Judgment and Decision Making, Vol. 5, No. 1, February 2010, pp. 1–10

You don't want to know what you're missing: When information about forgone rewards impedes dynamic decision making

A. Ross Otto* and Bradley C. Love
Department of Psychology, University of Texas at Austin

Abstract

When people learn to make decisions from experience, a reasonable intuition is that additional relevant information should improve their performance. In contrast, we find that additional information about foregone rewards (i.e., what could have gained at each point by making a different choice) severely hinders participants' ability to repeatedly make choices that maximize long-term gains. We conclude that foregone reward information accentuates the local superiority of short-term options (e.g., consumption) and consequently biases choice away from productive long-term options (e.g., exercise). These conclusions are consistent with a standard reinforcement-learning mechanism that processes information about experienced and forgone rewards. In contrast to related contributions using delay-of-gratification paradigms, we do not posit separate top-down and emotion-driven systems to explain performance. We find that individual and group data are well characterized by a single reinforcement-learning mechanism that combines information about experienced and foregone rewards.

Keywords: decision-making, delay of gratification, learning, self-control, dynamic environments, reinforcement learning.

1 Introduction

When immediate temptations conflict with long-term aspirations, immediate temptations often prevail and important goals remain unfulfilled (Loewenstein, 1996; Rachlin, 1995). Such failures of self-control are well documented in behavioral domains as diverse as dieting, smoking, and interpersonal conflict (Baumeister, Heatherton, & Tice, 1996). The ability to forego small immediate rewards in order to receive larger future rewards is viewed as a hallmark of effective self-control in both humans and animals (Ainslie, 1975; Rachlin & Green, 1972). In this report, we examine the impact of information about foregone (or fictive) outcomes on human decision-making behavior in situations in which short- and long-term rewards are in conflict. These forgone outcomes are counterfactual rewards that *could* have been obtained had one made alternate choices. Our task captures aspects of real-world tasks in which people face repeated choices with outcomes determined by past choices as well as current choices.

In a classic study of self-control in children, Mischel, Shoda, and Rodriguez (1989) found that preschoolers' ability to forego immediate gratification (e.g., one cookie immediately) in order to attain more valuable eventual outcomes (e.g., two cookies after a brief delay) is predictive of a number of key competencies in adolescence. Several studies suggest that the most effective strategy for delaying gratification is directing one's attention away from the reward-related stimuli during the waiting period (Eigsti et al., 2006; Mischel et al., 1989; Rodriguez, Mischel, & Shoda, 1989). Specifically, children who direct their attention away from the smaller immediate reward are better able to forego short-term gains, allowing them to maximize long-term gains.

The hot/cool systems framework (Metcalfe & Mischel, 1999) is one popular theory of performance in delay of gratification studies. According to this theory, one's level of self-control is dictated by interacting "hot" and "cool" systems. The hot system is fast, emotionally driven, and automatically triggered by reward-related stimuli. In contrast, the cool system is slow, reflective, and top-down-goal-driven. The interplay between these two systems can either undermine or support one's efforts to delay gratification. External (e.g., whether the immediate reward is visible) and internal factors (e.g., whether one's attention is focused on the immediate reward) that accentuate immediate rewards activate the hot system. As immediate rewards become more salient, one becomes more likely to succumb to the control of the hot system, which can lead to failure to delay gratification.

One criticism of previous self-control studies is that they involve explicit tradeoffs of intertemporal options

*This research was supported in part by AFOSR Grant FA9550–07–1-0178 and ARL Grant W911NF-09–2-0038 to Bradley C. Love. Address: A. Ross Otto, Department of Psychology, University of Texas, Austin, Texas 78712. E-mail: rotto@mail.utexas.edu.

that do not occur in real-life situations (e.g., Mischel et al., 1989; Rachlin & Green, 1972). In these paradigms, participants are told explicitly when the larger, delayed rewards will be received. In the real world, people are rarely given explicit information about the long-term consequences of immediate actions (Rick & Loewenstein, 2008).

For example, consider the conflict between short- and long-term goals facing the executive of an expanding firm. At each choice point, the executive can elect to continue investing in new equipment and training, thus increasing the firm's future output and boosting long-term profits, or the executive can instead cut costs in order to boost short-term profits, thus deferring the future benefits of the new investment. The executive's choices effectively influence the *state* of the environment, which affects future returns. Here, the precise long-term consequences of the executive's choices are not known until future time points. The long-term optimal choice must to some extent be learned experientially through interactive exploration of the decision space (Busemeyer, 2002), which, in our example, corresponds to observing the increasing returns resulting from continued investment in equipment and training.

The present work examines optimal long-term choice behavior in a dynamic task with recurring choices, providing a more realistic representation of many short- and long-term tradeoffs in daily life. Whereas the explicit tradeoff paradigm treats choices as static, one-shot decisions, many real-world decisions are often informed by past outcomes and one's current situation is determined by past choices (Busemeyer & Pleskac, 2009).

Notably, an experimental paradigm proposed by Herrnstein and colleagues (1993) affords examination of individuals' ability to maximize long-term rewards in dynamic choice environments. Consider the task reward structure depicted in Figure 1. The horizontal axis represents the current state of the decision environment, whereas the vertical axis represents the reward from selecting either choice. In every state, one option (which we refer to as the *Short-term* option) always yields a higher immediate reward than the other option (which we refer to as the *Long-term* option). The state of the environment is defined by the number of Long-term choices made over the last ten trials. Making a larger proportion of Long-term choices over time moves the current state of the environment rightwards on the horizontal axis, thus increasing the rewards for both choices. Effectually, choices that yield larger immediate rewards negatively affect future rewards (analogous to cutting costs in the above example), whereas options that are less immediately attractive lead to larger future rewards (analogous to continuing investment in new equipment and training). In this dynamic choice environment, maximizing long-term (i.e., global) rewards requires forgoing larger immediate (i.e., local) rewards guaranteed by the Short-term option and continually making Long-term choices. A growing literature has examined factors that bear on the decision-maker's ability to learn the reward-maximizing choice strategy in similar environments (Herrnstein et al., 1993; Otto, Gureckis, Markman, & Love, 2009; Tunney & Shanks, 2002).

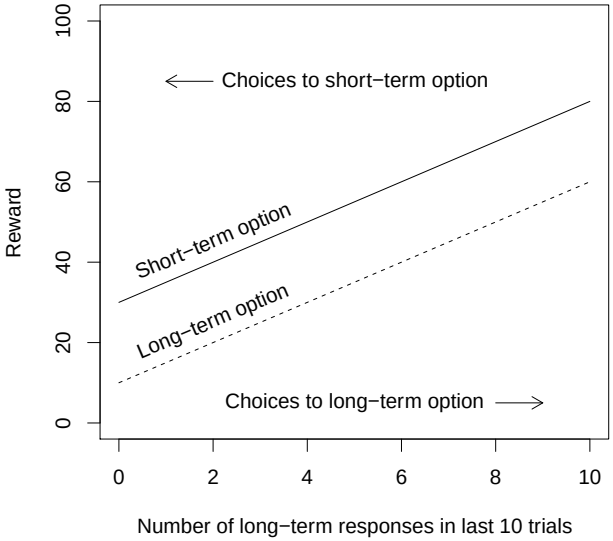

Figure 1: Reward (vertical axis) for the two choices as a function of response allocations over previous 10 trials (horizontal axis). Consider a participant who has made only Short-term choices for 10 trials in a row, making the state 0. The rewards from the Long- and Short-term choice rewards would be 10 and 30 respectively. If she makes one Long-term choice at this point, the task state would change to 1, as only 1 out of 10 of the last trials in her history were Long-term choices. Consequently, Long- and Short-term choices would result in rewards of 15 and 35 respectively. Selections to the Long-term choice effectively move the participant rightwards on the horizontal axis, while selections to the Short-term choice move the participant leftwards on the horizontal axis. Thus, the optimal strategy is to choose the Long-term option on every trial (assuming that the end of the sequences is unknown).

In this report, we consider dynamic decision making tasks in which people learn about rewards resulting from both chosen *and* unchosen (i.e., foregone) options. For example, an executive may be able to both observe the actual result of further investment and calculate the hypothetical cost savings that would have resulted from suspending investment. Although the neural substrates of a signal representing foregone rewards, distinct from directly experienced rewards, have been identified in both humans (Boorman et al., 2009) and primates (Hayden, Pearson, & Platt, 2009), it is unclear how these reward

https://doi.org/10.1017/S1930297500001984 Published online by Cambridge University Press

signals combine to shape behavior. Behavioral research in human decision making has yielded some insight into how knowledge of foregone rewards can affect choice behavior. In settings where rewards change over time, foregone rewards can support optimal choice when "chasing" foregone rewards is advantageous and likewise hinder optimal choice when "chasing" foregone rewards is disadvantageous. Further, sensitivity to foregone rewards appears to diminish over time in repeated-choice settings (Ert & Erev, 2007; Grosskopf, Erev, &Yechiam, 2006; Yechiam & Busemeyer, 2006).

The present work extends this research to dynamic decision-making environments. Unlike the static tasks which have been used to examine the effects of foregone rewards information, the rewards of performing actions in dynamic tasks are contingent upon past choice behavior. Namely, rewards on the present trial are determined by one's recent history of choices. We consider a particular form of sequential dependency in which short- and long-term rewards conflict. In this task, the strategy for maximizing long-term rewards is not obvious to the decision-maker at the outset. Thus a computational challenge arises — described in greater detail below — of crediting past decisions that have indirectly led to better or worse rewards in the present.

We hypothesize that, when short- and long-term rewards are in conflict, foregone rewards accentuate the local superiority of the Short-term option (e.g., cutting costs) and consequently bias choice away from the Long-term option (e.g., investment in equipment and training). Our predictions for dynamic decision tasks are consistent with findings from delay-of-gratification studies that find that focusing attention away from a short-term reward increases the likelihood of exercising self-control to gain a delayed, but larger reward (e.g., Rodriguez et al., 1989).

To investigate this possibility, we manipulate information presented about foregone rewards after each choice in the laboratory task outlined above (see Figure 1). For example, after a Long-term choice, a participant might be shown the 30 points that were actually earned as well as the 50 points that could have been earned had the participant made the Short-term choice. Our hypothesis is that the forgone information will accentuate the Short-term option's larger immediate payoff and discourage exploration of the decision space, leading to behavior that is suboptimal in the long run. To foreshadow our results, we find that inclusion of more information (i.e., about forgone payoffs) makes it less likely that the decision-maker will maximize long-term returns. This result, which is surprising on the surface, is anticipated by a standard reinforcement learning (RL: Sutton & Barto, 1998) mechanism that has a single information-processing stream, as opposed to separate hot and cool systems (Metcalf & Mischel, 1999).

According to the hot/cool systems view (Metcalf & Mischel, 1999), accentuating the greater immediate rewards associated with the Short-term option should increase the hot system's control over choice, leading to consistent selection of the globally inferior Short-term option. The RL mechanism explains this result without recourse to two systems with different computational properties. RL models learn from their interactions with the environment to maximize gains, making use of a reward signal that provides information about the "goodness" of actions. This framework has been used to model human decision-making behavior (Fu & Anderson, 2006; Gureckis & Love, 2009) as well as firing patterns in dopaminergic neurons in primates (Schultz, Dayan, & Montague, 1997).

Our RL model demonstrates that weighting of a fictive (i.e., forgone) reward signal for the action not taken impedes exploration of the decision space, because, in tasks where short- and long-term rewards conflict, adequate exploration is necessary to reach the optimal choice strategy. In the absence of specialized hot and cool components, this associative account provides a simple account of choice behavior for this and related tasks.

To rule out that the effect of foregone rewards does not arise from confusion stemming from additional information (Ert & Erev, 2007; Payne, Bettman, & Johnson, 1993), but rather arises from highlighting the local superiority of the Short-term option, we include an additional condition in which specious foregone rewards are provided. We sought a manipulation that would demonstrate that foregone reward information systematically biases choice and does not merely overload participants with information, hindering optimal choice behavior. To demonstrate this, we employ a control condition termed False Foregone Rewards (False-FR). In this condition, when the Short-term option is chosen, the foregone rewards from the Long-term option appear greater than the experienced rewards in order to promote a belief that the Long-term option is locally superior. Likewise, when the Long-term option is chosen, the foregone rewards from the Short-term option appear smaller than the experienced reward in order to promote a belief that the Short-term option is locally inferior.

To link to a real-world example, this manipulation is akin to a college student telling a friend who spent the night studying at the library that the party they had foregone (which was actually fun) was boring. In short, actual feedback for the chosen option is veridical, but information about the forgone option favors the Long-term option locally. If additional information about forgone rewards does not confuse participants, we should find that participants provided with specious information about forgone rewards should make more optimal choices than participants not provided with foregone rewards. In other

https://doi.org/10.1017/S1930297500001984 Published online by Cambridge University Press

words, we predict that "more is less" when local information emphasizes short-term gains, but "more is more" when local information emphasizes long-term gains. To foreshadow our results, this prediction holds and is anticipated by our RL model.

2 Method

2.1 Participants

Seventy-eight undergraduates at the University of Texas at Austin participated in this experiment for course credit plus a small cash bonus tied to performance on the task. Twenty-six participants were assigned to each of the three conditions: No Foregone Rewards (No-FR), True Foregone Rewards (True-FR), and False Foregone Rewards (False-FR).

2.2 Materials

The experiment stimuli and instructions were displayed on 17-inch monitors. The participants were told that their goal was to maximize overall long-term points by pressing one of two buttons each trial, and that each trial, they would be shown the number of points they earned from their choice. Crucially, the participants were not informed about the properties of each choice, which were simply labeled "Option A" and "Option B." With these minimal instructions, participants needed to use their experience with outcomes of previous choices in order to learn the optimal choice strategy.

Participants in the two foregone rewards conditions (True-FR and False-FR) were told that they would also see the number of points they could have earned from selecting the other option. Participants were also informed that a small cash bonus would be tied to their performance.

2.3 Procedure

A graphical depiction of the reward structure is shown in Figure 1. The number of points generated for selections of the Long-term option was 10+70*(h/10), while the reward for selecting the Short-term option was 30+70*(h/10), where h in both equations represents the number of Long-term choices made by the participant over the last 10 trials. Foregone rewards in the True-FR condition were determined by calculating the reward, given h, for the option not selected. In the False-FR condition, the foregone rewards displayed were defined by 5+70*(h/10) when the Long-term option was selected, and 35+70*(h/10) when the Short-term option was selected. A small amount of Gaussian noise (μ=0, σ=2) was added to the actual rewards as well as to the foregone rewards for participants in the True-FR and False-FR conditions.

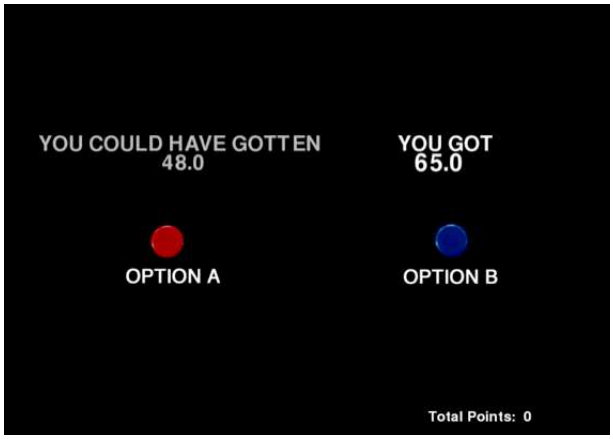

Figure 2: Screenshot of a trial in the True-FR condition. After the participant makes a selection, the immediate actual and foregone payoffs are presented, after which the participant is prompted to make a new selection.

The experiment consisted of 250 trials. At the start of the experiment, the number of Long-term responses over the last 10 trials (i.e., the state) was initialized to five. On each trial, participants were presented with two buttons, as shown in Figure 2. Using the mouse, participants clicked one of the buttons to indicate their choice. After each selection, the actual rewards from the chosen option (as well as the foregone rewards from the unchosen option, for participants in the True-FR and False-FR conditions) were presented above the chosen and unchosen choice buttons respectively. The mapping of response buttons to choices was counterbalanced across participants. At the end of the trials, participants were paid a cash bonus commensurate with their cumulative earnings.

3 Results

3.1 Performance measures

The main dependent measure was the proportion of trials for which participants made Long-term optimal responses, depicted in Figure 3A. A one-way ANOVA on this measure revealed a significant effect of foregone reward information, $F(2,77)=57.73$, $p<.001$, $\eta^2=.61$. More germane to our hypothesis, planned comparisons revealed that participants in the True-FR condition ($M=.21$, $SD=.06$) made significantly fewer Long-term optimal choices than participants in the No-FR condition ($M=.49$,$SD=.05$) [$t(50)=-3.92$, $p<.001$, $d=-5.07$].

https://doi.org/10.1017/S1930297500001984 Published online by Cambridge University Press

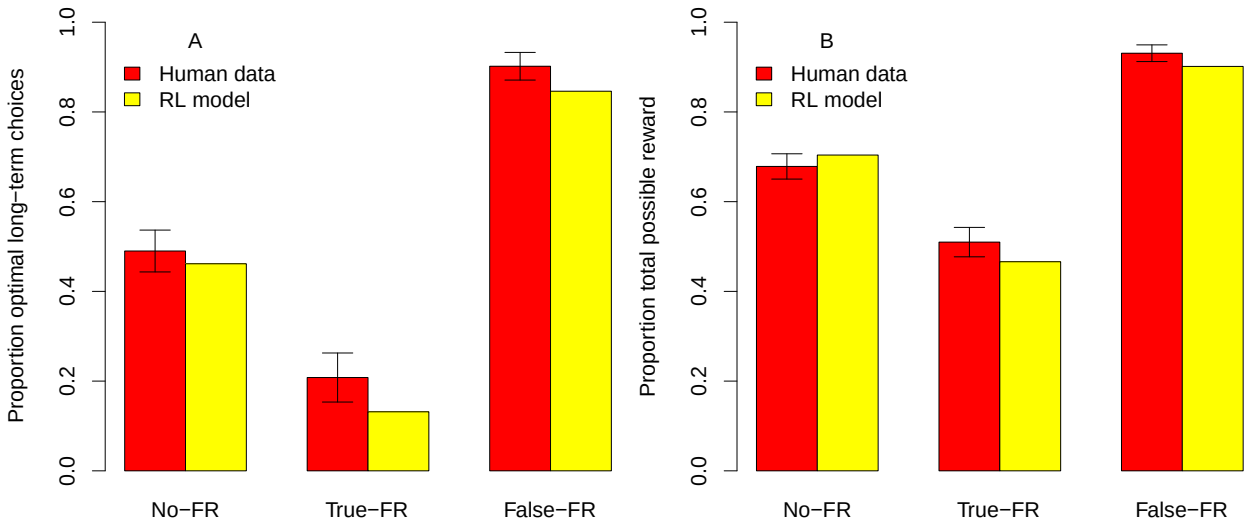

Figure 3: Comparison of human performance (red bars) with the RL model (yellow bars). The performance of each condition of the experiment is shown along with predicted overall responses proportions (Panel A) and proportion of possible cumulative rewards earned (Panel B) for the model using the best-fitting parameters. Error bars are standard errors.

Further, participants in the False-FR condition (M=.90, SD=.03) made significantly more Long-term choices than participants in the No-FR condition, $t(50)$=11.05, p<.001, d=9.94.

As another measure of optimal performance, we calculated each participant's cumulative rewards as a proportion of the maximum possible cumulative rewards available in the decision environment, depicted in Figure 3B. A one-way ANOVA revealed a significant effect of foregone rewards information, $F(2,77)$=60.78, p<.001, η^2=.62. Planned comparisons revealed that participants in the True-FR condition (M=.51, SD=.03) earned significantly fewer total points than participants in the No-FR condition (M=.67, SD=.03) [$t(50)$=-3.90, p<.001, d=-5.33]. Participants in the False-FR condition (M=.93, SD=.06) earned significantly more points than participants in the No-FR condition, $t(50)$=7.47, p<.001, d=5.06.

We also determined if information about immediate foregone rewards — particularly with respect to high immediate foregone rewards observed when making a Long-term choice — would permanently deter participants from making subsequent Long-term choices over the course of the experiment (akin to the "hot stove effect," see Denrell & March, 2001), precluding further exploration of the task environment. As a proxy, we ascertained whether each participant had made a Long-term choice at any point in the experiment after making a single Long-term choice. Strikingly, we found that 19% of participants in the True-FR condition never made a Long-term choice after choosing the Long-term option once, in contrast to 0% of participants in both the No-FR and False-FR conditions. This difference was statistically significant (p<.01, Fisher's exact test).

3.2 Model-based analysis

Rather than create a new model, we extend an existing RL (recognition learning) model that captures human behavior when short- and long-term rewards are in conflict (Bogacz et al., 2007). Our straightforward extensions allow the model to be applied to situations involving forgone rewards. Like the human learners in our task environment, the model begins with no prior expectations about the long- or short-term optimality of the two options. Following the general approach of RL, the model learns from its interactions with the environment to maximize gains, making use of a reward signal that provides information about the relative "goodness" of actions (Sutton & Barto, 1998). At each choice point, the model's policy dictates the probability of making a Long- or Short-term action. This policy is informed by continually updated estimates of the rewards associated with each action: more rewarding actions should be chosen over less rewarding actions. After each choice is made, the model updates its estimates of reward associated with action using a temporal-difference (TD) learning rule, whereby the size of the update is proportional to the difference between the model's predicted reward and actual obtained reward at that time.

Sequential decision-making tasks, where rewards from immediate actions are determined by the sequence of actions made in the recent past, pose a special problem for

https://doi.org/10.1017/S1930297500001984 Published online by Cambridge University Press

RL models. The model, like our participants, is provided with only the immediate rewards resulting from its actions. Thus the model faces the temporal credit assignment problem: an algorithm that learns to make optimal long-term choices must be able to assign credit to actions in the past that lead to rewards in the present. Our model solves this problem through the use of a partial memory for past actions. Under this approach, the model maintains separate memory traces for each action, which "remember" the frequency with which each action was made in the past. In the machine learning literature, these memory traces are called eligibility traces (ETs: Bogacz et al., 2007; Sutton & Barto, 1998).[1]

3.2.1 Model overview

Our model presents a straightforward extension to computational accounts of the impact of foregone rewards on choice behavior (e.g., Grosskopf et al., 2006) and computational accounts of learning in sequential decision-making environments (Bogacz et al., 2007). To model the impact of indirectly experienced (foregone) rewards in the True-FR and False-FR conditions, we propose an extension to standard RL by assuming that two learning episodes occur after an action is made: one update is made to the estimated reward associated with the *chosen* action, and a secondary update is made the estimated reward associated with the *foregone* action. By using separate learning rates for actual and foregone rewards, similar to the expectancy-weighted attraction model of strategy learning in economic games (Camerer & Ho, 1999), the model differentially weights foregone and directly experienced rewards. Psychologically, the claim is that people learn (perhaps to different degrees) from both experienced and forgone rewards within a unified system.

We demonstrate two principles in our model-based analysis. First, we reveal through simulations that information about foregone rewards affects our model's choice behavior in the same way it affects the choice behavior of human decision-makers. Second, we demonstrate through the use of separate learning rates that we can formally describe participants' "attention" to foregone rewards. Intuitively, we find that the larger the weight participants place on true foregone rewards, the more suboptimal their choice behavior is.

3.2.2 Formal model description

Under the RL framework, we assume that the goal of the model is to achieve the most reward possible by adapting its behavior based on its experienced reward with the two actions (option A and option B). The model maintains an estimate of the rewards associated with each action *i*, which we denote $Q(a_i)$. To generate responses, the model utilizes the "softmax" rule (Sutton & Barto, 1998) that transforms the rewards associated with each action into probabilities for executing each action (e.g., choosing the Short- or Long-term option). According to the softmax rule, the probability of selecting option *i* at trial *t* is given by the difference between the estimated rewards of the two options:

$$Pr(a_{\mathrm{i}}) = \frac{e^{\gamma \cdot Q(a_{\mathrm{i}},t)}}{\sum_{i=1}^{2} e^{\gamma \cdot Q(a_{\mathrm{i}},t)}} \qquad (1)$$

where γ is an exploitation parameter controlling the steepness of the rule's sensitivity to the difference in rewards, and $Q(a_i,t)$ is a current estimate of the reward associated with option a_i at trial *t*.

As a result of choosing action a_{chosen} on trial *t*, the model directly experiences reward $r_{obtained}(t)$. Similarly, the model has foregone reward $r_{foregone}(t)$ on trial *t* by not choosing the alternate action $a_{unchosen}$. These two reward sources provide the basis for updating the model's estimates of rewards associated for each action, $Q(a_{chosen})$ and $Q(a_{unchosen})$. To do so, the temporal-difference (TD) errors for both chosen and unchosen actions are calculated. This error term δ encapsulates the difference between experienced and predicted reward (Sutton & Barto, 1998). In the present model, TD errors are calculated for both directly experienced (obtained) and foregone rewards:

$$\delta_{\mathrm{obtained}}(t) = r_{\mathrm{obtained}}(t) - Q(a_{\mathrm{chosen}}, t) \qquad (2)$$

$$\delta_{\mathrm{foregone}}(t) = r_{\mathrm{foregone}}(t) - Q(a_{\mathrm{unchosen}}, t) \qquad (3)$$

where $Q(a_{chosen},t)$ denotes the model's estimated rewards for the chosen action on that trial, and $Q(a_{foregone},t)$ denotes the model's estimated rewards for the unchosen action that trial.

To handle temporal credit assignment, as explained above, the model maintains an ET for each action, denoted by *e,* representing the eligibility for updates to *Q* values. ETs scale the size of the update made to *Q* for each action. Thus, when action *i* has been chosen recently, the value of e_i will be large, and a correspondingly large update will be made to $Q(a_i)$. ETs can be seen as a kind of memory that enables linkage between past

[1] A different approach involves explicitly representing the state of the environment (in this case, the proportion of Long-term choices over the previous 10 trials) and learning the value of each state by combining the immediate rewards available in that state with the expected rewards for the next state the model transitions to. In essence, the model learns to value actions on the basis of estimates of future, discounted rewards associated with each action. In the present task the state is not fully observable in the environment and thus invalidates the assumption that the model can form an explicit state representation. Therefore, we incorporate ETs, rather than explicit state representations, in our model simulations. For extended discussion and psychological investigation of this matter, we refer the reader to Gureckis and Love (2009).

https://doi.org/10.1017/S1930297500001984 Published online by Cambridge University Press

choices and current rewards. Our implementation of ETs in this model is standard and straightforward.

At the beginning of the trials, the ET for each action *j* is initialized to 0. After each choice, both ETs are decayed by a constant term λ:

$$e_{\mathrm{j}}(t+1) = \lambda e_{\mathrm{j}}(t) \tag{4}$$

The decay process can be thought of as decaying memory or judged relevance of past actions: the farther in the past the action was last chosen, the less credit that action receives for present rewards.

To update expected rewards for each action, the model makes use of both experienced (Equation 2) and foregone (Equation 3) TD errors in updating Q values to incorporate the assumption that decision-makers learn from both experienced and foregone rewards. In order to accommodate learning from both sources of reward, the model temporarily increments the ETs for both the chosen and unchosen actions. This temporary increment of both ETs allows the model to treat both chosen and foregone actions as eligible for updating as if both were chosen. The update of the estimated reward associated with each action is governed by the ET for that action — which scales the size of the update — and a learning rate parameter α:

$$\begin{aligned} Q(a_{\text{chosen}}, t+1) &= Q(a_{\text{chosen}}, t) + \alpha_{\text{obtained}} \cdot \\ &\quad [e_{\text{chosen}}(t)+1] \cdot \delta_{\text{obtained}} \end{aligned} \tag{5}$$

$$\begin{aligned} Q(a_{\text{unchosen}}, t+1) &= Q(a_{\text{unchosen}}, t) + \alpha_{\text{foregone}} \cdot \\ &\quad [e_{\text{unchosen}}(t)+1] \cdot \delta_{\text{foregone}} \end{aligned} \tag{6}$$

In Equations 5 and 6, α_{chosen} and $\alpha_{foregone}$ are learning rate parameters for actual and foregone rewards respectively and e_{chosen} and $e_{foregone}$ are ETs for the chosen and foregone actions respectively. These updated Q values are used to inform the model's action selection (Equation 1) on the next trial. Equations 5 and 6 have the same form, consistent with the stance that experienced and forgone rewards are processed within a unified system.

Finally, following the update of both Q-values, a persistent increment is made to the chosen action's ET in order to update the model's memory of its choice history:

$$e_{\text{chosen}}(t+1) = e_{\text{chosen}}(t) + 1 \tag{7}$$

As a consequence, the chosen action's Q-value becomes more eligible and hence, receives larger updates in future trials per Equations 5 and 6 (Sutton & Barto, 1998).

3.2.3 Model fitting procedure

We employed two different methods of fitting the model to the participant data. First, in order to derive model predictions for performance between the three conditions (No-FR, True-FR, and False-FR), we conducted simulations where the model was given the same feedback as participants and made 250 choices in the dynamic task environment. Second, we fit model to participants individually using maximum likelihood estimation, allowing for recovery of parameter information describing individual differences in task performance.

Group simulations Performance for the model was measured in the same way as for participants, using proportion of Long-term choices over the course of the experiment. We found a single set of parameters (γ, $\alpha_{obtained}$, $\alpha_{foregone}$, and λ) for all three conditions by minimizing root-mean squared error (RMSE) between average model performance and participant performance averaged over 50-trial blocks. Note that participants in the No-FR condition did not have access to information about foregone rewards, so the value of $\alpha_{foregone}$ had no effect on model behavior. Our best-fitting parameter values were .12, .11, .49, and .70 for γ, $\alpha_{obtained}$, $\alpha_{foregone}$, and λ respectively, resulting in a RMSE of 0.233.

Figures 3A and 3B compare the performance of the model to the performance of participants in our experiment. The model exhibits the same ordinal pattern of results — namely, veridical information about foregone rewards hinders reward-maximizing choice while bogus foregone reward information facilitates optimal choice — given a single mechanism and set of parameters between the conditions. Intuitively, observations of veridical foregone rewards exacerbate the immediate superiority of the Short-term option in the model's estimates of choice values, biasing the model's future choices away from the Long-term option

Individual fits We also examined if individual differences in foregone learning rates could predict overall reward-maximizing behavior. We fit the model to each participant, letting each participant's foregone learning rate (parameterized by $\alpha^s_{foregone}$) vary on an individual basis. In doing so, we could evaluate the extent to which a participant's weighting of foregone rewards influences his or her ability to discover the globally optimal choice strategy. We fit the behavior of all participants using single values of the remaining model parameters (γ, $\alpha_{obtained}$, and λ; for a similar procedure, see Daw et al., 2006). To capture individual differences in attention to foregone rewards, we fit $\alpha^s_{foregone}$ separately for each participant s. Specifically, our model fitting procedure sought parameter values that maximized the likelihood of the observed choices:

$$L = \prod_s \prod_t P_{\text{c,s,t}} \tag{8}$$

https://doi.org/10.1017/S1930297500001984 Published online by Cambridge University Press

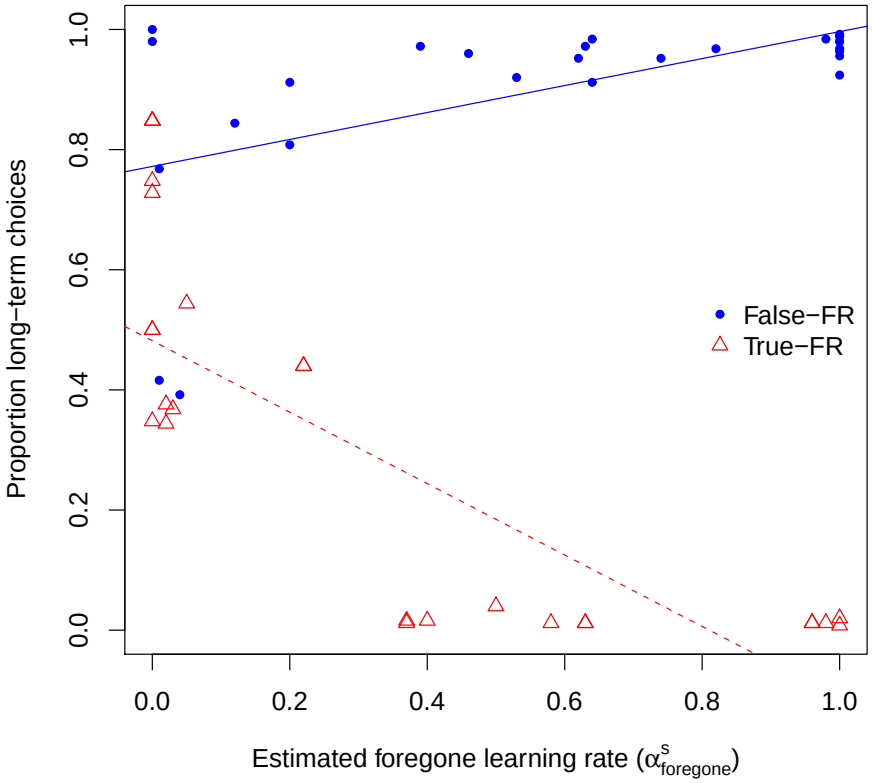

Figure 4: Maximum-Likelihood estimated participant foregone learning rate (horizontal axis) plotted against participant performance operationalized as proportion of Long-term choices (vertical axis) for True-FR and False-FR conditions. Line of regression is plotted for both conditions.

over each participant s and trial t, where $c_{s,t}$ reflects the choice made by subject s on trial t, informed by the participant's choice and outcome experience up to trial t. Note that, as in our above simulations, parameter $\alpha^s_{foregone}$ had no effect on choice likelihoods in the No-FR conditions. The best-fitting yoked parameter values were .12, .03, and .37 for γ, $\alpha_{obtained}$, and λ respectively.

Figure 4 depicts the relationship between values of $\alpha^s_{foregone}$ and choice performance among individuals in both foregone reward conditions. We predicted that, among participants in the True-FR condition, the more an individual weighted foregone rewards, the less likely he or she was able find the optimal long-term choice strategy. Indeed, we found that individual estimated values of $\alpha^s_{foregone}$ negatively predicted Long-term optimal responding in this condition — operationalized as the proportion of Long-term choices made, $r(24)= -.48$, $p<.01$. In contrast, we predicted that in the False-FR condition, the more weight individuals placed on the specious forgone rewards, the more the participants would be able to discover the optimal strategy. In this False-FR condition, we found that estimated $\alpha^s_{foregone}$ values positively predicted Long-term optimal responding, $r(24)= .56$, $p<.01$.

4 Discussion

More information, especially when capacity is available to process it, is usually considered a positive. However, as shown in this report, when additional information is included about what could have been (i.e., forgone rewards), people perform worse in a dynamic decision task in which short- and long-term rewards are in conflict. As in a number of real-world situations, maximization of long-term rewards in our study required that the decision-maker learn to forego larger immediate rewards guaranteed by one option and instead persist with a locally inferior option (Rachlin, 1995). Our study revealed that veridical information about foregone rewards holds deleterious effects for reward-maximizing choice in our task environment.

More specifically, we found that veridical information about foregone rewards hinders exploration of the decision space, which is necessary for discovery of the long-term optimal choice strategy in our dynamic environment. Our model-based analysis suggests that a simple reward-learning mechanism can explain the detrimental effects of veridical information about foregone rewards.

https://doi.org/10.1017/S1930297500001984 Published online by Cambridge University Press

Specifically, a second learning episode which updates the model's estimated value of the foregone option results in over-learning of the local rewards of both options, preventing the decision-maker from overcoming the difference in local rewards between Short- and Long-term options — which is necessary for exploration of the decision space. This mechanistic account offers an explanation of how individuals' weighting of indirectly experienced (i.e., foregone) rewards impacts long-term optimal behavior.

Unlike other proposals, such as the hot/cool framework (Metcalfe & Mischel, 1999), we demonstrate that a single, general, information-processing mechanism predicts the observed behavior. We do not posit separate emotional and deliberative mechanisms, but instead propose that behavior is governed by a single RL mechanism that seeks to maximize rewards by using information about both observed and forgone payoffs. The argument for our account relies on parsimony and computational considerations. We do not deny that hot and cool systems exist and that they could govern behavior in related tasks. Rather, we suggest that data from our task and related tasks naturally follow from basic RL principles and do not justify the added complexity of a theory in which multiple systems, organized along diametrically opposed principles, are required.

Identification of mechanisms supporting choice in situations where the availability of immediately available rewards subverts maximization of long-term returns has been a topic of central interest in developmental psychology (Mischel et al., 1989; Rodriguez et al., 1989), neuroscience (Hare, Camerer & Rangel, 2009; McClure et al., 2004), and studies of animal behavior (Ainslie, 1974; Rachlin & Green, 1972). The repeated-choice dynamic task used in the present study is posed to further elucidate choice behavior under conflicting long- and short-term rewards. Not only does the task afford quantification of individual differences, but it also relies on the learner to explore the changing reward contingencies of options, yielding insight into mechanisms of learning long-term optimal choice. Further, the individual differences revealed by our individual model fits — as instantiated by foregone learning rate parameter values — may have predictive bearing on self-control behavior in an explicit-tradeoff paradigm (e.g., McClure et al., 2004). Future work is needed to evaluate the extent to which individual differences in dynamic tasks predict choice behavior in other task settings.

The present study extends the existing literature about foregone rewards by examining the effect of information about foregone rewards in a dynamic decision-making task where short- and long-term rewards are in conflict. Using standard RL techniques, we have described a general-purpose psychological model that correctly predicts individuals' patterns of choice behavior in this task. The predictions of this model are not constrained to the dynamic choice task discussed in this report. Rather, our proposed mechanism for learning from both experienced and foregone rewards is quite general and accounts for key effects observed in previous studies on foregone rewards, such as Problems 5 and 6 reported in Grosskopf et al. (2006). In those studies, participants made repeated binary choices in environments with correlated and uncorrelated noisy rewards. The main finding was that foregone rewards were more helpful when noise, and therefore reward values, were correlated. Using the parameter values reported in the simulations above, our model correctly predicts the effect of foregone rewards exhibited by their human participants. Because the Grosskopf et al. studies do not involve dynamic decision tasks (according to our above definition), the ET mechanism in our model should not be necessary to account for the key effects. Indeed, when ETs are removed, the model still makes the correct predictions.

As detailed above, dynamic choice tasks bear a number of similarities to real-world decision-making situations where long- and short-term rewards conflict. For example, dieters find it difficult to forego the short-term rewards of eating palatable foods in the service of long-term weight control. One contemporary account of health behavior posits that weight control is problematic in an environment where consistent and salient presentation of palatable foods is detrimental to long-term dieting goals (Stroebe, 2008). Further, longitudinal research suggests that obesity — controlling for genetic ties and geographic proximity — spreads between peers through social influence (Christakis & Fowler, 2007). A possible mechanism underpinning social influence is that dieting individuals are frequently exposed to the palatable foods (which need be foregone) that their obese companions indulge in, resulting in a failure to fulfill long-term weight goals. Indeed, our simple model predicts that indirectly experienced rewards prevent individuals from overcoming the difference in immediate rewards between eating versus abstaining from palatable foods, resulting in nonfulfillment of long-term health goals.

The present study reveals that, under the circumstances observed here, withholding information about local rewards from decision-makers can actually facilitate long-term optimal choice. These results underscore the consequences of local feedback in situations, such as that facing the executive or dieter in our examples, where globally optimal behavior is not locally obvious.

https://doi.org/10.1017/S1930297500001984 Published online by Cambridge University Press

References

Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Research, 1153*, 111–21.

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron, 62,* 733–743.

Busemeyer, J. R. (2002). Dynamic Decision Making. In Smelser, N. J. & Baltes, P. B. (Eds.) *International Encyclopedia of the Social and Behavioral Sciences.* Oxford: Elsevier Press Vol. 6., 3903–3908.

Camerer, C., & Ho, T. (1999). Experienced-Weighted Attraction Learning in Normal Form Games. *Econometrica, 67,* 827–874.

Christakis, N. A., & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *The New England Journal of Medicine, 357,* 370–379.

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

Denrell, J., & March, J. G. (2001). Adaptation as Information Restriction: The Hot Stove Effect. *Organization Science, 12,* 523–538.

Ert, E., & Erev, I. (2007). Replicated alternatives and the role of confusion, chasing, and regret in decisions from experience. *Journal of Behavioral Decision Making, 20,* 305–322.

Grosskopf, B., Erev, I., & Yechiam, E. (2006). Foregone with the Wind: Indirect Payoff Information and its Implications for Choice. *International Journal of Game Theory, 34,* 285–302.

Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*, 293–313.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmpfc valuation system. *Science, 324*, 646–648.

Herrnstein, R. J., Loewenstein, G. F., Prelec, D., & Vaughn, W. (1993). Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making, 6,* 149–185.

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

Mischel, W., Shoda, Y., & Rodriguez, M. (1989). Delay of gratification in children. *Science, 244*, 933–938.

Otto, A. R., Gureckis, T. M., Markman, A. B., & Love, B. C. (2009). Navigating through abstract decision spaces: Evaluating the role of state generalization in a dynamic decision-making task. *Psychonomic Bulletin & Review, 16,* 957–963.

Payne, J., Bettman, J., & Johnson, E. (1993). *The Adaptive Decision Maker*. New York: Cambridge University Press.

Rachlin, H. (1995). Self-control: Beyond commitment. *Behavioral and Brain Sciences, 18,* 109–159.

Rick, S., & Loewenstein, G. (2008). Intangibility in intertemporal choice. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363,* 3813–3824.

Rodriguez, M. L., Mischel, W., & Shoda, Y. (1989). Cognitive person variables in the delay of gratification of older children at risk. *Journal of Personality and Social Psychology, 57,* 358–367.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science, 275,* 1593–1599.

Stroebe, W. (2008). *Dieting, overweight, and obesity: self-regulation in a food-rich environment.* Washington, D.C.: American Psychological Association.

Sutton, R., & Barto, A. G. (1998). *Reinforcement Learning.* Cambridge, MA: MIT Press.

Tunney, R. J., & Shanks, D. R. (2002). A re-examination of melioration and rational choice. *Journal of Behavioral Decision Making*, 15, 291–311.