Temporal-difference reinforcement learning with distributed representations

Zeb Kurth-Nelson, A. David Redish*,

Department of Neuroscience, University of Minnesota, Minneapolis MN 55455 USA.

* E-mail: redish@umn.edu

IN PRESS

PLoS ONE

Accepted 4 September 2009.

Abstract

Temporal-difference (TD) algorithms have been proposed as models of reinforcement learning (RL). We examine two issues of distributed representation in these TD algorithms: distributed representations of belief and distributed discounting factors. Distributed representation of belief allows the believed state of the world to distribute across sets of *equivalent states*. Distributed exponential discounting factors produce hyperbolic discounting in the behavior of the agent itself. We examine these issues in the context of a TD RL model in which state-belief is distributed over a set of exponentially-discounting "micro-Agents", each of which has a separate discounting factor (γ). Each μAgent maintains an independent hypothesis about the state of the world, and a separate value-estimate of taking actions within that hypothesized state. The overall agent thus instantiates a flexible representation of an evolving world-state. As with other TD models, the value-error (δ) signal within the model matches dopamine signals recorded from animals in standard conditioning reward-paradigms. The distributed representation of belief provides an explanation for the decrease in dopamine at the conditioned stimulus seen in overtrained animals, for the differences between trace and delay conditioning, and for transient bursts of dopamine seen at movement initiation. Because each μAgent also includes its own exponential discounting factor, the overall agent shows hyperbolic discounting, consistent with behavioral experiments.

Introduction

Temporal-difference (TD) learning algorithms have been proposed to model behavioral reinforcement learning (RL) [1–3]. The goal of reinforcement learning is to learn what actions to select in what situations by learning a value function of situations or "states" [4]. (As noted by Daw *et al.* [5], it is not necessarily true that the agent's estimate of the world-state always corresponds to the actual state of the world. We have already explored some of the potential consequences of this mismatch in another paper [6] and will not address it here.) In TD models, the value function is learned through the calculation of a value-prediction error signal (termed δ, [4, 7, 8]), calculated each time the agent changes world-states. δ reflects the difference between the value-estimate and the actual value (including immediate reward) observed on the transition. From δ, the value-estimate of the old state can be updated to approach the observed value. This δ signal appears at unexpected rewards, transfers with learning from rewards to anticipatory cue stimuli, and shifts with changes in anticipated reward [4, 8]. This algorithm is a generalization of the early psychological reward-error models [9, 10]. Components of these models have been proposed to correspond to neurophysiological signals [1, 2, 8, 11–14]. In particular, the firing of midbrain dopaminergic neurons closely matches δ.

TD RL models have been able to provide strong explanations for many neurophysiological observations, such as qualitative changes in dopamine firing [1, 5], including changes at first thought not to reflect prediction error (e.g. generalization and exploration [15]). More recent experiments have shown quantitative matches to the predictions of these models [16–22]. In addition, more recent models have been based on distributed representations of belief within those state-spaces [5, 23–26].

In this paper, we examine the effects of distributed state representation, distributed

value-representation, and distributed discounting rate in TD learning.

- Distributed discounting rates along with distributed value representation lead to hyperbolic discounting, matching the hyperbolic discounting experimentally observed in humans and animals.

- Distributed representations of state-belief allow the agent to divide its believed state across multiple *equivalent states*. This distributed state-representation can account for the slowing of learning rates across intertrial intervals and trace conditioning paradigms, and can account for dopamine signals seen at movement initiation in certain instrumental conditioning paradigms.

These two hypotheses are separable and produce separable predictions, but together they form a coherent and parsimonious description of a multi-micro-agent (μAgent) TD model of reinforcement learning that provides a good fit to the experimental data. We will make clear in the simulations below which components are necessary for which results, and in the discussion which predictions follow from which hypotheses.

This multiple micro-agents model is consistent with anatomical studies suggesting that the basal ganglia consist of separable "loops" that maintain their separation through the basal ganglia pathway [27–29]. The model is also consistent with recent fMRI studies suggesting that the striatum consists of functional "slices" reflecting a range of discounting factors [30, 31].

Methods

It is important to note that the theoretical consequences of distributed representation are independent of many of the methodological details. However, in order to implement

simulations, specific choices have to be made. Throughout the methods section, we will identify which simulation details are theoretically important and which are not.

The simulation comprised two entities: the world and the agent. The world consisted of a semi-Markov state space (M^W) with two additions. First, it provided *observations* and *rewards* to the agent; second, its current state could be changed by an *action* of the agent. The agent consisted of a set of μAgents, each of which contained a model of the world M_i^A, a hypothesis of the state of the world s_i, a value function of those states $V_i(\cdot)$, and an exponential discounting factor γ_i. On each time step, a value-prediction-error δ_i was calculated independently by each μAgent. The overall agent performed actions based on the state beliefs and value functions of the μAgents, and the δ signals of all μAgents could be averaged to represent an overall δ signal. The world and agent were simulated in discrete time-steps. The world provided an observation or null-observation to the agent on each time-step, and the agent provided an action or null-action to the world on each time-step. See Figure 1 and Table 1 for an overview of the model structure.

State-space/process-model

Both the world and the agent contain an internal state-space: M_W and M_A, respectively. In principle it is not necessary that $M_A = M_W$. In fact, it is quite possible for each μAgent to have an individual world-model M_i^A. In the simulations used, all μAgents used an identical state-space model M_A, defined as identical to the world-model M_W.

States corresponded to temporally extended circumstances salient to the agent, such as being located at an arm of a maze or waiting within an interstimulus interval. Transitions defined jumps from one state to another. On entry into a state, a random time was drawn from that state's dwell-time distribution, which determined how long the world would remain within that state before a transition occurred. Observations provided feed-

back from the world to the agent on each time-step and were drawn from the $P(O|S)$ distribution, dependent on the actual state of the world $S \in M_W$. Rewards were a special type of observation, which included a magnitude and were used in the calculation of δ.

The world

The world consisted of a semi-Markov state process, a current state $s_W(t)$, a dwell-time within that state $t_W(t)$, a current observation $O(t)$, and a current reward $R(t)$. Only observation ($O(t)$) and reward ($R(t)$) were provided to the agent.

A transition in the state of the world could occur due to a process inherent in the world or due to the action of the agent. For example, in our model of the adjusting-delay assay, the world will remain in the action-available state (providing an *observation* of two levers to the animal) until the agent takes an action. In contrast, once the agent has taken an action and the world has transitioned to one of the delay states (ISI-1, or ISI-2), the world will remain in that state for an appropriate number of time-steps and then transition to the reward state, irrespective of the agent's actions.

The macro-agent

The macro-agent corresponded to the animal or traditional "agent" in reinforcement learning models. The macro-agent interacted with the world and selected *actions*. Internal to the macro-agent were a set of n_μ μAgents, which instantiated the macro-agent's belief distribution of the state of the world. Smaller n_μ yielded noisier output. However, results were qualitatively unchanged down to n_μ=10. Results were stabler with explicitly uniform distributions of γ_i. The only simulation in which this made a noticeable difference was in the measure of hyperbolic discounting (because the hyperbolic function emerges

from the sum of many exponentials).

Individual μAgents

Each μAgent i was fully specified by a five-tuple $\langle s_i, t_i, \gamma_i, \delta_i, V_i(s)\rangle$, encoding the μAgent's currently believed state, s_i; the believed dwell-time, t_i (i.e., how long since the last state transition), the μAgent's internal discounting parameter γ_i, the current value-prediction-error signal δ_i, and the μAgent's value estimation function $V_i(s)$. Each μAgent contained its own individual discounting parameter γ_i, drawn from a uniform random distribution in the range $0 \leq \gamma_i \leq 1$.

The state, $s_i(t)$, and dwell-time, $t_i(t)$, of each μAgent are hypotheses of the actual state of the world, $s_W(t)$, and the actual dwell-time, $t_W(t)$ of the world within that state. Even if the μAgent knew the true initial state of the world, that hypothesis could diverge from reality over time. In order to maintain an accurate belief distribution, μAgents at each time-step computed the probability $P(s_i(t)|O(t))$, where $O(t)$ was the observation provided by the world at time t, and $s_i(t)$ was μAgent i's state at time t. μAgents with low $P(s_i(t)|O(t))$ updated their state belief by setting s_i to a random state s^* selected with probability $P(s^*|O(t))$. This is one of three mechanisms by which μAgents could change state (see below). An individual δ_i value error signal was computed at each μAgent state transition (see below).

Action selection

Actions can only occur at the level of the macro-agent because they are made by the organism as a whole. Because the state belief and value belief are distributed across the μAgents, a mechanism was required to select the best action given that belief distribution. In the model as implemented here, the macro-agent simply "took a vote" from the μAgents

as to which action to perform. Each μAgent provided an equally-weighted measure of the expected value for each action. The exact action selection algorithm is not crucial but must take account of the belief distribution and must balance exploration and exploitation.

Actions were selected based on an ϵ-greedy algorithm [4], with ϵ decreasing with each trial. This produces exploration early and exploitation later. At each time-step, a random number was drawn between 0 and 1. If that number was less than ϵ, then actions were taken based on the μAgents' vote on what actions were possible. If the number was greater than ϵ, then actions were taken based on the μAgents' vote on the expected values of the subsequent states. ϵ started at 1 and was multiplied by a factor of 0.95 each time reward was delivered, producing an exponential decrease in exploration with experience.

Exploration. If the macro-agent decided to explore, the action to be taken was drawn from a distribution based on which actions the μAgent population suggested was possible.

$$X(a_j) = \sum_{i \in \mu\text{Agents}} OK(a_j|s_i) \tag{1}$$

where $OK(a_j|s_i)$ was true (1) if action a_j was available from μAgent i's believed state s_i and false (0) otherwise. Actions were then selected linearly from the distribution of possible actions:

$$P(\text{select action } a_j) = \frac{X(a_j|s_i)}{\sum_j X(a_j|s_i)} \tag{2}$$

Exploitation. If the macro-agent decided to exploit the stored value functions, then actions were selected based on the normalized expected total value of the achieved state:

$$Q(a_j) = \sum_{i \in \mu\text{Agents}} (E[R(s_i')] + E[V(s_i')]) \tag{3}$$

where s'_i the state that would be achieved by taking action a_j given the current state s_i of μAgent i, $E[R(s'_i)]$ the expected reward in state s'_i, $E[V(s'_i)]$ the expected value of state s'_i. $E[R(s'_i)]$ was calculated from the internal world model M_i^A, and $E[V(s'_i)]$ was calculated from the internal value representation stored in μAgent i. If action a was not available from the current state of μAgent i, μAgent i was not included in the sum. Because our simulations only include reinforcement, only positive transitions were included, thus $Q(a_j)$ was rectified at 0. (Our simulations only include reinforcement primarily for simplicity. The mechanisms we describe here can be directly applied to aversive learning; however, because the extinction literature implies that reinforcement and aversion use separate, parallel systems [6], we have chosen to directly model reinforcement here.) Actions were then selected linearly between the possible Q functions:

$$P(\text{select action } a_j) = \frac{Q(a_j)}{\sum_j Q(a_j)} \tag{4}$$

Once an action was selected (either from $X(a_j)$ or from $Q(a_j)$), a decision was made whether to take the action or not based on the number of μAgents who believed the action was possible:

$$P(\text{take selected action } a_j) = \frac{X(a_j)}{n_\mu} \tag{5}$$

If the selected action was taken, the agent passed action a_j to the world. If the selected action was not taken, the agent passed the "null action" (which did not change the state and was always available) back to the world. If the macro-agent tried to take action a_j, but action a_j was incompatible with the actual world state s_W, no action was taken, and the "null action" was provided to the macro-agent.

When proportions of actions were measured (e.g. in the discounting experiments), proportions were only measured after 200 trials (by which time $\epsilon < 0.0001$).

μAgent transitions

There were three possible mechanisms by which μAgents could make transitions between hypothesized belief states $s_i \rightarrow s'_i$.

1. **Internal transitions.** On each time-step, each μAgent i decided whether to transition or not as a function of the dwell-time distribution, given its hypothesized state s_i and its hypothesized dwell-time t_i. If the μAgent took a transition, it followed the transition matrix stored within M_i^A.

2. **Taking an action.** If the macro-agent took action a_j, providing $A(t) = a_j$ to the world, all μAgents were then updated assuming the action occurred given the state-hypothesis of the μAgent s_i. If the action was incompatible with the μAgent's state belief, the μAgent's belief-state s_i was revised as described below.

3. **Incompatible observations.** On each time step, each μAgent i compared the observation provided by the world $O(t)$ with the observation expected given its internal hypothesized state $P(O|s_i)$. If $P(O|s_i)$ was 0 (meaning the observation was incompatible with s_i), the μAgent transitioned to a new state based on the probability of the state given the current observation $P(s'|O(t))$.

Calculating the error signal: δ

μAgents could experience a state transition as a consequence of the macro-agent taking an action, as a consequence of its dwell-time belief, or as a consequence of revising its state hypothesis due to low fitness. No matter how the μAgent changed its state hypothesis $s_i \rightarrow s'_i$, when μAgent i made a transition, it generated a δ contribution δ_i according to

$$\delta_i = {\gamma_i}^{t_i}(R(t) + V_i[s'_i]) - V_i[s_i] \tag{6}$$

where γ_i was the discounting parameter of the μAgent, t_i was the μAgent's hypothesized time since the last transition, $R(t)$ was the observed reward at time t, s'_i was the new state hypothesis to which the μAgent transitioned, and s_i was the old state hypothesis from which the μAgent transitioned. Of course, the process of transitioning set the μAgent's believed state to be s'_i and t_i to be 0. Note that $R(t)$ is not a function of s_i, but rather delivered to the agent from the world, based on the world state $s_W(t)$. Note that equation 6 is an exponential discounting function. Thus, each μAgent performed exponential discounting. The macro-agent showed hyperbolic discounting as an emergent process from the set of all the μAgents. Also, note that both the value of the new state and the current reward were discounted, as the sum of these quantities represents the total expected value of making a transition to a new state. Thus the sum $(R(t) + V_i[s'_i])$ must be discounted proportional to the time the agent remained in state s_i before reaching the new state s'_i.

On each μAgent state transition, the μAgent updated its internal estimation of the value of its hypothesized state s_i, using its individual δ_i:

$$V_i[s_i] \leftarrow V_i[s_i] + \alpha\delta_i \tag{7}$$

where α was the learning rate. The mean of the δ_i signals $\sum_i \delta_i/n_\mu$ from all μAgents conforms to the quantity reported in this paper as "the δ signal of the model" but never appeared explicitly within the simulation code. It is this total δ signal, however, which was compared to the population dopamine signal [13, 32–35].

Results

Hyperbolic discounting

Value, as defined in reinforcement learning models, is the integrated, expected reward, minus expected costs. The longer one must wait for a reward, the more likely it is for an unexpected event to occur, which could invalidate one's prediction [36, 37]. Agents, therefore, should discount future rewards: the more one must wait for the reward, the less valuable it should be. In addition, early rewards are more valuable than late rewards because early rewards can be invested (whether economically or ethologically) [36–38]. Any function that decreases with time could serve as a discounting function. In many situations, humans and other animals discount future rewards using a hyperbolic function [38–42] matching equation 12 rather than equation 11 (Figure 2).

TD algorithms incrementally learn an estimate of the value function, and thus require either a general analytical solution to the discounting function or an incremental calculation such that the value can be discounted with each timestep [8, 43, 44]. Because the discounting rate changes with time in hyperbolic discounting [38, 41], the calculation cannot be performed incrementally [8]. We suggest a possible mechanism for generating hyperbolic discounting via a multitude of exponential discounting factors. In the limit as the number of exponential discounters (having uniformly distributed discounting factors γ) approaches infinity, the average resultant discounting approaches hyperbolic. (See Supporting Information *Appendix S1* for mathematical proof.) In practice, having dozens or more of exponential discounters produces a close approximation to hyperbolic discounting.

Because each μAgent has an independent (exponential) discounting factor but actions are taken by the macro-agent based on a voting process of actions suggested by the

μAgents, the macro-agent will show a discounting curve that is the average of all the μAgent discounting curves. If the μAgent discounting curves are exponential functions with γ uniformly distributed over the range from 0 to 1, then the macro-agent will show approximately hyperbolic discounting in its behavior. The hypothesis that hyperbolic discounting arises from a (finite) set of exponential factors is consistent with recent fMRI observations [30,31] and suggests that the difference between this approximate hyperbolic and true hyperbolic discounting could be tested with sufficiently large data sets [45, 46].

Simulations. In order to measure the effective discounting function of our model, we modified the adjusting-delay assay of Mazur [39]. A five-state state-space was used to provide the macro-agent a choice between two actions, each of which led to a reward. In short, the agent was provided two choices (representing two levers): action a_1 brought reward r_1 after delay d_1 and action a_2 brought reward r_2 after delay d_2. For a given experiment, both rewards r_1, r_2 and one delay d_1 were held fixed, while the other delay d_2 was varied. For each set of $\langle r_1, r_2, d_1 \rangle$, the delay d_2 was found where the number of a_1 choices taken matched the number of a_2 choices taken in 300 trials. At this point, the actions indicate that the two discounting factors in the two delays exactly compensate for the difference in magnitudes of the two rewards. The delay d_2 at this equivalent action-selection point can be plotted against different fixed values of d_1. The slope of that curve indicates the discounting function used by the agent [39]. In the case of exponential discounting (γ^D where γ is the discounting factor, $0 \leq \gamma \leq 1$, and D is the delay), the slope will be 1, regardless of r_1 or r_2. In the case of reciprocal (R/D) discounting, the slope will equal to the ratio of rewards r_2/r_1, and the y-intercept will be 0. In the case of hyperbolic discounting ($R/(1 + kD)$, [39, 40, 47]), the slope will equal the ratio r_2/r_1, and in the case where $k = 1$, the y-intercept will be $r_2/r_1 - 1$. Simulations produced

a slope equal to the ratio of rewards r_2/r_1 (Figure 3) and a y-intercept approximating r_2/r_1-1, indicating that, even though each individual μAgent implemented an exponential discounting function, the macro-agent showed hyperbolic discounting, compatible with the behavioral literature [39–41, 47, 48].

Discounting across multiple steps. Temporal difference learning can use any function as a discounting function across a single state-transition. However, if hyperbolic discounting is implemented directly, a problem arises when discounting is measured over a sequence of multiple state transitions. This can be seen by comparing two state-spaces, one in which the agent remains in state $S0$ for ten timesteps and then transitions to state $S1$ (Figure 4A), and another in which the time taken between state $S0$ and $S1$ are divided into ten substates, with the agent remaining in each for one timestep (Figure 4H). These two statespaces encode equivalent information over equivalent time and (theoretically) should be discounted equivalently. If temporal discounting were implemented directly with equation 12, then the agent would show hyperbolic discounting across the first statespace, but not the second.

We tested this explicitly by comparing four simulations (see Figure 4):

1. Discounting is not distributed, and δ is calculated by

$$\delta = \frac{(R(t) + V[s'])}{1+t} - V[s] \tag{8}$$

In this condition, the measured discounting of the model was hyperbolic over a single-step state-space (Figure 4G). However, over an equivalent chained state-space (Figure 4N), the macro-agent discounted each state-jump hyperbolically. Since each state had a delay of D=1, the amount of discounting for each state-jump was $\frac{1}{1+D} =$

0.5, leading to exponential discounting (with $\gamma = 0.5$) over the chain of states. This occurred whether or not value representation was distributed (Figure 4D,K).

2. Discounting is not distributed, and δ is calculated by

$$\delta = \gamma^{\,t/\tau}(R(t) + V[s']) - V[s] \tag{9}$$

where $\gamma = 0.75$. In this condition, the measured discounting of the model was exponential over both the single-step state-space (Figure 4F) and the chained state-space (Figure 4M). This occurred whether or not value representation was distributed (Figure 4C,J).

3. Discounting is distributed (i.e., each μAgent has a different exponential discounting rate γ_i drawn uniformly at random from $(0, 1)$). δ is thus calculated using Eqn. 6 as specified in the Methods section. However, value representation is not distributed; all μAgents access the same value representation $V(s)$. Thus, Eqn. (7) was replaced with

$$V[s_i] \leftarrow V[s_i] + \frac{\alpha\delta_i}{n_\mu} \tag{10}$$

In this equation, although the μAgents could update different states based on their hypothesized state-beliefs, all values were united into a single universal value function $V(s)$. In this condition, the macro-agent reverted to the one-step hyperbolic equation in version 1 (Eqn 8), showing hyperbolic discounting in the single-step state-space (Figure 4E) but not the chained state-space (Figure 4L). In the chained state-space, the sum of distributed exponential discounting rates produces hyperbolic discounting across each state-jump, so across the chain of states discounting was exponential (with $\gamma = \frac{1}{1+1} = 0.5$).

4. Both discounting (Eqn. 6) and value (Eqn. 7) are distributed. This model showed hyperbolic discounting under both the single-step state-space (Figure 4B) and the chained state-space (Figure 4I). Because each μAgent has its own value representation for each state, the value decrease across each state-jump was exponential, with each μAgent having a different γ. Thus the average value of a state was the average of these exponentially-discounted values, which was hyperbolic.

It is still an open question whether real subjects show differences between single-step and chained state-space representations. Such an experiment would require a mechanism to change the internal representation of the subject (as one state lasting for ten seconds or as ten states lasting for one second each). This could be tested by concatenating multiple delays. Simulation 1, using explicit hyperbolic discounting, predicts that discounting across a chained state-space will be much faster than discounting across a single-step. Whether this occurs remains a point of debate [49]. The model of distributed discounting and distributed values best fits the data that discounting is hyperbolic even across multiple delays.

Non-uniform distributions of discounting rates. So far in exploring distributed discounting, we have selected γ_i uniformly from $(0, 1)$. Using this γ distribution, the overall agent exhibits hyperbolic discounting as $\frac{1}{1+d}$. However, different γ distributions should produce different overall discounting functions.

We tested this by altering the γ distribution of the μAgents and measuring the resulting changes in discounting of the overall agent. In the uniform distribution (which was also used for all other simulations in this paper), $P(\gamma < x) = x$, $x \in (0, 1)$ (Figure 5A). As was also shown in Figure 4B, this results in hyperbolic discounting for the overall agent (Figure 5B). Fitting the function $\frac{1}{1+d}$ to this curve gives an R^2 of 0.9999 (using 200

μAgents; the fit improves as n_μ increases). To bias for slow discounting rates, we used the distribution $P(\gamma < x) = x^2$ (Figure 5C). The measured discounting of the overall agent using this γ distribution was slower (Figure 5D) and was well-fit by the function $\frac{1}{1+0.5d}$. To bias for fast discounting rates, we used the distribution $P(\gamma < x) = \sqrt{x}$ (Figure 5E). The measured discounting of the overall agent using this γ distribution was faster (Figure 5F) and was well-fit by the function $\frac{1}{1+2d}$. These results match theoretical predictions for the effect of biased γ distributions on discounting [37]. Mathematically, it can also be shown that non-hyperbolic discounting can result from γ distributions that do not follow $P(\gamma < x) = x^a$; for example if the γ distribution is bimodal with a relative abundance of very slow and very fast discounting μAgents.

Smokers, problem gamblers, and drug abusers all show faster discounting rates than controls [48, 50–53]. Whether discounting best-fit by different time-constants is exactly hyperbolic or not is still unknown (see, for example, [48, 51, 54], in which the hyperbolic fit is clearly imperfect). These differences could be tested with sufficiently large data sets, as the time-courses of forgetting have been: although forgetting was once hypothesized to follow hyperbolic decay functions, forgetting is best modeled as a sum of exponentials, not as hyperbolic or logistic functions [45,46]. Similar experiments could differentiate the hyperbolic and multiple-exponential hypotheses.

All subsequent experiments used a uniform distribution of γ_i.

Distributed belief

Because each μAgent instantiates an independent hypothesis about the state of the world, the macro-agent can maintain a distributed belief of world-state. We describe two consequences of distributed belief that explain experimental data.

First, some situations contain readily identifiable cues which allow those times when

the agent is in those situations to be separated from times when the agent is not. For example, during delay conditioning, there is a specific stimulus (e.g. a light or tone) that is played continuously through the delay. Separating "tone-on" situations from "tone-off" situations readily identifies the inter-stimulus-interval. Other situations are not as readily identifiable. For example, during inter-trial intervals and during the inter-stimulus interval in trace conditioning, there is a gap in which the agent does not know what cues to attend to. Our model simulates this cue ambiguity by representing the gap with a set of identical *equivalent states*. These equivalent states slow value learning because each state only holds a fraction of the μAgent state-belief distribution and therefore only receives a fraction of the total δ produced by a state-transition. We suggest that equivalent-states explain the well-established slower learning rates of trace compared to delay conditioning [55], and explain the slow loss of dopamine signal at conditioned stimuli with overtraining [32].

Second, distributed belief allows TD to occur in ambiguous state-spaces [5, 6], which can explain the generalization responses of dopamine [15, 34] and the transient burst of dopamine observed at movement initiation [56, 57].

Trace and delay-conditioning

In delay conditioning, the CS remains on until the reward is delivered, while in trace conditioning there is a gap between the CS and US — the CS disappears before the US appears [55, 58]. This simple change produces dramatic effects: trace conditioning takes much longer to learn than delay conditioning, and requires the hippocampus, unlike delay conditioning [55, 59, 60]. One possible explanation for the difference is that, because there is no obvious cue for the animal to pay attention to, the intervening state representation during the gap in trace conditioning is spread out over many multiple "equivalent states". (There is new evidence that trace conditioning requires hippocampus only under aversive

training conditions [61], which may suggest that other structures can bridge the gap in appetitive trace conditioning. This does not change our primary hypothesis — that trace conditioning entails an "equivalent states" representation of the gap between CS and US.)

Because the μAgents model can represent distributed belief, we can model trace conditioning by placing a collection of equivalent states between the cue and the reward. As noted above, because value learning is distributed across those equivalent states, value is learned more slowly than in well-identified states.

Simulations. In order to test the effect of a collection of equivalent states in the interstimulus time, we simulated a Pavlovian conditioning paradigm, under two conditions: with a single state intervening between CS and US, or with a collection of 10 or 50 equivalent states between the CS and US. As can be seen in Figure 6, the value of the initial ISI state (when the CS turns on) $V(CS)$ increases more quickly under delay than under trace conditioning. This value function is the amount of expected reward given receipt of the CS. Thus in trace conditioning, the recognition that the CS implies reward is delayed relative to delay conditioning. Increasing the number of equivalent states in the ISI from 10 to 50 further slows learning of $V(CS)$ (Figure 6).

Discussion and implications. Sets of equivalent states can be seen as a model of the attention the agent has given to a single set of identified cues. Because the stimulus remains on during delay conditioning, the stimulus may serve to focus attention, which differentiates the Stimulus-on state from other states. Because there is no obvious attentional focus in the interstimulus interval in trace conditioning, this may produce more divided attention, which can be modeled as a large collection of equivalent intervening states in the ISI period. Levy [62] has explicitly suggested that the hippocampus may play a role in finding single states with which to fill in these intervening gaps, which may

explain the hippocampal-dependence of trace-conditioning [55, 59]. Consistent with this, Pastalkova *et al.* [63] have found hippocampal sequences which step through intervening states during a delay period. Levy's theory predicted that it should take some time for that set of intervening states to develop [62]; before the system has settled on a set of intervening states, μAgents would distribute themselves among the large set of potential states, producing an equivalent-set-like effect. This hypothesis predicts that it should be possible to create intermediate versions of trace and delay conditioning by filling the gap with stimuli of varying predictive usefulness, thus effectively controlling the size of the set of equivalent states. The extant data seem to support this prediction [55, 64].

The disappearance of CS-related dopamine signals with overtraining

During classical conditioning experiments, dopamine signals occur initally at the delivery of reward (which is presumably unexpected). With experience, as the association between the predictive cue stimulus (CS) and the reward (unconditioned stimulus, US) develops, the dopamine signal vanishes from the time of delivery of the US and appears at the time of delivery of the CS [34]. However, with extensive overtraining with very regular intertrial intervals, the dopamine signal vanishes from the CS as well [32].

Classical conditioning can be modeled in one of two ways: as a sequence of separate trials, in which the agent is restarted in a set $S0$ state each time or as a loop with an identifiable inter-trial-interval (ITI) state [5,8,14,24]. While this continuous looped model is more realistic than trial-by-trial models, with the inclusion of the ITI state, an agent can potentially see across the inter-trial gap and potentially integrate the value across all future states. Eventually, with sufficient training, an agent would not show any δ signal to the CS because there would be no unexpected change in value at the time the CS was delivered. We have found that this decrease happens very quickly with standard TD

simulations (tens to hundreds of trials, data not shown). However, Ljungberg *et al.* report that monkeys required >30,000 movements to produce this overtraining effect. This effect is dependent on strongly regular intertrial intervals (W. Schultz, personal communication).

The μAgents model suggests one potential explanation for the slowness of the transfer of value across the ITI state in most situations: Because the ITI state does not have a clearly identifiable marker, it should be encoded as a distributed representation over a large number of equivalent states. Presumably, in a classical conditioning task, the inter-stimulus interval is indicated by the presence of a strong cue (the tone or light). However, the appropriate cue to identify the inter-trial-interval (ITI) is not obvious to the animal, even though there are presumably many available cues. In our terminology, the ITI state forms a collection of *equivalent states*. Because all of these ITI states provide the same observation, the agent does not know which state the world entered and the μAgents distribute over the many equivalent ITI states. The effect of this is to distribute the δ signal (and thus the change in value) over those many equivalent states. Thus the value of the ITI states remains low for many trials, and the appearance of an (unexpected) CS produces a change in value and thus a positive δ signal.

Simulations. In order to test the time-course of overtraining, we simulated a standard classical conditioning task (Figure 7A). Consistent with many other TD simulations, the value-error δ signal transferred from the reward to the CS quickly (on the order of 25 trials) (Figure 7B,C,E). This seemingly steady-state condition (δ in response to CS but not reward) persists for hundreds of trials. But as the learned value-estimates of the equivalent ITI states gradually increase over thousands of trials, the δ signal at the CS gradually disappears (Figure 7D,E). The ratio of time-to-learn to time-to-overlearn is compatible with the data of Ljungberg *et al.* [32]. Increasing the number of equivalent

states in the ITI further slows abolition of δ at the CS (Figure 7E).

Discussion and implications. The prediction that the inability of the delta signal to transfer across ITI states is due to the ITI state's lack of an explicit marker suggests that it should be possible to control the time course of this transfer by adding markers. Thus, if explicit, salient markers were to be provided to the ITI state, animals should show a faster transfer of delta across the ITI gap, and thus a faster decrease in the delta signal at the (no-longer-unexpected) CS. This also suggests that intervening situations without markers should show a slow transfer of the delta signal, as was proposed for trace conditioning above.

Transient dopamine bursts at uncued movement initiation.

Dopamine cues occurring at cue-stimuli associated with expected reward have been well-studied (and well-modeled) in Pavlovian conditioning paradigms. However, dopaminergic signals also appear just prior to uncued movements in instrumental paradigms [56, 65] and can appear even without external signals [57]. One potential explanation is that this dopamine signal is indicative of an internal transition occurring in the agent's internal world-model, perhaps from a state in which an action is unavailable to a state in which an action is available, thus providing a change in value and thus providing a small δ signal. Only a few μAgents would have to make this transition in order to produce such a signal and initiate an action. Once the action was initiated, the other μAgents would be forced to update their state belief in order to remain compatible with the ensuing world observations.

Simulations. In order to test the potential existence of dopaminergic signals just prior to movement appearing with no external cues, we built a state-space which contained an

internally- but not externally-differentiated GO state (Figure 8A). That is, the GO-state was not identifiably different in the world, but actions were available from it. μAgents in the ITI state would occasionally update their state belief to the GO state due to the similarity in the expected observations in the GO and ITI states. If a sufficient number of μAgents were present in the GO state, the agent could take the action. Because the GO state was temporally closer to the reward than the ITI state, more value was associated with the GO state than with the ITI state. Thus, a μAgent transitioning into the GO state would produce a small δ signal. Taking an action requires the overall agent to believe that the action is possible. However, there is no external cue to make the μAgents all transition synchronously to the GO state, so they instead transition individually and probabilistically, which produces small pre-movement δ signals. In the simulations, μAgents gradually transitioned to the GO state until the action was taken (Figure 8B, top panel). During this time immediately preceding movement, small probabilistic δ signals were observed (Figure 8B, middle panel). When these signals were averaged over trials, a small ramping δ signal was apparent prior to movement (Figure 8B, bottom panel).

Discussion and implications. As can be seen in Figure 8, there is a ramping of delta signals as μAgents transfer from the ITI state to the GO state. A similar ramping has been seen in dopamine levels in the nucleus accumbens preceding a lever press for cocaine [56, e.g. Figure 2, p. 615]. This signal has generally been interpreted as a causative force in action-taking [65]. The signal in our simulation is not causative; instead it is a read-out of an internal shift in the distributed represented state of the macro-agent — the more μAgents there are in GO state, the more likely the macro-agent is to take action. Whether this ramping δ signal is a read-out or is causative for movement initiation is an open-question that will require more detailed empirical study.

Other TD simulations

The μAgents model proposed here enabled novel explanations and models for (a) hyperbolic discounting, (b) differences between trace- and delay-conditioning, (c) effects of overtraining, and (d) the occurrence of dopamine signals prior to self-initiated movement.

However, TD models have been shown in the past to be able to accommodate a number of other critical experiments, including (e) that unsignaled reward produces a positive dopamine signal ($\delta > 0$) [5, 8, 18, 24, 32, 34, 66, 67], (f) that phasic dopamine signals ($\delta > 0$) transfer from the time of an unconditioned stimulus to the time of the corresponding conditioning stimulus [1, 2, 8, 18, 19, 21, 32, 34], (g) that dopamine neurons pause in firing (δ decreases) with missing, but expected, rewards [5, 8, 18, 24, 32–34], (h) that early reward produces a positive dopamine signal ($\delta > 0$) with no corresponding decrease at the expected reward time [5,8,24,33], (i) that late reward produces a negative dopamine signal ($\delta < 0$) at the expected time of reward and a positive dopamine signal ($\delta > 0$) at the observed (late) reward [5,8,24,33]. Finally, TD models have been able to explain (j) dopamine responses to changing probabilities of receiving reward [5,8,68], and (k) generalization responses [15, 34].

Extensive previous work already exists on how TD models capture these key experimental results. Some of these cases occur due to the basic identification of the phasic dopamine signal with δ [1,2,11]. Some occur due to the use of semi-Markov models (which allows a direct simulation of time) [5,8,44]. Others occur due to the distributed representation of belief (e.g. *partially observability* [5,8,15,44]). Because our μAgents model is an implementation of all of these, it also captures these basic results. Although the results included in this supplemental section do not require μAgents, the inclusion of μAgents does not lose them, which we briefly illustrate here.

Unsignaled reward produces a positive δ signal. When presented with an unexpected reward signal, dopamine neurons fire a short phasic burst [32, 34, 69]. Following Daw [8], this was modeled by a simple two state state-space: after remaining within the ITI state for a random time (drawn from a normal distribution, $\mu = 15, \sigma = 1$ time-steps), the world transitioned to a *reward-state*, during which time a reward was delivered, at the completion of which, the world returned to the ITI state (Figure 9A). On the transition to the reward state, a positive δ signal occurred (Figure 9B). Standard TD algorithms produce this result. Using sets of equivalent states to represent the ITI extends the time that the US will continue to cause a dopamine surge. Without this set of equivalent ITI states, the dopamine surge to the US would diminish within a number of trials much smaller than observed in experimental data.

δ transfers from the unconditioned reward to conditioned stimuli. With unexpected reward, dopamine cells burst at the time of reward. However, when an expected reward is received, dopamine cells do not change their firing rate [32, 34]. Instead, the dopamine cells fire a burst in response to the conditioned stimulus (CS) that predicts reward [32,34]. Following "the dopamine as δ" hypothesis, this transfer of δ from reward to anticipatory cues is one of the keys to the TD algorithm [1, 2, 8, 34]. We modeled this with a three-state state-space (ITI, ISI, and Rwd; Figure 7A). As with other TD models, δ transferred from US to CS (Figure 7B,C,E). We modeled the ITI state as a set of equivalent states to extend the time that the CS will continue to cause a dopamine surge. In previous looped models, the dopamine surge to the CS would diminish within a small number of trials, giving a learning rate incompatible with realistic CS-US learning. As with other TD models living within a semi-Markov state-space [5, 8], the delta signal shifted back from the reward state to the previous anticipatory stimulus without

progressing through intermediate times [70].

Missing, early, and late rewards. When expected rewards are omitted, dopamine neurons pause in their firing [18,32,33]. When rewards are presented earlier or later than expected, dopamine neurons show an excess of firing [33]. Importantly, late rewards are preceded by a pause in firing at the expected time of reward [33]. With early rewards, the data is less clear as to the extent of the pause at the time of expected reward (see Figure 6 of Hollerman *et al.* [33]). As noted by Daw *et al.* [5] and Bertin *et al.* [24], these results are explicable as consequences of semi-Markov state-space models.

In semi-Markov models, the expected time distribution of the ISI state is explicitly encoded. μAgents will take that transition with the expected time distribution of the ISI state. These μAgents will find a decrease in expected value because no actual reward is delivered. The δ signal can thus be decomposed into two components: a positive δ signal arising from receipt of reward and a negative signal arising from μAgents transitioning on their own. These two components can be separated temporally by providing reward early, late, or not providing it at all (missing reward).

After training with a classical conditioning task, a δ signal occurs at the CS but not the US (Figure 10A). When we delivered occasional probe trials on which reward arrived early, we observed a δ signal at the US (Figure 10B). This is because the value of the CS state accounts for a reward that is discounted by the normal CS-US interval. If the reward occurs early, it is discounted less. On the other hand, when we delivered probe trials with late reward arrival, we observed a negative δ signal at the expected time of reward followed by a positive δ signal at the actual reward delivery (Figure 10C). The negative δ signal occurs when μAgents transition to the reward state but receive no actual reward. The observation of the ISI state is incompatible with μAgents' belief that they

are in the reward state, so μAgents transition back to the ISI state. When reward is then delivered shortly afterwards, it is discounted less than normal and thus produces a positive δ signal.

If reward fails to arrive when expected (missing reward), then the μAgents will transition to the reward state anyway due to their dwell-time and state hypotheses, at which point, value decreases unbalanced by reward. This generates a negative δ signal (Figure 10D). The signal is spread out in time corresponding to the dwell-time distribution of the ISI state.

δ transfers proportionally to the probability of reward. TD theories explain the transfer seen in Figure 7 through changes in expected value when new information is received. Before the occurrence of the CS, the animal has no reason to expect reward (the value of the ITI state is low); after the CS, the animal expects reward (the value of the ISI state is higher). Because value is dependent on expected reward, if reward is given probabilistically, the change in value at the CS should reflect that probability. Consistent with that hypothesis, Fiorillo *et al.* [68] report that the magnitude of the dopamine burst at the CS is proportional to the probability of reward-delivery. In the μAgents model, a high probability of reward causes δ to occur at the CS but not US after training (Figure 11; also see Figure 7C and Figure 10A). As the probability of reward drops toward zero, δ shifts from CS to US (Figure 11). This is because the value of the ISI state is less when it is not a reliable predictor of reward.

Generalization responses. When provided with multiple similar stimuli, only some of which lead to reward, dopamine neurons show a phasic response to each of the stimuli. With the cues that do not lead to reward, this positive signal is immediately followed by a negative counterbalancing signal [34]. As suggested by Kakade and Dayan [15],

these results can arise from partial observability: on the observation of the non-rewarded stimulus, part of the belief distribution transfers inappropriately to the state representing a stimulus leading to a rewarding pathway. When that belief distribution transfers back, the negative δ signal is seen because there is a drop in expected value. This explanation is compatible with the μAgents model presented here in that it is likely that some μAgents would shift to the incorrect state producing a generalization δ signal which would then reverse when those μAgents revise their state-hypothesis to the correct state.

To test the model's ability to capture the generalization result, we designed a state-space that contained two CS stimuli, both of which provided a "cue" observation. However, after one time-step, the CS- returned to the ITI state, while the CS+ proceeded to an ISI state, which eventually led to reward. Because (in this model), both the CS's provided similar observations, when either CS appeared, approximately half the μAgents entered each CS state, providing a positive δ signal. In the CS- case, the half that incorrectly entered the CS+ state updated their state belief back to the ITI state after one time-step, providing a negative signal. In the CS+ case, the half that incorrectly entered the CS- state updated their state belief back to the ISI state after one time-step, providing a lengthened positive signal. See Figure 12.

Discussion

In this paper, we have explored distributing two parameters of the standard temporal difference (TD) algorithm for reinforcement learning (RL): the discounting factor γ and the belief state s. We implemented these distributed factors in a unified semi-Markov temporal-difference-based reinforcement learning model using a distribution of μAgents, the set of which provide a distributed discounting factor and a distributed representation

of the believed state. Using distributed discounting produced hyperbolic discounting consistent with the experimental literature [40, 41]. The distributed representation of belief, along with the existence of multiple states with equivalent observations (i.e. *partial observability*), provided for the simulation of collections of "equivalent-states", which explained the effects of overtraining [32], and differences between trace and delay conditioning [55]. Distributed state-belief also provided an explanation for transient dopamine signals seen at movement initiation [56, 57], as well as generalization effects [15].

Although the μAgents model we presented included both distributed discounting and distributed belief states (in order to show thorough compatibility with the literature), the two hypotheses are actually independent and have separable consequences.

Distributed discounting

The mismatch between the expected exponential discounting used in most TD models and the hyperbolic discounting seen in humans and other animals has been recognized for many years [5, 8, 37, 38, 41, 71, 72].

Although hyperbolic discounting will arise from a uniform (and infinite) distribution of exponential functions [37, 73, see also Supporting Information *Appendix S1*], as the number of exponential functions included in the sum decreases, the discounting function deviates from true hyperbolicity. Changing the uniformity of the distribution changes the impulsivity of the agent (Figure 5). We also found that because the product of hyperbolic functions is not hyperbolic, it was necessary to maintain the separation of the discounting functions until action-selection, which we implemented by having each μAgent maintain its own internal value function $V_i(s)$ (Figure 4).

Other models. In addition to the suggestion that hyperbolic discounting could arise from multiple exponentials proposed here, three explanations for the observed behavioral hyperbolic discounting have been proposed [37]: (1) maximizing average reward over time [5, 71, 74], (2) an interaction between two discounting functions [75–77], and (3) effects of errors in temporal perception [8, 37].

While the assumption that animals are maximizing average reward over time [5, 71, 74] does produce hyperbolic discounting, assumptions have to be made that animals are ignoring intertrial intervals during tasks [37, 74]. Another complication with the average-reward theory is that specific dopamine neurons have been shown to match prediction error based on exponential discounting when quantitatively examined within a specific task [18]. In the μAgents model, this could arise if different dopamine neurons participated in different μAgents, thus recording from a single dopamine neuron would produce an exponential discounting factor due to recording from a single μAgent within the population.

The two-process model is essentially a two-μAgent model. While it has received experimental support from fMRI [76, 77] and lesion [78] studies, recent fMRI data suggest the existence of intermediate discounting factors as well [30]. Whether the experimental data is sufficiently explained by two exponential discounting functions will require additional experiments on very large data sets capable of determing such differences [45, see discussion in *Predictions*, below].

There is a close relationship between the exponential discounting factor and the agent's perception of time [8, 79, 80]. Hyperbolic discounting can arise from timing errors that increase with increased delays [79, 81, 82]. The duality between time perception and discounting factor suggests the possibility of a μAgent model in which the different μAgents are distributed over time perception rather than discounting factor. Whether such a model is actually viable, however, will require additional work and is beyond the scope of

this paper.

Distributed Belief

The concept of a distributed representation of the believed state of the world has also been explored by other researchers [5,8,23,24,83]. In all of these models (including ours), action-selection occurs through a probabilistic voting process. However, the δ function differs in each model. In the Doya *et al.* [23] models, a single δ signal is shared among multiple models with a "responsibility signal". In the Daw [5] models, belief is represented by a partially-observable Markov state process, but is collapsed to a single state before δ is calculated. Our distributed δ signal provides a potential explanation for the extreme variability seen in the firing patterns of dopaminergic neurons and in the variability seen in dopamine release in striatal structures [84], in a similar manner to that proposed by Bertin *et al.* [24].

Distributed attention. A multiple-agents model with distributed state-belief provides for the potential for situations represented as collection of equivalent states rather than as a single state. This may occur in situations without readily identifiable markers. For example, during inter-trial-intervals, there are many available cues (machinery/computer sounds, investigator actions, etc.) Which of these cues are the reliable differentiators of the ITI situations from other situations is not necessarily obvious to the animal. This leads to a form of divided attention, which we can model by providing the μAgents with a set of equivalent states to distribute across. While the μAgents model presented here requires the user to specify the number of equivalent states for a given situation, it does show that under situations in which we might expect to have many of these equivalent states, learning occurs at a slower rate than over situations in which there is only one

state. Other models have suggested hippocampus may play a role in identifying unique states across these unmarked gaps [62, 63, 85]. While our model explains why learning occurs slowly across such an unmarked gap, the mechanisms by which an agent identifies states is beyond the scope of this paper.

The implementation of state representations used by many models are based on distributed neural representations. Because these representations are distributed, they can show variation in internal self-consistency — the firing of the cells can be consistent with a single state, or they can be distributed across multiple possibilities. The breadth of this distribution can be seen as a representation in the inherent uncertainty of the information represented [86–90]. This would be equivalent to taking the distribution of state belief used in the μAgents model to the extreme in which each neuron represents an estimate of a separate belief. Ludvig *et al.* [25,26] explicitly presented such a model using a distributed representation of stimuli ("microstimuli").

Markov and semi-Markov state-spaces

Most reinforcement-learning models live within Markov state spaces (e.g. [1,2,67,91,92]), which do not enable the direct simulation of temporally-extended events. Semi-Markov models represent time explicitly, by having each state represent a temporally-extended event [5, 93–95].

In a Markov chain model, each state represents a single time-step, and thus temporally extended events are represented by a long sequence of states [93, 94, 96]. Thus, as a sequence is learned, the δ signal would step back, state by state. This backwards stepping of the δ signal can be hastened by including longer eligibility traces [19] or graded temporal representations [25,26], both of which have the effect of blurring time across the multiple intervening states. In contrast, in a semi-Markov model, each state contains within it a

(possibly variable) dwell-time [5, 8, 93, 97, 98]. Thus while the δ signal still jumps back state-by-state, the temporal extension of the states causes the signal to jump back over the full inter-stimulus time without proceeding through the intervening times. As noted by Wörgötter and Porr [70], this is more compatible with what is seen by Schultz and colleagues [13, 32, 34, 35, 99–101]: the dopamine signal appears to jump from reward to cue without proceeding through the intermediate times.

Semi-Markov state spaces represent intervening states (ISI states) as a single situation, which presumably precludes responding differently within the single situation. In real experiments, animals show specific time-courses of responding across the interval as the event approaches, peaking at the correct time [102]. The temporal distribution of dopamine neuron firing can also change across long delays [103]. Because our model includes a distribution of belief across the semi-Markov state space (the t_i terms of the μAgent distribution), the number of μAgents that transition at any given time step can vary according to the distribution of expected dwell times. While matching the distributions of specific experiments is beyond the scope of this paper, if the probability of responding is dependent on the number of μAgents (Equation (5)), then the macro-agent can show a similar distribution of behavior (see Figure 8).

Anatomical instantiations

The simulations and predictions reported here are based on behavioral observations and on the concept that dopamine signals prediction error. However, adding the hypotheses that states are represented in the cortex [5, 6, 104], while value functions and action selection are controlled by basal ganglia circuits [104–107] would suggest that it might be possible to find multiple μAgents within striatal circuits. Working from anatomical studies, a number of researchers have hypothesized that the cortical-striatal circuit con-

sists of multiple separable pathways [27, 28, 108, 109]. Tanaka *et al.* [30] explicitly found a gradient of discounting factors across the striata of human subjects. This suggests a possible anatomical spectrum of discounting factors which would be produced by a population of μAgents operating in parallel, each with a preferred exponential discounting factor γ_i. Many researchers have reported that dopamine signals are not unitary (See [8] for review). Non-unitary dopamine signals could arise from different dopamine populations contributing to different μAgents. Haber *et al.* [29] report that the interaction between dopamine and striatal neural populations shows a regular anatomy, in a spiral progressing from ventral to dorsal striatum. The possibility that Tanaka *et al.*'s slices may correspond to Haber *et al.*'s spiral loops, and that both of these may correspond to μAgents is particularly intriguing.

Predictions

Hyperbolic discounting. The hypothesis that hyperbolic discounting arises from multiple exponential processes suggests that with sufficient data, the actual time-course of discounting should be differentiable from a true hyperbolic function. While the fit of real data to hyperbolic functions are generally excellent [39, 40, 48, 110], there are clear departures from hyperbolic curves in some of the data (e.g. [51, 54]). Rates of forgetting were also once thought to be hyperbolic [45], but with experiments done on very large data sets, rates of forgetting have been found, in fact, to be best modeled as the sum of multiple exponential processes [45, 46]. Whether discounting rates will also be better modeled as the sum of exponentials rather than as a single hyperbolic function is still an open question.

True hyperbolicity only arises from an infinite sum of exponentials drawn from a distribution with $P(\gamma < x) = x^a$, $x \in (0, 1)$. Under this distribution, the overall hyperbolic

discounting is described by $\frac{1}{1+kD}$, where $k = 1/a$. Changing the parameter a can speed up or slow down discounting while preserving hyperbolicity; changing the γ distribution to follow a different function will lead to non-hyperbolic discounting.

Serotonin precursors (tryptophan) can change an individual's discount rate [111,112]. These serotonin precursors also changed which slices of striatum were active [112]. This suggests that the serotonin precursors may be changing the selection of striatal loops [29], slices [30], or μAgents. If changing levels of serotonin precursors are changing the selection of μAgents and the μAgent population contains independent value estimates (as suggested above), then learning under an excess of serotonin precursors may have to be relearned in the absence of serotonin precursors and vice-versa due to the change in the population of μAgents occurring with the change in serotonin levels.

In addition, in tasks structured such that exponential discounting maximizes the reward, subjects can shift their discounting to match the exponential to the task [113]. Drug-abusers [48, 50], smokers [51, 52], and problem gamblers [53] all show faster discounting rates than matched control groups. One possible explanation is that these altered overall discounting rates reflect differences in the distribution of μAgent discounting factors. As shown in Figure 5, biasing the μAgent γ distribution can speed or slow overall discounting. Further, while a γ distribution following $P(\gamma < x) = x^a$ exhibits hyperbolic discounting, other distributions lead to non-hyperbolic discounting. Model comparison could be used on human behavioral data to determine if subsets of subjects show such patterns of discounting. However, this may require very large data sets [45].

Distributed belief and collections of equivalent states. The hypothesis that the slow development of overtraining [32] and the differences between trace- and delay conditioning [55] occur due to the distribution of attention across collections of equivalent

states implies that these effects should depend on the ambiguity of the state given the cues. Thus, value should transfer across a situation proportionally to the identifiability of the that situation. Decreasing cue-ambiguity during inter-trial-intervals should speed up the development of overtraining (observable as a faster decrease in dopamine signal at the CS). Increasing cue-ambiguity during inter-stimulus-intervals should slow down learning rates of delay-conditioning. As the cues become more ambiguous and less salient, delay-conditioning should become closer and closer to trace conditioning. The extant data seem to support this prediction [55, 64].

Summary/Conclusion

In this paper, we explored distributing two parameters of temporal difference (TD) models of reinforcement learning (RL): distributed discounting and distributed representations of belief. The distributed discounting functions provide a potential mechanistic explanation for hyperbolic discounting. The distributed representations of belief provide potential explanations for the decrease in dopamine at the conditioned stimulus seen in overtrained animals, for the differences in learning rate between trace and delay conditioning, and for transient dopamine at movement initiation. These two hypotheses, although separable, together provide a unified model of temporal difference reinforcement learning capable of explaining a large swath of the experimental literature.

Acknowledgements

We thank Nathaniel Daw, Jadin Jackson, Steve Jensen, Adam Johnson, Daniel Smith, Kenji Doya, Warren Bickel, Jim Kakalios, Reid Landes, Matthijs van der Meer, and Neil Schmitzer-Torbert for helpful discussions.

References

1. Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience 16: 1936-1947.

2. Schultz W, Dayan P, Montague R (1997) A neural substrate of prediction and reward. Science 275: 1593-1599.

3. Doya K (2000) Metalearning, neuromodulation, and emotion. In: Hatano G, Okada N, Tanabe H, editors, Affective Minds, Elsevier.

4. Sutton RS, Barto AG (1998) Reinforcement Learning: An introduction. Cambridge MA: MIT Press.

5. Daw ND, Courville AC, Touretzky DS (2006) Representation and timing in theories of the dopamine system. Neural Computation 18: 1637-1677.

6. Redish AD, Jensen S, Johnson A, Kurth-Nelson Z (2007) Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. Psychological Review 114: 784-805.

7. Sutton RS, editor (1992) Special issue on reinforcement learning, volume 8(3/4) of *Machine Learning*. Boston: Kluwer Academic Publishers.

8. Daw ND (2003) Reinforcement learning models of the dopamine system and their behavioral implications. Ph.D. thesis, Carnegie Mellon University.

9. Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokesy

WF, editors, Classical Conditioning II: Current Research and Theory, New York: Appleton Century Crofts. pp. 64-99.

10. Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: Expectation and prediction. Psychological Review 88: 135-170.

11. Barto AG (1995) Adaptive critics and the basal ganglia. In: Houk JC, Davis JL, Beiser DG, editors, Models of Information Processing in the Basal Ganglia, Cambridge MA: MIT Press. pp. 215-232.

12. Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. Annual Review of Neuroscience 23: 473-500.

13. Schultz W (2002) Getting formal with dopamine and reward. Neuron 36: 241-263.

14. Redish AD (2004) Addiction as a computational process gone awry. Science 306: 1944-1947.

15. Kakade S, Dayan P (2002) Dopamine: generalization and bonuses. Neural Networks 15: 549-599.

16. O'Doherty JP, Peter Dayan and KF, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38: 329-337.

17. O'Doherty JP (2004) Reward representations and reward-related learning in the human brain: insights from neuroimaging. Current Opinion in Neurobiology 14: 769-776.

18. Bayer HM, Glimcher P (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47: 129-141.

19. Pan WX, Schmidt R, Wickens JR, Hyland BI (2005) Dopamine Cells Respond to Predicted Events during Classical Conditioning: Evidence for Eligibility Traces in the Reward-Learning Network. J Neurosci 25: 6235-6242.

20. Stuber GD, Wightman RM, Carelli RM (2005) Extinction of cocaine self-administration reveals functionally and temporally distinct dopaminergic signals in the nucleus accumbens. Neuron 46: 661-669.

21. Day JJ, Roitman MF, Wightman RM, Carelli RM (2007) Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nature Neuroscience 10: 1020-1028.

22. Bayer HM, Lau B, Glimcher PW (2007) Statistics of midbrain dopamine neuron spike trains in the awake primate. J Neurophysiol 98: 1428-1439.

23. Doya K, Samejima K, Katagiri KI, Kawato M (2002) Multiple model-based reinforcement learning. Neural Computation 14: 1347-1369.

24. Bertin M, Schweighofer N, Doya K (2007) Multiple model-based reinforcement learning explains dopamine neuronal activity. Neural Networks 20: 668-675.

25. Ludvig EA, Sutton RS, Kehoe EJ (2008) Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural Computation 20: 3034-3054.

26. Ludvig EA, Sutton RS, Verbeek E, Kehoe EJ (2009) A computational model of hippocampal function in trace conditioning. In: Koller D, Schuurmans D, Bengio Y, Bottou L, editors, Advances in Neural Information Processing Systems 21. pp. 993–1000.

27. Alexander GE, DeLong MR, Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annual Reviews Neuroscience 9: 357-381.

28. Strick PL, Dum RP, Picard N (1995) Macro-organization of the circuts connecting the basal ganglia with the cortical motor areas. In: Houk JC, Davis JL, Beiser DG, editors, Models of Information Processing in the Basal Ganglia, MIT Pres. pp. 117-130.

29. Haber SN, Fudge JL, McFarland NR (2000) Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. Journal of Neuroscience 20: 2369–2382.

30. Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, et al. (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nature Neuroscience 7: 887-893.

31. Schweighofer N, Bertin M, Shishida K, Okamoto Y, Tanaka SC, et al. (2008) Low-serotonin levels increase delayed reward discounting in humans. Journal of Neuroscience 28: 4528-4532.

32. Ljungberg T, Apicella P, Schultz W (1992) Responses of monkey dopamine neurons during learning of behavioral reactions. Journal of Neurophysiology 67: 145-163.

33. Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. Nature Neuroscience 1: 304-309.

34. Schultz W (1998) Predictive reward signal of dopamine neurons. Journal of Neurophysiology 80: 1-27.

35. Schultz W (2004) Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. Current Opinion in Neurobiology 14: 139-147.

36. Stephens DW, Krebs JR (1987) Foraging Theory. Princeton.

37. Redish AD, Kurth-Nelson Z (2010) Neural models of temporal discounting. In: Madden G, Bickel W, editors, Impulsivity: The Behavioral and Neurological Science of Discounting, APA books. pp. 123-158.

38. Ainslie G (1992) Picoeconomics. Cambridge Univ Press.

39. Mazur J (1997) Choice, delay, probability and conditioned reinforcement. Animal Learning and Behavior 25: 131-147.

40. Mazur JE (2001) Hyperbolic value addition and general models of animal choice. Psychological Review 108: 96-112.

41. Ainslie G (2001) Breakdown of Will. Cambridge Univ Press.

42. Madden G, Bickel W, Critchfield T, editors (in press) Impulsivity: Theory, Science, and Neuroscience of Discounting. APA books.

43. Bellman R (1958) On a routing problem. Quarterly Journal of Applied Mathematics 16: 87-90.

44. Si J, Barto AG, Powell WB, Wuncsch II D, editors (2004) Handbook of learning and approximate dynamic programming. Wiley: IEEE Press.

45. Rubin DC, Wenzel AE (1996) One hundred years of forgetting: A quantitative description of retention. Psyhcological Review 103: 734-760.

46. Rubin DC, Hinton S, Wenzel A (1999) The precise time course of retention. Journal of Experimental Psychology: Learning, Memory, and Cognition 25: 1161-1176.

47. Herrnstein RJ (1997) The Matching Law. Harvard Univ Press.

48. Madden GJ, Bickel WK, Jacobs EA (1999) Discounting of delayed rewards in opioid-dependent outpatients exponential or hyperbolic discounting functions? Experimental and Clinical Psychopharmacology 7: 284-293.

49. Read D (2001) Is time-discounting hyperbolic or subadditive? Journal of Risk and Uncertainty 23: 5-32.

50. Petry NM, Bickel WK (1998) Polydrug abuse in heroin addicts: a behavioral economic analysis. Addiction 93: 321-335.

51. Mitchell SH (1999) Measures of impulsivity in cigarette smokers and non-smokers. Psychopharmacology 146: 455-464.

52. Odum AL, Madden GJ, Bickel WK (2002) Discounting of delayed health gains and losses by current, never- and ex-smokers of cigarettes. Nicotine and Tobacco Research 4: 295-303.

53. Alessi SM, Petry NM (2003) Pathological gambling severity is associated with impulsivity in a delay discounting procedure. Behavioural Processes 64: 345-354.

54. Reynolds B (2006) A review of delay-discounting research with humans: relations to drug use and gambling. Behavioural Pharmacology 17: 651-667.

55. Shors TJ (2004) Memory traces of trace memories: neurogenesis, synaptogenesis and awareness. Trends in Neurosciences 27: 250-256.

56. Phillips PEM, Stuber GD, Heien MLAV, Wightman RM, Carelli RM (2003) Subsecond dopamine release promotes cocaine seeking. Nature 422: 614-618.

57. Roitman MF, Stuber GD, Phillips PEM, Wightman RM, Carelli RM (2004) Dopamine operates as a subsecond modulator of food seeking. Journal of Neuroscience 24: 1265-1271.

58. Pavlov I (1927) Conditioned Reflexes. Oxford Univ Press.

59. Solomon PR, Schaaf ERV, Thompson RF, Weisz DJ (1986) Hippocampus and trace conditioning of the rabbit's classically conditioned nictitating membrane response. Behavioral Neuroscience 100: 729-744.

60. Beylin AV, Gandhi CC, Wood GE, Talk AC, Matzel LD, et al. (2001) The role of the hippocampus in trace conditioning: Temporal discontinuity or task difficulty? Neurobiology of Learning and Memory 76: 447-461.

61. Thibaudeau G, Potvin O, Allen K, Dore FY, Goulet S (2007) Dorsal, ventral, and complete excitotoxic lesions of the hippocampus in rats failed to impair appetitive trace conditioning. Behavioural Brain Research 185: 9-20.

62. Levy WB, Sanyal A, Rodriguez P, Sullivan DW, Wu XB (2005) The formation of neural codes in the hippocampus: trace conditioning as a prototypical paradigm for studying the random recoding hypothesis. Biol Cybern 92: 409-426.

63. Pastalkova E, Itskov V, Amarasingham A, Buzsaki G (2008) Internally generated cell assembly sequences in the rat hippocampus. Science 321: 1322-1327.

64. Kaplan PS (1984) Bridging temporal gaps between cs and us in autoshaping: A test of a local context hypothesis. Animal Learning and Behavior 12: 142-148.

65. Self D (2003) Dopamine as chicken and egg. Nature 422: 573-574.

66. Mirenowicz J, Schultz W (1996) Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. Nature 379: 449-451.

67. Suri RE, Schultz W (2001) Temporal difference model reproduces anticipatory neural activity. Neural Computation 13: 841-862.

68. Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299: 1898–1902.

69. Mirenowicz J, Schultz W (1994) Importance of unpredictability for reward responses in primate dopamine neurons. Journal of Neurophysiology 72: 1024-1027.

70. Wörgötter F, Porr B (2005) Temporal sequence learning, prediction, and control - a review of different models and their relation to biological mechanisms. Neural Computation 17: 245-319.

71. Daw ND, Kakade S, Dayan P (2002) Opponent interactions between serotonin and dopamine. Neural Networks 15: 603-616.

72. Ainslie G, Monterosso J (2004) Behavior: A marketplace in the brain? Science 306: 421-423.

73. Sozou PD (1998) On hyperbolic discounting and uncertain hazard rates. The Royal Society London B 265: 2015-2020.

74. Kacelnik A (1997) Normative and descriptive models of decision making: time discounting and risk sensitivity. In: Bock GR, Cardew G, editors, Characterizing Human Psychological Adaptations, Chichester UK: Wiley, volume 208 of *Ciba Foundation Symposia*. pp. 51-66. Discussion 67-70.

75. Laibson DI (1996) An economic perspective on addiction and matching. Behavioral and Brain Sciences 19: 583-584.

76. McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. Science 306: 503-507.

77. Sanfey AG, Loewenstein G, McClure SM, Cohen JD (2006) Neuroeconomics: cross-currents in research on decision-making. Trends in Cognitive Sciences 10: 108-116.

78. Cardinal RN, Pennicott DR, Sugathapala CL, Robbins TW, Everitt BJ (2001) Impulsive choice induced in rats by lesion of the nucleus accumbens core. Science 292: 2499-2501.

79. Staddon JER, Cerutti DT (2003) Operant conditioning. Annual Reviews of Psychology 54: 115-144.

80. Kalenscher T, Pennartz CMA (2008) Is a bird in the hand worth two in the future? the neuroeconomics of intertemporal decision-making. Progress in Neurobiology 84: 284-315.

81. Gibbon J, Church RM, Fairhurst S, Kacelnik A (1988) Scalar expectancy theory and choice between delayed rewards. Psychological Review 95: 102-114.

82. Gallistel CR, Gibbon J (2000) Time, rate, and conditioning. Psychological Review 107: 289-344.

83. Samejima K, Doya K, Kawato M (2003) Inter-module credit assignment in modular reinforcement learning. Neural Networks 16: 985-994.

84. Wightman RM, Heien MLAV, Wassum KM, Sombers LA, Aragona BJ, et al. (2007) Dopamine release is heterogeneous within microenvironments of the rat nucleus accumbens. European Journal of Neuroscience 26: 2046-2054.

85. Levy WB (1996) A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. Hippocampus 6: 579-591.

86. Zemel RS, Dayan P, Pouget A (1998) Probabilistic interpretation of population codes. Neural Computation 10: 403-430.

87. Dayan P, Abbott LF (2001) Theoretical Neuroscience. MIT Press.

88. Jackson JC, Redish AD (2003) Detecting dynamical changes within a simulated neural ensemble using a measure of representational quality. Network: Computation in Neural Systems 14: 629-645.

89. Johnson A, Seeland KD, Redish AD (2005) Reconstruction of the postsubiculum head direction signal from neural ensembles. Hippocampus 15: 86-96.

90. Johnson A, Jackson J, Redish AD (2008) Measuring distributed properties of neural representations beyond the decoding of local variables — implications for cognition. In: Hölscher C, Munk MHJ, editors, Mechanisms of information processing in the Brain: Encoding of information in neural populations and networks, Cambridge University Press. pp. 95-119.

91. Dayan P (2002) Motivated reinforcement learning. In: Dietterich TG, Becker S, Ghahramani Z, editors, Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press.

92. Suri RE, Schultz W (1999) A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience 91: 871-890.

93. Norris JR (1997) Markov Chains. New York: Cambridge University Press.

94. Brémaud P (1999) Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues. New York: Springer.

95. Daw ND, Courville AC, Touretzky DS (2002) Dopamine and inference about timing. Proceedings of the Second International Conference on Development and Learning .

96. Niv Y, Duff MO, Dayan P (2005) Dopamine, uncertainty, and TD learning. Behavioral and Brain Functions 1: 6.

97. Badtke SJ, Duff MO (1995) Reinforcement-learning methods for continuous-time Markov decision problems. In: Tesauro G, Touretzky D, Leen T, editors, Advances in Neural Information Processing 7, MIT Press.

98. Das T, Gosavi A, Mahadevan S, Marchalleck N (1999) Solving semi-markov decision problems using average reward reinforcement learning. Management Science 45: 575-596.

99. Schultz W, Romo R, Ljungberg T, Mirenowicz J, hollerman JR, et al. (1995) Reward-related signals carried by dopamine neurons. In: Houk JC, Davis JL, Beiser DG, editors, Models of Information Processing in the Basal Ganglia, Cambridge MA: MIT Press. pp. 233-248.

100. Fiorillo CD, Tobler PN, Schultz W (2005) Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropogating TD errors. Behavioral and Brain Functions 1: 7.

101. Cheer JF, Aragona BJ, Heien MLAV, Seipel AT, Carelli RM, et al. (2007) Coordinated accumbal dopamine release and neural activity drive goal-directed behavior. Neuron 54: 237-244.

102. Mackintosh NJ (1974) The Psychology of Animal Learning. Academic Press.

103. Fiorillo CD, Newsome WT, Schultz W (2008) The temporal precision of reward prediction in dopamine neurons. Nature Neuroscience 11: 966-973.

104. Doya K (1999) What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex? Neural networks 12: 961-974.

105. Doya K (2000) Complementary roles of basal ganglia and cerebellum in learning and motor control. Current Opinion in Neurobiology 10: 732-739.

106. Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310: 1337-1340.

107. Kawato M, Samejima K (2007) Efficient reinforcement learning: computational theories, neuroscience and robotics. Current Opinion in Neurobiology 17: 205-212.

108. Alexander GE, Crutcher MD (1990) Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. Trends in Neurosciences 13: 266-271.

109. Graybiel AM, Flaherty AW, Giménez-Amaya JM (1991) Striosomes and matrisomes. In: Bernardi G, Carpenter MB, Di Chiara G, editors, The Basal Ganglia III, Plenum.

110. Vuchinich RE, Simpson CA (1998) Hyperbolic temporal discounting in social drinkers and problem drinkers. Experimental and Clinical Psychopharmacology 6: 292-305.

111. Schweighofer N, Tanaka SC, Doya K (2007) Serotonin and the evaluation of future rewards. theory, experiments, and possible neural mechanisms. Annals of the New York Academy of Sciences 1104: 289-300.

112. Tanaka SC, Schweighofer N, Asahi S, Okamoto Y, Doya K (2004) An fMRI study of the delay discounting of reward after tryptophan depletion and loading. 2: reward-expectation. Society for Neuroscience Abstracts .

113. Schweighofer N, Shishida K, Han CE, Yamawaki YOSCTS, Doya K (2006) Humans can adopt optimal discounting strategy under real-time constraints. PLoS Computational Biology 2: e152.

Figures

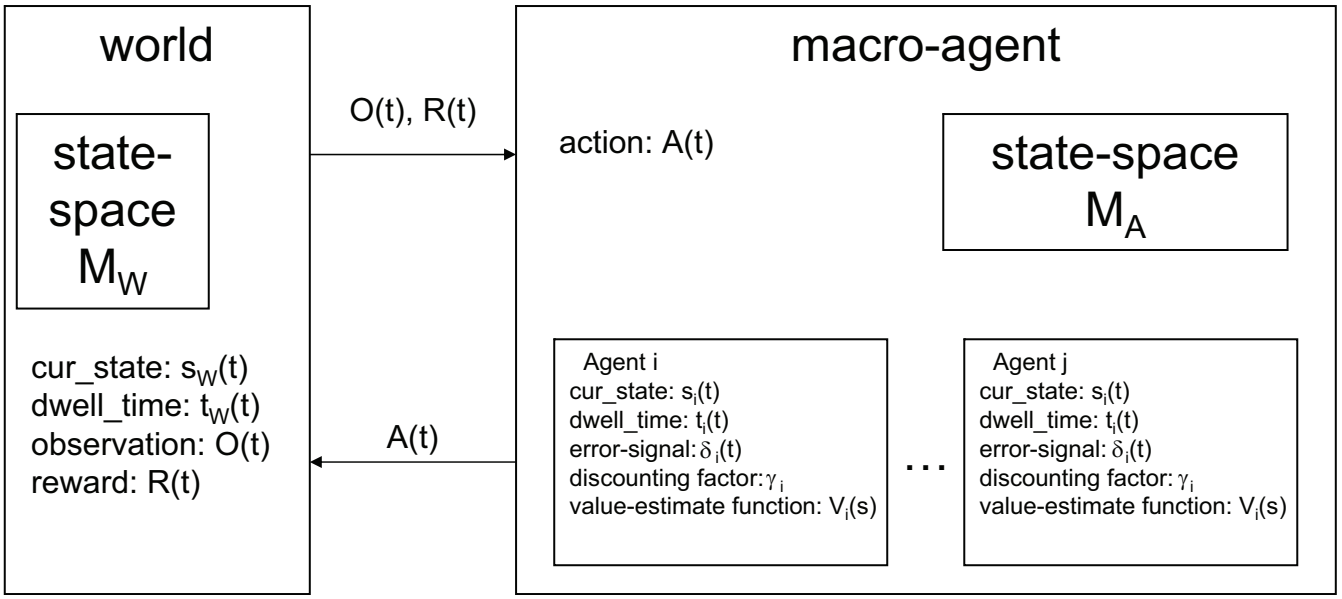
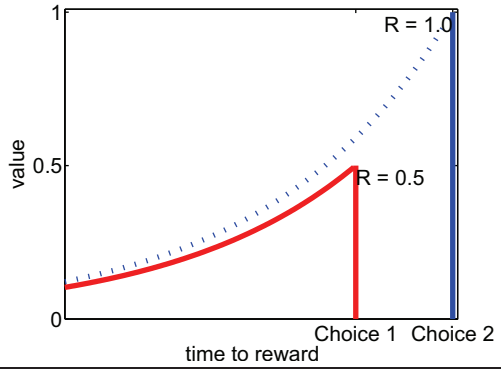


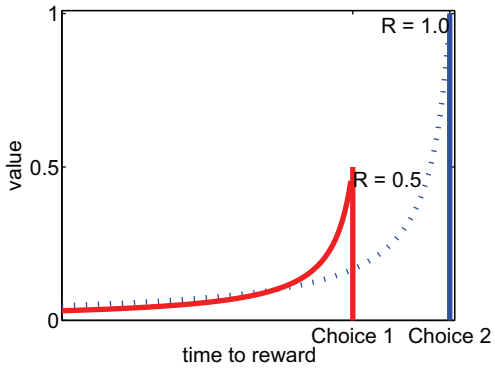
Figure 1. Model overview. The world communicates with the agent by sending observations and rewards and receiving actions. The world maintains its own "true" state and dwell time in that state. The agent is composed of independent μAgents that each maintain a belief of the world's state and dwell time. Each μAgent has its own value estimate for each state and its own discounting factor, and generates an independent δ signal. The μAgents' belief is integrated for action selection by a voting process.



With an exponential discounting function, rate of discounting does not change with time.

$$V(t) = \int_t^\infty \gamma^{\tau - t} E[R(\tau)] d\tau \qquad (11)$$

Figure shows expected value of each reward with a discounting parameter $\gamma = 0.9$.



With a hyperbolic discounting function, rate of discounting changes with time to reward.

$$V(t) = \int_t^\infty \frac{E[R(\tau)]}{1 + k(\tau - t)} d\tau \qquad (12)$$

Figure shows expected value of each reward with a discounting parameter $k = 1.0$.

Figure 2. Discounting functions. (A) Exponential discounting reduces value by a fixed percentage over any time interval. Therefore the relative preference of two future rewards does not change as the time to these rewards approaches. (B) In hyperbolic discounting, a later/larger reward may be preferred over a sooner/smaller reward until the rewards draw closer, at which point choice preference can reverse so the sooner/smaller reward is impulsively preferred. After Ainslie [38, 41].

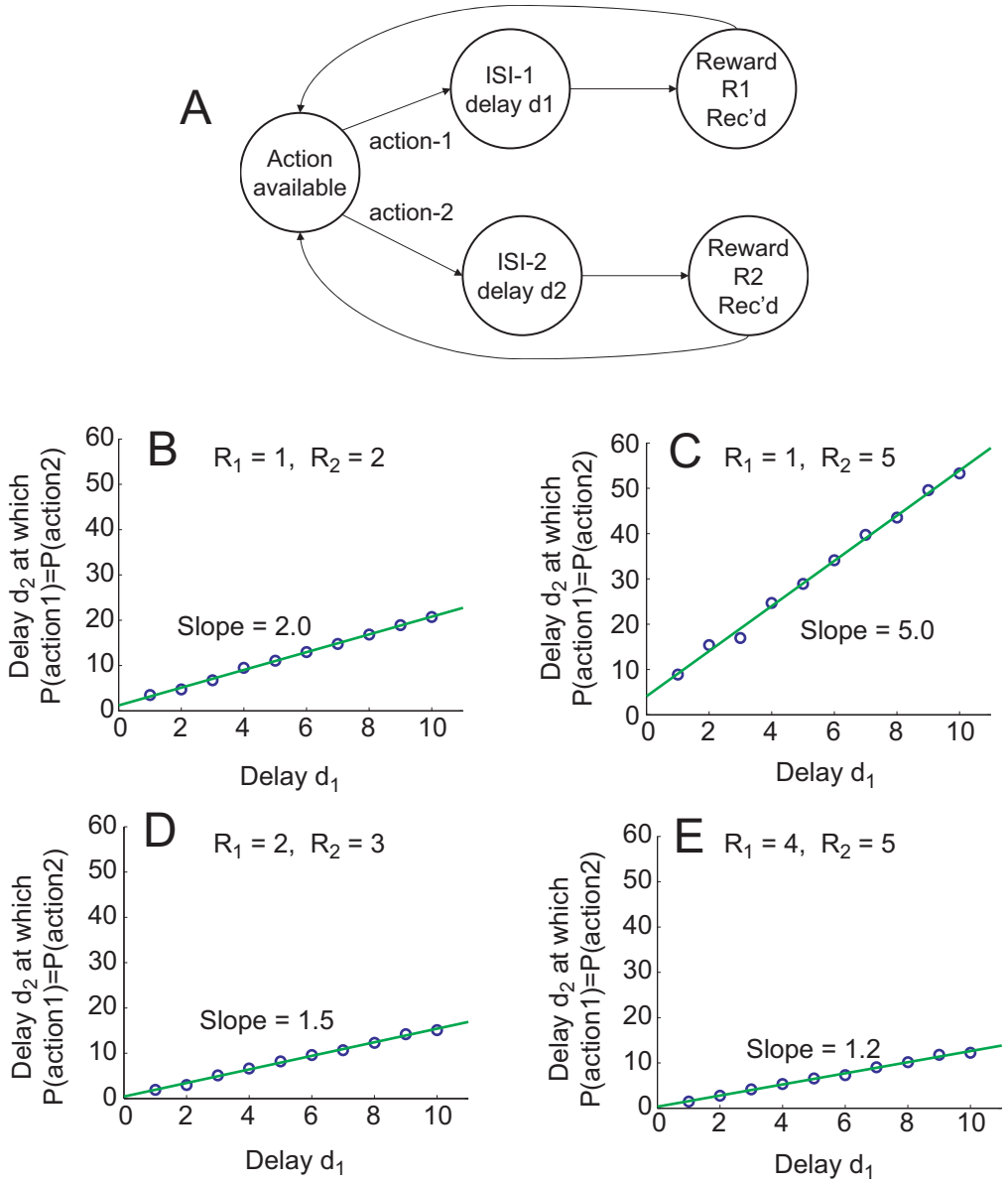


Figure 3. Hyperbolic discounting. (A) State-space used. (B-E) Mazur-plots. These plots show the delay d_2 at the *indifference point* where actions a_1 and a_2 are selected with equal frequency, as a function of the delay d_1. The ratio of actions a_1:a_2 is an observable measure of the relative values of the two choices. Blue circles represent output of the model, and green lines are least-squares fits. For hyperbolic discounting, the slope of the line will equal the ratio r_2/r_1, with a non-zero y-intercept. Compare [39, 40].

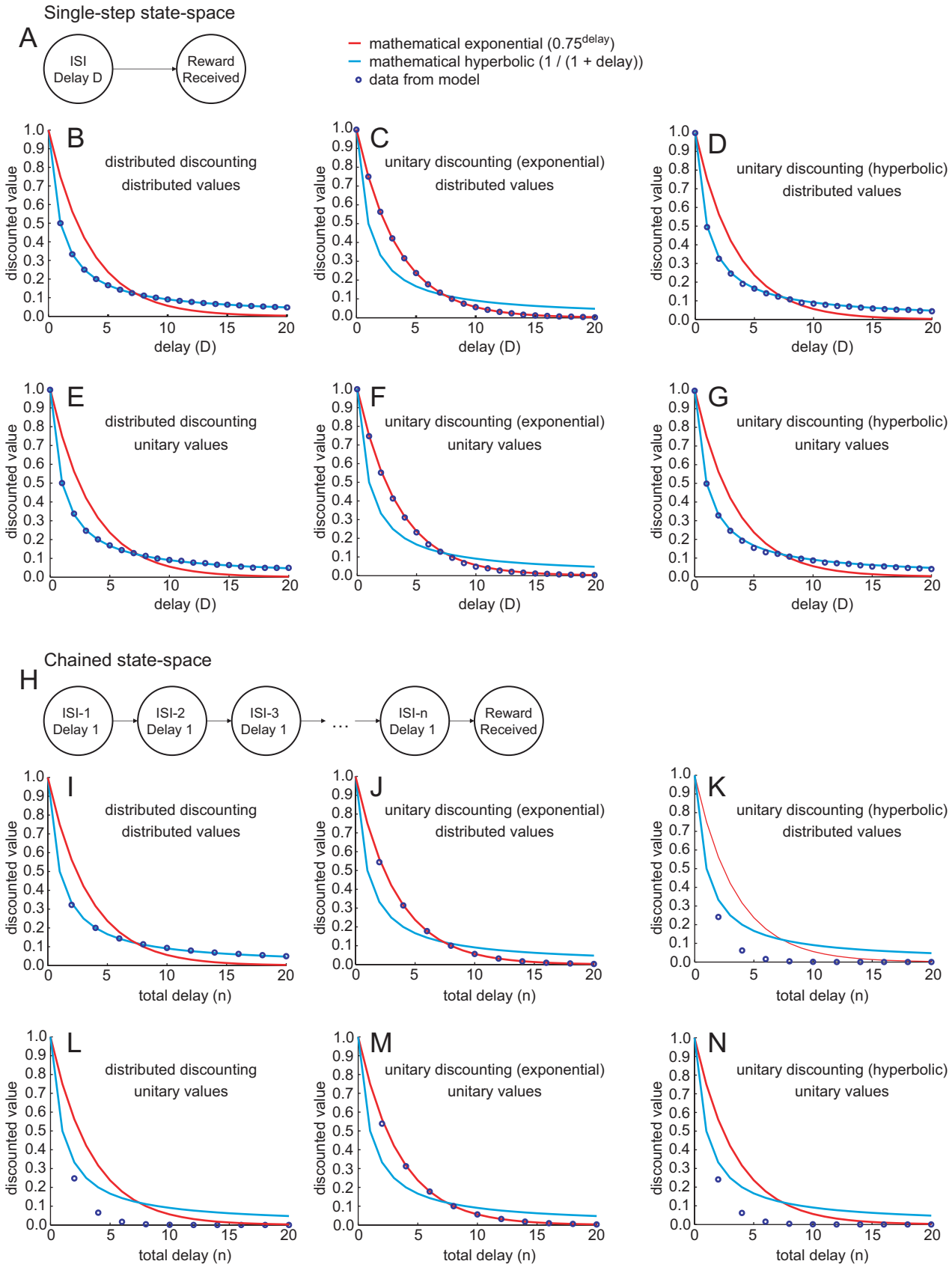

Single-step state-space
A
ISI
Delay D
Reward
Received
mathematical exponential (0.75^delay)
mathematical hyperbolic (1 / (1 + delay))
data from model
B
distributed discounting
distributed values
C
unitary discounting (exponential)
distributed values
D
unitary discounting (hyperbolic)
distributed values
E
distributed discounting
unitary values
F
unitary discounting (exponential)
unitary values
G
unitary discounting (hyperbolic)
unitary values
discounted value
delay (D)
Chained state-space
H
ISI-1
Delay 1
ISI-2
Delay 1
ISI-3
Delay 1
...
ISI-n
Delay 1
Reward
Received
I
distributed discounting
distributed values
J
unitary discounting (exponential)
distributed values
K
unitary discounting (hyperbolic)
distributed values
L
distributed discounting
unitary values
M
unitary discounting (exponential)
unitary values
N
unitary discounting (hyperbolic)
unitary values
discounted value
total delay (n)

Figure 4. Discounting across state-chains. (A) Single-step state-space used for B-G. (B,E) When the model consists of a set of exponential discounters with γ drawn uniformly from $(0, 1)$, the measured discounting closely fits the hyperbolic function. (C,F) When the model consists of a single exponential discounter with $\gamma = 0.75$, the measured discounting closely fits the function $V = 0.75^D$ (exponential). (D,G) When the model consists of a single hyperbolic discounter, the measured discounting closely fits the function $V = \frac{1}{1+D}$ (hyperbolic). (H) Chained state-space used for I-N. (I) If values are distributed so each exponential discounter has its own value representation, the result is hyperbolic discounting over a chained state space. (J,M) A single exponential discounter behaves as in the single-step state space, because multiplying exponentials gives an exponential. (K,N) A single hyperbolic discounter now behaves as an exponential discounter with $\gamma = 0.5$, because each step is discounted by $\frac{1}{1+D}$, where $D = 1$. (L) Likewise, a set of exponential discounters with shared value representation behave as an exponential discounter with $\gamma = 0.5$, for the same reason.

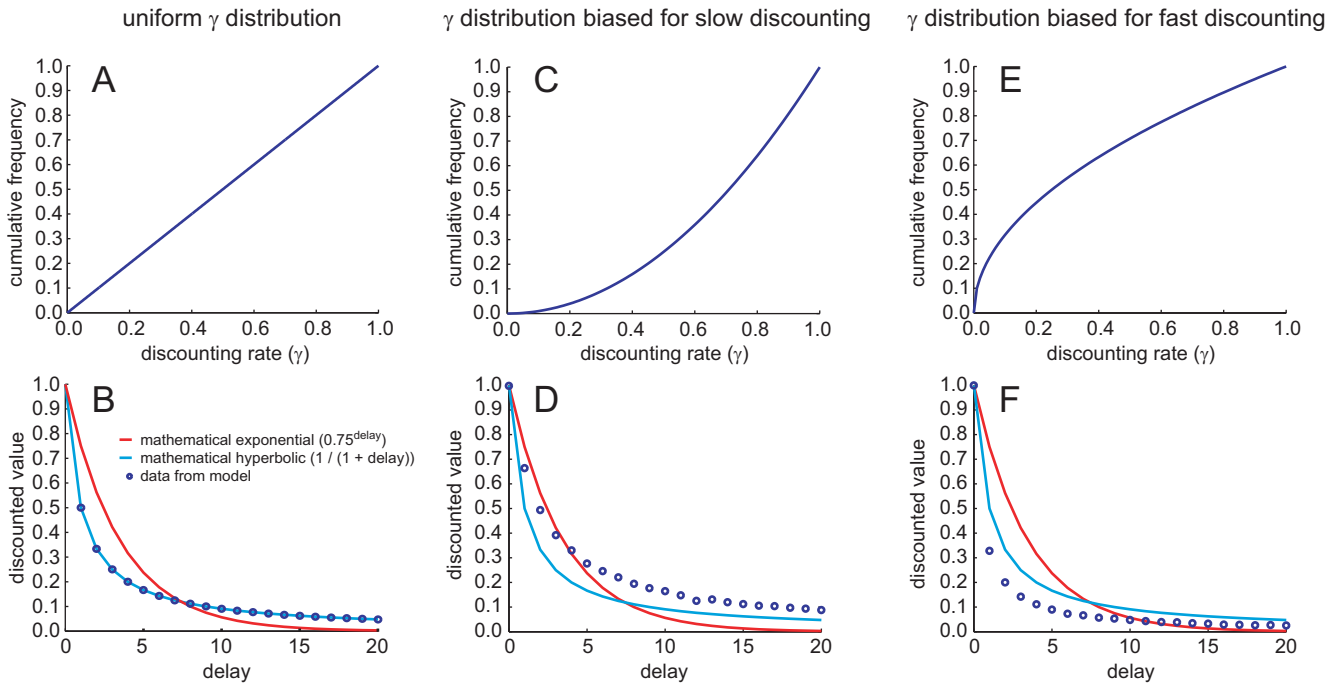


Figure 5. Rate of discounting depends on γ distribution. (A) The uniform distribution of exponential discounting rates used in all other figures. (B) As shown in Figure 4, the overall discounting is hyperbolic. (C) A distribution of exponential discounting rates containing a higher proportion of slow discounters. (D) Overall discounting is slower. (Note that it is now fit by the function $\frac{1}{1+0.5D}$.) (E) A distribution of exponential discounting rates containing a higher proportion of fast discounters. (F) Overall discounting is faster. (It is now fit by the function $\frac{1}{1+2D}$.)

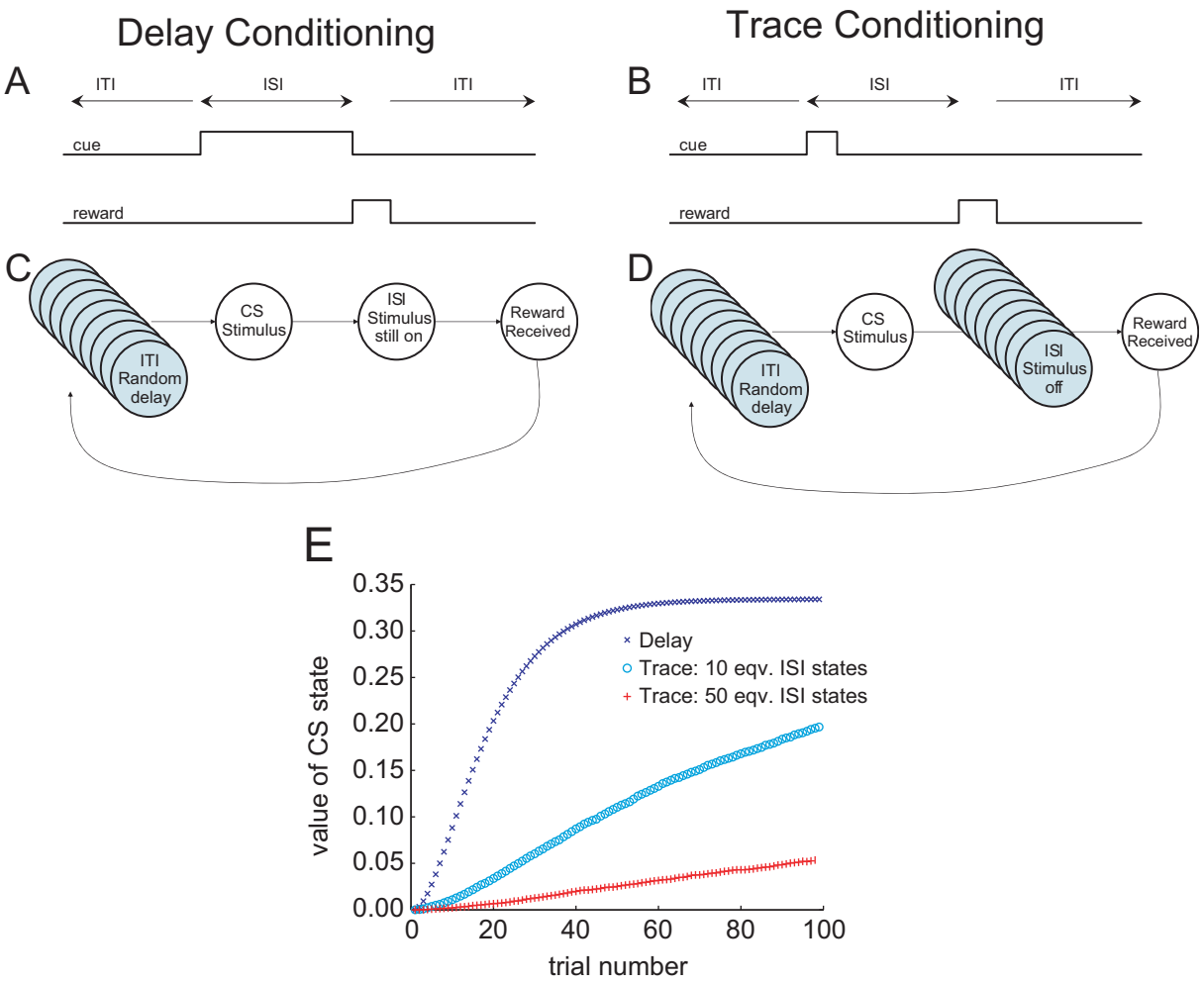


Figure 6. Trace and Delay conditioning paradigms. (A,B) Explanation of delay (A) and trace (B) conditioning. In delay conditioning, the cueing stimulus remains on until the reward appears. In trace conditioning, the cueing stimulus turns back off before the reward appears. (C,D) State spaces for delay-conditioning (C) and trace-conditioning (D). In delay conditioning, the presence of the (presumably salient) stimulus produces a single, observationally-defined state. In trace conditioning the absence of a salient stimulus produces a collection of equivalent states. (E) Simulations of trace vs. delay conditioning. Value learning at the CS state is slower under trace conditioning due to the intervening collection of equivalent states. Larger sets of equivalent states lead to slower value-growth of the CS state.

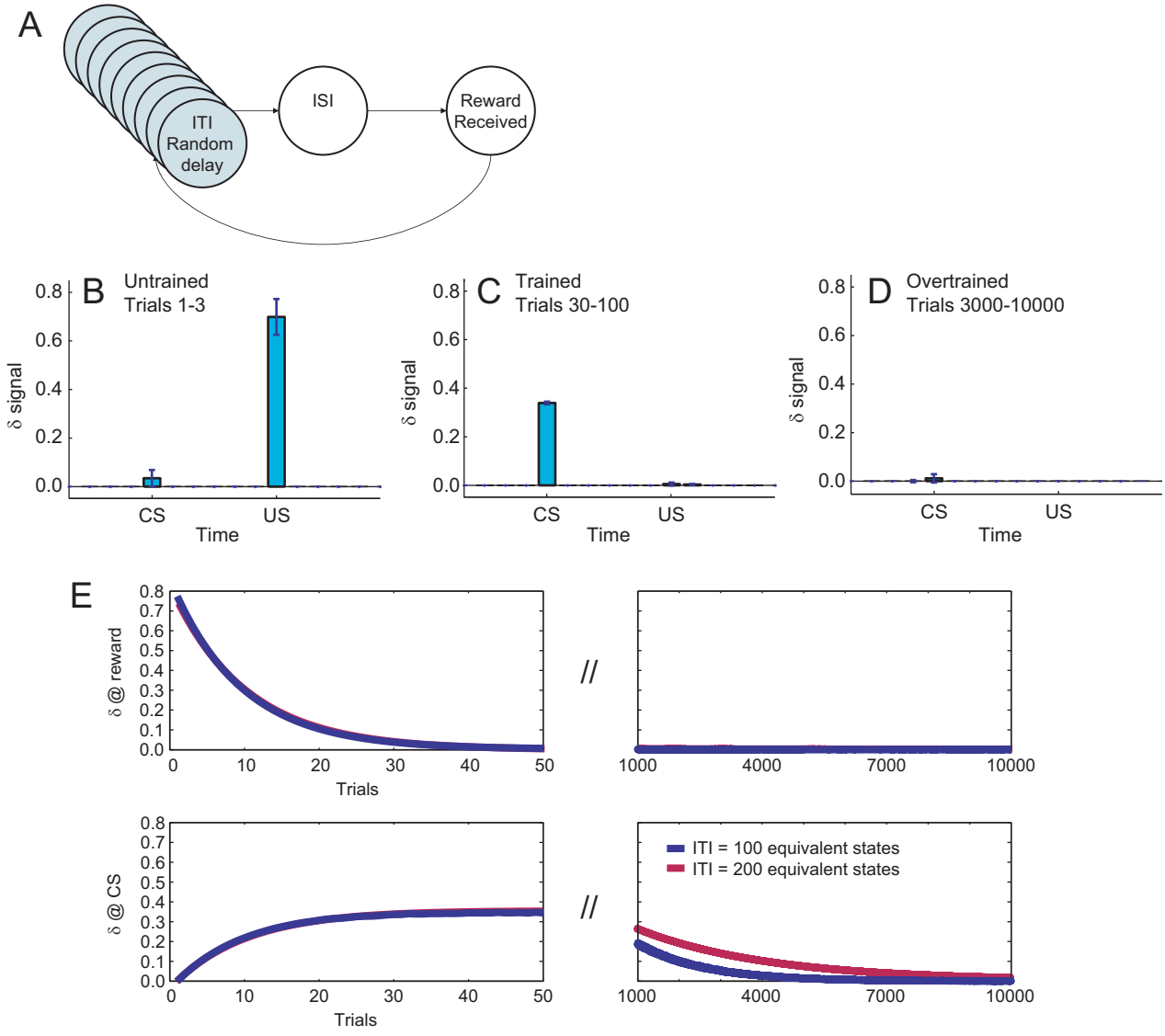


Figure 7. Effect of equivalent ITI states on δ signals at conditioned stimuli. (A) A state-space for classical conditioning. (B, C, D) Learning signaled reward delivery. (B) Untrained: δ occurs at US but not CS. (C) Trained: δ occurs at CS but not US. (D) Overtrained: δ occurs at neither CS nor US. (E-H) Transfer of value-error δ signal. Left panels show the first 50 trials, while right panels show trials 1000 to 10,000. Y-axes are to the same scale, but x-axes are compressed on the right panels. Increasing the number of equivalent ITI states increases the time to overtraining. Compare [32].

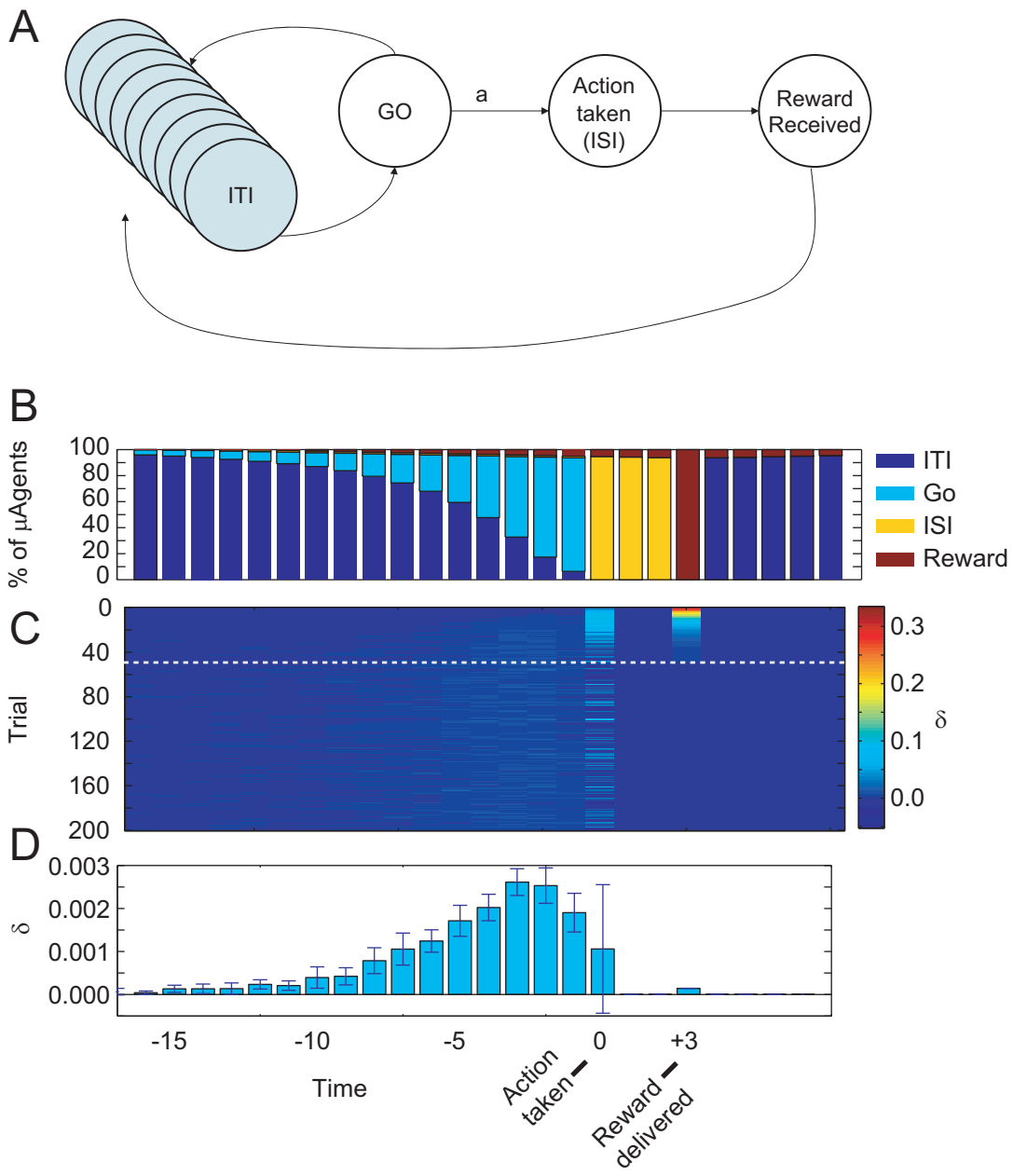


Figure 8. Modeling dopaminergic signals prior to movement. (A) State space used for simulations. The GO state has the same observation as the ITI states, but from GO an action is available. (B) Due to the expected dwell-time distribution of the ITI state, μAgents begin to transition to the GO state. When enough μAgents have their state-belief in the GO state, they select the action a, which forces a transition to the ISI state. After a fixed dwell time in the ISI state, reward is delivered and μAgents return to the ITI state. (C) As μAgents transition from ITI to GO, they generate δ signals because V(GO) > V(ITI). These probabilistic signals are visible in the time steps immediately preceding the action. Trial number is represented on the y-axis; value learning at the ISI state leads to quick decline of δ at reward. (D) Average δ signal at each time step, averaged across 10 runs, showing pre-movement δ signals. These data are averaged from trials 50-200, illustrated by the white dotted line in C. B, C, and D share the same horizontal time axis. Compare to [56].

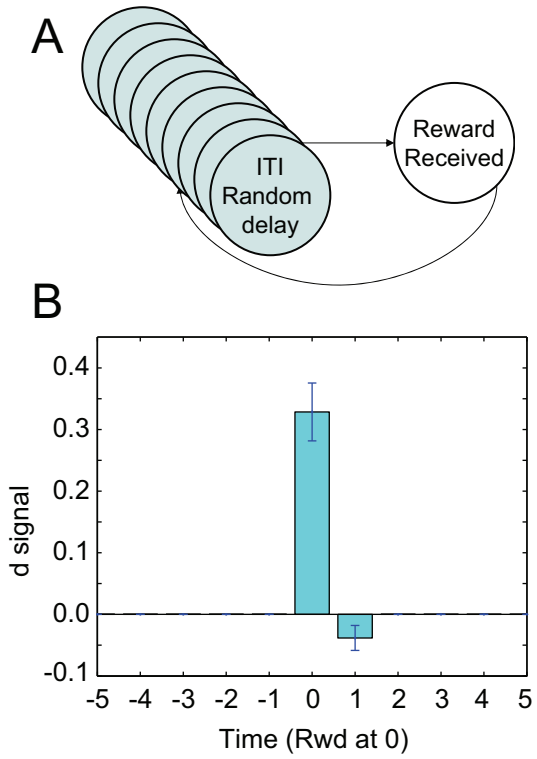


Figure 9. Unsignalled reward modulates δ. (A) State-space used for unsignaled reward. (B) δ increases at unexpected rewards.

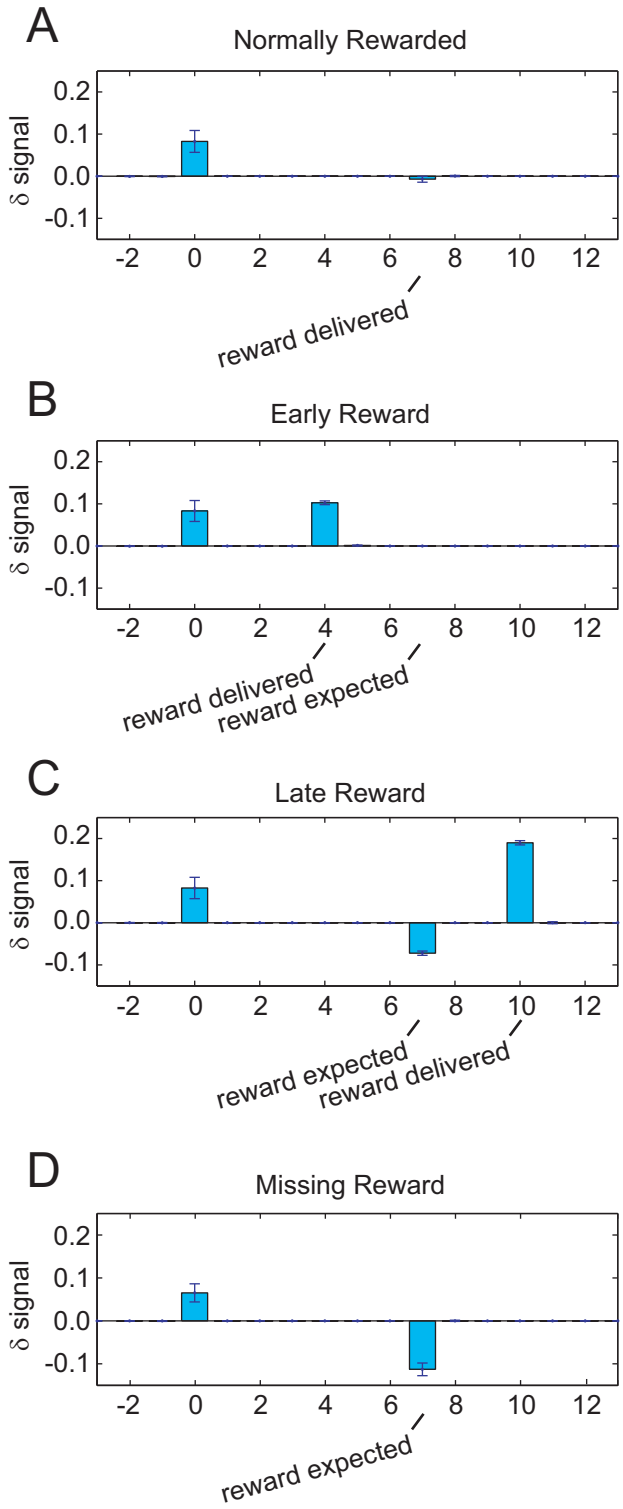


Figure 10. Early, late, and missing rewards modulate δ. (A) After training, δ is seen at CS but not US. (B) If reward is delivered early, δ appears at US. (C) If reward is delivered late, negative δ appears at the time when reward was expected, and positive δ occurs when reward is actually delivered. (D) If reward is omitted, negative δ occurs when reward was expected.

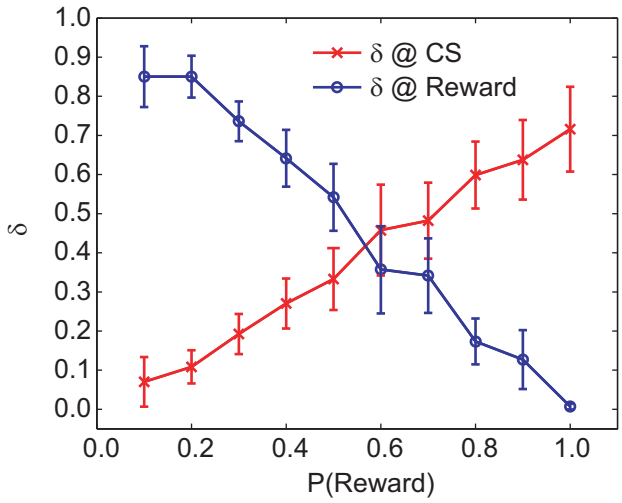


Figure 11. Probabilistic reward delivery modulates δ at CS and US. As the probability of reward drops, the δ signal shifts proportionately from the CS to the US. All measurements are taken after training for 100 trials.

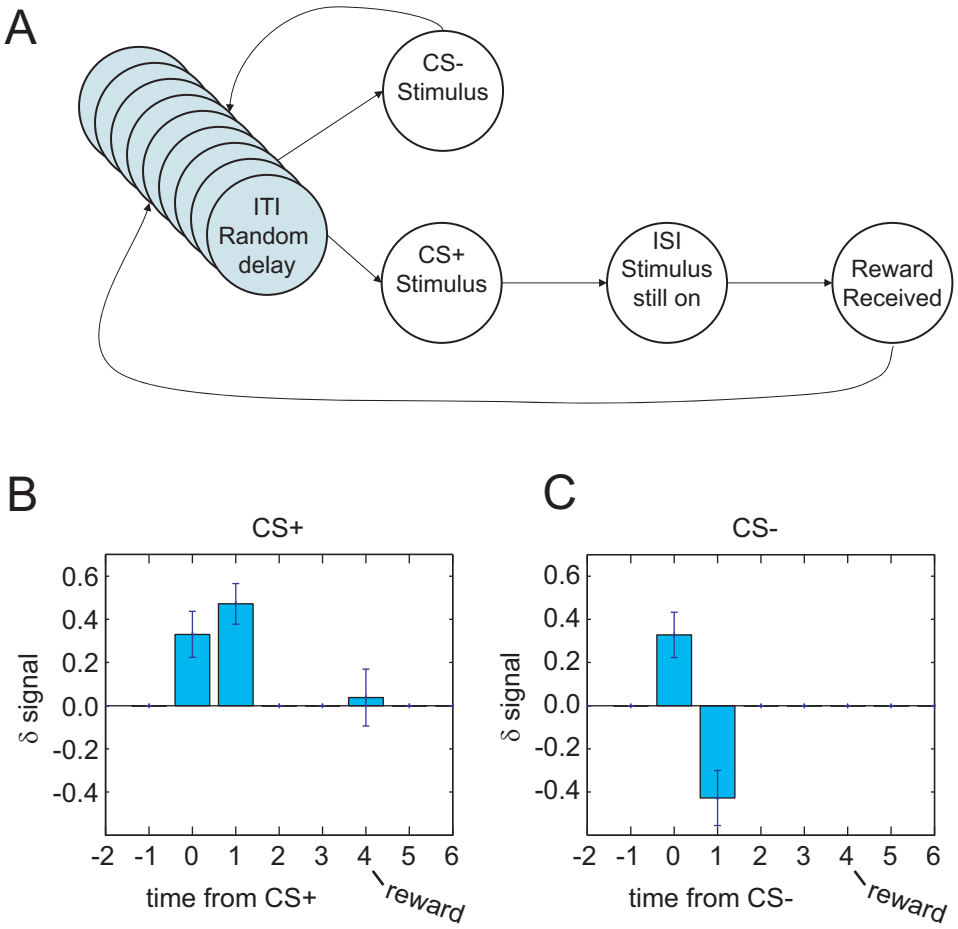


Figure 12. Effects of generalization on δ signals. (A) State-space used for measuring generalization. (B,C) Either CS+ or CS- produces a δ signal at time 0. (B) With CS+, the positive δ signal continues as µAgents transition to the ISI state, but (C) with CS-, the (incorrect) positive δ signal is counter-balanced by a negative δ correction signal when µAgents in the CS+ state are forced to transition back to the unrewarded ITI state.

Tables

Variables	
M^W	world model
$s_W(t)$	current world state
$t_W(t)$	current dwell time in $s_W(t)$
$P(O\|s)$	probability of observing observation O given state S
$P(s\|O)$	calculated from $P(O\|s)$
$O(t)$	observation passed from world to macro-agent at time t
$A(t)$	action passed from macro-agent to world at time t
γ_i	discounting factor, $\in (0, 1)$ for μAgent i
$\delta_i(t)$	value-prediction-error for μAgent i
M_i^A	μAgent world model $(= M^W)$ for μAgent i
$s_i(t)$	hypothesized state for μAgent i
$t_i(t)$	hypothesized dwell time in s_i for μAgent i
$V_i(s)$	value function for μAgent i

Parameters		
n_μ	number of μAgents	100
α	learning rate	0.1
τ	time-step compression factor	1.0
ϵ	exploration/exploitation	$1.0 * (0.95)^{[\text{N rwds}]}$

Table 1. Variables and parameters used in the simulations.

Temporal-difference reinforcement learning with distributed representations

Zeb Kurth-Nelson, A. David Redish

Appendix S1

Hyperbolic discounting from a sum of exponentials. The summed effect of these exponential discounting functions provides the overall agent with hyperbolic discounting:

$$\int_0^1 \gamma^x d\gamma = \frac{1}{1+x} \tag{13}$$

By the standard integration power law,

$$\int_0^1 \gamma^x d\gamma = \lim_{\gamma \to +0} \frac{-(\gamma^{x+1} - 1)}{x+1} \tag{14}$$

which, if $x > 0$, approaches $1/(1+x)$ as γ approaches 0 from $\gamma > 0$.