Retrospective value bias in setting temporally extended goals in humans

Sneha Aenugu (saenugu@caltech.edu)

John P. O'Doherty (jdoherty@caltech.edu)
Division of Humanities and Social Sciences, California Institute of Technology
1200 E California blvd Pasadena, CA 91125 USA

Abstract

A choice to stay committed to a temporally-extended goal or to switch away from it entails weighing of retrospective value — how much has been accomplished so far — against prospective value — how much further till the finish line. In a novel task where an option needs to be persistently executed till a set target to earn rewards, we demonstrate an undue bias in favor of retrospective value in human behavior resulting in sub-optimal performance. We further propose computational hypotheses embedded in framework of reinforcement learning to account for human performance in the task.

Keywords: reinforcement learning; goal-directed decision-making; prospective planning

Introduction

The conditions halfway through your PhD might be different from when you started. A lot of things could go differently than your expectations. Amidst shifting contingencies, do you decide to stay on the current goal or decide to drop out?

We encounter many such scenarios where a goal needs to be set for a temporally extended course of time. It seems to be the case that people often stick with an option longer than it is viable for them, a phenomenon documented as a retrospective cost or sunk cost in decision-making (Arkes & Blumer, 1985), (Arkes & Ayton, 1999). On the other hand, contemplating the prospective cost — future costs and benefits of further investment — is considered a rational approach.

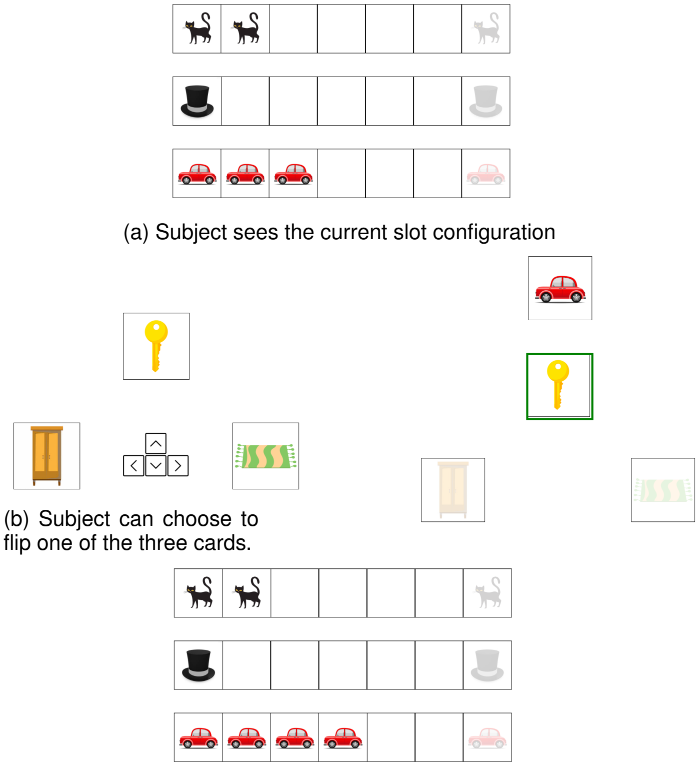
How do humans weigh contributions from retrospective and prospective value computations to guide selection and switching of temporally extended goals? The rich framework for temporal abstractions in reinforcement learning can offer tremendous insights into the computational mechanisms guiding decision-making over extended courses of actions (Sutton & Barto, 2016), (Sutton, Precup, & Singh, 1999).

We now introduce a novel task where an option or goal needs to be executed through extended courses of actions to reach a target and earn rewards. Subjects were allowed to switch between different goals with no partial or intermittent rewards. We formulate several computational strategies to solve the task and study how they can potentially account for human behavior.

Experiment

Suit collection paradigm The objective of the task is to collect suits of tokens (7 tokens of the same kind constitutes a suit). There are three different tokens in the task — CAT, HAT, CAR — each endowed with 7 empty slots to collect the tokens. When one collects 7 of the same kind, points are awarded.

Subjects can collect tokens by flipping appropriate cards. There are 6 cards in the game (2 cards for each of the 3 tokens). We chose the card images to be semantically related to the tokens to ensure strong CARD–TOKEN (action–goal) associations.



(a) Subject sees the current slot configuration

(b) Subject can choose to flip one of the three cards.

(c) Subject sees the updated slot configuration

Figure 1: Suits task: a sample round

In every round of the task (Fig 1), subjects can choose amongst 3 different cards to collect tokens and fill desired slots. Every third round we explicitly probed the subjects of their current goal. Once subjects completed a suit, points are awarded and slots for the token were refreshed. Subjects were awarded a bonus that is commensurate to the total number of suits collected.

Subjects participated in 540 rounds which are grouped into 18 blocks. Each block is one of three types: 80-20, 70-30, 60-40. For instance, in the 80-20 condition one token is received with 80% chance upon flipping its cards and the other two are available with 20%. The most abundant token switches across adjacent blocks. Subjects were explicitly made aware of the token probabilities in a block, though they were not told exactly which token has the highest chance.

Modeling

Models of temporally extended goal-setting entail valuation of different goals admist shifting contingencies and formulation of policies to guide action selection and adaptive goal-switching to accrue rewards in the game.

The goals in the task are C, H, R: **C**AR, **H**AT, CA**R**. State space of the markov decision process is the slot configuration. Value of choosing goal G at time t is denoted by $Q^G(S_t^G)$.

TD Agent The first model avoids explicitly encoding the token probabilities, instead reactively performs the task. Goal values are updated by temporal difference (TD) learning rule. Value is formulated as a function approximator of the state space.

$$Q^G(S_t^G = g) = w_t^G g^\alpha + b_t^G$$

where g is the number of tokens of the goal G collected and α is the non-linearity in valuation of the token count.

$$\begin{aligned}\delta &= \gamma Q^G(S_t^G) - Q^G(S_{t-1}^G)\\ w_t^G &= w_{t-1}^G + \alpha\delta g^\alpha\\ b_t^G &= b_{t-1}^G + \alpha\delta\end{aligned}$$

The values of goals with greater number of slots filled loom larger over others up until their parameter values are sufficiently decremented through learning. This model, therefore, demonstrates retrospective bias.

Prospective This model explicitly encodes the token contingencies and uses it to prospectively estimate the value of staying committed to a given goal G at every instant. Token probabilities p_t^G are updated at every round by assuming a beta distribution.

$$Q^G(S_t^G) = p_t^G \gamma Q^G(S_t^G + 1) + (1 - p_t^G)\gamma Q^G(S_t^G)$$

Hybrid This model assumes a weighted combination of values from the above two modes of decision-making (Lee, Shimojo, & O'Doherty, 2014).

$$Q^G(S_t^G) = wQ^G_{\text{pros}}(S_t^G) + (1-w)Q^G_{\text{TD}}(S_t^G)$$

Policy for goal selection is a softmax over all goal values.

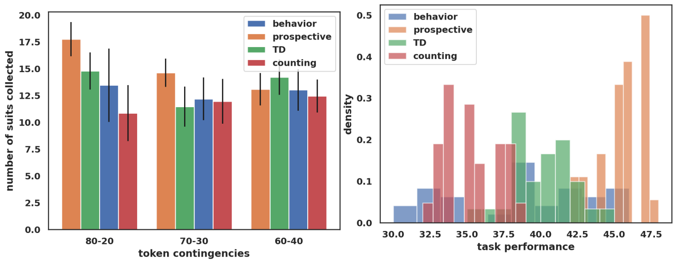
$$\pi(S_t^G, G) = \text{softmax}_\sigma\{Q^C(S_t^C), Q^H(S_t^H), Q^R(S_t^R)\}$$

Counting As a baseline, we included a purely retrospective model which values different goals solely based on the number of tokens of the kind already collected.

Results

$N = 30$ subjects recruited through Prolific for the task (median time – 35 min; base pay – 10.50\$/ hour; bonus – upto 2\$). All model simulations are run over 30 random seeds.

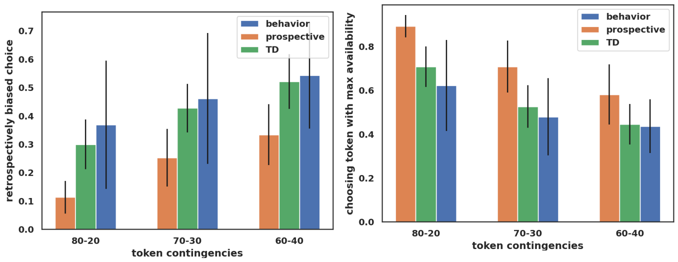
Human performance shows influences of prospective and retrospective value computations There is a natural gradation in task performance from the three models (prospective, TD, counting). Prospective model provides overall best performance largely powered by outcomes in the 80-20, 70-30 contingencies where prospective valuation of tokens has the highest benefit (Fig 2a, 2b). Human behavior falling short of the prospective model in the said contingencies indicates influences of higher retrospective valuation.



(a) Performance per token contingency. Prospective model outperforms in 80-20 and 70-30 contingencies.

(b) Density histogram of total number of suits collected. Behavior shows a spread indicative of a mixture of strategies.

Figure 2: Performance in the task



(a) Probability of choosing the token with the maximum number of slots filled (high retrospective value) but not the most available one (low prospective value).

(b) Probability of choosing the token with maximum availability in the contingency. Prospective model leveraging token contingencies chooses optimally.

Figure 3: Evidence of retrospective value bias

Retrospective bias in goal selection is evident in human behavior Human task performance takes a hit when tokens with higher number of slots filled (high retrospective value) but lower future availability (low prospective value) are chosen as goals. Choice data analyzed from subjects indicates that this is the case (Fig 3a). Overall, probability of humans choosing the token with maximum availability in a contingency as the goal falls short of the prospective model (Fig 3b).

Weight on prospective value correlates with task performance Hybrid model accounts for human choices better than pure prospective model or pure TD model: 20/30 prefer hybrid over prospective only, 23/30 over TD only (loglikelihood ratio test at $p < 0.05$). Fitted weight on prospective value in the hybrid model correlates with task performance $(R = 0.68, p < 0.0001)$.

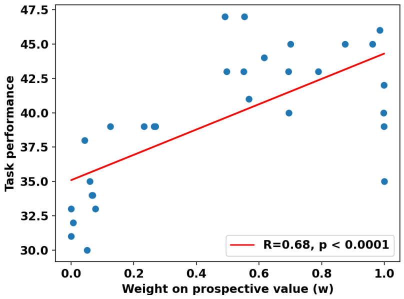


Figure 4: Prospective value weight on task performance

Acknowledgments

JOD is supported by grants from the National Institute of Mental Health, the Army Research Office, and the T&C Chen Center for Social and Decision Neuroscience.

References

Arkes, H. R., & Ayton, P. (1999). The sunk cost and concorde effects: Are humans less rational than lower animals? *Psychological Bulletin*, *125*, 591-600.

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *BEHAVIOR AND HUMAN DECISION PROCESSES*, *35*, 124-140.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014, 2). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*, 687-699.

Sutton, R. S., & Barto, A. G. (2016). Reinforcement learning: An introduction second edition, in progress.

Sutton, R. S., Precup, D., & Singh, S. (1999, 8). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*, 181-211.