Neural Models of Temporal Discounting

A. David Redish*

Zeb Kurth-Nelson†

Acknowledgements: We thank Nathanial Daw, John Ferguson, Adam Johnson, Steve Jensen, and Matthijs van der Meer as well as Warren Bickel, Reid Landes, and Jim Kakalios for helpful discussions. We thank Adam Johnson and Steve Jensen for comments on an early draft of the manuscript.

Corresponding author address

A. David Redish
Department of Neuroscience
6-145 Jackson Hall
321 Church St. SE
University of Minnesota
Minneapolis MN 55455

* **email**: redish@umn.edu (Corresponding author)

† **email**: zeb@zebk.com

In this chapter, we address the question of temporal discounting from the perspective of computational neuroscience. We first review why agents must discount future rewards in order to make reasoned decisions and then discuss the role of temporal discounting in the context of the *temporal difference reinforcement learning* family of decision-making algorithms. These algorithms require exponential discounting functions in order to achieve mathematical stability, but as noted in the other chapters in this book, humans and other animals show hyperbolic discounting functions. In the second half of the chapter, we review four theories for this discrepancy: (1) competition between two decision-making systems, (2) interactions between multiple exponential discounting functions, (3) normalization by estimates of average reward, and (4) effects of errors in temporal perception. All four theories are likely to contribute to the effect of hyperbolic discounting.

1 Introduction

The necessity of discounting arises from the recognition of uncertainty and risk — something may happen that precludes receiving the prima facie value of a delayed reward. Technically, the value of each choice is the expected reward integrated over all future possibilities. Thus if the expected reward achieved from a decision is not going to be delivered for 24 hours, one has to integrate over all the possible events that could happen within that 24 hours, including starving to death, being hit by a bus, global thermonuclear war, money raining down from space, and all the other possibilities. While many of these possibilities are so rare as to be ignorable in the first approximation, any agent[1] attempting to actually calculate this integral would face an inordinate calculation with nearly infinite unknown variables. Additionally, this integration over future possibilities must include all the consequences of selecting an option, carried out for the infinite future. In the artificial intelligence and robotics literatures, this problem is known as the *infinite*

horizon problem (Sutton and Barto, 1998). In practice, the calculation would be extremely computationally expensive. Additionally, the calculation would require estimates of a large number of unknown variables (such as the actual probability of thermonuclear war happening in the next 24 hours). A much simpler process is to approximate the uncertainty and risk in a discounting function which reduces the value of delayed rewards. Similarly, immediately-delivered rewards are more valuable than they appear on the surface because one can invest those rewards (whether in terms of monetary investments (Frederick et al., 2002) or in terms of energy resources for increasing offspring and improving evolutionary success (Stephens and Krebs, 1987; Rogers, 1997)). As above, this could be calculated explicitly by integrating over all possible futures following the choice. Again, this is a computationally expensive calculation with many unknown variables. A much simpler process is to approximate the lost investment of waiting by a discounting function which reduces the value of delayed rewards.

Technically, any function which monotonically decreases with time (Eq. 1) will meet the primary criteria laid out above (accommodate uncertainty, risk, and lost investments).

$$\begin{aligned} &V_d\left(r\right) = f\left(r,d\right) \\ &f\left(r,d_1\right) < f\left(r,d_2\right) \text{ iff } d_1 < d_2 \end{aligned} \tag{1}$$

where $V_d\left(r\right)$ is the estimated value of receiving expected reward r after expected delay d, and *iff* indicates an *if-and-only-if* relationship. However, for many reasons, an exponential discounting function (Eq. 2) is a logically sound choice (Samuelson, 1937).

$$V_d\left(r\right) = r \cdot \gamma^{d/\tau} = r \cdot e^{-kd/\tau} \tag{2}$$

where r is the expected reward and d is the expected delay before receiving the reward; τ is a constant that defines the time-scale of the delay. The rate of exponential discounting can be expressed either in terms of a temporal constant $k > 0$ (usually used in the animal and human

discounting literature, e.g. Mazur, 1997, 2001, Ainslie, 1992, 2001, Myerson and Green, 1995, Madden et al., 1997, Bickel and Marsch, 2001) or in terms of a γ discounting factor ($0 < \gamma < 1$, usually used in the artificial intelligence and robotics literatures, e.g. Sutton and Barto, 1998, Daw, 2003). Under simple assumptions of compound interest with no uncertainty, exponential discounting is the most logical choice for a discounting function because the discounting rate is a constant over time (Samuelson, 1937; Frederick et al., 2002), however, as noted below, there is an ongoing debate as to whether exponential discounting remains a logical choice under more realistic conditions of uncertainty and measurement error (Sozou, 1998; Gallistel and Gibbon, 2000; Frederick et al., 2002). Nevertheless, because of its underlying regularity, exponential discounting allows a simple iterative calculation of value through experience, which simplifies the learning algorithms (Bellman, 1958; Sutton and Barto, 1998; Daw, 2003).

The Bellman equation. The Bellman equation is most easily seen in the discrete formulation (Sutton and Barto, 1998), but it is easily translatable into a temporally continuous formulation (Doya, 2000b; Daw, 2003; Daw et al., 2006a). Starting from an exponentially discounted value function

$$V\left(t_0\right) = \sum_{t=t_0}^{\infty} \gamma^{t-t_0} E\left(r\left(t\right)\right) \tag{3}$$

where $V(t_0)$ is the value at time t_0 (that is, the total integrated expected reward over the future from t_0). This is

$$\begin{aligned} V\left(t_0\right) &= \gamma^0 r_0 + \gamma^1 r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \cdots \\ V\left(t_0\right) &= r_0 + \gamma \cdot \left(r_1 + \gamma^1 r_2 + \gamma^2 r_3 + \cdots\right) \end{aligned} \tag{4}$$

But since value at one time step later $t_1 = t_0 + 1$ is also

$$V\left(t_1\right) = \gamma^0 r_1 + \gamma^1 r_2 + \gamma^2 r_3 + \cdots \tag{5}$$

we can rewrite value at time t_0 as a function of value at time $t_1 = t_0 + 1$.

$$V(t_0) = r_0 + \gamma V(t_1) \tag{6}$$

This provides a mechanism with which one can select actions within a given situation[2] by estimating the value of taking an action within a given situation

$$\hat{V}(s,a) = E(r) + \gamma \hat{V}(s') \tag{7}$$

where *E(r)* is the estimated reward to be received immediately upon taking action *a* in situation *s*, *s′* is the estimated new situation one expects to be in (at time *t + 1*), and $\hat{V}(s')$ the estimated value of being in situation *s′* (i.e. the maximum estimated value taken over all available actions leading from situation *s′*).

Even more importantly, this equation provides a way of updating one's estimate of value upon taking an action by calculating the *value prediction error* δ as the difference between the expected value-estimate $\hat{V}(s,a)$ and observed values (based on the actually-received observed reward and the actually identified new situation).

$$\begin{aligned} \delta(s,a) &= V(s,a) - \hat{V}(s,a) \\ &= \left(r(t) + \gamma \hat{V}(s')\right) - \hat{V}(s,a) \end{aligned} \tag{8}$$

where s' is now the actual new situation one has achieved. The value estimate $\hat{V}(s,a)$ can be easily updated from this δ term.

$$\hat{V}(s,a) \leftarrow \hat{V}(s,a) + \eta\delta \tag{9}$$

where η is a constant that controls the learning-rate.

Equations 8 and 9 can be extended to the continuous formulation easily by moving from a discrete-time state-space to a continuous-time state-space (Doya, 2000b; Daw, 2003; Daw et al., 2006a; Redish, 2004). In both the discrete and continuous models, all information about the

agent's history is assumed to be contained in the discrete state s that describes the agent's understanding of the world. In the discrete-time models, the agent is assumed to take an action a (potentially the null-action) after each discrete time step Δt, taking the agent from state $s(t)$ to state $s(t + \Delta t)$. In the continuous-time model, the agent is assumed to remain in state $s(t)$ for a given amount of time d. When the agent's hypothesis of the world changes (either through action taken by the agent or events in the world), the state changes to a new state s'. Value of a given state is identified with entry into that state (Daw, 2003), and thus the value update must take into account the time d that the agent spent in state s before transitioning to s'. Thus in the continuous-time model, Eq. 8 becomes

$$\delta(s,a) = \gamma^d \left(r_t + \hat{V}(s')\right) - \hat{V}(s,a) \tag{10}$$

where d is the time spent in situation s before taking action a to get to situation s'. Because the reward was also delayed by time d, it needs to be incorporated into the discounting factor.

Under specific conditions of stationarity, observability, and sufficient exploration, the Bellman exponential update equations can be proven to converge on the actual value of taking an action a in situation s: $V(s, a)$ (Sutton and Barto, 1998; Daw, 2003).[3]

1.1 Non-exponential discounting functions

As appealing as the exponential discounting model is, extensive evidence has shown that neither humans facing monetary decisions nor animals (including humans) facing more direct reward (e.g. food, water) decisions discount future choices with a constant discounting rate (Strotz, 1955; Mazur, 1985, 1997; Ainslie, 1992, 2001; Green and Myerson, 2004; Frederick et al., 2002). Qualitatively, experimental data show choice-reversal, and quantitatively, the data are better fit by regression to non-exponential functions. There are three methods that have been

used to measure discounting functions: questionnaires (Myerson and Green, 1995; Bickel and Marsch, 2001; Bickel et al., 2007), titrated delivery of real rewards after a delay (the *adjusting-delay assay*, Mazur, 1985, 1997, 2001; Richards et al., 1997), and correlations with decisions as they were made (Tanaka et al., 2004, 2007; Sugrue et al., 2004).

In questionnaire assays, humans are given a set of choices between receiving a set amount r_1 at a given time t_1 (often "now") with a set amount r_2 at a later time t_2 ($t_2 > t_1$, $r_2 > r_1$). From the set of choices made for a given delay t_2-t_1 at a given time t_1, it is possible to derive an *indifference point*, defined as the time t_2 at which $V_{t_1}(r_1) = V_{t_2}(r_2)$. From the set of indifference points, one can calculate the expected value of a given reward r after a given delay d (Myerson and Green, 1995; Ainslie, 1992; Bickel et al., 2007). While there have been concerns about potential confounds of real versus hypothetical choices (Kirby, 1997; Kacelnik, 1997), experiments have found qualitatively similar results under both conditions (Kirby, 1997; Johnson and Bickel, 2002). Usually, questionnaires are given in a random order and analyses are done post-experiment, but some recent experiments have used a titration method in which intervals are narrowed until the indifference point is found (Wittmann et al., 2007). This allows questionnaire techniques to achieve a block design capable of being used with fMRI. Although it is obviously impossible to directly ask animals questionnaires about hypothetical choices, it is possible to signal the values and delays of available choices before an animal acts, providing it with a questionaire-like behavior (Sohn and Lee, 2007).

In the adjusting-delay assay, agents are given two choices a_1, a_2, leading to two rewards r_1, r_2, delivered after two delays d_1, d_2. Action a_1 brings reward r_1 after delay d_1; action a_2 brings reward r_2 after delay d_2. For a given experiment, both reward (r_1, r_2) and the first delay (d_1) variables are held fixed, and delay d_2 is titrated until the probability of choosing each action is

equal: if the agent chooses action a_1, the delay d_2 is reduced on the next trial, while if the agent chooses action a_2, the delay d_2 is increased on the next trial. At the point where the two actions are chosen with equal probability, we can say that the agent's estimate of the values of the two choices are equal $V_{d_1}(r_1) = V(a_1) = V(a_2) = V_{d_2}(r_2)$. The slope of the curve of titrated d_2 delays as a function of fixed d1 delays indicates the discounting function used by the agent (Mazur, 1997). In the case of exponential discounting (Eq. 2), the slope will be 1, regardless of r_1 or r_2. In the case of hyperbolic discounting, the slope will reflect the ratio of rewards r_2/r_1 (Mazur, 1997). Experiments consistently show slopes significantly different from 1, and generally consistent with the ratio of rewards r_2/r_1 and with hyperbolic discounting (Mazur, 1985, 1997; Richards et al., 1997). Because these experiments require actual choices, actual rewards, and actual delays, these experiments are limited to fast time courses (seconds). Because these experiments are based on repeated trials, one may need to take into account the actual reward sequence achieved (or potentially available) to the animal (Daw, 2003), including the inherent variability of that sequence (Kacelnik and Bateson, 1996). Such procedures can be used in both animal (Mazur, 1997) and human (McClure et al., 2007) experiments.

The third option is to calculate the expected value from an agent given a long sequence of decision choices with a complex reward structure (e.g. Tanaka et al., 2004, 2007, or Sugrue et al., 2004). In particular, these sequences include changes in the value delivered to the agent. For example, Tanaka et al. (2004) tested subjects in an experiment in which they had to continuously alternate between a task in which the optimal solution was to select immediate rewards (SHORT condition) and a task in which the optimal solution was to select delayed rewards (LONG condition). From each subject's actual selections, Tanaka et al. calculated the estimated value (based on an exponential discounting function) at each moment in time. This function is, of

course, dependent on the discounting factor γ. This calculation gave Tanaka et al. two time series: one of the value at time *t* which was a function of the discounting factor used, and the other the fMRI BOLD signal. They then correlated the two signals to determine if there were any significant relationships between value estimates and the BOLD signal. Similar procedures have been used in animal decision-making tasks (Sugrue et al., 2004; Bayer and Glimcher, 2005; Bayer et al., 2007).

These experiments measure discounting functions at different timescales (questionnaires: days to weeks to years; titrated delay: seconds; decision choices: seconds to minutes) and with different substances (money, food, drugs). Although analogous procedures to all three experiments can be used on both humans and animals, for obvious reasons, questionnaires tend to be used with humans, while titrated delay experiments tend to be used with animals. Thus, some of the differences in timescales may be due to differences in subjects rather than the procedures themselves.

Discounting rates in humans have been found to change with both size of reward offered (e.g. $1000 vs. $10000 (Myerson and Green, 1995; Green et al., 1997; Kirby, 1997)) and with substance (e.g. food vs. money, (Odum and Rainaud, 2003; Estle et al., 2007)). Titration experiments in animals (rats, pigeons) have not found a similar effect of size of reward on discounting rate (Grace, 1999; Green et al., 2004; Ong and White, 2004). Recent evidence, in fact, has found that reward size and delay to reward receipt are encoded in different populations of neurons within the rodent orbitofrontal cortex (Roesch et al., 2006). Although experiments comparing valuation of different substances have been done in several animal species (Tremblay and Schultz, 1999; Kelley and Berridge, 2002; Padoa-Schioppa and Assad, 2006, 2008), these experiments have not directly examined the dependence of delay on valuation. However,

lexigraphic experiments in multiple species have consistently found differences in ability to inhibit responding and ability to wait (related to discounting rate) between lexigraphic rewards (in which rewards are indicated by symbols) and directly-given rewards (in which the rewards are directly visible) (Mischel and Underwood, 1974; Boysen and Berntson, 1995; Metcalfe and Mischel, 1999; Evans and Beran, 2007), which may indicate the importance of linguistic abilities for long delays (Beran et al., 1999; Metcalfe and Mischel, 1999).

It is unclear at this point whether these various paradigms access the same systems and mechanisms or whether there are systems and mechanisms specifically aligned to specific time-courses or specific rewards. However, all of the available neural models are based on the concept that all three experimental paradigms are measuring the same phenomena. Data, such as the recent fMRI data from McClure et al. (2003, 2004, 2007) and Tanaka et al. (2004, 2007) suggest that the same neural structures are involved in the discounting seen by all three methods. However, fMRI data from humans in a titrated questionnaire paradigm suggest that there may be different structures involved with medium-time-scale (< 1 year) and very long time-scale (> 1 year) discounting functions (Wittmann et al., 2007).

The changing discount function is usually modeled by a *hyperbolic* discounting function, as suggested by Ainslie (1975, 1992, 2001) and Mazur (1985, 1997)

$$V_d(r) = \frac{r}{1 + kd} \tag{11}$$

where r is the expected reward and d is the expected delay before receiving the reward.[4] This function fits the animal experimental data at fast time scales (seconds) significantly better than exponential functions (Mazur, 1985, 1997) and has been found to explain a large percentage of the variance as evidenced from questionnaires (addressing long time scales, days to weeks to years, Myerson and Green, 1995; Madden et al., 1997; Reynolds, 2006; Bickel et al., 2007).

However, there is some deviation of the animal experimental data from Eq. 11. Similarly, indifference points measured by questionnaires show consistent deviations from Eq. 11, particularly, at longer time-scales (Myerson and Green, 1995; Madden et al., 1999; Mitchell, 1999; Reynolds, 2006; Bickel et al., 2007).

While the issue of whether a hyperbolic discounting function is a more valid normative accounting of decision making than an exponential function is still being debated (Ainslie, 1992, 2001; Kacelnik, 1997; Rogers, 1997; Sozou, 1998; Frederick et al., 2002), there is little doubt that it is a more valid descriptive account than an exponential function (Myerson and Green, 1995; Mazur, 1997; Bickel et al., 2007). Although there are still some researchers who argue that the hyperbolic discounting is a consequence of the specific research methods designed to study the question (e.g. Rubenstein, 2003), if animals (including humans) did, in fact, use an exponential discounting function to discount future choices, one would still require an explanation for choice reversal. Any non-exponential discounting function must produce changing choices with changing delays — a decision which prefers option B delayed by two weeks over option A delayed by one week can switch when option A is offered immediately and option B offered in a week (Strotz, 1955; Ainslie, 1992, 2001; Frederick et al., 2002; Ainslie and Monterosso, 2004). Because the discount rate for exponential discounting does not change with time, choice reversal cannot occur. However, delay-dependent choice reversal is well-established at all time-scales (Mazur, 1997; Ainslie, 1992, 2001; Bickel et al., 2007).

A number of other functions have also been proposed (see Rodriguez and Logue, 1988, for review), most notably the "extended hyperbolic" equation

$$V_d(r) = \frac{r}{(1+kd)^b} \tag{12}$$

where b is an additional constant, which Myerson and Green and colleagues (Myerson and

Green, 1995; Green and Myerson, 2004; Green et al., 2005) report provides a better fit to the data than Eq. 11. Including the *b* term generalizes the standard hyperbolic discounting function to a more general power law. Whether another function can better describe the data remains an open question.

Unfortunately, hyperbolic discounting has several computational difficulties. First, because the discounting rate changes with each time step, there is no analytical solution to Equation 11, nor can the calculation be performed incrementally analogous to the Bellman equation (Daw, 2003). One can substitute the hyperbolic discounting function into the Bellman equation (Eq. 10) anyway

$$\delta(s,a) = \frac{\left(r(t) + \hat{V}(s')\right)}{1+kd} - \hat{V}(s,a) \qquad (13)$$

where *k* is the discounting factor and *d* is the time spent in situation *s* before taking action *a*. This equation is equivalent to that used in the addiction simulations of Redish (2004, see also Kurth-Nelson and Redish (2004)). A similar proposal has been made recently by Kalenscher and Pennartz (2008). Action-selection based on this equation leads to generally reasonable behavior (unpublished data, Kurth-Nelson and Redish), but this equation is intrinsically inconsistent, because the discounting rate depends on the number of subparts identified within a task. A situation identified as a single part (situation s_0 proceeds to situation s') that lasts for a given time before an action is taken is discounted hyperbolically, but if the same situation is identified by a set of subparts (situation s_0 leads to situation s_1 leading to situation s_2 ... eventually leading to situation s'), then the discounting function deviates from the predicted hyperbolic function (Eq. 11) dramatically.

For example, take a Pavlovian experiment (with no actual choices being made), in which a cue is followed some set number of seconds later by a reward. At the appearance of the cue, the

animal can predict the subsequent appearance of a cue. The expected value of that cue should take into account the delay before that cue. If the neural representation encodes this as two situations (an interstimulus interval [*ISI*] lasting for the set number of seconds followed by a reward-received state, Daw, 2003; Redish, 2004; Daw et al., 2006a), Equation 13 only performs a single step of hyperbolic discounting. In contrast, if the neural representation encodes each second as a different situation (ISI-1, ISI-2, ISI-3, etc., Daw (2003); Niv et al. (2005)), then Equation 13 runs through ten steps. Since a composition of hyperbolic terms is no longer hyperbolic, the effective valuation of the first state is no longer hyperbolic in time. Whether dopamine signals in the primate brain imply that time is encoded by two-situation or by chained-state-space representations is still debated (Fiorillo et al., 2005; Niv et al., 2005; Wörgötter and Porr, 2005).

Simulations demonstrating this effect are shown in Figure 1 which compares a simulated agent that remains in the inter-stimulus interval *ISI* situation through the entire delay before transitioning to a reward-delivery situation (panels A,B), and another in which the inter-stimulus interval is represented by multiple one-second sub-situations (panels C,D). If the temporal difference reinforcement learning algorithm is implemented directly with Eq. 13, then the agent shows hyperbolic discounting across the first state-space, but not the second. The multiple-exponentials model (see below) shows hyperbolic discounting across both by maintaining multiple independent exponential discounting "micro"-agents. Each micro-agent (μAgent) applies a complete temporal-difference reinforcement learning agent (Sutton and Barto, 1998; Daw, 2003), with independent situation s_i, value-estimation[5] $\hat{V}_i(s,a)$, value-prediction error (δ_i), and action-selection components (Kurth-Nelson and Redish, 2004; Redish, 2004).

This analysis shows that the one-step hyperbolic equation (Eq. 13) is inconsistent. Different

conceptual representations of the time during an inter-stimulus interval produces different discounting functions. This means that if temporal difference reinforcement learning algorithms are implemented with a one-step hyperbolic equation (Eq. 13), it may be possible to change the discounting function by providing more or less information during ISI delays (which may drive a subject to represent the intervening interval by a different number of sub-intervals). Whether real discounting functions used by real animals are actually subadditive remains a point of debate (Read, 2001).

[Figure 1 about here.]

2 Neural models

Because hyperbolic discounting functions are computationally difficult to work with, several neural models have been proposed that use computationally tractable mathematics internally but show behavioral discounting rates that change with time. We will review four of these models and the data supporting each in turn.

2.1 Normalization by estimates of average reward

Following the rate conditioning literature (see Gallistel, 1990, for review), and the observations by Kacelnik (1997) that applying discounting factors in the titrated delay task (e.g. Mazur, 1997) ignores the effects of rewards expected in future trials, Daw (2003, see also Daw and Touretzky, 2000 and Daw et al., 2006a) has suggested an alternative to the discounting concept based on the concept of average reward. This model assumes that agents have evolved to optimize the total expected reward, integrated over many multiple trials (Stephens and Krebs, 1987; Kacelnik, 1997). Kacelnik (1997) notes that Mazur's titration experiments are, in fact, repeated trials, and notes that humans answering questionnaires (e.g. Myerson and Green, 1995;

Bickel and Marsch, 2001; Bickel et al., 2007) may be treating the choices as elements of a system considering time in terms of repeated trials (but see Mazur, 2001, 2006). In this model, decisions are assumed to be made based on the *rate of reward* rather than as a single decision between two immediate values (Gallistel, 1990; Kacelnik, 1997; Daw and Touretzky, 2000). In this model (see Daw, 2003), value estimates are updated by

$$\delta = r(t) - \rho(t) + \hat{V}\left(s(t+1)\right) - \hat{V}\left(s(t)\right) \tag{14}$$

which maximizes the function

$$V(t_0) = \sum_{t=t_0}^{\infty}\left[r(t) - \rho(t)\right] \tag{15}$$

where ρ(t) is the estimate of the average reward available to the animal over long time scales.

[Figure 2 about here.]

The problem with these models is that titration experiments have shown that when the inter-trial-interval is increased after the small reward (thus matching the total time between small rewards and between large rewards), animals can still show impulsive choices (Mazur, 2001, 2006). One possible explanation for this is that animals ignore the inter-trial-interval and only make decisions based on the time between the cueing stimulus and the reward (Stephens and Krebs, 1987; Gallistel and Gibbon, 2000; Daw, 2003). Daw and Touretzky have shown that an average decay model that takes into account only the time between the cueing stimulus and the reward can show hyperbolic-like behavior (Figure 2).

2.2 Temporal perception

Daw (2003, p. 98) notes that there is a strong relationship between an exponential discounting factor γ and the agent's perception of the delay *d*. A similar proposal has been made by Staddon and Cerutti (2003, see also Kalenscher and Pennartz, 2008) that hyperbolic-like

discounting can arise from timing errors due to Weber's law applied to timing.

Because the discounting applied to a given delay depends not on the actual delay, but rather on the perceived delay, variability in delay perception combined with a set exponential discounting function would be mathematically equivalent to a distribution of exponentials (Daw, 2003; Staddon and Cerutti, 2003), which would lead to an approximation of hyperbolic discounting (Kacelnik, 1997; Sozou, 1998, or at least to a power-law Staddon et al., 2002). A number of researchers have noted that the perceived delay $\hat{d}$ of an actual delay d follows a Gaussian distribution with mean d and standard deviation proportional to d (Gibbon et al., 1988; Gallistel and Gibbon, 2000; Staddon and Cerutti, 2003). This is, of course, the expectation of the distribution that would arise if delay perception were driven by a clock firing with Poisson statistics (Gallistel and Gibbon, 2000).[6] Daw (2003) has shown that this simple assumption about perceived delays leads to indifference functions compatible with those found by Mazur (1997, 2001). See Figure 3.

As noted by Daw (2003), there is a duality between the exponential discounting rate γ and the delay to reward receipt d (see Equation 2). A uniform distribution of discounting rates $\gamma \in (0,1)$ (which produces a hyperbolic discounting function when summed) can be rewritten as an (admittedly complex) distribution of delays. However, even within this model, slight differences from hyperbolic discounting are seen at very small and very large delays (Daw, 2003, see Figure 3). It is not yet known whether those differences occur in actual subjects. Nor is it yet known whether the actual errors in delay estimation (producing a distribution of delays over trials) are compatible with the complex functions needed to produce realistic discounting functions.

If this delay perception hypothesis were true, then one would expect to see hyperbolic

functions arising in other delay tasks, such as in memory recall tasks. Hyperbolically decreasing functions are better fits to memory recall probabilities than exponentially decreasing functions (Rubin and Wenzel, 1996; Rubin et al., 1999; Wixted and Ebbesen, 1997); power laws and sums of exponentials provide even better fits than hyperbolic functions (Wixted and Ebbesen, 1997; Rubin et al., 1999). The possibility that the power laws that fit the memory recall data may arise from different forgetting factors (mathematically equivalent to discounting factors) in different subjects has been addressed (Anderson and Tweney, 1997). Even when looking at individuals, power laws are better fits to the memory recall data than exponential functions (Wixted and Ebbesen, 1997). Staddon and Higa (1999) explicitly proposed a theory in which interval timing arises from multiple components, which when added together produce non-exponential timing (and thus discounting) functions.

An interesting consequence of the temporal-perception hypothesis would be that agents with faster discounting functions (such as addicts) would have show a similar over-emphasis on local time-perception preferences. When asked to speculate about their future, normal subjects included details ranging up to 4.5 years in the future. In contrast, addicts only included details about the next several days (Petry et al., 1998).

[Figure 3 about here.]

2.3 Competition between two systems

Even some of the early economics literature suggested that the observed changing discounting rate with time may be due to interactions between two systems, each with different discount preferences. Generally, these proposals have entailed an *impulsive* subsystem preferring immediate rewards and a second subsystem willing to delay gratification.

Mathematically, this has been primarily studied from the perspective of the $\beta\delta$ hypothesis

(Laibson, 1996; McClure et al., 2004, 2007) in which discounting is assumed to be

$$V_d(r) = r\beta\delta^d \tag{16}$$

where r is the expected reward and d is the expected delay before receiving the reward. β encodes the willingness to wait for later rewards (i.e. $1/\beta$ encodes the impulsivity), while δ is an exponential discounting component (equivalent to γ, Eq. 2, above). McClure et al. (2007) notes that Equation 16 can be decomposed into two components:

$$V(t_0) = \left(\frac{1}{\beta} - 1\right) r(t_0) + \sum_{t=t_0}^{\infty} \delta^t r(t_0 + t) \tag{17}$$

The first component (the β-system, impulsive) emphasizes immediate rewards ($r(t0)$), while the second component (the δ-system) shows exponential discounting (compare Eq. 2). This model essentially entails an exponential discounting function with an added (impulsive) preference for immediate rewards. This would predict that an agent that could ignore or inactivate the impulsive system would show exponential discounting.

While many questionnaire experiments are based on a choice of immediate rewards versus delayed reward, Wittmann et al. (2007) used a task in which both options entailed reward-receipt after delays and found equivalent hyperbolic discounting functions. Green et al. (2005) have found that their more general hyperbolic equation (Eq. 12) fits both situations in which an immediate reward is contrasted with a delayed reward and situations in which one delayed reward is contrasted with a later more-delayed reward. While many animal delay reward experiments are based on the choice between immediate and delayed reward (e.g. Cardinal et al., 2001), the classic titrated experiments of Mazur (1985, 1997) are based on situations in which both rewards are delayed. The βδ equation cannot accommodate the non-exponential discounting seen in these paired delay experiments.

However, the fundamental hypothesis that changing discount rates are due to competing neural systems is more general than Equation 16. All that is required is that discounting can be written as the sum of two functions

$$V_d(r) = \alpha \cdot f_0(r,d) + (1-\alpha) \cdot f_1(r,d) \qquad (18)$$

where f_0 has a fast (impulsive) decay function and f_1 is slower (more willing to wait); α controls the balance between the two. The underlying neural hypothesis is that these two discounting functions arise from different neural structures competing for behavioral control.

fMRI data has found positive correlations between hemodynamic activity (the BOLD signal) in specific structures (including ventral striatum, medial orbitofrontal cortex, medial prefrontal cortex, and posterior cingulate cortex) and the availability of imminent rewards (McClure et al., 2004). While direct correlations between other structures and longer delays were not found, McClure et al. (2004, 2007) suggest that the "δ-system" may be engaged in all conditions, while the impulsive system is only engaged when immediate rewards are selected. Supporting this, they found that decisions were related to the ratio of hemodynamic activity in other structures (such as lateral prefrontal cortex) and the impulsive-related structures listed above (McClure et al., 2004). Whether this is due to lack of activity in "impulsive" structures or increased activity in delay-preferring structures is unknown, but does suggest a competition between the two systems. McClure et al. (2007) have recently extended these results to direct (e.g. juice) rewards with actual delays on the order of minutes rather than hypothetical delays on the order of days and found similar structures involved in the impulsive (β) component (anterior cingulate cortex, nucleus accumbens [ventral striatum], medial orbitofrontal cortex). Different structures were found to be involved in delayed rewards, including lateral frontal cortical structures and posterior parietal structures. These results imply a competition between neural subsystems, one of which

drives a preference for immediate, impulsive choices and one of which drives a willingness to wait for delayed rewards.

[Figure 4 about here.]

Lesion data has also provided support for a competition between systems hypothesis, but, again, which structures are involved in which systems is still unclear. For example, lesions of the ventral prefrontal cortex are correlated with an increase in impulsive decisions, particularly in cases of reversals and developing negative consequences (Grant et al., 2000; Bechara, 2005; Bechara and van der Linden, 2005; Torregrossa et al., 2008, however, see also Clark et al., 2003). However, the lack of consensus on definitions of substructure within the ventral prefrontal cortex has made the comparison of studies difficult (Barbas, 2007; Murray et al., 2007; Price, 2007). Lesions of the ventral striatum have been found to make animals more impulsive and less capable of delaying responses to receive rewards (Cardinal et al., 2001). Interestingly, there is very little data in which lesions drive animals to be less impulsive. What little data there is suggests a role of orbitofrontal cortex in re-evaluating the discounted delayed rewards (Winstanley et al., 2004), particularly the likelihood of its delivery (Mobini et al., 2002), which may suggest a role for orbitofrontal cortex in between-session changes and reversals (Murray et al., 2007; Schoenbaum and Shaham, 2008; Torregrossa et al., 2008). Neural recordings from orbitofrontal cortex suggest that some OFC neurons signal the discounted value of rewards (Roesch et al., 2007), and anticipate future rewards (Ramus et al., 2007). These neurons can change their responses under changing reward conditions (Tremblay and Schultz, 1999; Schoenbaum et al., 2006; Padoa-Schioppa and Assad, 2008).

Many neuroscientists have suggested that these two systems may reflect two more general decision-making systems, one of which (often referred to as the *impulsive* system) reacts quickly

to specific stimuli while the other is capable of considering longer-term possibilities (see O'Keefe and Nadel, 1978, Squire, 1987, Metcalfe and Mischel, 1999, Redish, 1999, Poldrack and Packard, 2003, Cardinal, 2006, and Redish and Johnson, 2007, for reviews). Bernheim and Rangel (2004) explicitly suggested that agents switch between two modes ("hot" and "cold"), in which agents reacted to the highest value most immediately available reward when under the influence of the "hot" system but considered consequences (under appropriate slow discounting functions) when under the influence of the "cold" system (see also Metcalfe and Mischel, 1999). In Bernheim and Rangel's model, the presence of drug-related cues forced an agent into the "hot" mode. Many experiments have shown that when faced with drug-related cues, addicts become highly impulsive and unable to inhibit drug-related responses (Tiffany, 1990; Lubman et al., 2004; Noël et al., 2007)

One of the few papers to build a working model of the two systems is that of Daw et al. (2005), in which the impulsive system is assumed to be a slowly-learned "habit" system in which values are stored and only changed with experience, while the other (cognitive) system is a flexible "planning" system in which values are calculated on the fly from expectations. Daw et al. suggest that which system controls behavior is dependent on underlying uncertainty — the more uncertain a situation is, the more the agent should rely on the flexible, cognitive system. Although not phrased in terms of impulsivity (the cached-value system Daw et al. model also incorporates a slow discounting factor), the two systems in this model react very differently to changes in reward delivery probabilities — the cached-value system can only react slowly (if at all), while the planning system is more flexible. However, neither Bernheim and Rangel (2004) nor Daw et al. (2005) consider whether the average behavior of such an agent would match the discounting functions seen in the human or animal literatures.

This literature is related to the very large impulsivity (Evenden, 1999; Zermatten et al., 2005; Glimcher et al., 2007; Torregrossa et al., 2008) and behavioral response inhibition literature (Gray, 1982a,b; Gray and McNaughton, 2000). In response inhibition experiments, a subject is faced with one simulus (S1) after which taking an action (*go*) leads to reward and a second, similar stimulus (S2) after which not taking that action (*no-go*) leads to reward. Because S1 is shown much more often than S2, the subject expects S1 and prepares for S1. In order to get reward after S2, the subject has to inhibit the prepared response. Response inhibition is now known to require the anterior cingulate cortex (Braver et al., 2001; Botvinick et al., 1999; Walton et al., 2007; Rushworth et al., 2007) and other aspects of frontal cortices (such as the supplementary motor area, Isoda and Hikosaka, 2007, and the dorsomedial prefrontal cortex, Brass and Haggard, 2007). Anterior cingulate cortex is currently thought to monitor conflict (Amiez et al., 2005) or to integrate historical trends (Walton et al., 2007; Rushworth et al., 2004; Kennerley et al., 2006), while supplementary motor, dorsomedial prefrontal, and ventral frontal cortices override prepotent actions stored in direct sensory-motor connections (Okano and Tanji, 1987; Crutcher and Alexander, 1990; Tanji, 2001; Rushworth et al., 2004; Bechara, 2005; Bechara and van der Linden, 2005; Brass and Haggard, 2007; Chamberlain and Sahakian, 2007; Isoda and Hikosaka, 2007). Response inhibition can be envisioned as a flexible system overriding an impulsive, more habitual system (Gray and McNaughton, 2000; Daw et al., 2005; Redish et al., ress).

While there is strong evidence for a competition between systems, it is not completely clear which structures are involved in which systems. This may be due, in part, to the available resolution in fMRI, lesion, and recording experiments.

2.4 Multiple exponential discounting systems

While the sum of two exponential discounting functions leads to changing discount rates with delay and thus to preference reversals (Laibson, 1996; McClure et al., 2004, 2007), it does not closely approximate the hyperbolic discounting function (Eq. 11) reported in much of the literature (e.g. Ainslie, 1992; Mazur, 1985, 1997; Vuchinich and Simpson, 1998). A larger, uniform distribution of exponential distributions would, however, match the hyperbolic discounting seen experimentally (Kacelnik, 1997; Sozou, 1998; Daw, 2003; Kurth-Nelson and Redish, 2004; Redish, 2004). We (Kurth-Nelson and Redish, 2004, see also Redish, 2004) recently suggested a model in which multiple "micro-agents" compete to make decisions. Each of these μAgents instantiates a hypothesis about the state of the world (the current situation s_i and the time t_i spent within that situation), maintains a value estimate of that state $\hat{V}_i(s,a)$, and independently carries out an individual temporal difference reinforcement learning algorithm (thus requiring an individual value-prediction-error-term δ_i, Sutton and Barto, 1998; Bertin et al., 2007) with exponential discounting $0 < \gamma_i < 1$ drawn from a uniform random distribution. The hypothesized state, $s_i(t)$, and dwell-time, $t_i(t)$, of each μAgent instantiated a hypothesis 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 initial state correctly, that hypothesis could diverge from actuality. 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 the belief to a random hypothesis consistent with the current observation by setting s_i to a random state s^* selected with probability $P(s^*(t)|O(t))$, and setting t_i to 0. If the μAgent made a transition that

entailed a change in estimated value, it delivered a value prediction error signal (δ_i). Actions were selected based on the normalized, expected total value of the predicted state that would occur should an action be selected *Q(a)*, determined from the probability distribution over predicted states

$$Q\left(a_j\right) = \frac{1}{n_\mu}\sum_i\left[\text{is\_possible}\left(a_j \mid s_i\right)\cdot\left(E\left[r\left(s_i'\right)\right] + E\left[V\left(s_i'\right)\right]\right)\right] \tag{19}$$

where s_i was the state-hypothesis of μAgent i, s_i' the state that would be achieved by taking action a_j from state s_i, $E\left[r\left(s_i'\right)\right]$ the expected reward in state s_i', $E\left[V\left(s_i'\right)\right]$ the expected value of state s_i', and $\text{is\_possible}\left(a_j \mid s_i\right) \in \{0,1\}$ was a binary variable indicating whether action a_j was available from state s_i.

In order to determine the discounting function produced by our model, we modified the adjusting-delay assay of Mazur (1997, see above). See Figure 5. A five-state state-space was used to provide the macro-agent a choice between two actions, each of which led to a reward after a given delay. We ran the experiment for five agents (each of which consisted of 1000 μAgents) in this state-space for reward ratios of 2:1, 1:1, 3:2, and 3:1. As can be seen in Figure 5, the slopes of the indifference lines approximate the reward ratios, with a non-zero intercept. As reviewed above, this implies a hyperbolic discounting function like Equation 11 (Mazur, 1997). Thus, if each μAgent has a specific, different, exponential discounting function γ_i and maintains an independent estimate of the value of taking specific actions in a given situation, then the overall, behaviorally-observable "macro-Agent" will show hyperbolic discounting.

[Figure 5 about here.]

Working from anatomical studies, a number of researchers have hypothesized that the striatum consists of multiple separable pathways (Alexander et al., 1986; Alexander and

Crutcher, 1990; Graybiel et al., 1991; Strick et al., 1995). This suggests a possible anatomical spectrum of discounting factors which would be produced by a population of μAgents operating in parallel. Many researchers have reported that dopamine signals are not unitary (See Daw, 2003, for review). Non-unitary dopamine signals could arise from different dopamine populations contributing to different μAgents. Haber et al. (2000) report that the interaction between dopamine and striatal neural populations shows a regular anatomy, in a spiral progressing from ventral to dorsal striatum. Recently, Tanaka et al. (2004) explicitly found a gradient of discounting factors across the striata of human subjects.

In their recent fMRI experiment, Tanaka et al. (2004) found strong correlations between BOLD signals in striatum and different γ discounting factors (Eq. 2). They trained subjects to perform two tasks, each containing three states. Each state was identifiable by a clearly-differentiable cue. In both tasks, action a_1 led a transition from state s_1 to s_2 to s_3, while action a_2 led through a transition from state s_3 to s_2 to s_1. In the first task (the SHORT condition), positive rewards were given for all three transitions produced by action a_1 and negative rewards (punishments) were given for all three transitions produced by action a_2. In the second task (the LONG condition), positive rewards were given for two of the three transitions produced by action a1 and one of the transitions produced by action a_2, and punishments were given for two of the three transitions produced by action a_2 and one of the transitions produced by action a_1. The effect of this was that in the SHORT condition, action a1 was the optimal choice, which could be determined from a short horizon, while in the LONG condition action a_2 was the optimal choice, but required a longer horizon to determine. SHORT and LONG conditions were interspersed in a blocked format.

For each timestep in the task, for a given sequence of choices, for a given hypothesized γ, the

value at that moment could be calculated from the rewards delivered over the subsequent timesteps. This produced a family of functions $V^{\gamma}(t)$, which could then be correlated with the BOLD signals measured in the subjects. Fast discounting factors $\gamma \to 0$ were more strongly correlated with BOLD signals in ventral-anterior aspects of striatum; slower discounting factors $\gamma \to 1$ were more strongly correlated with BOLD signals in dorsal-posterior aspects of striatum. Tanaka et al. found a continuous distribution of best-correlated γ factors along the ventral-anterior to dorsal-posterior axis.

Since BOLD activity is more highly correlated with local field potentials (Logothetis, 2002) and local field potentials are more closely related to synaptic activity than local neural firing (Buzsáki, 2006), it is likely that the functional slices observed by Tanaka et al. (2004) reflect differential inputs rather than direct changes in striatal activity. Nevertheless, 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.

Reinforcement learning with multiple models (sometimes called multiple experts) has a long history (Doya et al., 2002; Bertin et al., 2007). The suggestion that the basal ganglia consist of multiple separable loops also has a long history (Alexander et al., 1986; Alexander and Crutcher, 1990; Haber et al., 2000; Middleton and Strick, 2000), yet remains controversial (Parthasarathy et al., 1992; Graybiel, 2000). The suggestion that these separate loops are indicative of separate discounting factors (Tanaka et al., 2004; Kurth-Nelson and Redish, 2004) is, however, novel. More work needs to be done to confirm or reject that hypothesis. In any case, it is likely that the Tanaka et al. (2004) data reflect patterns of activity that could correspond to a parallel computation based on a continuum of discounting factors.

[Figure 6 about here.]

One important consequence of this multiple-exponential hypothesis is that shifts in the distribution of included exponentials would produce discounting functions that deviate from Equation 11. While the hyperbolic fit for the animal behavior literature is often excellent (Mazur, 1985, 1997; Richards et al., 1997), the fit for the human decision literature is more variable (sometimes excellent (Vuchinich and Simpson, 1998), and sometimes less so (Reynolds, 2006), particularly for drug-users (Madden et al., 1999; Mitchell, 1999)). Schweighofer et al. (2006) report that under specific conditions in which a single exponential discounting rate is optimal, humans can learn to match that factor and show an exponential discounting function.

Changes in serotonin levels have long been associated with impulsivity [with lower levels of serotonin correlating with more impulsivity] (Chamberlain et al., 2006; Carver and Miller, 2006; Chamberlain and Sahakian, 2007). Rats with dorsal raphe (serotonin) lesions showed earlier indifference points on Mazur's adjusting-delay paradigm (Mobini et al., 2000; Wogar et al., 1993). These rats were still able to accurately time delays when no contrast was involved, so the change was not due to loss of temporal recognition (Morrissey et al., 1993). Changing levels of serotonin precursors (e.g. tryptophan) can change the measured discounting rates in human subjects (Tanaka et al., 2007). Doya (2000a, 2002) has explicitly suggested that serotonin may control the discounting rate used in an exponential discounting module. Alternatively, serotonin may control the distribution of exponential components contributing to the behavior. These proposals still constitute a controversial hypothesis and there is little direct evidence to support it; however, recent experiments in the Doya laboratory (Tanaka et al., 2007) have found that changes in serotonin precursors (e.g. tryptophan) can reduce activity in certain of the discounting slices seen by Tanaka et al. (2004) while enhancing activity in others.

A sum of internal exponential discounting functions will only produce hyperbolic

discounting in the case of a uniform distribution of exponentials covering the entire available range. In general, a sum of internal exponential discounting functions will produce a power law behaviorally.

$$V_d(r) = r\int_{\kappa=0}^{\infty} g(\kappa) e^{-\kappa d} d\kappa \qquad (20)$$

where g(κ) is the distribution of exponential discounting factors $0 < \kappa < \infty$. For simplicity, we assume $g(\kappa) = \kappa^{\beta}$, where β is a constant that controls the distribution of components. (In this formulation, β = 0 implies a flat distribution of exponentials, β > 0 implies more high κ, faster discounting components, while β < 0 implies more low κ, slower discounting components.) Under these assumptions, the integrated value function can be written analytically as

$$V_d(r) = r \cdot \frac{\alpha}{d^{(1+\beta)}} \qquad (21)$$

where α is a constant term and β is derived from the g(κ) distribution. When β = 0, this corresponds to a hyperbolic 1/*d* discounting function. As β increases this function deviates from hyperbolic to become more impulsive, and as β decreases this function deviates from hyperbolic to become less impulsive (see Figure 7). Both group data and individual data are well-fit by power-laws such as Eq. 21 (Redish, Landes, Bickel, unpublished observations), although it is not clear yet whether Eq. 21 provides any better fit to the experimental data than standard hyperbolic discounting analyses (e.g. Eq. 11). Of course, the actual g(κ) distribution could be any mix of exponential functions, and could potentially be variable under pharmacological or experimental control (Tanaka et al., 2007; Schweighofer et al., 2007), including becoming a single exponential function under the right conditions (Schweighofer et al., 2006).

[Figure 7 about here.]

3 Summary/Conclusion

In making a decision between multiple choices, a complete description of the values of the two choices would require specification and integration over all potential possibilities, taking into account the uncertainty, risk, and investment opportunities with each decision. This infinite calculation is, of course, impossible to do with a finite decision process. A reasonable method of solving this problem is to discount delayed rewards. Humans and animals discount delayed rewards with functions better described as hyperbolic or power-law functions (with changing discount rates over time) than as exponential functions (with constant discount rates). Four sets of neural models have been proposed to explain this discrepancy: (1) that agents are actually maximizing rates of reward, normalizing observed rewards by current expectations of average rates of reward, (2) that changes in time perception produce variations in underlying estimates of delays, leading to a spreading out of the exponential discounting function, which leads to a hyperbolic-like power law, and (3,4) that neural systems include two (or more) subsystems which discount future rewards at different rates. The extensive neural data support the multiple subsystem hypothesis quite strongly; however, the number of subsystems and the identity of the specific components making up each subsystem remain unresolved. Although a number of specific algorithms have been proposed to underlie these subsystems, these proposals remain controversial. These multiple subsystems may alse underlie complexities in temporal perception, as well as general memory, and behavioral processes. More work testing specific neural hypotheses under conditions that change discounting factors is needed.

Notes

1. For simplicity, we use the term “agent” to refer to any decision-making system (including

humans and other animals as well as simulations). Agency is used without any prejudice or presumption re free will.

2. The term situation refers to the agent's classification of the state of the world and the agent from which the agent can reason about decisions. We prefer "situation" over the psychology term "stimulus" so as to include context, cue, and interactions between cues, all of which are critical for appropriate behavior. Similarly, we prefer "situation" over the robotics term "state" to prevent confusion with internal parameters of the agent (e.g. "motivation states"). See Redish et al. (2007) and Zilli and Hasselmo (2008) for further discussion of these issues.

3. New models have begun to explore the limitations of these assumptions, including relaxing assumptions of stationarity (e.g. Courville et al., 2006; Redish et al., 2007), assumptions of observability (e.g. Daw et al., 2002b, 2006a; Yu and Dayan, 2005), and assumptions of exploration (e.g. Kakade and Dayan, 2002; Daw et al., 2006b). Others have begun incorporating the potential effects of working and episodic memory (Zilli and Hasselmo, 2008). However, these issues are not immediately relevant to this review and will not be pursued further here.

4. In other chapters in this book, the term A is used for *amount of reward*. This chapter uses the term r for reward to avoid confusion with *action* a.

5. This model maintains independent value-estimations across all the μAgents. If the μAgents instead maintain a shared value-estimation, the model reverts to be equivalent to the direct implementation of hyperbolic discounting (Eq. 13), showing hyperbolic discounting across only a single state-transition.

6. This timing model is still controversial (Staddon and Higa, 1999; Kalenscher and Pennartz, 2008).

References

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, 82(4):463–496.

Ainslie, G. (1992). *Picoeconomics*. Cambridge Univ Press. Ainslie, G. (2001). *Breakdown of Will*. Cambridge Univ Press.

Ainslie, G. and Monterosso, J. (2004). Behavior: A marketplace in the brain? *Science*, 306(5695):421–423.

Alexander, G. E. and Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in Neurosciences*, 13(7):266–271.

Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Reviews Neuroscience*, 9:357–381.

Amiez, C., Joseph, J.-P., and Procyk, E. (2005). Anterior cingulate error-related activity is modulated by predicted reward. *European Journal of Neuroscience*, 21:3447–3452.

Anderson, R. B. and Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, 25(5):724–730.

Barbas, H. (2007). Specialized elements of orbitofrontal cortex in primates. *Annals of the New York Academy of Sciences*, 1121:10–32.

Bayer, H. M. and Glimcher, P. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47:129–141.

Bayer, H. M., Lau, B., and Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *J Neurophysiol*, 98(3):1428–1439.

Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: a

neurocognitive perspective. *Nature Neuroscience*, 8(11):1458–1463.

Bechara, A., Dolan, S., and Hindes, A. (2002). Decision-making and addiction (part ii): myopia for the future or hypersensitivity to reward? *Neuropsychologia*, 40(10):1690–1705.

Bechara, A., Nader, K., and van der Kooy, D. (1998). A two-separate-motivational-systems hypothesis of opioid addiction. *Pharmacology Biochemistry and Behavior*, 59(1):1–17.

Bechara, A. and van der Linden, M. (2005). Decision-making and impulse control after frontal lobe injuries. *Cuerrent Opinion in Neurology*, 18:734–739.

Bellman, R. (1958). On a routing problem. *Quarterly Journal of Applied Mathematics*, 16(1):87–90.

Beran, M. J., Savage-Rumbaugh, E. S., Pate, J. L., and Rumbaugh, D. M. (1999). Delay of gratification in chimpanzees (pan troglodytes). *Developmental Psychobiology*, 34(2):119–127.

Bernheim, B. D. and Rangel, A. (2004). Addiction and cue-triggered decision processes. *The American Economic Review*, 94(5):1558–1590.

Bertin, M., Schweighofer, N., and Doya, K. (2007). Multiple model-based reinforcement learning explains dopamine neuronal activity. *Neural Networks*, 20(6):668–675.

Bickel, W. K. and Marsch, L. A. (2001). Toward a behavioral economic understanding of drug dependence: delay discounting processes. *Addiction*, 96:73–86.

Bickel, W. K., Miller, M. L., Yi, R., Kowal, B. P., Lindquist, D. M., and Pitcock, J. A. (2007). Behavioral and neuroeconomics of drug addiction: Competing neural systems and temporal discounting processes. *Drug and Alcohol Dependence*, 90(Suppl. 1):S85–S91.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., and Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, 402(6758):179–

181.

Boysen, S. T. and Berntson, G. G. (1995). Responses to quantity: Perceptual versus cognitive mechanisms in chimpanzees (pan troglodytes). *Journal of Experimental Psychology: Animal Behavior Processes*, 21(1):82–86.

Brass, M. and Haggard, P. (2007). To do or not to do: The neural signature of self-control. *J. Neurosci.*, 27(34):9141–9145.

Braver, T. S., Barch, D. M., Gray, J. R., Molfese, D. L., and Snyder, A. (2001). Anterior cingulate cortex and response conflict: Effects of frequency, inhibition and errors. *Cereb. Cortex*, 11(9):825–836.

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford.

Cardinal, R. N. (2006). Neural systems implicated in delayed and probabilistic reinforcement. *Neural Networks*, 19(8):1277–1301.

Cardinal, R. N., Pennicott, D. R., Sugathapala, C. L., Robbins, T. W., and Everitt, B. J. (2001). Impulsive choice induced in rats by lesion of the nucleus accumbens core. *Science*, 292:2499– 2501.

Carver, C. S. and Miller, C. J. (2006). Relations of serotonin function to personality: Current views and a key methodological issue. *Psychiatry Research*, 144(1):1–15.

Chamberlain, S. R., Muller, U., Robbins, T. W., and Sahakian, B. J. (2006). Neuropharmacological modulation of cognition. *Current Opinion in Neurology*, 19(6):607–612.

Chamberlain, S. R. and Sahakian, B. J. (2007). The neuropsychiatry of impulsivity. *Current Opinion in Psychiatry*, 20(3):255–261.

Clark, L., Manes, F., Antoun, N., Sahakian, B. J., and Robbins, T. W. (2003). The contributions

of lesion laterality and lesion volume to decision-making impairment following frontal lobe damage. *Neuropsychologia*, 41:1474–1483.

Courville, A. C., Daw, N. D., and Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10:294–300.

Crutcher, M. D. and Alexander, G. E. (1990). Movement-related neuronal activity selectively coding either direction or muscle parttern in three motor areas of the monkey. *Journal of Neurophysiology*, 64(1):151–163.

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

Daw, N. D., Courville, A. C., and Touretzky, D. S. (2002a). Dopamine and inference about timing. *Proceedings of the Second International Conference on Development and Learning*.

Daw, N. D., Courville, A. C., and Touretzky, D. S. (2002b). Timing and partial observability in the dopamine system. *NIPS*, 15:99–106.

Daw, N. D., Courville, A. C., and Touretzky, D. S. (2006a). Representation and timing in theories of the dopamine system. *Neural Computation*, 18:1637–1677.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8:1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006b). Cortical substrates for exploratory decisions in humans. *Nature*, 441:876–879.

Daw, N. D. and Touretzky, D. S. (2000). Behavioral considerations suggest an average reward TD model of the dopamine system. *Neurocomputing*, 32-33:679–684.

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

Doya, K. (2000b). Reinforcement learning in continuous time and space. *Neural Computation*, 12:219–245.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, 15:495–506.

Doya, K., Samejima, K., Katagiri, K.-I., and Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, 14(6):1347–1369.

Estle, S. J., Green, L., Myerson, J., and Holt, D. D. (2007). Discounting of monetary and directly consumable rewards. *Psychological Science*, 18(1):58–63.

Evans, T. A. and Beran, M. J. (2007). Delay of gratification and delay maintenance by rhesus macaques (macaca mulatta). *Journal of General Psychology*, 134(2):199–216.

Evenden, J. L. (1999). Varieties of impulsivity. *Psychopharmacology*, 146(4):348–361.

Fiorillo, C. D., Tobler, P. N., and 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(1):7.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic literature*, 40(2):351–401.

Gallistel, C. R. (1990). *The Organization of Learning*. MIT Press, Cambridge, MA.

Gallistel, C. R. and Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, 107(2):289–344.

Gibbon, J., Church, R. M., Fairhurst, S., and Kacelnik, A. (1988). Scalar expectancy theory and choice between delayed rewards. *Psychological Review*, 95(1):102–114.

Glimcher, P. W., Kable, J., and Louie, K. (2007). Neuroeconomic studies of impulsivity: Now or just as soon as possible? *American Economic Review*, 97.

Grace, R. C. (1999). The matching law and amount-dependent exponential discounting as

accounts of self-control choice. *Journal of the Experimental Analysis of Behavior*, 1:27–44.

Grant, S., Contoreggi, C., and London, E. D. (2000). Drug abusers show impaired performance in a laboratory test of decision making. *Neuropsychologia*, 38(8):1180–1187.

Gray, J. and McNaughton, N. (2000). *The Neuropsychology of Anxiety*. Oxford.

Gray, J. A. (1982a). *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System*. Oxford University Press, New York.

Gray, J. A. (1982b). Précis of The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System. Behavioral and Brain Sciences, 5:469–484. See also commentary and response, pages 484–534.

Graybiel, A. (2000). The basal ganglia. *Current Biology*, 10(14):R509–R511.

Graybiel, A. M., Flaherty, A. W., and Gim´enez-Amaya, J.-M. (1991). Striosomes and matrisomes. In Bernardi, G., Carpenter, M. B., and Di Chiara, G., editors, *The Basal Ganglia III*. Plenum.

Green, L. and Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130(5):769–792.

Green, L., Myerson, J., Holt, D. D., Slevin, J. R., and Estle, S. J. (2004). Discounting of delayed food rewards in pigeons and rats: Is there a magnitude effect? *Journal of the Experimental Analysis of Behavior*, 81:39–50.

Green, L., Myerson, J., and Macaux, E. W. (2005). Temporal discounting when the choice is between two delayed rewards. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 31(5):1121–1133.

Green, L., Myerson, J., and McFadden, E. (1997). Rate of temporal discounting decreases with amount of reward. *Memory & Cognition*, 25(5):715–723.

Haber, S. N., Fudge, J. L., and McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience*, 20(6):2369–2382.

Isoda, M. and Hikosaka, O. (2007). Switching from automatic to controlled action by monkey medial frontal cortex. *Nature Neuroscience*, 10:240–248.

Johnson, M. W. and Bickel, W. K. (2002). Within-subject comparison of real and hypothetical money rewards in delay discounting. *J Exp Anal Behav*, 77(2):129–146.

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

Kacelnik, A. and Bateson, M. (1996). Risky theories — the effects of variance on foraging decisions. *Amer. Zool.*, 36(4):402–434.

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

Kalenscher, T. and Pennartz, C. M. A. (2008). Is a bird in the hand worth two in the future? the neuroeconomics of intertemporal decision-making. *Progress in Neurobiology*, 84:284–315.

Kelley, A. E. and Berridge, K. C. (2002). The Neuroscience of Natural Rewards: Relevance to Addictive Drugs. *J. Neurosci.*, 22(9):3306–3311.

Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J., and Rushworth, M. F. S. (2006). Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience*, 9:940–947.

Kirby, K. N. (1997). Bidding on the future: Evidence against normative discounting of delayed

rewards. *Journal of Experimental Psychology: General*, 126(1):54–70.

Kurth-Nelson, Z. and Redish, A. D. (2004). µagents: action-selection in temporally-dependent phenomena using temporal difference learning over a collective belief structure. *Society for Neuroscience Abstracts*.

Laibson, D. I. (1996). An economic perspective on addiction and matching. *Behavioral and Brain Sciences*, 19(4):583–584.

Logothetis, N. (2002). The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. *Philosophical Transactions of the Royal Society London B*, 357:1003–1037.

Lubman, D. I., Y ¨ucel, M., and Pantelis, C. (2004). Addiction, a condition of compulsive behaviour? neuroimaging and neuropsychological evidence of inhibitory dysregulation. *Addiction*, 99:1491–1502. Commentary following.

Madden, G. J., Bickel, W. K., and Jacobs, E. A. (1999). Discounting of delayed rewards in opioid-dependent outpatients exponential or hyperbolic discounting functions? *Experimental and Clinical Psychopharmacology*, 7(3):284–293.

Madden, G. J., Petry, N. M., Badger, G. J., and Bickford, W. K. (1997). Impulsive and self-control choices in opioid-dependent patients and non-drug-using control patients: Drug and monetary rewards. *Experimental and Clinical Psychopharmacology*, 5(3):256–262.

Mazur, J. (1997). Choice, delay, probability and conditioned reinforcement. *Animal Learning and Behavior*, 25(2):131–147.

Mazur, J. E. (1985). Probability and delay of reinforcement as factors in discrete-trial choice. *J Exp Anal Behav*, 43(3):341–351.

Mazur, J. E. (2001). Hyperbolic value addition and general models of animal choice.

Psychological Review, 108(1):96–112.

Mazur, J. E. (2006). Choice between single and multiple reinforcers in concurrent-chains schedules. *J Exp Anal Behav*, 86(2):211–222.

McClure, S. M., Berns, G. S., and Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, 38(2):339–346.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., and Cohen, J. D. (2007). Time discounting for primary rewards. *J. Neurosci.*, 27(21):5796–5804.

McClure, S. M., Laibson, D. I., Loewenstein, G., and Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, 306(5695):503–507.

Metcalfe, J. and Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, 106(1):3–19.

Middleton, F. A. and Strick, P. L. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Research Reviews*, 31:236–250.

Mischel, W. and Underwood, B. (1974). Instrumental ideation in delay of gratification. *Child Development*, 45:1083–1088.

Mitchell, S. H. (1999). Measures of impulsivity in cigarette smokers and non-smokers. *Psychopharmacology*, 146(4):455–464.

Mobini, S., Body, S., Ho, M.-Y., Bradshaw, C. M., Szabadi, E., Deakin, J. F. W., and Anderson, I. M. (2002). Effects of lesions of the orbitofrontal cortex on sensitivity to delayed and probabilistic reinforcement. *Psychopharmacology*, 160:290–298.

Mobini, S., Chiang, T. J., Al-Ruwaitea, A. S., Ho, M. Y., Bradshaw, C. M., and Szabadi, E. (2000). Effect of central 5-hydroxytryptamine depletion on inter-temporal choice: a quantitative analysis. *Psychopharmacology*, 149(3):313–318.

Morrissey, G., Wogar, M. A., Bradshaw, C. M., and Szabadi, E. (1993). Effect of lesions of the ascending 5-hydroxytryptaminergic pathways on timing behaviour investigated with an interval bisection task. *Psychopharmacology*, 112(1):80–85.

Murray, E. A., O'Doherty, J. P., and Schoenbaum, G. (2007). What we know and do not know about the functions of the orbitofrontal cortex after 20 years of cross-species studies. *J. Neurosci.*, 27(31):8166–8169.

Myerson, J. and Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *J Exp Anal Behav*, 64(3):263–276.

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

Noël, X., Van der Linden, M., d'Acremont, M., Bechara, A., Dan, B., Hanak, C., and Verbanck, P. (2007). Alcohol cues increase cognitive impulsivity in individuals with alcoholism. *Psychopharmacology*, 192(2):291–298.

Odum, A. L. and Rainaud, C. P. (2003). Discounting of delayed hypothetical money, alcohol, and food. *Behav Processes*, 64(3):305–313.

Okano, K. and Tanji, J. (1987). Neuronal activities in the primate motor fields of the agranular frontal cortex preceding visually triggered and self-paced movement. *Experimental Brain Research*, 66(1):155–166.

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Clarendon Press, Oxford.

Ong, E. L. and White, K. G. (2004). Amount-dependent temporal discounting? *Behavioural Processes*, 66:201–212.

Padoa-Schioppa, C. and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode

economic value. *Nature*, 441:223–226.

Padoa-Schioppa, C. and Assad, J. A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nature Neuroscience*, 11(1):95–102.

Parthasarathy, H., Schall, J., and Graybiel, A. (1992). Distributed but convergent ordering of corticostriatal projections: analysis of the frontal eye field and the supplementary eye field in the macaque monkey. *J. Neurosci.*, 12(11):4468–4488.

Petry, N. M., Bickel, W. K., and Arnett, M. (1998). Shortened time horizons and insensitivity to future consequences in heroin addicts. *Addiction*, 93(5):729–738.

Poldrack, R. A. and Packard, M. G. (2003). Competition among multiple memory systems: Converging evidence from animal and human studies. *Neuropsychologia*, 41:245–251.

Price, J. L. (2007). Definition of the orbital cortex in relation to specific connections with limbic and visceral structures and other cortical regions. *Annals of the New York Academy of Sciences*, 1121:54–71.

Ramus, S. J., Davis, J. B., Donahue, R. J., Discenza, C. B., and Waite, A. A. (2007). Interactions between the orbitofrontal cortex and hippocampal memory system during the storage of long-term memory. *Annals of the New York Academy of Sciences*, 1121:216–231.

Read, D. (2001). Is time-discounting hyperbolic or subadditive? *Journal of Risk and Uncertainty*, 23(1):5–32.

Redish, A. D. (1999). Beyond the Cognitive Map: From Place Cells to Episodic Memory. MIT Press, Cambridge MA.

Redish, A. D. (2004). Addiction as a computational process gone awry. *Science*, 306(5703):1944– 1947.

Redish, A. D., Jensen, S., and Johnson, A. (in press). A unified framework for addiction:

vulnerabilities in the decision process. *Behavioral and Brain Sciences*.

Redish, A. D., Jensen, S., Johnson, A., and Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, 114(3):784–805.

Redish, A. D. and Johnson, A. (2007). A computational model of craving and obsession. *Annals of the New York Academy of Sciences*, 1104(1):324–339.

Reynolds, B. (2006). A review of delay-discounting research with humans: relations to drug use and gambling. *Behavioural Pharmacology*, 17(8):651–667.

Richards, J. B., Mitchell, S. H., de Wit, H., and Seiden, L. S. (1997). Determination of discount functions in rats with an adjusting-amount procedure. *J Exp Anal Behav*, 67(3):353–366.

Rodriguez, M. L. and Logue, A. W. (1988). Adjusting delay to reinforcement: comparing choice in pigeons and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, 14(1):105–117.

Roesch, M. R., Calu, D. J., Burke, K. A., and Schoenbaum, G. (2007). Should i stay or should i go?. transformation of time-discounted rewards in orbitofrontal cortex and associated brain circuits. *Annals of the New York Academy of Sciences*, 1104(1):21–24.

Roesch, M. R., Taylor, A. R., and Schoenbaum, G. (2006). Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron*, 51(4):509–520.

Rogers, A. R. (1997). Evolution and human choice over time. In Bock, G. R. and Cardew, G., editors, *Characterizing Human Psychological Adaptations*, volume 208 of *Ciba Foundation Symposia*, pages 231–248. Wiley, Chichester UK. Discussion 249-252.

Rubenstein, A. (2003). ”economics and psychology”? the case of hyperbolic discounting. *International Economic Review*, 44(4):1207–1216.

Rubin, D. C., Hinton, S., and Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5):1161–1176.

Rubin, D. C. and Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psyhcological Review*, 103(4):734–760.

Rushworth, M. F. S., Buckley, M. J., Behrens, T. E. J., Walton, M. E., and Bannerman, D. M. (2007). Functional organization of the medial frontal cortex. *Current Opinion in Neurobiology*, 17(2):220–227.

Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., and Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, 8(9):410–417.

Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies*, 4(2):155–161.

Schoenbaum, G., Setlow, B., Saddoris, M. P., and Gallagher, M. (2006). Encoding changes in orbitofrontal cortex in reversal-impaired aged rats. *J Neurophysiol*, 95(3):1509–1517.

Schoenbaum, G. and Shaham, Y. (2008). The role of orbitofrontal cortex in drug addiction: A review of preclinical studies. *Biological Psychiatry*, 63:256–262.

Schweighofer, N., Shishida, K., Han, C. E., Yamawaki, Y. O. S. C. T. S., and Doya, K. (2006). Humans can adopt optimal discounting strategy under real-time constraints. *PLoS Computational Biology*, 2(11):e152.

Schweighofer, N., Tanaka, S. C., and 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(1):289–300.

Sohn, J.-W. and Lee, D. (2007). Order-dependent modulation of directional signals in the supplementary and presupplementary motor areas. *Journal of Neuroscience*, 27(50):13655–

13666.

Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. *The Royal Society London B*, 265:2015–2020.

Squire, L. R. (1987). *Memory and Brain*. Oxford University Press, New York.

Staddon, J. E. and Higa, J. J. (1999). The choose-short effect and trace models of timing. *Journal of the Experimental Analysis of Behavior*, 72(3):473–478.

Staddon, J. E. R. and Cerutti, D. T. (2003). Operant conditioning. *Annual Reviews of Psychology*, 54:115–144.

Staddon, J. E. R., Chelaru, I. M., and Higa, J. J. (2002). Habituation, memory and the brain: the dynamics of interval timing. *Behavioural Processes*, 57(2-3):71–88.

Stephens, D. W. and Krebs, J. R. (1987). *Foraging Theory*. Princeton.

Strick, P. L., Dum, R. P., and Picard, N. (1995). Macro-organization of the circuts connecting the basal ganglia with the cortical motor areas. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, pages 117–130. MIT Pres.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3):165–180.

Sugrue, L. P., Corrado, G. S., and Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, 304(5678):1782–1787.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An introduction*. MIT Press, Cambridge MA.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., and Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, 7:887–893.

Tanaka, S. C., Schweighofer, N., Asahi, S., Shishida, K., Okamoto, Y., Yamawaki, S., and Doya, K. (2007). Serotonin differentially regulates short-and long-term prediction of rewards in the ventral and dorsal striatum. *PLoS ONE*, 2(12):e1333.

Tanji, J. (2001). Sequential organization of multiple movements: Involvement of cortical motor areas. *Annual Review of Neuroscience*, 24:631–651.

Tiffany, S. T. (1990). A cognitive model of drug urges and drug-use behavior: Role of automatic and nonautomatic processes. *Psychological Review*, 97(2):147–168.

Torregrossa, M. M., Quinn, J. J., and Taylor, J. R. (2008). Impulsivity, compulsivity, and habit: The role of orbitofrontal cortex revisited. *Biological Psychiatry*, 63:253–255.

Tremblay, L. and Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708.

Vuchinich, R. E. and Simpson, C. A. (1998). Hyperbolic temporal discounting in social drinkers and problem drinkers. *Experimental and Clinical Psychopharmacology*, 6(3):292–305.

Walton, M. E., Rudebeck, P. H., Bannerman, D. M., and Rushworth, M. F. S. (2007). Calculating the cost of acting in frontal cortex. *Annals of the New York Academy of Sciences*, 1104(1):340– 356.

Winstanley, C. A., Theobald, D. E. H., Cardinal, R. N., and Robbins, T. W. (2004). Contrasting roles of basolateral amygdala and orbitofrontal cortex in impulsive choice. *J. Neurosci.*, 24(20):4718–4722.

Wittmann, M., Leland, D. S., and Paulus, M. P. (2007). Time and decision making: differential contribution of the posterior insular cortex and the striatum during a delay discounting task. *Experimental Brain Research*, 179(4):643–653.

Wixted, J. T. and Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative

analysis of individual subject forgetting functions. *Memory & Cognition*, 25(5):731–739.

Wogar, M. A., Bradshaw, C. M., and Szabadi, E. (1993). Effect of lesions of the ascending 5-hydroxytryptaminergic pathways on choice between delayed reinforcers. *Psychopharmacology*, 111(2):239–243.

Wörgötter, F. and 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.

Yu, A. J. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681– 692.

Zermatten, A., Van der Linden, M., d'Acremont, M., Jermann, F., and Bechara, A. (2005). Impulsivity and decision making. *J Nerv Ment Dis*, 193(10):647–650.

Zilli, E. A. and Hasselmo, M. E. (2008). Modeling the role of working memory and episodic memory in behavioral tasks. *Hippocampus*, 18(2):193–209.

Figures

FIGURE 1: Discounting across state-chains. (A) Single-step state space. (B) Discounting over a single-step state space as a function of the delay *D* in state *S*0. Both value functions derived from Equation 13 and a sum of exponentials model (section 2.4) show hyperbolic discounting. (C) Chained state-space. (D) Discounting over a chained-state as a function of total delay from state S0 to reward state. The single-step hyperbolic model (Eq. 13) no longer shows hyperbolic discounting, while the multiple-exponentials model continues to do so.

FIGURE 2: Calculation of indifference points as a function of delay in the indifference task. Reprinted from Daw and Touretzky (2000) with permission from author and publisher (Elsevier, © 2000). The indifference points predicted by the average reward model. Indifference points are

shown as dots, and the line of best fit is also shown.

FIGURE 3: Calculation of indifference points using an exponential decay model with Poisson-like time-estimation (mean of the estimated delay = actual delay, variance of estimated delay proportional to actual delay). Reprinted from Daw (2003) with permission from author (© N. D. Daw, 2003).

FIGURE 4: Consistent brain areas are activated for intertemporal choices across reward modality, but different brain areas are activated for β-related (impulsive) and δ-related (more general decision-related) areas. Reprinted fromMcClure et al. (2007) with permission from author and publisher (Society for Neuroscience, © 2007). Original figure in color.

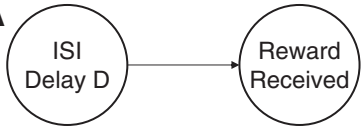
FIGURE 5: Discounting with multiple exponentials. (A) State-space used. (B-E) Mazur-plots. These plots show the delay $d2$ needed to make an agent choose actions $a1$ and $a2$ equally for a given delay $d1$. The ratio of actions $a1$:$a2$ is an observable measure of the relative values of the two choices. For hyperbolic discounting, the slope of the line will equal the ratio of $r2/r1$, with a non-zero y-intercept. A sum-of-exponentials model produces near-perfect hyperbolic discounting.

FIGURE 6: Correlations between reward prediction $V(t)$ and BOLD signal are significantly most correlated to different discounting factors γ. Reprinted from Tanaka et al. (2004) with permission from author and publisher (Macmillan Publishers Ltd: Nature Neuroscience, © 2004). Original figure in color.

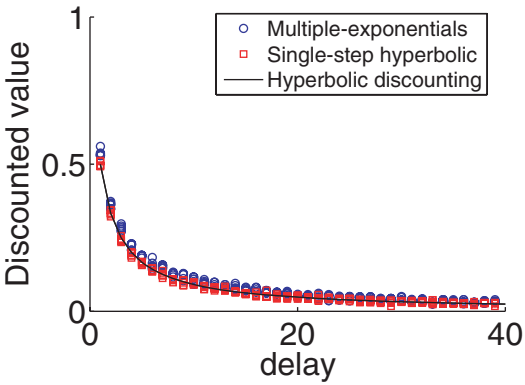
FIGURE 7: (top) Distributions of exponentials as β changes. (bottom) The resulting value-discounting functions can become more or less impulsive with changing distributions of exponentials. A uniform distribution of exponential discounting functions (characterized by β=0) produces a very close match to hyperbolic discounting.

Single-step state-space

A



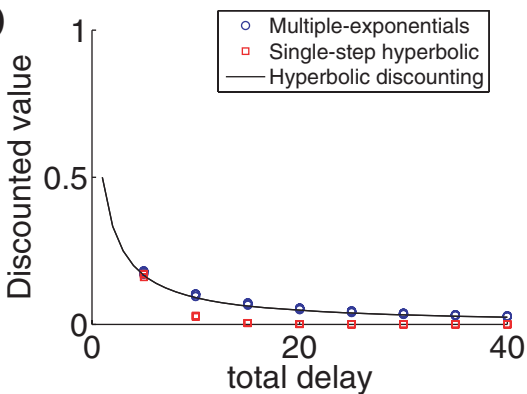
B

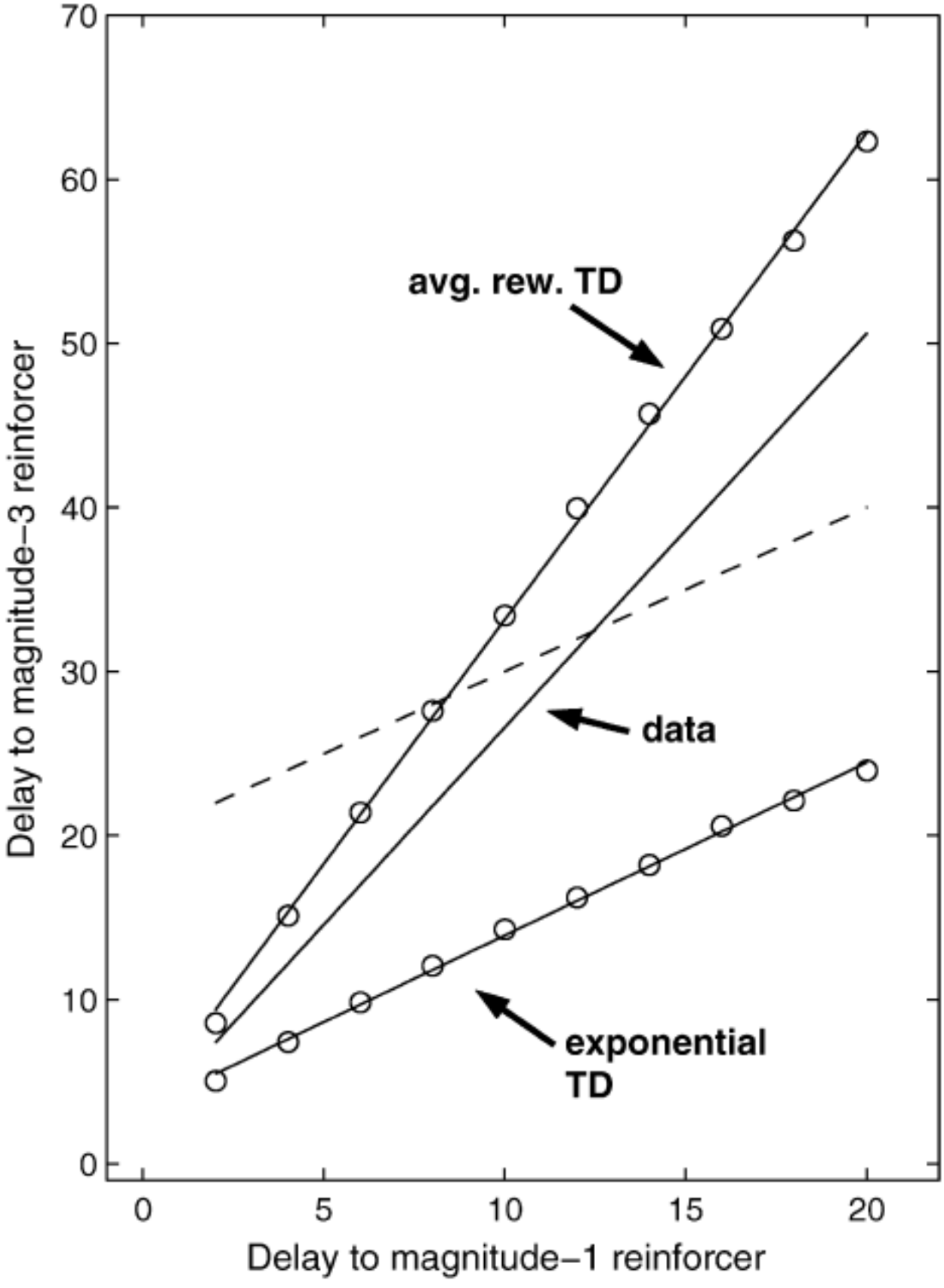


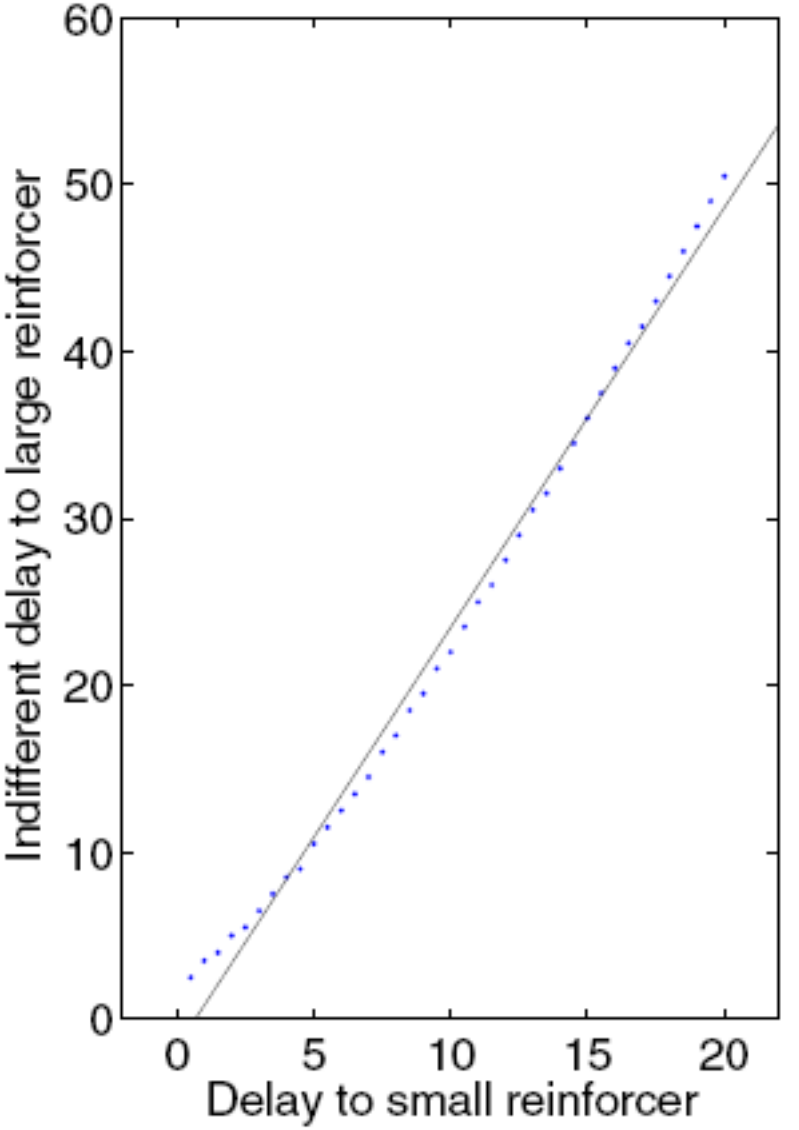
Chained state-space

C

D

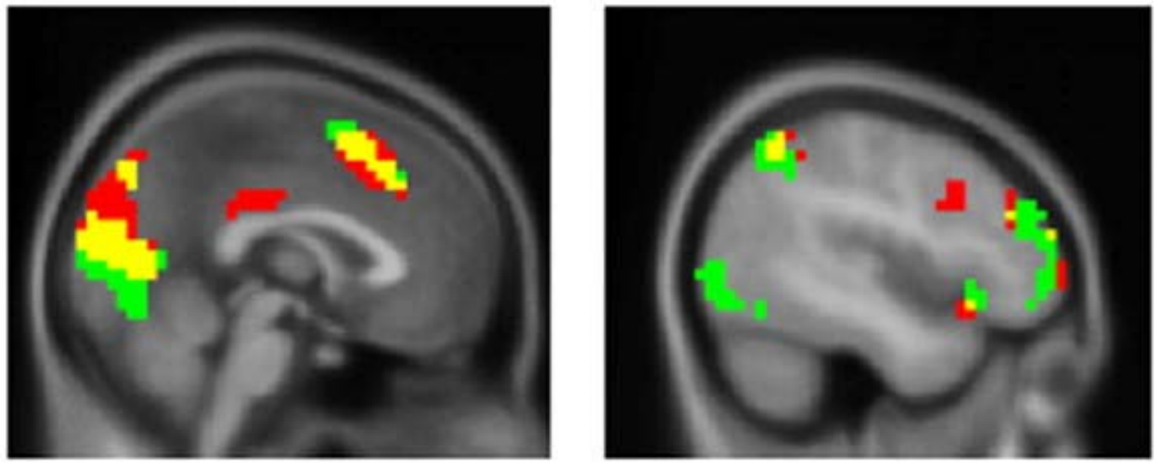



70
60
50
40
30
20
10
0
0
5
10
15
20
Delay to magnitude−3 reinforcer
Delay to magnitude−1 reinforcer
avg. rew. TD
data
exponential
TD


60
50
40
30
20
10
0
0
5
10
15
20
Indifferent delay to large reinforcer
Delay to small reinforcer

A

δ areas (p<0.001)

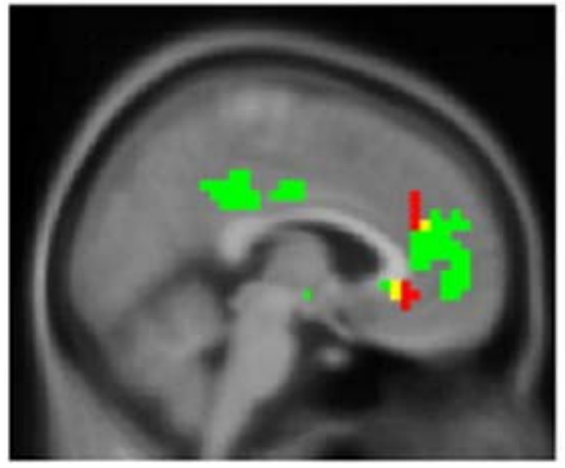


x = 0mm

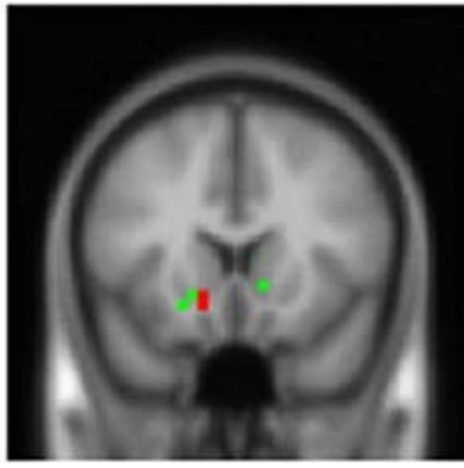
x = -44mm

B

β areas (p<0.001)

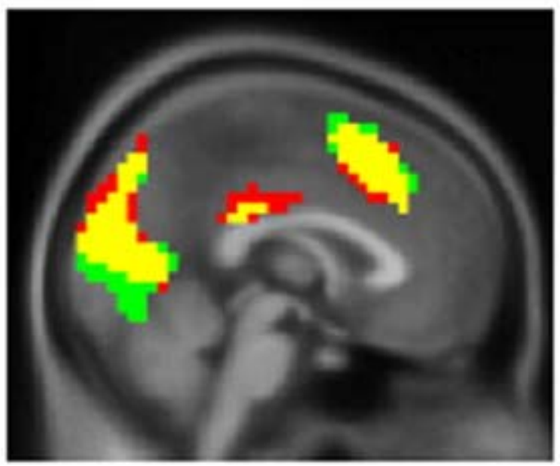


x = 4mm

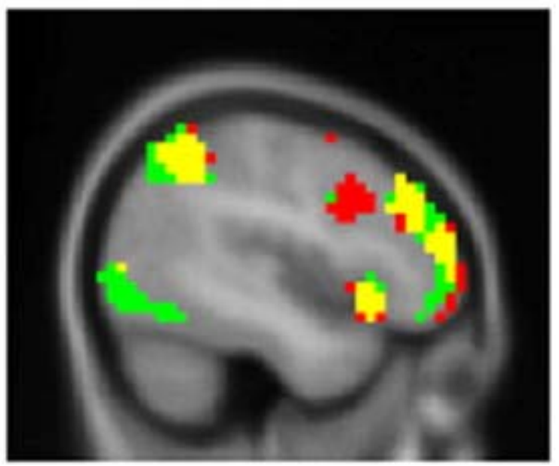


y = 16mm

δ areas (p<0.01)

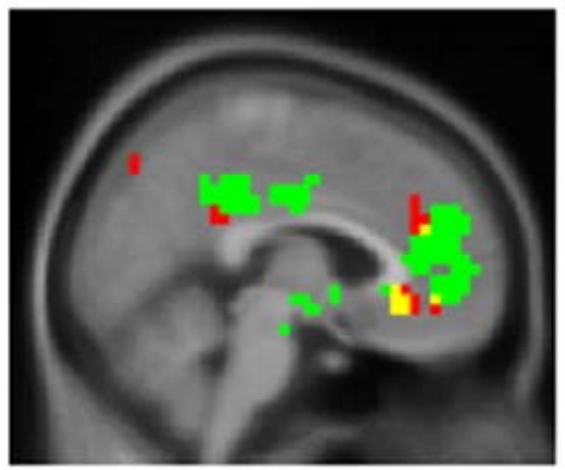


x = 0mm

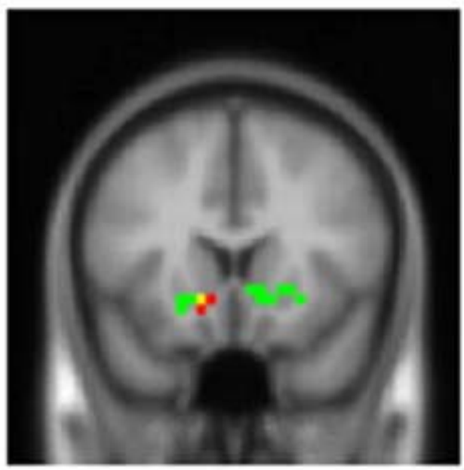


x = -44mm

β areas (p<0.01)



x = 4mm



y = 16mm

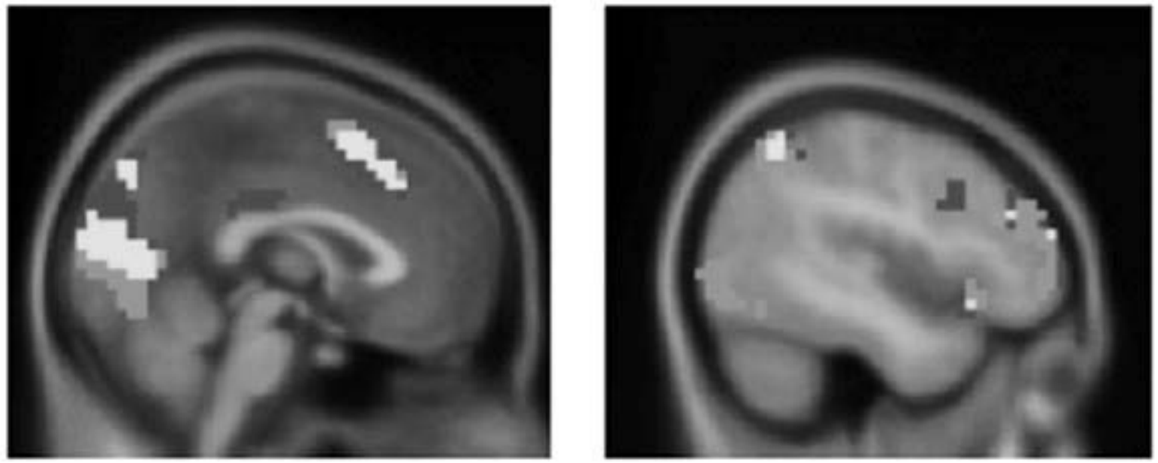
Juice

Money

Both

A

δ areas (p<0.001)

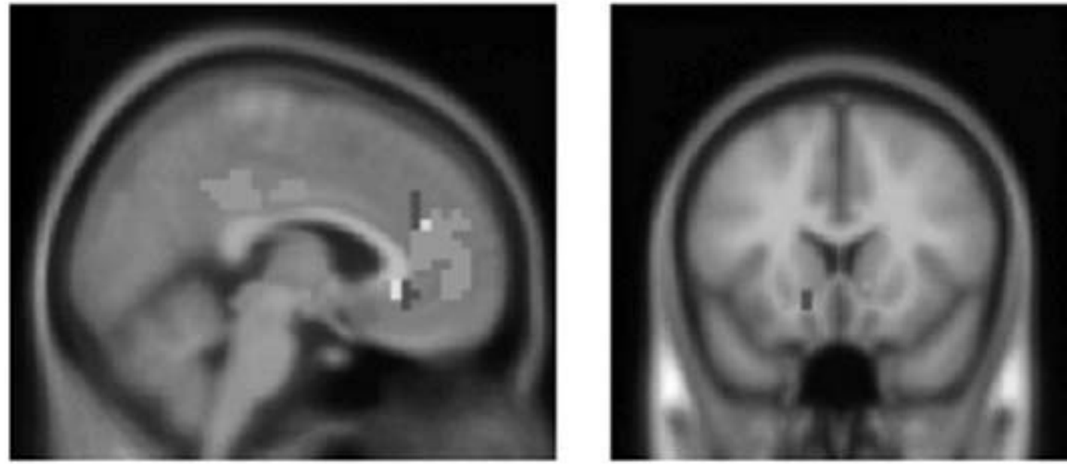


x = 0mm

x = -44mm

B

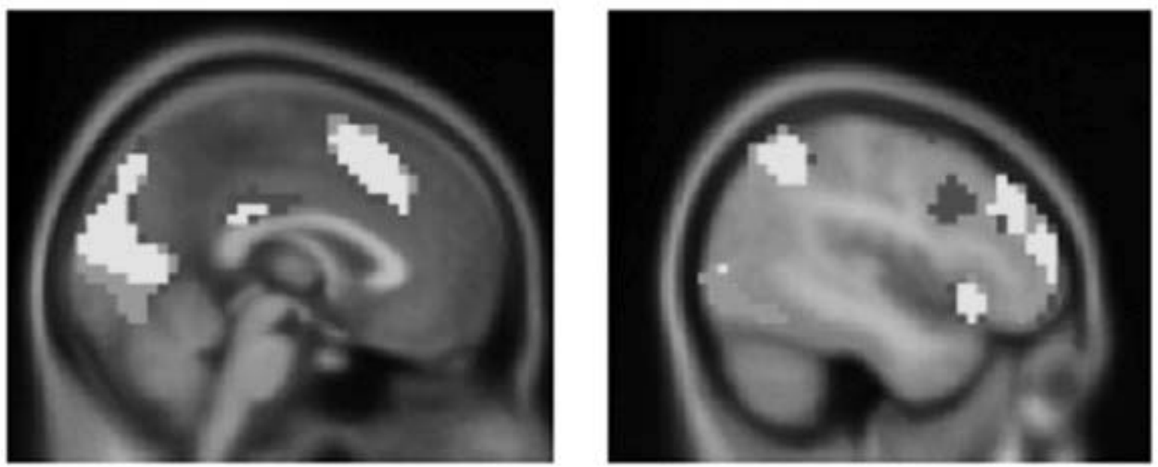
β areas (p<0.001)



x = 4mm

y = 16mm

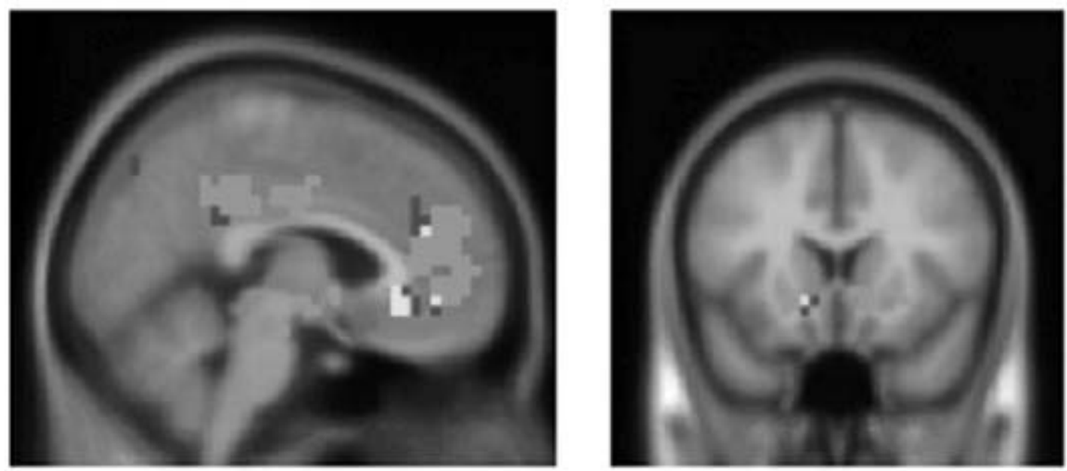
δ areas (p<0.01)



x = 0mm

x = -44mm

β areas (p<0.01)



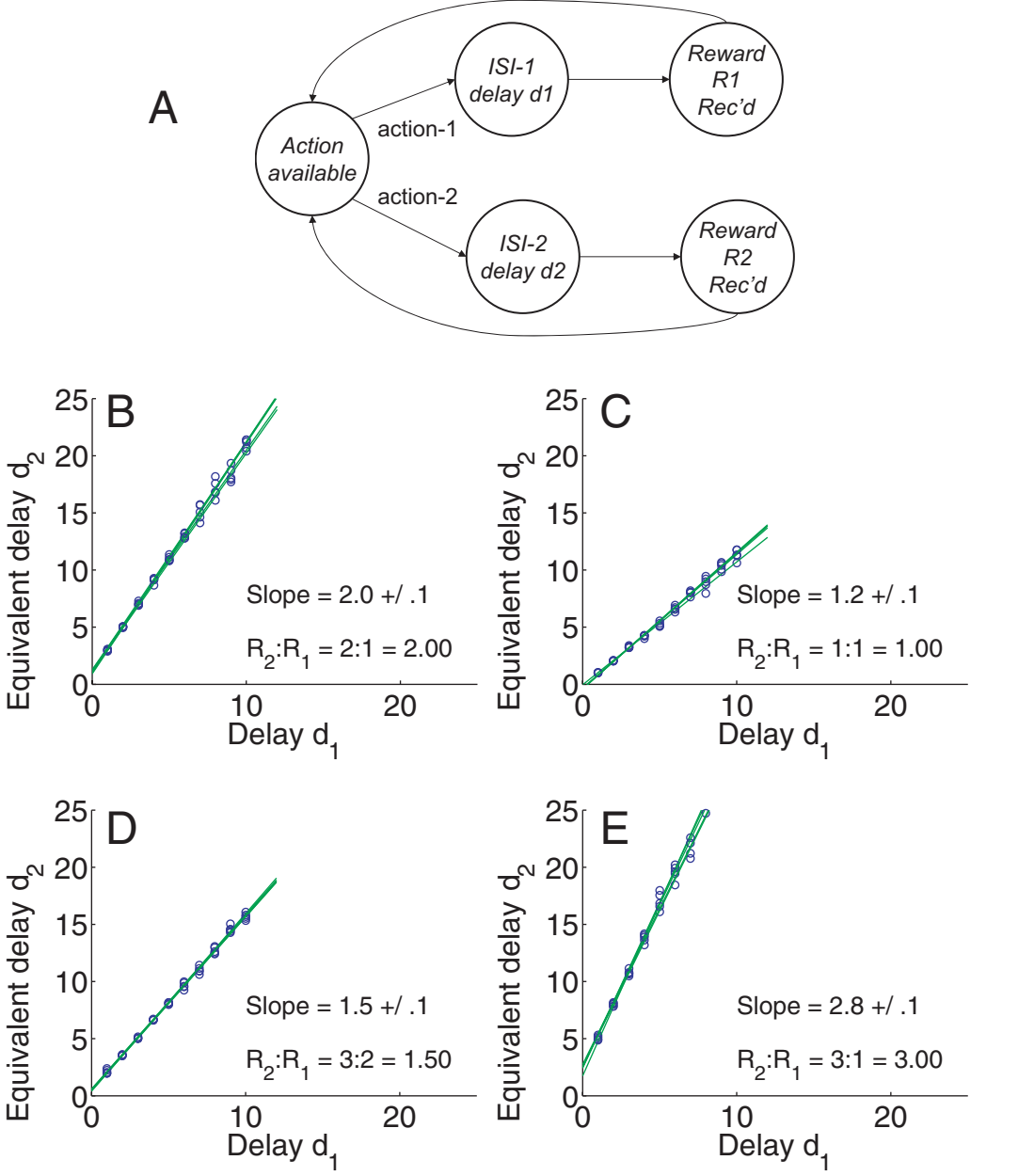
x = 4mm

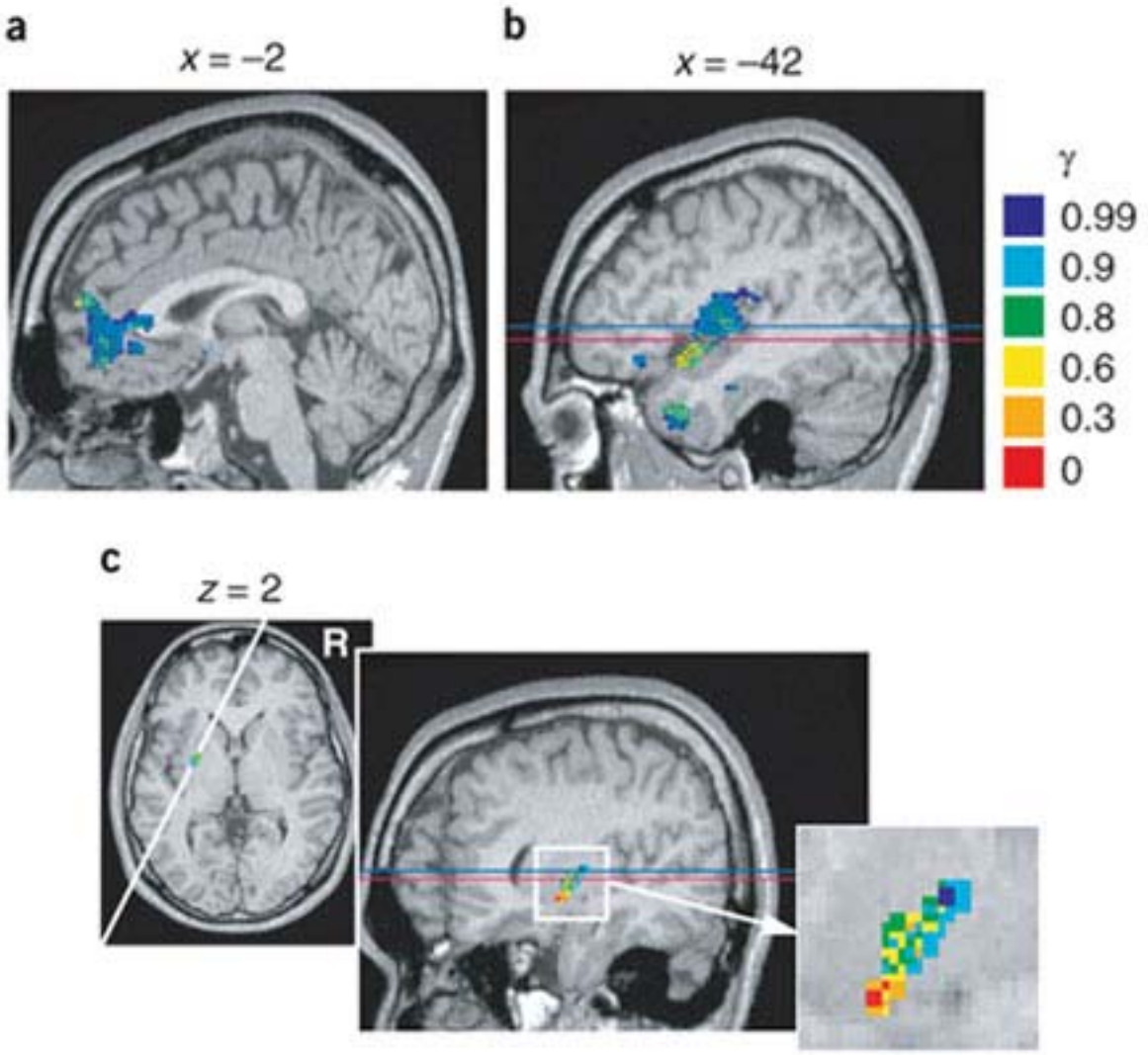
y = 16mm

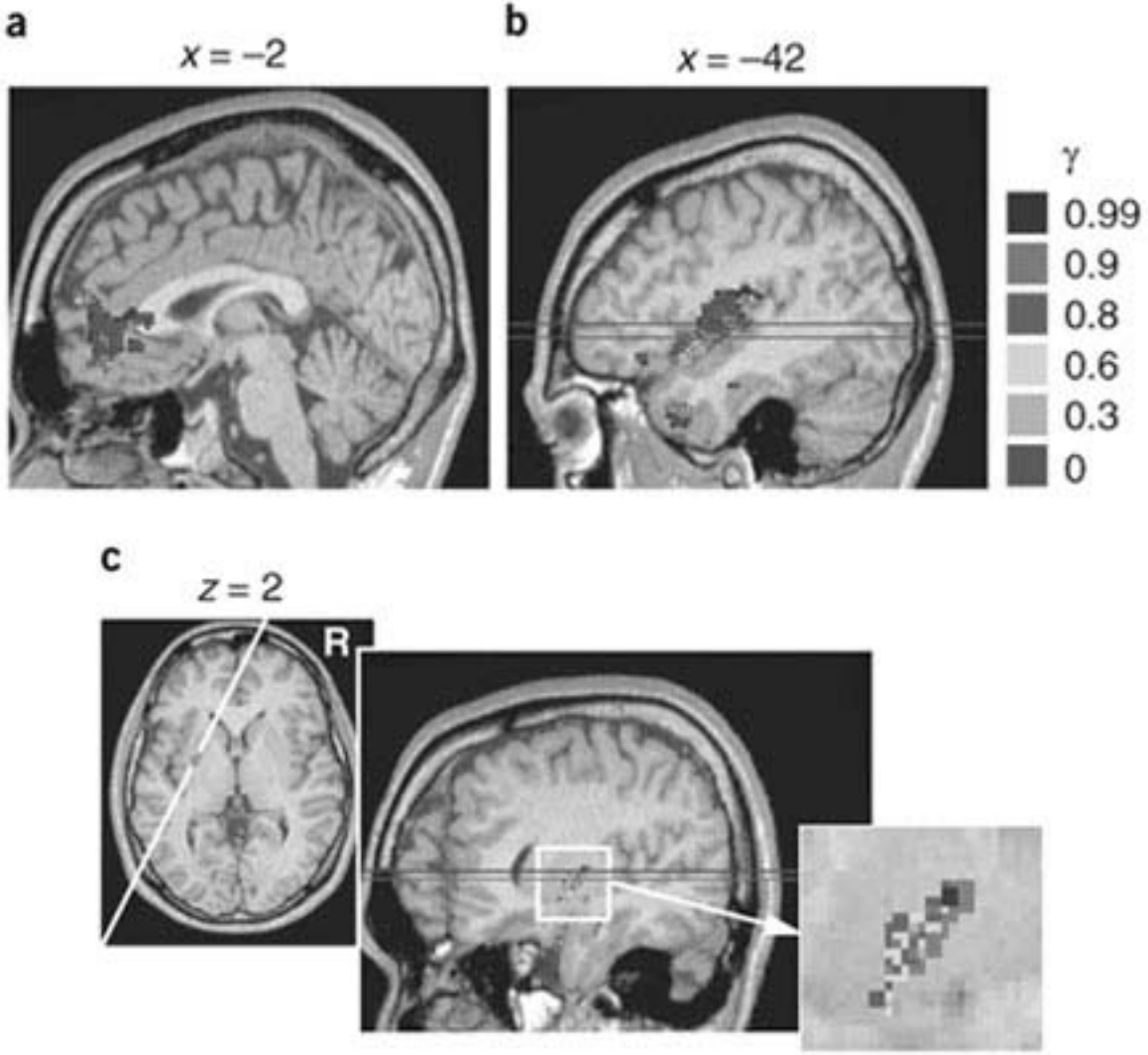
Juice

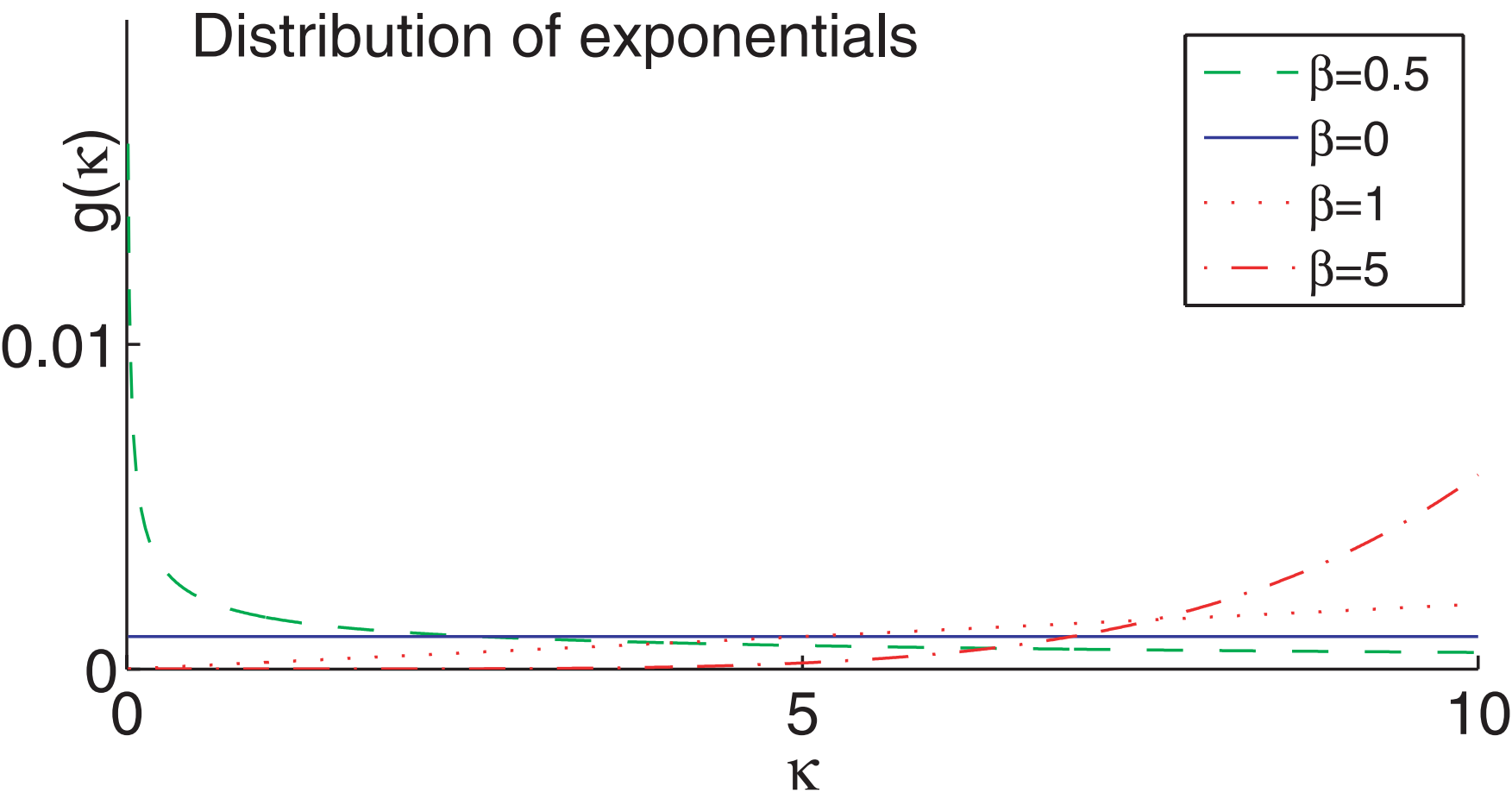
Money

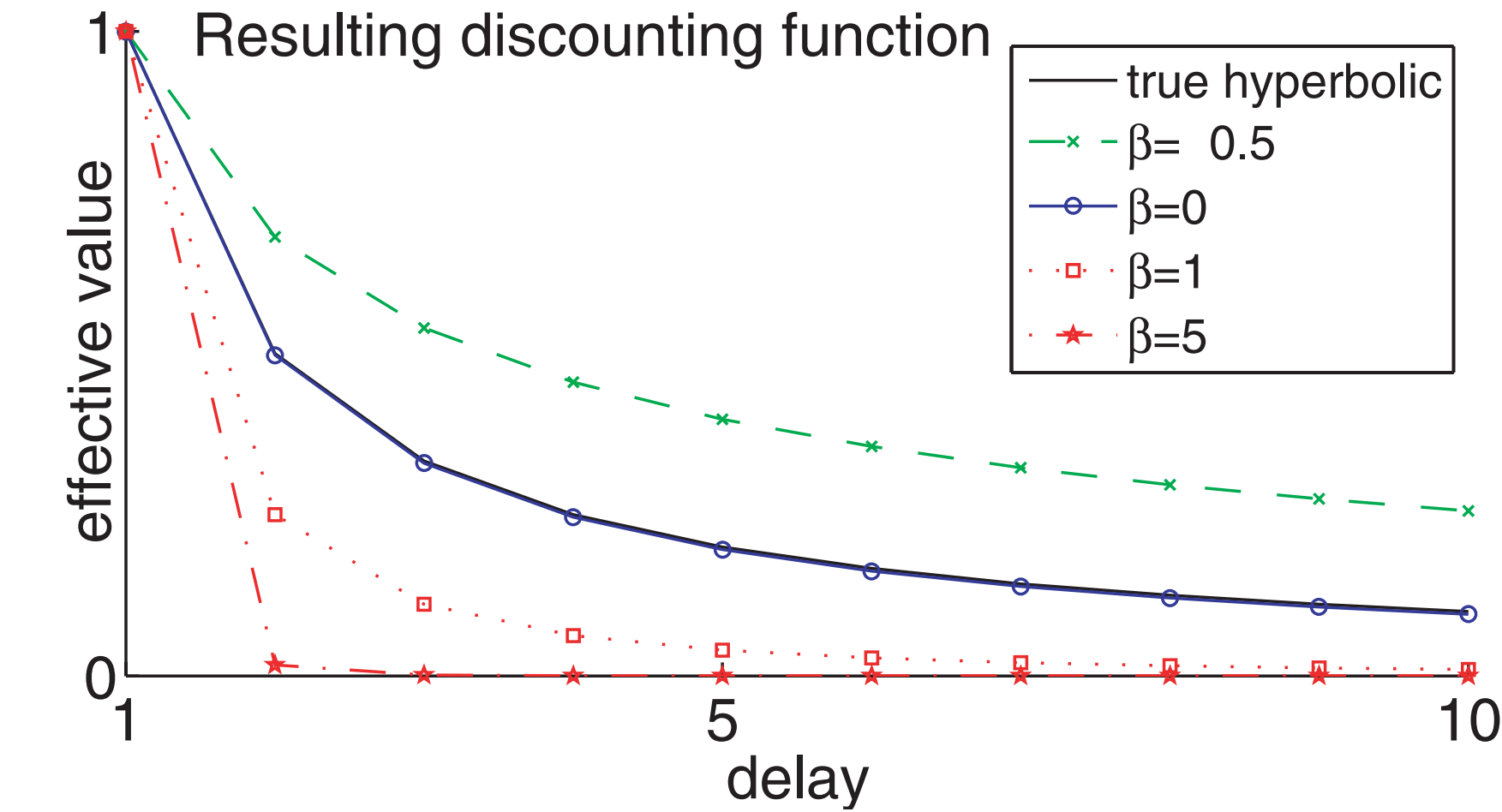
Both


A
Action available
action-1
action-2
ISI-1 delay d1
ISI-2 delay d2
Reward R1 Rec'd
Reward R2 Rec'd
B
Slope = 2.0 +/ .1
R2:R1 = 2:1 = 2.00
C
Slope = 1.2 +/ .1
R2:R1 = 1:1 = 1.00
D
Slope = 1.5 +/ .1
R2:R1 = 3:2 = 1.50
E
Slope = 2.8 +/ .1
R2:R1 = 3:1 = 3.00
Equivalent delay d2
Delay d1
0
5
10
15
20
25


a
x = −2
b
x = −42
γ
0.99
0.9
0.8
0.6
0.3
0
c
z = 2
R


a
x = −2
b
x = −42
γ
0.99
0.9
0.8
0.6
0.3
0
c
z = 2
R


Distribution of exponentials
g(κ)
0.01
0
0
5
10
κ
β=0.5
β=0
β=1
β=5


Resulting discounting function
effective value
1
0
1
5
10
delay
true hyperbolic
β= 0.5
β=0
β=1
β=5