A computational account optimism and pessimism as means of self-control

Boaz Rosenberg (Boaz.Rosenberg@huji.mail.ac.il)
Psychology Department, The Hebrew University of Jerusalem, Mount Scopus
Jerusalem, Israel

Eran Eldar (Eran.Eldar@mail.huji.ac.il)
Psychology and Cognitive & Brain Sciences Departments, The Hebrew University of Jerusalem, Mount Scopus
Jerusalem, Israel

Abstract:

From a decision-making perspective, self-deception doesn't seem to be a beneficial behavioral characteristic. Nevertheless, in this article, we demonstrate how, under certain viable constraints, agents learn that biasing their perceived state can be advantageous. This demonstration is based on recognizing the value of self-control in cases where cached and prospection-based evaluations result in conflicting action selection. When this conflict arises from evaluations of resource states, self-biasing could serve as an effective means of self-control, diminishing the influence of potentially incorrect cached values, and thus improving decision-making processes.

Keywords: Model-Based; Cached value**;** Biased perception; Reinforcement Learning

Introduction

Prospection-based & cached values conflict

Traditionally, reinforcement learning distinguishes between two approaches to making evaluations: a prospection-based (PB) approach, whereby a model of the environment is used to compute the value of a potential state or action based on its future consequences, and a caching approach whereby previously learned values are used without considering specific future consequences (Daw et al., 2011).

While it has been shown that PB and cached evaluations can complement each other, allowing the agent to adapt its actions based on the most accurate evaluation at hand (Daw, Niv, & Dayan, 2005), significant differences in evaluations can lead to conflict regarding the agent's choice of actions. This internal conflict is evident in scenarios such as dietary choices and procrastination (Story et al., 2014), where individuals may struggle to prioritize between the predicted future effects of their actions and their immediate, cached-value-driven urges.

In such cases, PB evaluation should value any behavior that promotes acting based on the PB-evaluation-preferred action instead of the actions urged by the cached evaluation. Such behaviors can be understood as a form of self-control.

Evaluating prospective resource states

In this article, our focus will be on the evaluations of different states to which an action could lead. This involves using a model of the environment to estimate the probability of reaching different possible states following different actions and assigning these states values either in a PB manner, based on the estimated future consequences of reaching each state, or using a cached value associated with the state. The latter evaluation is similar to the concept of "plan-until-habit" introduced by Keramati et al. (2016), where actions are evaluated based on a limited depth of planning, and the estimated outcomes are then assessed using cached values.

Similarly to the simple PB-cached evaluation conflict discussed earlier, discrepancies in action evaluations stemming from different state evaluations may also lead to internal conflict. For instance, let's consider a person who has experienced torment in kindergarten. As a result, even as an adult, they might harbor a strongly negative cached value associated with being socially unpopular. As they mature, they might learn that no one in their office will pull their hair or take their toys even if they aren't popular, and therefore their PB evaluations of social rejection might be considerably less negative. This difference in evaluations directly impacts the person's attitude towards actions that risk reaching a state of social rejection. According to cached evaluations, even a slight risk of ending up unpopular should lead to a negative evaluation of the risky action, even when it offers a good chance for positive outcomes. On the other hand, the PB evaluation, which has a more moderate evaluation of reaching a state of unpopularity, may determine that the opportunities that the action allow outweigh the possible negative outcome. Therefore, the PB evaluation might advocate for risky and profitable actions that the cached evaluations strongly oppose, leading to internal conflict.

When the states we are evaluating are continuous resource states (social or economic status, for example), we could refer to the state evaluations as

value functions – mapping different points along the state axis to a PB or cached value. When the PB and cached value functions differ in their rate of curvature – we will experience different rates of risk aversion from each evaluation, leading to an internal conflict akin to the previous example.

In this case, though, the agent has an efficient way to exert self-control and promote choosing the PB favored action. Since in many curved value functions the degree of implied risk aversion changes according to position along the x axis[1], any bias to the agent's perceived state could alter its degree of risk aversion. If this bias could be controlled, it could serve as a valuable means of self-control.

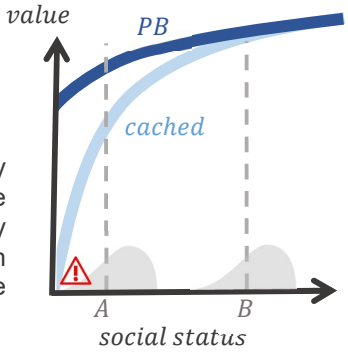


Figure 1: an example for PB and cached value functions

The same risk (gray distributions) would be evaluated differently by the cached value function when the agent is in state A or in state B.

Simulation

We implemented this logic in a simple computer simulation. The simulation included an agent that evaluates different possible actions based on their given probability distribution functions and its estimated current state ($\hat{S}$). The expected value of the agent's possible states after each action was evaluated using a value function (vf_1) and the highest valued action was selected. The value function used for this decision represents the combination of the PB and cached evaluations of the expected outcomes.

Additionally, the agent made a separate decision about the amount of bias it should exhibit in its state estimation. This decision was implemented by adjusting the value of the τ parameter in an asymmetric learning rate algorithm used for updating the agent's estimated state:

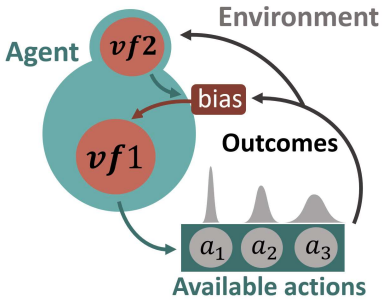
$$\hat{S}_{t+1} = \hat{S}_t + \eta(o_t - \hat{S}_t)$$

$$\eta = \begin{cases} \eta_0 \boldsymbol{\tau} & if\ (o_t - \hat{S}_t) > 0 \\ \eta_0(1-\boldsymbol{\tau}) & if\ (o_t - \hat{S}_t) \leq 0 \end{cases}$$

Where $\hat{S}_t$ represents the agent's state estimation at time t, o_t a noisy observation, η_0 the baseline learning rate and τ the agent's chosen degree of bias ($\in [0,1]$).

[1] This characteristic could be mathematically deduced from the assumption that a value function evaluating a resource axis should be monotonically increasing. This implies that any curvature should diminish as the function approaches one of its limits in order to prevent the slope from becoming negative, and therefore the resulting risk aversion should reduce / increase respectively.

Figure 2: a visualization of the simulation model



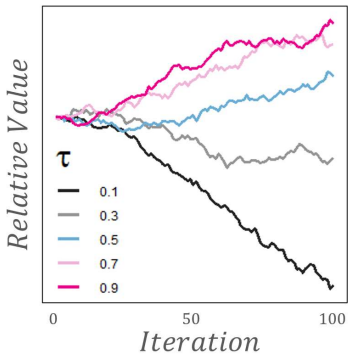
The agent selected various values of τ and evaluated the effect its choices under each corresponding bias had on its state. These effects were evaluated using a separate value function (vf_2), representing the PB evaluation of the different states.

Results

We conducted the simulation using various value functions for the PB and cached evaluations. Here, we present the outcomes for a diminishing marginal value function, similar to those shown in figure 1.

$$vf_1(x) = x - 1.05^{-x} \qquad vf_2(x) = x - 1.04^{-x}$$

Figure 3: simulation results



As illustrated in figure 3, the agents learnt to prefer focusing on positive outcomes ($\tau > 0.5$), which positively biases their state estimation. In such conditions the decision made by the agent were riskier and led to improved performance. In another simulation involving value functions with increasing marginal value, we observed the opposite behavior - agents preferred negative biases which reduced their risk-taking.

Summary

The theory outlined in the article, along with the simulations that put it into practice, provide a novel perspective on various behavioral phenomena that have largely remained unexplained. These include the general tendency towards optimistic biases (Sharot, 2011) as well as mystical and fantastical notions which might serve as catalysts for such beliefs.

In a full report of this work, we will discuss the boundaries of these biases, explore potential causes for individual variations in optimism and pessimism, and explore the implications of state deterioration and inadequate control of these biases and their possible link to disorders such as MDD, BD, addiction, and OCD.

References

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, *113*(45), 12868–12873.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941–R945.

Story, G. W., Vlaev, I., Seymour, B., Darzi, A., & Dolan, R. J. (2014). Does temporal discounting explain unhealthy behavior? A systematic review and reinforcement learning perspective. *Frontiers in Behavioral Neuroscience*, *8*, 76.