Overcoming Temptation: Incentive Design For Intertemporal Choice

Michael C. Mozer
Department of Computer Science
University of Colorado
Boulder, CO 80309-0430
mozer@colorado.edu

Shruthi Sukumar
Department of Electrical and Computer Engineering
University of Colorado
Boulder, CO 80309
Shruthi.Sukumar@colorado.edu

Camden Elliott-Williams
Department of Computer Science
University of Colorado
Boulder, CO 80309-0430
Camden.ElliottWilliams@colorado.edu

Shabnam Hakimi
Institute of Cognitive Science
University of Colorado
Boulder, CO 80309-0345, USA
Shabnam.Hakimi@colorado.edu

Adrian F. Ward
McCombs School of Business
University of Texas
Austin, TX 78705, USA
Adrian.Ward@mccombs.utexas.edu

Abstract

Individuals are often faced with temptations that can lead them astray from long-term goals. We're interested in developing interventions that steer individuals toward making good initial decisions and then maintaining those decisions over time. In the realm of financial decision making, a particularly successful approach is the prize-linked savings account: individuals are incentivized to make deposits by tying deposits to a periodic lottery that awards bonuses to the savers. Although these lotteries have been very effective in motivating savers across the globe, they are a one-size-fits-all solution. We investigate whether customized bonuses can be more effective in tasks involving delayed gratification. We formalize a delayed-gratification task as a Markov decision problem in which the agent must repeatedly choose between two actions: *defecting*, which obtains a small immediate reward, and *persisting*, which eventually obtains a large but delayed reward. We characterize individuals as rational agents subject to temporal discounting and stochastic fluctuations in *willpower*. Willpower is conceived of as modulating the subjective value of the small immediate reward. Our theory is able to explain key behavioral findings in intertemporal choice, including finish-line effects and the dependence of behavior on relative value of immediate and delayed rewards. We created an online delayed-gratification game in which players score points by selecting a queue to wait in and effortfully advancing to the front. Data collected from the game is fit to the model, and the instantiated model is then used to optimize predicted player performance over a space of incentives. We demonstrate that customized incentive structures can improve goal-directed decision making.

Keywords: intertemporal choice, delayed gratification, cognitive modeling, incentive design, mechanism design

Acknowledgements

This research was supported by NSF grants DRL-1631428, SES-1461535, SBE-0542013, SMA-1041755, and seed summer funding from the Institute of Cognitive Science at the University of Colorado. We thank Ian Smith and Brett Israelson for design and coding of the experiments.

Should you go hiking today or review your RLDM papers? Individuals are regularly faced with temptations that lead them astray from long-term goals. These temptations all reflect an underlying challenge in behavioral control that involves choosing between actions leading to small but immediate rewards and actions leading to large but delayed rewards. We introduce a formal model of this *delayed-gratification (DG)* decision task, extending the Markov decision framework to incorporate the psychological notion of willpower, and using formal models to optimize behavior by designing incentives to assist individuals in achieving long-term goals.

In the classic marshmallow test (Mischel and Ebbesen, 1970), children are seated at a table with a single marshmallow. They are allowed to eat the marshmallow, but if they wait while the experimenter steps out of the room, they will be offered a second marshmallow when the experimenter returns. In this DG task, children must continually contemplate whether to eat the marshmallow or wait for two marshmallows. Their behavior depends not only on the hypothetical discounting of future rewards but on an individual's moment-to-moment *willpower*—their ability to maintain focus on the larger reward and not succumb to temptation before the experimenter returns. Defection at any moment eliminates the possibility of the larger reward. We focus on modeling the temporal dynamics of behavior during the delay period.

Nearly all previous conceptualizations of *intertemporal choice*—situations involving decisions that produce gains and losses at different points in time—have focused on the shape of discounting functions and the initial 'now versus later' decision, not the time course. One exception is the work of McGuire and Kable (2013) who frame failure to postpone gratification as a rational, utility-maximizing strategy when the time at which future outcomes materialize is uncertain. Our theory is complementary in providing a rational account in the known time horizon situation. (Some DG tasks have a known time horizon, e.g., retirement planning or dieting for a wedding.) There is a rich literature on treating human decision making from the framework of Markov decision processes (*MDP*s; e.g., Shen et al., 2014; Niv et al., 2012), but this research does not directly address intertemporal choice. Kurth-Nelson and Redish (2010, 2012) have explored a reinforcement learning framework to model precommitment in decision making as a means of preventing impulsive defections. This interesting work focuses on the initial decision whether to precommit rather than the ongoing possibility of defection. To the best of our knowledge, we are the first to adopt an MDP perspective on intertemporal choice, a field which has relied primarily primarily on verbal, qualitative accounts.

1 Formalizing Delayed-Gratification Tasks as a Markov Decision Problem

We formalize a DG task as a Markov decision problem, which we will refer to as the *DGMDP*. We assume time to be quantized into discrete steps and we focus on situations with a known or assumed time horizon, denoted τ. At any step, the agent may DEFECT and collect a small reward, or the agent may PERSIST to the next step, eventually collecting a large reward at step τ. We use μ_{SS} and μ_{LL} to denote the *smaller sooner (SS)* and *larger later (LL)* rewards. Figure 1a shows a finite-state representation of a *one-shot* task with terminal states LL and SS that correspond to resisting and succumbing to temptation, respectively, and states for each time step between the initial and final times, $t \in \{1, 2, ..., \tau\}$. Rewards are associated with state transitions. The possibility of obtaining *intermediate* rewards during the delay period is annotated via $\boldsymbol{\mu}_{1:\tau-1} \equiv \{\mu_1, ..., \mu_{\tau-1}\}$, which we return to later.

Given the DGMDP, an optimal decision sequence is trivially obtained by value iteration. However, this sequence is a poor characterization of human behavior. With no intermediate rewards ($\boldsymbol{\mu}_{1:\tau-1} = \mathbf{0}$), it takes one of two forms: either the agent defects at $t = 1$ or the agent persists through $t = \tau$. In contrast, individuals will often persist some time and then defect, and when placed into the same situation repeatedly, behavior is nondeterministic. For example, replicability on the marshmallow test is quite modest ($\rho < 0.30$: Mischel et al., 1988).

The discrepancy between human DG behavior and the optimal decision-making framework might indicate an incompatibility. However, we prefer a *bounded rationality* perspective on human cognition according to which behavior is cast as optimal but subject to cognitive constraints. We claim two specific constraints.

1. Individuals exhibit moment-to-moment fluctuations in *willpower* based on factors such as sleep, hunger, mood, etc. Low willpower causes an immediate reward to seem more tempting, and high willpower, less tempting. We characterize willpower as a one-dimensional Gaussian process, $W = \{W_t\}$, with $w_1 \sim \text{Gaussian}(0, \sigma_1^2)$ and $w_t \sim \text{Gaussian}(w_{t-1}, \sigma^2)$. We suppose that willpower modulates an individual's subjective value of defecting at step t:

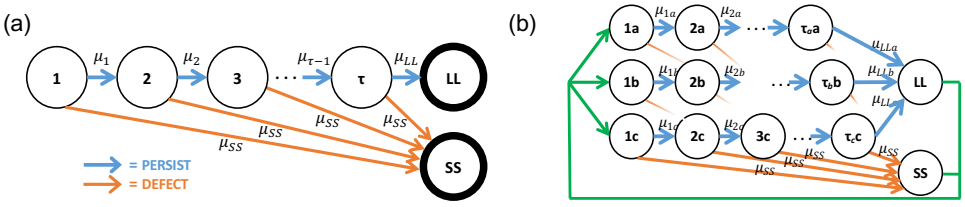

Figure 1: Finite-state environments formalizing (a) a one-shot DG task, and (b) an iterated DG task with variable delays and LL outcomes

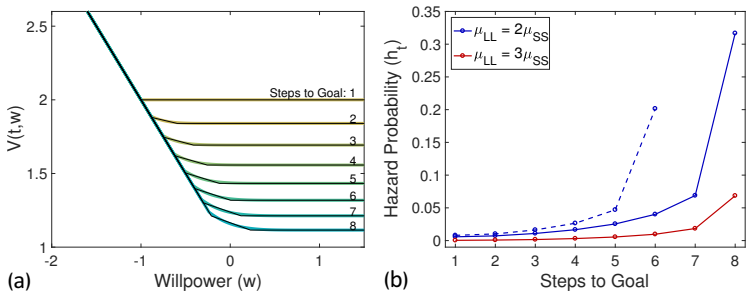

Figure 2: (a) Value function for a DGMDP with $\tau = 8$, $\sigma = .25$, $\sigma_1 = .50$, $\gamma = .92$, $\mu_{\text{E}} = \mu_t = 0$, $\mu_{\text{LL}} = 2$, $\mu_{\text{SS}} = 1$, exact (colored curves) and piecewise linear approximation (black lines). (b) Hazard functions for the parameterization in (a) (solid blue curve), with a higher level of LL reward (red curve), and with a shorter delay period, $\tau = 6$ (dashed blue curve).

$$Q(t, w; \text{DEFECT}) = \mu_{\text{SS}} - w, \tag{1}$$

where $Q(s; a)$ denotes the value associated with performing action a in state s, and the state space consists of the discrete step t and the continuous willpower level w.

2. Behavioral, economic, and neural accounts of decision making suggest that *effort* carries a cost, and that rewards are weighed against the effort required to obtain it (e.g., Kivetz, 2003). This notion is incorporated into the model via an effort cost, μ_{E} associated with persevering:

$$Q(t, w; \text{PERSIST}) = \begin{cases} \mu_{\text{E}} + \mu_t + \gamma\, \mathbb{E}_{W_{t+1}|W_t=w} V(t+1, w_{t+1}) & \text{for } t < \tau \\ \mu_{\text{LL}} & \text{for } t = \tau \end{cases} \tag{2}$$

Although the continuous willpower variable precludes an analytical solution for $V(t, w) \equiv \max_a Q(t, w; a)$, the shape of $V(t, w)$ permits a piecewise linear approximation (PLA) over w for each step t. Figure 2a shows that the PLA closely matches a solution obtained by fine discretization of the willpower space. Using the value function, we can characterize the agent's behavior in the DGMDP via the likelihood of defecting at various steps. With D denoting the defection step, we can compute the *hazard probability* $h_t \equiv P(D = t | D \geq t)$ for various scenarios (Figure 2b).

Taken as a rational account of human cognition, our theory explains two key phenomena in the literature. First, failure on a DG task is sensitive to the relative magnitudes of the SS and LL rewards (Mischel, 1974). Figure 2b presents hazard functions for two reward magnitudes. The probability of obtaining the LL reward is greater with $\mu_{\text{LL}}/\mu_{\text{SS}} = 3$ than with $\mu_{\text{LL}}/\mu_{\text{SS}} = 2$. Figure 2b can also accommodate the finding that environmental reliability and trust in the experimenter affect outcomes in the marshmallow test (Kidd et al., 2012): in unreliable or nonstationary environments, the expected LL reward is lower than the advertised reward, and the DGMDP is based on reward expectations. Second, a reanalysis of data from a population of children performing the marshmallow task shows a declining hazard rate over time (McGuire and Kable, 2013). The rapid initial drop in the empirical curve looks remarkably like the curves in Figure 2b.

2 Optimizing Incentives

With a computational theory of the DG task in hand, we explored a mechanism-design approach (Nisan and Ronen, 1999) aimed at steering individuals toward improved long-term outcomes. We asked whether we can provide incentives to rational value-maximizing agents that will increase their expected reward subject to constraints on the incentives. We explored two incentive structures. The first is related to *prize-linked savings accounts* (Kearney et al., 2010). The idea is to pool a fraction of the interest earned from all deposits to savings accounts fund a prize awarded by periodic lotteries. Although the account yields a lower interest rate to fund the lottery, the PLSA increases the net expected account balance due to greater commitment to participation. In our work, we show that we can search over a space lottery allocations (via the μ_t which specify timing and amount of awards, cast in terms of subjective value from prospect theory) to significantly increase total payout an individual receives, and that the ideal lottery allocation depends on the agent's discount rate. The second incentive structure we explored is based on experiments we conducted with human subjects, and involves offering a small portion of μ_{LL} at earlier points in time (via the μ_t).

3 Experiments

To collect and model behavioral data, we created a simple online delayed-gratification game in which players score points by waiting in a queue (Figure 3a). The upper queue is short, having only one position, and delivers a 100 point reward when the player is serviced. The lower queue is long, having τ positions, and delivers a $100\tau\rho$ point reward when the player is serviced. The *reward-rate ratio*, ρ, is either 1.25 or 1.50 in our experiments. Points awarded per queue are displayed left of the queue. The player (red icon) starts in a vestibule (right side of screen) and selects a queue with the up and down arrow keys. The game updates are clocked, at which time the player's request is processed and the queues advance (from

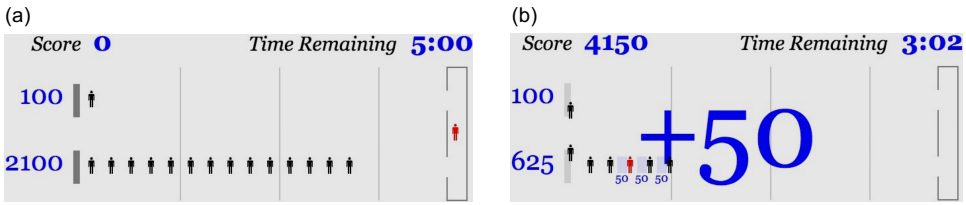

Figure 3: The queue-waiting game. (a) Initial game state. (b) A snapshot of the game, taken while the queue advances. In this condition, bonuses are offered at certain positions in the long queue. Point increments are flashed as they are awarded.

right to left). After choosing a queue, the player must hit the left-arrow key to advance or the up- or down-arrow keys to switch queues. If the player takes no action, the simulated participants behind the player jump past. When points are awarded, the screen flashes the points and a cash register sound is played, and the player returns to the vestibule and a new *episode* begins. In our experiments, the long-queue length τ varies from episode to episode. This iterative version of the DGMDP is described by Figure 1b. The vestibule in Figure 3a corresponds to state 1 in Figure 1b and lower queue position closest to the service desk to state τ. Note the left-to-right reversal of the two Figures, which has often confused the authors of this work.

We conducted a series of 4 experiments. Experiments 1 and 2 were designed to constrain parameters of the model, and Experiments 3 and 4 were designed to test the fully-constrained model. Here, we discuss only Experiments 1 and 4. In each experiment, participants were recruited to play the game for five minutes via Amazon Mechanical Turk. In our analyses of player behavior, we remove the first and last thirty seconds of play.

In Experiment 1, we tested queues of lengths $\tau \in \{4, 6, 8, 10, 12, 14\}$ and relative reward rates of $\rho \in \{1.25, 1.50\}$. Figure 4a shows empirical and theoretical hazard functions—dashed and solid lines, respectively—for the 6×2 conditions. The model has four free parameters which were fit to the data. The model shows key qualitative properties of human behavior, including sensitivity to τ, ρ, and the relative position in the long queue, and obtains good quantitative fits to the data. This Figure shows data from a population of 40 participants. We divided the participants into low and high performers based on a median split of their overall scores, and obtained model fits for each subpopulation (not shown). The discount rate for the two subpopulations are $\gamma_{\text{strong}} = 0.999$ and $\gamma_{\text{weak}} = 0.875$.

Our next experiment explored customized incentives with the fully constrained model. The type of incentives—or *bonuses* as we call them—we allowed consist of the allocation of up to 200 points at various positions in the long queue, awarded in 50 point increments and subtracted from the front-of-queue payoff. We constrained the search such that no mid-queue defection strategy could lead to $\rho > 1$. Figure 3b shows an example of such bonuses, depicted by the points awarded in a given queue position. Although we ultimately wish to customize incentives to individuals, we settled for customizing incentives to subpopulations, specifically, the low and high performers in Experiment 1. We then searched over the discrete space of bonuses to determine the bonus structure that—according to the model parameterized by data from the Experiment 1 subpopulation—would optimize performance of each subpopulation. As shown in Figure 4b, the optimization yields A brute-force optimization yields bonuses *early* in the queue for the weak group, and bonuses *late* in the queue for the strong group.

Experiment 4 tested participants on three line lengths—6, 10, and 14—and three bonus conditions—early, late, and no bonuses. The 54 participants who completed Experiment 4 were median split into a weak and a strong group based on their reward rate on no-bonus episodes only. Consistent with the model-based optimization, the weak group performs better on early bonuses and the strong group on late bonuses (the yellow and blue bars in Figure 4c). Importantly, there is a 2×2 interaction between group and early versus late bonus ($F(1, 51) = 11.82$, $p = .001$) indicating a differential effect of bonuses on the two groups. Figure 4c also shows model predictions based the parameterization determined from Experiment 1. The model has a perfect rank correlation with the data, and correctly predicts that both bonus conditions will facilitate performance, despite the objectively equal reward rate in the bonus and no-bonus conditions. That bonuses should improve performance is nontrivial: the persistence induced by the bonuses must overcome the tendency to defect because the LL reward is lower (as we observed in Experiment 1 with $\rho = 1.25$ versus $\rho = 1.50$).

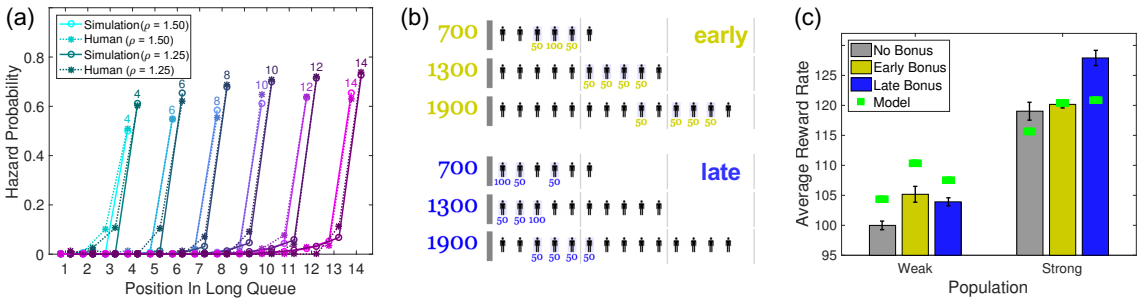

Figure 4: (a) Experiment 1: empirical and model-fit hazard curves, (b) Experiment 4: customized bonus structures, (c) Experiment 4: reward rates by subpopulation and bonus condition

4 Discussion

In our work, we developed a formal theoretical framework to modeling the dynamics of intertemporal choice. We hypothesized that the theory is suitable to modeling human behavior. We obtained support for the theory by demonstrating that it explains key qualitative behavioral phenomena and predicts quantitative outcomes from a series of behavioral experiments. Although our Experiment 1 merely suggests that the theory has the flexibility to fit behavioral data post hoc, each following experiment used parametric constraints from the earlier experiments, leading to strong predictions from the theory that match behavioral evidence. The theory allows us to design incentive mechanisms that steer individuals toward better outcomes, and we showed that this idea works in practice for customizing bonuses to subpopulations playing our queue-waiting game. The theory and the behavioral evidence both show a non-obvious and non-intuitive statistical interaction between the subpopulations and various incentive schemes. Because the theory has just four free parameters, it is readily pinned down to make strong, make-or-break predictions. Furthermore, it should be feasible to fit the theory to individuals as well as to subpopulations. With such fits comes the potential for maximally effective, truly individualized approaches to guiding intertemporal choice.

References

Melissa Schettini Kearney, Peter Tufano, Jonathan Guryan, and Erik Hurst. Making savers winners: An overview of prize-linked savings products. Working Paper 16433, National Bureau of Economic Research, October 2010. URL `http://www.nber.org/papers/w16433`.

C. Kidd, H. Palmeri, and R. N. Aslin. Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, 126:109–114, 2012. doi: doi:10.1016/j.cognition.2012.08.004.

R Kivetz. The effects of effort and intrinsic motivation on risky choice. *Marketing Science*, 22:477–502, 2003.

Z. Kurth-Nelson and A. D. Redish. A reinforcement learning model of precommitment in decision making. *Frontiers in Behavioral Neuroscience*, 4, 2010. doi: http://doi.org/10.3389/fnbeh.2010.00184.

Z. Kurth-Nelson and A. D. Redish. Don't let me do that! models of precommitment. *Frontiers in Neuroscience*, 6, 2012. doi: http://doi.org/10.3389/fnins.2012.00138.

J. T. McGuire and J. W. Kable. Rational temporal predictions can underlie apparent failure to delay gratification. *Psychological Review*, 120:395–410, 2013.

W. Mischel. Processes in delay of gratification. *Advances in Experimental Social Psychology*, 7:249–292, 1974. doi: doi: 10.1016/S0065-2601(08)60039-8.

W. Mischel and E. B. Ebbesen. Attention in delay of gratification. *Journal of Personality and Social Psychology*, 16:329–337, 1970.

W Mischel, Y Shoda, and P K Peake. The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of Personality & Social Psychology*, 54:687–696, 1988.

Noam Nisan and Amir Ronen. Algorithmic mechanism design (extended abstract). In *Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing*, STOC '99, pages 129–140, New York, NY, USA, 1999. ACM. ISBN 1-58113-067-8. doi: 10.1145/301250.301287. URL `http://doi.acm.org/10.1145/301250.301287`.

Y. Niv, J.A. Edlund, P. Dayan, and J.P. O'Doherty. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience*, 32:551–562, 2012.

Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer. Risk sensitive reinforcement learning. *Neural Computation*, 26:1298–1328, 2014.