rewards). **(b):** Deck choices during training for Q-learning (typical of all agents). **(c)–(f):** Like (a), but with learned 0.5-expectiles from EDRL agents of varying imputation strategy ((c), (d): optimisation-based; (e), (f): alternative) and asymmetry distribution ((c), (e): full; (d), (f): optimistic).

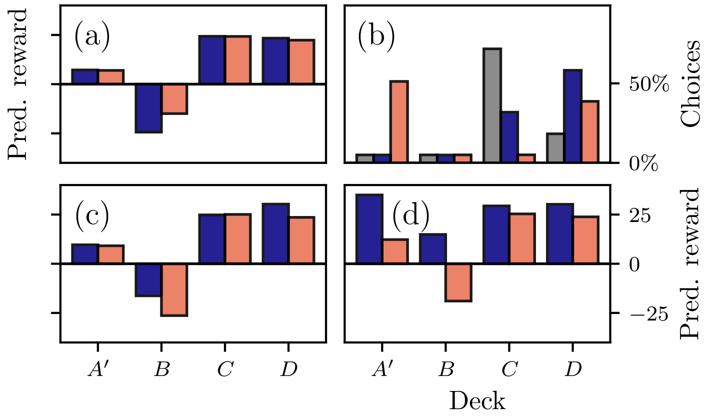

Figure 7: Learning the modified IGT. **(a), (c), (d):** Like figure 6’s (a), (e), (f). **(b):** Deck choices during training (gray: Q-learning; purple: full-asymmetry EDRL; orange: optimistic-asymmetry EDRL).

The divergence shown in our fidelity experiments affected learning and behaviour in our IGT experiments. All agents successfully demonstrated a learned preference for decks C and D in the original task (figure 6b). Comparing pairs of predictions, we see most agents ‘backed up’ most predictions to the decision-making state accurately (figure 6a,c–e). One agent (6f: EDRL, alternative imputation, optimistic asymmetry distribution) learned *systematically higher predictions in the decision-making state*, but still acted optimally. The shift affected decks A and B (with higher positive rewards) most. The modified task (table 2, figure 7), with A_2’s value closer to (but still below) C_2’s and D_2’s, induced a behavioural change in this agent. The agent accurately learned A_2’s, but not A_1’s, value (figure 7d), and this lead to altered decision-making (figure 7b).

4 Discussion

Our results represent a *proof-of-concept* of the ability for the asymmetry distribution used by an expectile-based distributional RL algorithm to lead to a distortion at the level of behaviour. We have established, through our fidelity experiments, that an efficient and otherwise faithful imputation strategy can have its resulting distribution distorted when faced with an 'incomplete' distribution of asymmetry parameters. We have further established, through our IGT experiments, that this imputation distortion can lead to a distorted learning dynamic, corrupting the predictions learned by the algorithm, and, ultimately, the agent's behaviour. We note that these findings may benefit from a well-chosen task and distribution—future work could investigate the prevalence of these effects over a wider range of distributions and learning tasks, and investigate the dynamics of distributions during learning in greater detail. Moreover, future experiments could investigate the asymmetry distribution of dopaminergic neurons by estimating reversal points and asymmetric response slopes [4] to inform both computational distributional RL and links between asymmetry distributions and mental disorders.

We propose that the overvaluation of drug-related stimuli stems from an impairment of pessimistic neurons, leaving an overly optimistic action selection system that favors immediate, drug-related rewards regardless of the negative consequences. Findings had shown that cocaine-addicted individuals continue showing responses to previously rewarding stimuli even when their reward-related predictive value is not given anymore, thus indicating that circuits encoding negative RPE (corresponding to the Stop system decribed in section 1.2.3) are impaired [82], backing up our hypothesis on a systematic distortion of distributional information in pathological individuals.

Our hypothesis of unbalanced asymmetry parameters distorting behaviors can be reconciled with the biological findings. Repeated omission of reward seen in gambling can potentiate the dopamine response for negative outcomes [83], and the concomitant increasing activity in emotional pain related area further depresses efficient allocentric evaluation of the loss [84]. Reward predictive cue itself also places extra emphasis on positive signals [85]. In addition, habenula lesions may potentially explain reducing inhibitory to reward omission and biased learning since it impairs learning more severely from negative RPE than the positive [86]. However, the clear evidence of the involvement of imputation is insufficient to our knowledge. We find an alternative model using a shifted spectrum of asymmetry parameters without explicit use of imputation is able to simulate gambling disorder as well, see appendix B for details.

As [68] has provided the neuroimaging evidence regarding, with dopamine agonist medication, an enhancement in reward learning, or gain learning, in Parkinson's Disease (PD) patients with Impulsive Control Disorder (ICD) and impairment in punishment learning, or loss learning, in PD patients without ICD were observed. From that, the PD patients with ICD case has offered a rigid support to our hypothesis asymmetry cognition of the reward distribution in reward processing systems, which the experimental results are consistent with both the hypothesis and the findings, to a degree—the experimental results can be seen as a mixture of increased gain learning of the agents and impaired punishment learning. However, it would be interesting to model and explore, using distributional RL, the computational agent performance in with a task that has hierarchical reward structure as seen in [64]. The performance on the task infers upon their gain and punishment learning, and along with that, further studies should involve an uncertainty component, which is preferable in experimental designs involved with gambling studies, in their experimental design. In addition to the proposed experimental design to separate reward and punishment learning components, by taking into account that only a certain subtype of dopamine neurons are affected in PD, and that there is a high diversity between dopamine neurons even *within* the same anatomical structures [75], we speculate that the specific dopamine neurons affected in PD patients may correspond to the pessimistic population in our distributional model.

Anhedonia (reduced pleasure in previously enjoyable activities) is commonly reported as a symptom of Major Depressive Disorder (MDD) [87]. Existing literature suggests that anhedonia is likely caused by dysfunctional processing of reward information [88, 89]. Furthermore, individuals with MDD reportedly have reduced sensitivity to rewards [90], increased sensitivity to punishments [91], and increased pessimistic views [43]. We speculate that these behavioural symptoms stem from a pessimistic asymmetry distribution of dopaminergic neurons. Further experiments are required to establish a link between our model and the biological mechanisms of anhedonia.

Our alternative expectile imputation strategy is a step toward more efficient and biologically plausible expectile-based distributional RL. Compared to the optimisation-based strategy, our strategy can impute larger sets of expectiles into larger samples with less computational expense. Since learning more expectiles may improve agent performance [3], we consider this a significant advantage. Future simulations could quantify the fidelity of the method given a full asymmetry distribution over a range of return distributions. Further discussion in appendix A.4. Note that the efficiency of our strategy is linked both to its susceptibility to imputation distortion, and to its biological plausibility. If the brain uses imputation, its strategy may be similarly susceptible to a shifted asymmetry distribution.

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[2] Samuel J Gershman and Nathaniel D Daw. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology*, 68:101–128, 2017.

[3] Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. *arXiv preprint arXiv:1902.08102*, 2019.

[4] Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, pages 1–5, 2020.

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[7] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[8] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[9] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, 5:834–846, 1983.

[10] Dennis J Aigner, Takeshi Amemiya, and Dale J Poirier. On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review*, pages 377–396, 1976.

[11] Whitney K Newey and James L Powell. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847, 1987.

[12] Linda Schulze Waltrup, Fabian Sobotka, Thomas Kneib, and Göran Kauermann. Expectile and quantile regression—david and goliath? *Statistical Modelling*, 15(5):433–456, 2015.

[13] M Chris Jones. Expectiles and m-quantiles are quantiles. *Statistics & Probability Letters*, 20(2):149–153, 1994.

[14] Bradley Efron. Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, pages 93–125, 1991.

[15] Thomas Kneib. Beyond mean regression. *Statistical Modelling*, 13(4):275–303, 2013.

[16] James W Taylor. Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, 6(2):231–252, 2008.

[17] Sabine K Schnabel and Paul HC Eilers. A location-scale model for non-crossing expectile curves. *Stat*, 2(1):171–183, 2013.

[18] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.

[19] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return density estimation for reinforcement learning. In *27th international conference on machine learning (ICML)*, pages 21–25, 2010.

[20] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 449–458. JMLR. org, 2017.

[21] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. *arXiv preprint arXiv:1806.06923*, 2018.

[23] Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva Tb, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*, 2018.

[24] Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew J. Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

[25] James Olds and Peter Milner. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of comparative and physiological psychology*, 47(6):419, 1954.

[26] Pragathi P Balasubramani, V Srinivasa Chakravarthy, Balaraman Ravindran, and Ahmed A Moustafa. An extended reinforcement learning model of basal ganglia to understand the contributions of serotonin and dopamine in risk-based decision making, reward prediction, and punishment learning. *Frontiers in computational neuroscience*, 8:47, 2014.

[27] Shankar J Chinta and Julie K Andersen. Dopaminergic neurons. *The international journal of biochemistry & cell biology*, 37(5):942–946, 2005.

[28] P Read Montague, Peter Dayan, and Terrence J Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, 16(5):1936–1947, 1996.

[29] Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[30] William R Stauffer, Armin Lak, and Wolfram Schultz. Dopamine reward prediction error responses reflect marginal utility. *Current biology*, 24(21):2491–2500, 2014.

[31] Wolfram Schultz and Ranulfo Romo. Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of neurophysiology*, 63(3):607–624, 1990.

[32] Wolfram Schultz, Paul Apicella, and Tomas Ljungberg. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of neuroscience*, 13(3):900–913, 1993.

[33] Richard S Sutton and Andrew G Barto. Toward a modern theory of adaptive networks: expectation and prediction. *Psychological review*, 88(2):135, 1981.

[34] Helen M Nasser, Donna J Calu, Geoffrey Schoenbaum, and Melissa J Sharpe. The dopamine prediction error: contributions to associative models of reward learning. *Frontiers in psychology*, 8:244, 2017.

[35] Isabel García-García, Yashar Zeighami, and Alain Dagher. Reward prediction errors in drug addiction and parkinson's disease: from neurophysiology to neuroimaging. *Current neurology and neuroscience reports*, 17(6):46, 2017.

[36] Michael J Frank. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive

deficits in medicated and nonmedicated parkinsonism. *Journal of cognitive neuroscience*, 17(1):51–72, 2005.

[37] Yael Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, jun 2009.

[38] Yuji Takahashi. Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. *Frontiers in Neuroscience*, 2(1):86–99, jul 2008.

[39] Daphna Joel, Yael Niv, and Eytan Ruppin. Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, 15(4-6):535–547, jun 2002.

[40] Robert A McCutcheon, Anissa Abi-Dargham, and Oliver D Howes. Schizophrenia, dopamine and the striatum: from biology to symptoms. *Trends in neurosciences*, 42(3):205–220, 2019.

[41] Roy A Wise and Mykel A Robble. Dopamine and addiction. *Annual Review of Psychology*, 71:79–106, 2020.

[42] Guillaume Drui, Sébastien Carnicella, Carole Carcenac, Matthieu Favier, Anne Bertrand, Sabrina Boulet, and Marc Savasta. Loss of dopaminergic nigrostriatal neurons accounts for the motivational and affective deficits in parkinson's disease. *Molecular psychiatry*, 19(3):358–367, 2014.

[43] Lauren B Alloy and Anthony H Ahrens. Depression and pessimism for the future: biased use of statistically relevant information in predictions for self versus others. *Journal of personality and social psychology*, 52(2):366, 1987.

[44] Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3, 2014.

[45] Mehdi Keramati, Serge H Ahmed, and Boris S Gutkin. Misdeed of the need: towards computational accounts of transition to addiction. *Current opinion in neurobiology*, 46:142–153, 2017.

[46] A. David Redish. Addiction as a computational process gone awry. *Science*, 306(5703):1944–1947, 12 2004.

[47] Wolfram Schultz. Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 80(1):1–27, 1998.

[48] George F Koob and Nora D Volkow. Neurobiology of addiction: a neurocircuitry analysis. *The Lancet Psychiatry*, 3(8):760–773, 2016.

[49] George F. Koob and Michel Le Moal. Drug abuse: Hedonic homeostatic dysregulation. *Science*, 278(5335):52–58, oct 1997.

[50] Laura Jean Bierut, Stephen H. Dinwiddie, Henri Begleiter, Raymond R. Crowe, Victor Hesselbrock, John I. Nurnberger, Bernice Porjesz, Marc A. Schuckit, and Theodore Reich. Familial transmission of substance dependence: Alcohol, marijuana, cocaine, and habitual smoking: A report from the collaborative study on the genetics of alcoholism. *Archives of General Psychiatry*, 55(11):982–988, nov 1998.

[51] Martin Zack, Ross St. George, and Luke Clark. Dopaminergic signaling of uncertainty and the aetiology of gambling addiction. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 99(December 2019):109853, 2020.

[52] Mark W Howe, Patrick L Tierney, Stefan G Sandberg, Paul E M Phillips, and Ann M Graybiel. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature*, 500(7464):575–579, 2013.

[53] Christopher D. Fiorillo, Philippe N. Tobler, and Wolfram Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902, 2003.

[54] Candice Biback and Martin Zack. The Relationship Between Stress and Motivation in Pathological Gambling: a Focused Review and Analysis. *Current Addiction Reports*, 2(3):230–239, 2015.

[55] Anna Render and Petra Jansen. Dopamine and sense of agency: Determinants in personality and substance use. *PLOS ONE*, 14(3):1–19, 03 2019.

[56] W. Spencer Murch and Luke Clark. Games in the Brain: Neural Substrates of Gambling Addiction. *The Neuroscientist*, 22(5):534–545, 2016.

[57] Juliette Tobias-Webb, Eve H. Limbrick-Oldfield, Claire M. Gillan, James W. Moore, Michael R. F. Aitken, and Luke Clark. Let me take the wheel: Illusory control and sense of agency. *Quarterly Journal of Experimental Psychology*, 70(8):1732–1746, 2017.

[58] Florent Wyckmans, A. Ross Otto, Miriam Sebold, Nathaniel Daw, Antoine Bechara, Mélanie Saeremans, Charles Kornreich, Armand Chatard, Nemat Jaafari, and Xavier Noël. Reduced model-based decision-making in gambling disorder. *Scientific Reports*, 9(1):1–10, 2019.

[59] A David Redish, Steve Jensen, Adam Johnson, and Zeb Kurth-Nelson. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological Review*, 114(3):784–805, 2007.

[60] A. Wiehler, K. Chakroun, and J. Peters. Attenuated directed exploration during reinforcement learning in gambling disorder. *bioRxiv*, page 823583, 2019.

[61] William Dauer and Serge Przedborski. Parkinson's disease: mechanisms and models. *Neuron*, 39(6):889–909, 2003.

[62] Dennis W Dickson. Neuropathology of parkinson disease. *Parkinsonism & related disorders*, 46:S30–S33, 2018.

[63] Joshua M Shulman, Philip L De Jager, and Mel B Feany. Parkinson's disease: genetics and pathogenesis. *Annual Review of Pathology: Mechanisms of Disease*, 6:193–222, 2011.

[64] Michael J Frank, Lauren C Seeberger, and Randall C O'reilly. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, 306(5703):1940–1943, 2004.

[65] Daniel Weintraub. Impulse control disorders in parkinson's disease: a 20-year odyssey. *Movement Disorders*, 34(4):447–452, 2019.

[66] Alain Dagher and Trevor W Robbins. Personality, addiction, dopamine: insights from parkinson's disease. *Neuron*, 61(4):502–510, 2009.

[67] Payam Piray, Yashar Zeighami, Fariba Bahrami, Abeer M Eissa, Doaa H Hewedi, and Ahmed A Moustafa. Impulse

control disorders in parkinson's disease are associated with dysfunction in stimulus valuation but not action valuation. *Journal of Neuroscience*, 34(23):7814–7824, 2014.

[68] Valerie Voon, Mathias Pessiglione, Christina Brezing, Cecile Gallea, Hubert H Fernandez, Raymond J Dolan, and Mark Hallett. Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. *Neuron*, 65(1):135–142, 2010.

[69] Paul W Glimcher. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement 3):15647–15654, 2011.

[70] Matteo Colombo. Deep and beautiful. the reward prediction error hypothesis of dopamine. *Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences*, 45:57–67, 2014.

[71] Neir Eshel, Michael Bukwich, Vinod Rao, Vivian Hemmelder, Ju Tian, and Naoshige Uchida. Arithmetic and local circuitry underlying dopamine prediction errors. *Nature*, 525(7568):243–246, 2015.

[72] Christopher D Fiorillo, Philippe N Tobler, and Wolfram Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902, 2003.

[73] Masayuki Matsumoto and Okihide Hikosaka. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248):837–841, 2009.

[74] Peter Dayan. Twenty-five lessons from computational neuromodulation. *Neuron*, 76(1):240–256, 2012.

[75] Daniela Maria Vogt Weisenhorn, Florian Giesert, and Wolfgang Wurst. Diversity matters–heterogeneity of dopaminergic neurons in the ventral mesencephalon and its relation to parkinson's disease. *Journal of neurochemistry*, 139:8–26, 2016.

[76] Baihan Lin, Djallel Bouneffouf, and Guillermo Cecchi. Split q learning: reinforcement learning with two-stream rewards. *arXiv preprint arXiv:1906.12350*, 2019.

[77] Baihan Lin, Guillermo A Cecchi, Djallel Bouneffouf, Jenna Reinen, and Irina Rish. A story of two streams: Reinforcement learning models from human behavior and neuropsychiatry. In *AAMAS*, pages 744–752, 2020.

[78] Djallel Bouneffouf, Irina Rish, and Guillermo A Cecchi. Bandit models of human behavior: Reward processing in mental disorders. In *International Conference on Artificial General Intelligence*, pages 237–248. Springer, 2017.

[79] Antoine Bechara, Hanna Damasio, Daniel Tranel, and Antonio R Damasio. Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304):1293–1295, 1997.

[80] Antoine Bechara, Antonio R Damasio, and Hanna Damasio. Insensitivity to future consequences following damage to human prefrontal. *The Science of Mental Health: Personality and personality disorder*, 50:287, 2001.

[81] Helen Steingroever, Daniel J Fridberg, Annette Horstmann, Kimberly L Kjome, Veena Kumari, Scott D Lane, Tiago V Maia, James L McClelland, Thorsten Pachur, Preethi Premkumar, et al. Data from 617 healthy participants performing the iowa gambling task: A "many labs" collaboration. *Journal of Open Psychology Data*, 3(1):340–353, 2015.

[82] Karen D. Ersche, Claire M. Gillan, P. Simon Jones, Guy B. Williams, Laetitia H.E. Ward, Maartje Luijten, Sanne De Wit, Barbara J. Sahakian, Edward T. Bullmore, and Trevor W. Robbins. Carrots and sticks fail to change behavior in cocaine addiction. *Science*, 352(6292):1468–1471, jun 2016.

[83] Aurelijus Burokas, Javier Gutiérrez-Cuesta, Elena Martín-García, and Rafael Maldonado. Operant model of frustrated expected reward in mice. *Addiction biology*, 17(4):770–782, 2012.

[84] Birgit Abler, Henrik Walter, and Susanne Erk. Neural correlates of frustration. *Neuroreport*, 16(7):669–672, 2005.

[85] Shelly B Flagel, Huda Akil, and Terry E Robinson. Individual differences in the attribution of incentive salience to reward-related cues: Implications for addiction. *Neuropharmacology*, 56:139–148, 2009.

[86] Ju Tian and Naoshige Uchida. Habenula lesions reveal that multiple mechanisms underlie dopamine prediction errors. *Neuron*, 87(6):1304–1316, 2015.

[87] Jessica A Cooper, Amanda R Arulpragasam, and Michael T Treadway. Anhedonia in depression: biological mechanisms and computational models. *Current opinion in behavioral sciences*, 22:128–135, 2018.

[88] Chong Chen, Taiki Takahashi, Shin Nakagawa, Takeshi Inoue, and Ichiro Kusumi. Reinforcement learning in depression: a review of computational research. *Neuroscience & Biobehavioral Reviews*, 55:247–267, 2015.

[89] Roee Admon and Diego A Pizzagalli. Dysfunctional reward processing in depression. *Current Opinion in Psychology*, 4:114–118, 2015.

[90] Lou Safra, Coralie Chevallier, and Stefano Palminteri. Depressive symptoms are associated with blunted reward learning in social contexts. *PLoS computational biology*, 15(7):e1007224, 2019.

[91] Joana V Taylor Tavares, Luke Clark, Maura L Furey, Guy B Williams, Barbara J Sahakian, and Wayne C Drevets. Neural basis of abnormal response to negative feedback in unmedicated mood disorders. *Neuroimage*, 42(3):1118–1126, 2008.

[92] Luke Clark, Bruno Averbeck, Doris Payer, Guillaume Sescousse, Catharine A. Winstanley, and Gui Xue. Pathological choice: The neuroscience of gambling and gambling addiction. *Journal of Neuroscience*, 33(45):17617–17623, 2013.

[93] Steve Sharman, Michael Aitken, and Luke Clark. Dual effects of 'losses disguised as wins' and near-misses in a slot machine game. *International Gambling Studies*, 15:1–12, 03 2015.

[94] Karl J Friston, Tamara Shiner, Thomas FitzGerald, Joseph M Galea, Rick Adams, Harriet Brown, Raymond J Dolan, Rosalyn Moran, Klaas Enno Stephan, and Sven Bestmann. Dopamine, affordance and active inference. *PLoS Computational Biology*, 8(1):e1002327, 2012.

Appendices

A Exploring Expectiles

A.1 Efficiently computing sample expectiles

To our knowledge, an efficient method of computing the expectiles of a sample is not available within standard software libraries. Here we derive a method for computing the expectiles of a sample without resorting to standard iterative optimisation routines, based on the observation that the sample expectile optimisation target has a piece-wise-linear continuous gradient for which the root may be easily computed[12].

Given a sample $\vec{x} = x_1, \dots, x_N$ and an asymmetry parameter $\tau \in (0, 1)$, define the *τ-sample-expectile* $\epsilon_{\vec{x}}(\tau)$:[4]

$$\epsilon_{\vec{x}}(\tau) = \arg\min_{\epsilon} \frac{1}{N} \sum_{i=0}^{N} [\![x_i > \epsilon]\!]_{1-\tau}^{\tau} (x_i - \epsilon)^2. \tag{7}$$

Differentiating the optimisation target of (7) with respect to ϵ yields a necessary condition for $\epsilon = \epsilon_{\vec{x}}(\tau)$:

$$0 = \frac{-2}{N} \sum_{i=0}^{N} [\![x_i > \epsilon]\!]_{1-\tau}^{\tau} (x_i - \epsilon) \qquad \text{or, equivalently,}$$

$$0 = (1-\tau) \sum_{i:x_i \le \epsilon} \frac{1}{N}(x_i - \epsilon) + \tau \sum_{i:x_i > \epsilon} \frac{1}{N}(x_i - \epsilon). \tag{8}$$

Define the *sample cumulative distribution function* $F_{\vec{x}}(\epsilon)$ and the *sample partial moment function* $M_{\vec{x}}(\epsilon)$

$$F_{\vec{x}}(\epsilon) = \sum_{i:x_i \le \epsilon} \frac{1}{N} \qquad M_{\vec{x}}(\epsilon) = \sum_{i:x_i \le \epsilon} \frac{1}{N} x_i \tag{9}$$

and note their ‘complement’ identities

$$1 - F_{\vec{x}}(\epsilon) = \sum_{i:x_i > \epsilon} \frac{1}{N} \tag{10}$$

$$\mu_{\vec{x}} - M_{\vec{x}}(\epsilon) = \sum_{i:x_i > \epsilon} \frac{1}{N} x_i \tag{11}$$

where $\mu_{\vec{x}} = \sum_{i=1}^{N} \frac{1}{N} x_i = M_{\vec{x}}(\max_i x_i)$ is the sample mean.

With these functions, we may restate the right-hand side of (8) as a function, $G_{\vec{x}}(\epsilon)$, for which we seek a root:

$$\begin{aligned} G_{\vec{x}}(\epsilon) = &-((1-\tau)F_{\vec{x}}(\epsilon) + \tau(1 - F_{\vec{x}}(\epsilon))) \cdot \epsilon \\ &+ (1-\tau)M_{\vec{x}}(\epsilon) + \tau(\mu_{\vec{x}} - M_{\vec{x}}(\epsilon)) \end{aligned} \tag{12}$$

Note that $F_{\vec{x}}(\epsilon)$ and $M_{\vec{x}}(\epsilon)$ are piece-wise constant (with a discontinuity at each x_i). Therefore, $G_{\vec{x}}(\epsilon)$ is piece-wise linear. Each piece has a negative slope. Moreover, $G_{\vec{x}}(\epsilon)$ is continuous[13]. This leads to the following algorithm for finding the root of $G_{\vec{x}}(\epsilon)$ and thus computing the τ-sample-expectile of $\vec{x}$.

Algorithm 3 (sample expectile)**.** Given asymmetry parameter $\tau \in (0, 1)$ and sample $\vec{x} = x_1, \dots, x_N$:

1. Sort $\vec{x}$. (Below, assume that $x_1 \le \dots \le x_N$.)
2. Compute $F_{\vec{x}}(x_i)$ and $M_{\vec{x}}(x_i)$ for $i = 1, \dots, N$. Since $\vec{x}$ is sorted, these take particularly simple forms:

   $$F_i \leftarrow F_{\vec{x}}(x_i) = \frac{i}{N}, \quad M_i \leftarrow M_{\vec{x}}(x_i) = \sum_{j=1}^{i} \frac{x_j}{N}.$$

   Note that $M_N = \mu_{\vec{x}}$ and $F_N = 1$.
3. Compute $G_{\vec{x}}(x_i)$ for $i = 1, \dots, N$:

   $$\begin{aligned} G_i \leftarrow &-((1-\tau)F_i + \tau(F_N - F_i)) \cdot x_i \\ &+ (1-\tau)M_i + \tau(M_N - M_i). \end{aligned}$$

4. The i^{th} segment of $G_{\vec{x}}(\epsilon)$ runs from (x_i, G_i) down to (x_{i+1}, G_{i+1}). Find the i with $G_i \ge 0 > G_{i+1}$ and compute this segment's root:

   $$\frac{(1-\tau)M_i + \tau(M_N - M_i)}{(1-\tau)F_i + \tau(F_N - F_i)}.$$

   This is both $G_{\vec{x}}(\epsilon)$’s root and the τ-sample-expectile.

Finally, note that omitting the $\frac{1}{N}$ factor in the computation of F_i and M_i may assist with numerical stability. It’s for this reason that we use F_N to replace 1 in steps 3 and 4.

A.2 Estimating expectiles online

For the purposes of exploring expectiles further, we demonstrate a method for learning expectile estimates *online* (that is, maintaining an estimate given a sequence of sample points *one at a time*). This method resembles the dynamics of an expectile’s estimation during an imputation-based RL algorithm, and provides a simplified context in which to understand the hypothetical role of dopamine neurons under the *distributional hypothesis* of the brain’s learning system.

Given an asymmetry parameter $\tau \in (0, 1)$ and an i.i.d. sequence of T sample points $\vec{x} = x_1, \dots, x_T$, consider estimating the τ-expectile, $\epsilon_{\vec{x}}(\tau)$, of $\vec{x}$’s underlying distribution: In response to each sample point x_i, update the estimate, $\hat{\epsilon}_{\vec{x}}(\tau)$, according to equation (13) where α is a positive *learning rate*.[3,4]

$$\hat{\epsilon}_{\vec{x}}(\tau) \overset{+}{\leftarrow} \alpha [\![x_i > \hat{\epsilon}_{\vec{x}}(\tau)]\!]_{1-\tau}^{\tau} (x_i - \hat{\epsilon}_{\vec{x}}(\tau)) \tag{13}$$

As is characteristic of expectiles statistics, this update features an asymmetric response to positive and negative differences. With $\hat{\epsilon}_{\vec{x}}(\tau)$ understood as a *prediction* of values drawn from $\vec{x}$’s distribution, we can understand the difference $x_i - \hat{\epsilon}_{\vec{x}}(\tau)$ as a *prediction error*. Equation (13) moves $\hat{\epsilon}_{\vec{x}}(\tau)$ towards x_i in ‘proportion’ to this prediction error, except with different proportionality constants for positive and negative errors. Refer to these two proportionality constants as *positive-error* and *negative-error learning rates*—for fixed asymmetry parameter τ and *base learning rate* α, they act like constant multipliers (denoted α_+ and α_-) as per equation (14):

$$\alpha_+ = \alpha\tau \qquad \alpha_- = \alpha(1-\tau) \tag{14}$$

Indeed, any pair of constants $\alpha_+, \alpha_- > 0$ determines an update scheme for *some* asymmetry parameter τ and base learning rate

[12] We also provide a Python/NumPy implementation, available at https://github.com/matomatical/expectiles.

[13] This is despite discontinuities in $F_{\vec{x}}(\epsilon)$ and $M_{\vec{x}}(\epsilon)$. One may verify that $\lim_{\epsilon \uparrow x_i} G_{\vec{x}}(\epsilon) = G_{\vec{x}}(x_i)$ for $i = 2, \dots, N$ because $x_i F_{\vec{x}}(x_i) - x_i F_{\vec{x}}(x_{i-1}) = x_i/N = M_{\vec{x}}(x_i) - M_{\vec{x}}(x_{i-1})$. We omit a full proof.

α. The relationship between τ, α, and the asymmetric learning rates α_+ and α_- is equation (15):

$$\alpha = \alpha_+ + \alpha_- \qquad \tau = \frac{\alpha_+}{\alpha_+ + \alpha_-} \tag{15}$$

Note that it is with this relationship that [4] interprets the asymmetric slopes of dopaminergic neuron response curves as encoding an asymmetry parameter for each neuron.

The asymmetry of update equation (13) causes the estimate to 'balance' at the τ-expectile after successive updates (figure 8). This convergence arises because the update scheme is an instance of *stochastic gradient descent*. That is, it corresponds to updating our expectile estimate according to a loss function's estimated gradient[14]. The loss function is the optimisation target defining the τ-expectile of a distribution, the so-called *expectile regression loss* $\mathrm{L_{ER}}(\epsilon; X, \tau)$, given by equation (16).

$$\mathrm{L_{ER}}(\epsilon; X, \tau) = \mathbb{E}\left[[\![X > \epsilon]\!]_{1-\tau}^{\tau} (X - \epsilon)^2 \right] \tag{16}$$

Stochastic gradient descent steers estimates towards minima of its loss function. In this case, the loss function defines the expectiles of the distribution. This link underpins the convergence of update equation (13) to the true τ-expectile of the distribution from which each x_i is drawn.

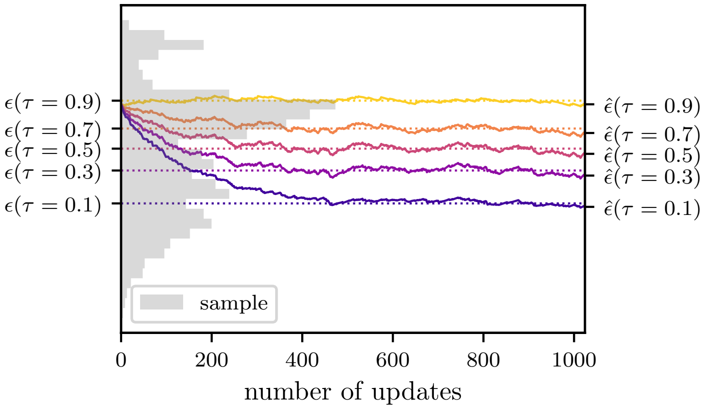

Figure 8: Update scheme (13) for several τ values, on a sample of 1,000 points from a toy distribution (gray histogram). Dotted lines show sample expectiles. Solid lines trace expectile estimates. The estimates 'balance' where the asymmetrically-scaled positive and negative prediction errors cancel in expectation—at the expectile.

The *distributional hypothesis* posits that the phasic activity of midbrain dopaminergic neurons signals the asymmetrically-scaled prediction errors of equation (13). Accordingly, each neuron is imbued with its own degree of positive outlook (a τ-value, determining the neurons response slopes α_+ and α_-). Moreover, each neuron plays the role of comparing a reward prediction (akin to a τ-expectile estimate based on a stimulus) to a sample of the discounted future reward (akin to x_i).

A.3 On scaling asymmetric learning rates in mice

When learning a suite of expectile estimates at varying asymmetry parameters, it's desirable to control the speed and stability of each estimate's convergence. The base learning rate α in equation (13) plays an important role in stabilising the learning process. In order to achieve efficient and stable updates for the entire suite of estimates, we consider the possibility of varying the chosen α with the estimate's asymmetry parameter. Here we explore this idea, and a potential connection to the neural recordings analysed in [4].

We may use the same base learning rate α for each estimate. Using a constant α for a range of asymmetry parameters corresponds to holding the arithmetic mean of the asymmetric learning rates α_+ and α_- fixed with τ. However, an unweighted arithmetic average may not be appropriate, since highly asymmetric expectiles will produce more error in one direction than the other. For example estimating the 0.9-expectile entails producing many and/or higher negative prediction errors than positive prediction errors, and so α_- will intuitively contribute more to the speed and stability of learning than α_+.

Rather than keeping the arithmetic mean asymmetric learning rate constant, it seems more desirable to maintain a fixed speed and stability. The speed and stability are a function of both α_+ and α_-, as well as the proportion and size of updates utilising each of them. Since asymmetric learning rates contribute multiplicatively to updates, we conjecture that holding the geometric mean of these rates constant will lead to uniformly efficient and stable updates across a wide range of asymmetry parameters. To do so, we may determine the base learning rate as the following function of τ:

$$\alpha_\tau = \frac{\alpha_0}{2\sqrt{\tau(1-\tau)}} \tag{17}$$

where α_0 is an *effective base learning rate* controlling the speed and stability of all estimates' convergence. This learning rate-setting scheme leads to the following asymmetric learning rates, which still satisfy equation (15) but do so with constant geometric mean α_0 for varying τ:

$$\alpha_+ = \frac{\alpha_0}{2}\sqrt{\frac{\tau}{1-\tau}} \qquad \alpha_- = \frac{\alpha_0}{2}\sqrt{\frac{1-\tau}{\tau}}. \tag{18}$$

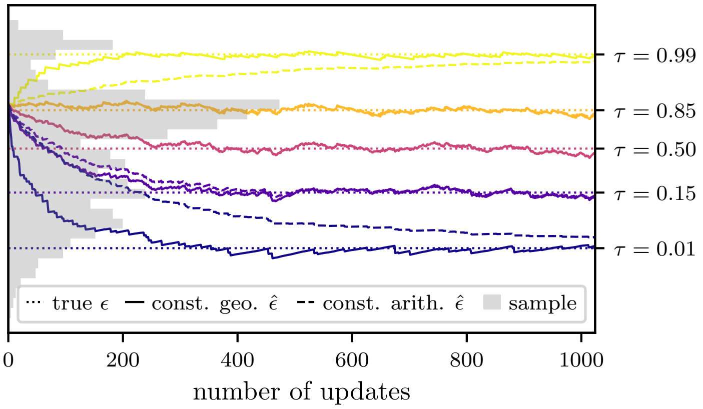

Figure 9: Online expectile estimation. Solid trajectories: Constant geometric mean of α_+ and α_- by equation (18). Dashed: Constant α.

Figure 9 contrasts the schemes. Note that for $\tau = 0.5$, $\alpha_\tau = \alpha_0$. Therefore, the updates coincide for the 0.5-expectile. While the convergence speed for highly asymmetric expectiles appears dampened under the constant learning rate scheme, all estimates behave qualitatively similarly in terms of their convergence speed and stability under equation (18).

[14] In the case of distributional RL, the 'sample point' x_i may have a non-zero gradient with respect to our estimate, and so we have rather an instance of *stochastic semi-gradient descent*, cf. [1].

Finally, we compare the schemes against neural data. In their analysis of neural recording data prepared in [71], [4] estimate asymmetry parameters and asymmetric learning rates for 40 dopaminergic neurons across two mice. The resulting asymmetric learning rates are plotted against their corresponding asymmetry parameters in figure 10.

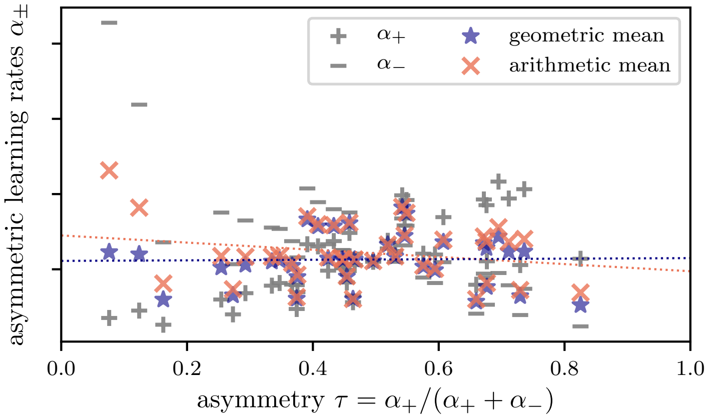

Figure 10: Each vertical quadruple $(+, \star, \times, -)$ represents one neuron. Horizontal position: Estimated asymmetry. Gray: Estimated asymmetric learning rates. Coloured: Geometric and arithmetic means of the neuron's learning rates. Estimates from [4], one neuron excluded.

We note that both the geometric means and the arithmetic means of the estimated rates seem roughly constant across the range of asymmetries. Thus, both rate-setting schemes seem appear consistent with the neural data. Further data and analysis is necessary before a definitive conclusion can be drawn about the asymmetric-learning-rate-setting scheme hypothetically used by these neurons. In particular, the schemes differ most in their predictions for the asymmetric learning rates of highly asymmetric neurons. Moreover, it will be important to assess whether any preprocessing steps performed by [4], including measurement of asymmetric learning rates in an estimated utility space, influence this analysis.

A.4 Towards an efficient expectile imputation strategy

In principle, algorithm 2 represents a compelling alternative to optimisation-based imputation for its efficiency and simplicity, but in practice, its approximation suffers due to the need extrapolate beyond outlying coordinates, and the method shows a lack of robustness to non-monotonicity in its input. We discuss some advantages and limitations of this strategy.

Our implementation of step 4, in particular extrapolation, introduces visible biases into imputed samples. Observe, for example, the small outlying peaks in figure 5g, which become large under the optimistic asymmetry distribution (5h, left purple peak) due to the significant reliance on extrapolation. In fact, this bias seems to drive all effects observed in our experiments. While optimisation-based imputation is less susceptible to this extrapolation problem, the method faces the same fundamentally under-determined problem of recovering a distribution from an insufficient statistical summary, and so must inject its own assumptions. Limiting the sample to K points ensures solution uniqueness but may underpin the distributional divergence of this method demonstrated in our fidelity experiments.

Furthermore, the method is not robust to noisy, approximate, or otherwise non-monotonic input coordinates. Algorithm 2 requires that its input coordinates follow a monotonically increasing function. This may not always hold. In tabular learning algorithms, distributional updates seem to preserve this monotonicity. However, in the context of function approximation and in the brain it may be possible for expectile prediction curves to *cross*—a known problem in the related context of *expectile regression* [17]. Moreover, in the analysis of neural data involving noisy measurements, even a truly increasing expectile function may not appear so. In contrast, [4] applied optimisation-based imputation to this kind neural data with impressive results our strategy is unable to replicate.

Despite these fidelity and robustness issues, our alternative strategy is more simple and computationally efficient than the optimisation-based strategy. In our experiments, we were easily able to train using orders of magnitude more expectiles and sample points in an order of magnitude less time. Since [3] has linked increasing the number of expectiles learned per stimulus to increased performance gains, we consider this a significant practical advantage of the approach. Moreover, the approach achieves these efficiency gains through the reliance on primarily *local* computation: Each estimated coordinate of the CDF is derived from a small number of nearby input coordinates, regardless of the number of expectiles learned, leading to a simple *parallelisable* algorithm. Intuitively, this style of computation seems more biologically plausible than global synchronous root-finding.

Future work may lead to an approach based on theorem 1 overcoming some of these limitations. We explored alternative interpolation and extrapolation techniques including fitting a polynomial to an inverse logistic transformation of $\tau_X(\epsilon)$. This method displayed comparable fidelity, but clearly introduces its own assumptions, and harmed the efficiency and locality of the computation. It may be possible to find a more principled way of working with the finite number of coordinates. For example, Bayesian inference may help explicitly handle assumptions. Finally, an 'implicit expectile network' representation of the expectile function, along the lines of work in quantile-based distributional RL [22], may lead to an improved biologically plausible imputation strategy if combined with theorem 1.

B Alternative model of gambling disorder

The experiments above demonstrates how the incomplete τ range affects the behavior acquired after the learning process involves expectile imputation and sampling. However, this may not be the only plausible explanation for gambling addiction. We have also examined the traditional actor-critic model without explicit imputation, and found that an unbalanced τ range can result in biased strategy as well.

B.1 Model

In the traditional actor-critic model, the actor is updated by the observation s, the action α chosen and the temporal difference error of the critic ϵ. The objective function is

$$-\log(\Pr_s(a)) \cdot \epsilon \tag{19}$$

which is maximized when the probability of picking the action whose real reward is much higher than expected is low.

For the distributional actor-critic model, we rewrite the function as follows

$$-\log(\Pr_s(a)) \cdot \frac{1}{k} \sum_{m=1}^{k} w_{\tau_m} \epsilon_{\tau_m} \tag{20}$$

$$w_\tau = 1\{\epsilon_\tau \geq 0\} \cdot \tau + 1\{\epsilon_\tau < 0\} \cdot (1 - \tau) \tag{21}$$

where w_τ is the τ-weight, k is the number of τ value, τ_m is the m_{th} τ-value and ϵ_{τ_m} is the temporal difference error corresponds to τ_m at the critic.

When the τ-range is symmetric along $\tau = 0.5$, we call the range unbiased. Thus a non-distributional critic is a model with one atom and unbiased τ-range. From our experience, symmetric τ-range always results in same strategy as non-distributional agent with proper tuning after convergence, even there may be more fluctuation during the process of training when the atoms are sparse but more than one.

B.2 Experiments

In the first experiment, we simulate a gambling machine that generates the reward from a mixture of gaussian distribution. The mean value of the reward is controlled around -2. The player is offered three choices at each round: 'no bet', 'small bet' and 'large bet', which indicate that the returned reward or penalty will be multiplied by a factor 0, 1 or 2. In this sense, a tendency towards 'large bet' could be interpreted as a more risky strategy that holds an optimistic perspective on the outcome, while 'no bet' implies conservativeness and pessimism. Since the true mean value of the outcome is negative, we expect that the agent should adopt the conservative strategy after properly trained.

Here we assume the game is as simple as 1draw-then-check', where the player decides his or her next move, check the result and go on next round. So the only available observation that the player may refer to is the history of win or loss. Therefore, the observation is encoded in three states: winning (more win than loss in the memory window), losing (more loss than win in the memory window) and neutral (no significant difference between the chance of wining and losing in the past). Since the reward is generated out of complete randomness, the current state in fact has little impact on the reward or the next state.

As shown in fig. 11, the strategy the agent adopts after stabilized varies across different τ-ranges for distribution model a and b, while remains unchanged for model c. Though the expected returns under distribution model a, b, c are the same, their preferred strategies are qualitatively different. This simple experiment shows how the distribution itself might affect the behavior.

The result also coincides with our knowledge about gambling. Among the three models, model b is the most similar to the outcome of gambling game, where we witness a higher chance of loss and also fairly notable chance of wins despite the overall monetary reward is negative; as a result, its resulting behavior under a optimistic τ-range is most identical to 'addiction'. Model c is least like the case of gambling, where a single peak around losing money definitely would not appeal to the players; in such case, the shift of τ-range cannot twist the action preference. The above suggests that the biased strategy may be the consequence of both the biased τ-range and a specific distribution.

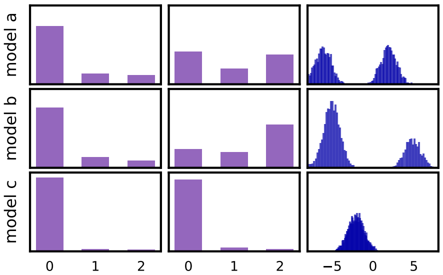

Figure 11: The result of testing the agent with 21 atoms after learning under different τ ranges and distributions for 3000 steps. The left column shows the strategies learned under the unbiased range $\tau = (0.4, 0.6)$. The middle column shows the strategies learned under $\tau = (0.8, 1.0)$. The right column shows the distributions of reward generation. All three distributions have the expected reward of -2.

Studies suggest that the addiction could be attributed to its other characteristics related to cognitive distortions, including 'near miss' or 'losses disguised as wins' [56, 92, 93]. So we design the second experiment where we take the visual effect and the subjective reward into consideration. The prototype model is the Electronic Gaming Machine (EGM) [93, 94], where after the player places the bet, the reel starts rolling. The player only gets rewarded when all slots display the winning icon on the payline when they stop spinning. We define the states and features as follows. With the setting as table. 3, we formulate the reward at each round as:

$$W = \prod_{k=1}^{n_{\text{slot}}} 1\{s_k = s_w\} \tag{22}$$

$$\text{reward} = W \cdot r_m + (1 - W) \cdot p_m + r_s \sum 1\{s_k = s_w\} \tag{23}$$

Where we introduce the idea of subjective reward, the sense of satisfaction that the player gains under visual stimulation even without monetary reward will contribute to the player's enjoyment and willingness to lay higher bets.

n_{slot}	number of slots
size_w	size of the mem-window
r_m	monetary reward
p_m	monetary penalty
r_s	subjective reward for each wining icon
p_r	the chance each slot stops at wining icon
s_k	the icon at k_{th} slot
s_w	the winning icon

Table 3: EGM Setting

w_k	if $s_k = s_w$
n_w	count of winning icons
R	monetary reward obtained
h_w	history of wins

Table 4: EGM Features

Using the features listed in table 4 as observations further generalize the traditional tabular actor-critic model. See fig.12 for the result. The agent always manages to learn to adopt the conservative strategy when the expected reward is negative, but its preference only twists under some settings. It is obvious that the subjective reward will disguise the negative monetary reward as positive and thus change the preferred strategy. But what is interesting is that when the expected reward is controlled, the subjective reward is still able to impair the ability of learning with the joint effect of biased τ-range by altering the distribution of the apparent reward.

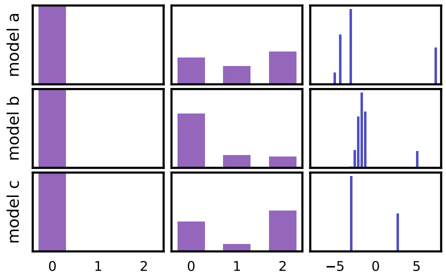

Figure 12: The result of testing the agent with 21 atoms after learning under different τ-ranges and distributions for 3000 steps. The left column shows the strategies learned under unbiased $\tau = (0.4, 0.6)$. The middle column shows the strategies learned under $\tau = (0.8, 1.0)$. The right column shows the distributions of reward generated by the EGM with different parameter settings. All three distributions has the expected reward of -1. The settings of the EGM to yield the distribution models a, b and c are available in table 5

Note that when the critic passes the loss to the actor without the τ-weight, that is, we rewrite the formula (20) as

$$-\log(\Pr_s(a)) \cdot \frac{1}{k} \sum_{m=1}^{k} \epsilon_{\tau_m}.$$

The strategy preference shift is no longer observable, as shown in fig. 13. Besides, there is no qualitative discrepancy between the strategy acquired with a non-distributional and distributional critic. In fact, compared to batch normalization, increasing the number of atoms helps little to no significant effect on smoothing out the errors or speeding up training in our cases.

	a	b	c
n_{slot}	3	5	5
size_w	3	3	3
r_m	5	3	3
p_m	-5	-3	-3
r_s	1	0.5	0
p_r	0.6	0.6	0.8

Table 5: EGM Features

Therefore, we may conclude that the biased τ-weight is the only direct caused of the biased strategy.

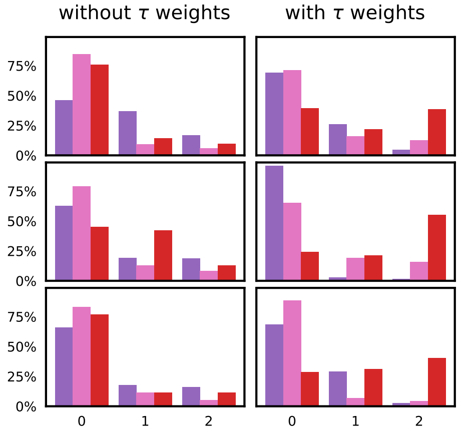

Figure 13: Compare the result strategy with and without τ weight. Top: trained with one atom (non-distributional); Middle: trained with 21 atoms; Bottom: trained with 41 atoms. The purple, pink and red bars are for τ-range [0, 0.2], τ-range [0.4, 0.6] and the τ-range [0.8, 1.0], respectively.

B.3 Discussion

When the τ-range is unbiased, we find no significant difference between the strategy adopted by the non-distributional and distributional learner under all gaming machine settings. The only factor that proved to distort the learned preference is the unbalanced learning rate for samples above and below the expectation.

The effect of illusion of control, near-miss or loss-disguised-as-wins should be much more complicated than our model. Though it is assumed that the player may wrongly take visual or sound effects as the sign of reward, it remains unclear to us when and how the conditioning of visual stimulus gets involved [51, 94]. Other conceivable hypothesis related to reward proximity or stress neurocircuitry are not captured by our model.

Other behavioral studies of animals demonstrate the phenomena of 'sign-tracking', the gambling disorder is viewed as a motivated chasing of the stimulus regardless of the reward [51]. However, we did not succeed in linking biased preference to distributional learning with the sensitization and evaluation separated.