A potential reset mechanism for the modulation of decision processes under uncertainty

Krista M. Bond
kbond@andrew.cmu.edu

Alexis Porter
alexisp@andrew.cmu.edu

Timothy Verstynen
timothyv@andrew.cmu.edu

Department of Psychology and Carnegie Mellon Neuroscience Institute
Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA

Abstract

Humans and other mammals flexibly select actions in noisy, uncertain contexts, quickly using feedback to adapt their decision policies to either explore other options or to exploit what they know. Drawing inspiration from the plasticity of cortico-basal ganglia-thalamic circuitry, we recently developed a cognitive model of decision-making that uses both a value-driven learning signal to update an internal estimate of state action-value (i.e., conflict in the probability of reward between two choices) and a change-point-driven learning signal that adapts to changes in reward contingencies (i.e., a previously high value target becoming devalued). In this work, we expand on previous results from our group (Bond, Dunovan, & Verstynen, 2018) to more carefully detail how these environmental signals drive changes in the decision process. Across nine separate behavioral testing sessions, we independently manipulated the level of value-conflict and volatility in action-outcome contingencies. Using a hierarchical drift diffusion model, we found that the belief in the value difference between options had the greatest influence on decision processes, impacting drift rate, while estimates of environmental change had a smaller, but detectable influence on the decision threshold. Taken together, these findings bolster our previous work showing how separate environmental signals impact different aspects of the decision algorithm.

Keywords: change point detection; decision-conflict; corticobasal ganglia networks; adaptive decision-making

Introduction

In natural contexts, successful behavior in a dynamic environment requires making fast, accurate decisions and updating those decisions based on an internal model of the state of the environment. Drawing inspiration from the computational architecture of cortico-basal ganglia-thalamic circuitry (Dunovan & Verstynen, 2016), we previously proposed a cognitive model that 1) updates the rate of evidence accumulation using estimates of value differences between possible actions and 2) updates the threshold of decision processes using estimates of change point probability. Using an adaptive-decision-making algorithm that unifies drift diffusion models and reinforcement learning (Dunovan & Verstynen, 2017; Pedersen, Frank, & Biele, 2017), we modeled decision processes under more expansive conditions of value-conflict, or the proximity of the probability of reward between two choices, and feedback volatility, or the instability of action-value associations. We sought to replicate our previously observed effects showing that value-conflict decreases the rate of evidence accumulation and that volatility in action-value associations decreases the amount of evidence needed to make a decision. For this replication, we adopted a high-power, within-subject design ($N = 4$ subjects, 3600 trials/subject) where we independently manipulated the degree of conflict in reward values and volatility in action-outcome contingencies.

Methods

Task

Participants Four participants were recruited from the Paid Psychology Subject Pool and the local community. They were paid $10 per session in addition to a performance bonus. These experiments were approved by the Institutional Review Board at Carnegie Mellon University.

Stimuli and procedure Each participant completed nine sessions composed of 400 trials each, generating 3600 trials per subject. Data were collected from four participants in accordance with a replication-based design, with each participant serving as a replication experiment 1. Participants completed these sessions across three weeks in randomized order. Each trial presented a male and female greeble (Gauthier & Tarr, 1997), with the goal of selecting the sex identity of the greeble which was most profitable. Individual greeble identities were resampled on each trial; thus, the task of the participant was to choose the sex identity rather than the individual identity of the greeble which was most rewarding (Figure 1). Probabilistic reward feedback was given in the form of points drawn from the normal distribution $N(\mu = 3, \sigma = 1)$ and these points were displayed at the center of the screen. Participants

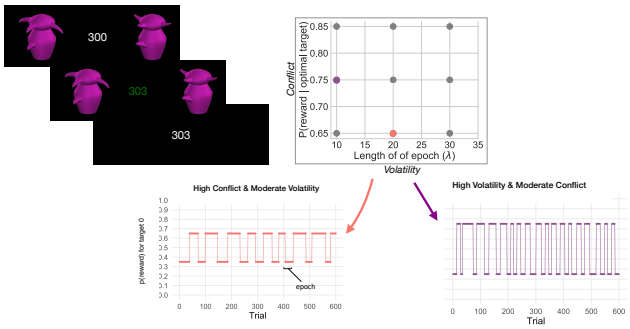

Figure 1: The behavioral task. Participants chose one of two greebles and received probabilistic reward. The total number of points earned was displayed at the center of the screen. Three levels of conflict and volatility were independently manipulated, yielding nine total sessions. Each session was composed of 400 trials with multiple change points within a session, as determined by the level of volatility. Sample reward sessions are shown at the bottom of the figure.

began with 200 points and lost one point for each incorrect decision. To promote incentive compatibility, participants earned a cent for every point earned. If participants responded in $< .1s$, $> 1s$, or failed to respond altogether, the point total turned red and decreased by 5 points. Each trial lasted 1.5 s and reward feedback for a given trial was displayed from the participant response to the end of this window. Reaction time was constrained such that participants were required to respond within 0.1 and 0.75 s from stimulus presentation.

To manipulate change point probability, the sex identity of the most rewarding greeble was switched probabilistically. To manipulate the belief in the value of the optimal target, the probability of reward for the optimal target was manipulated. Further, the position of the high-value target was pseudo-randomized on each trial to prevent prepotent response selections on the basis of location.

Throughout the task, the head-stabilized diameter of the left pupil was measured with an Eyelink 1000 at 1000 Hz from within a custom-built booth designed to eliminate the influence of ambient sources of luminance. Because the dynamic range of the pupillary response is known to be highly sensitive to a variety of influences (Sirois & Brisson, 2014), participants were exposed to a sinusoidal variation in luminance prior to the reward-learning task to establish the dynamic range of the pupillary response for that session. During the reward-learning task, all stimuli were rendered isoluminant with the background of the display to further prevent luminance-related confounds of the task-evoked pupillary response. To minimize the convolution of the task-evoked pupillary response from trial to trial, the inter-trial interval was sampled from a truncated exponential distribution with a minimum of 4 s, a maximum of 16 s, and a rate parameter of 2.

The pupillometry data are not presented at this time, but will be used in follow-up analyses.

Cognitive Model

Here we propose that the drift rate (v) and the decision threshold (a) are modulated on a trial-by-trial basis according to two estimates of uncertainty from an ideal observer.

Updating action-values To model how learners update action-values, we calculate an estimate of how often the same action will give a *different* reward. We call this learning signal change point probability (Ω). The change point probability will be close to 1 as the probability of a sample coming from a uniform distribution, relative to a Gaussian distribution, increases:

$$\Omega_t = \frac{U(r_{\Delta_t})H}{U(r_{\Delta_t})H + N(r_{\Delta_t}|B_{\Delta_t}, \sigma_t^2)(1-H)} \tag{1}$$

H refers to the hazard rate, or the global probability of a change point:

$$H = \frac{\sum_{cp_0}^{cp_n}}{n_{trials}} \tag{2}$$

Model confidence [ϕ] is a function of the change point probability [Ω] and the variance of the generative distribution [σ_n^2], both of which form an estimate of relative uncertainty (RU):

$$RU_t = \frac{\Omega_t\sigma_n^2 + (1-\Omega_t)(1-\phi_t)\sigma_n^2 + \Omega_t(1-\Omega_t)(\delta_t\phi_t)^2}{\Omega_t\sigma_n^2 + (1-\Omega_t)(1-\phi_t)\sigma_n^2 + \Omega_t(1-\Omega_t)(\delta_t\phi_t)^2 + \sigma_n^2} \tag{3}$$

Thus [ϕ] is determined as:

$$\phi_{t+1} = 1 - RU \tag{4}$$

Relative action-value Along with estimates of the stability of action-value contingencies, feedback signals also drive the belief in the reward of an action. We call this signal B, and it is learned separately for each action target. Given that c = the chosen target and u = the unchosen target, the belief in the mean of the distribution of reward differences on the next trial is calculated as:

$$B_{t+1,c} = B_{t,c} + \alpha_t\delta_t \tag{5}$$

The unchosen target value decays to the pooled expected value of both targets, $E(r)$:

$$B_{t+1,u} = B_{t,u}(1-\Omega_t) + \Omega_t E(r) \tag{6}$$

$$E(r) = \frac{\bar{r}_{t_0} + \bar{r}_{t_1}}{2} \tag{7}$$

The signed belief in the reward difference between targets is calculated as the difference in belief for targets 0 and 1:

$$B_{\Delta_{t+1}} = B_{t,1} - B_{t,0} \tag{8}$$

Update rules The learning rate of the model [α] is determined by the change point probability [Ω] and the model confidence [ϕ]. Here, the learning rate will be high if either 1) a change in the mean of the distribution of the difference in expected values is likely [Ω is high] or 2) the estimate of the mean is highly imprecise [σ_n^2 is high]:

$$\alpha_t = \Omega_t + (1-\Omega)(1-\phi_t) \tag{9}$$

The prediction error, δ, is the difference between the model belief and the reward difference observed:

$$\delta_t = r_t - B_{t,c} \tag{10}$$

And the estimated variance, σ^2, is calculated as:

$$\sigma_t^2 = \sigma_n^2 + \frac{(1-\phi_t)\sigma_n^2}{\phi_t} \tag{11}$$

We propose that the belief in the relative reward for the two choices, B, updates the drift rate, v, or the speed of evidence accumulation:

$$v_{t+1} = \hat{\beta}_v \cdot B_{\Delta_t} + v_t \tag{12}$$

and that the change point probability, Ω decreases the decision threshold, a, or the amount of evidence needed to make a decision:

$$a_{t+1} = \hat{\beta}_a \cdot \Omega_t - a_0 \tag{13}$$

We adapted the above ideal observer calculations from a previous study (Vaghi et al., 2017).

Results

As change point probability increased, accuracy decreased ($p < 0.01$ in 3/4 replicates, $\bar{\beta} = -0.55 \pm 0.24$, Figure 2). As the belief in the value of the optimal target increased, accuracy increased (Figure 2, $p < 0.03$ in all replicates, $\bar{\beta} = 0.15 \pm 0.1$). Reaction times decreased as change point probability increased in the majority of cases ($p < 0.03$ in 3/4 replicates, $\bar{\beta} = -0.02 \pm 0.01$, Figure 3). The belief in the value of the optimal target had minimal impact on reaction times ($\bar{\beta} = 0.00 \pm 0$ in 4/4 subjects).

The RT distributions generated from each participant were then submitted to hierarchical drift diffusion model regression (Wiecki, Sofer, & Frank, 2013). For these regressions, we evaluated the fit of either our hypothesized update rule or the inverse model to the data, with Ω and B as predictors of either a or v. Consistent with our hypothesis, we found strong evidence for the model that mapped drift-rate updates onto trial-wise changes in the belief of the value of the optimal target and decision threshold updates onto changes in change point probability (hypothesized model best accounted for the data in 3/4 cases; $DIC\Delta = -31$ points; Figure 4).

Using the posterior probability distributions of the regression coefficients, we found that the drift-rate increased with the belief in the value of the optimal target (Figure 5; observed $p(\hat{\beta}_v < 0) < .01$ in all cases; mean $p(\hat{\beta}_v < 0) = 0 \pm 0$). We found a weak effect of change point probability on the decision threshold (mean observed $p(\hat{\beta}_a > 0) = 0.27 \pm 0.11$).

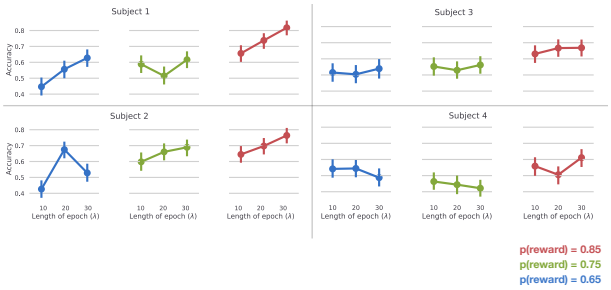

Figure 2: Accuracy. The mean accuracy in selecting the most probably rewarding option is plotted for each participant across varying levels of conflict and volatility. Error bars represent bootstrapped 95% confidence intervals.

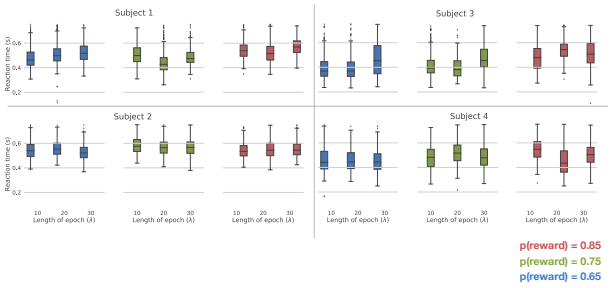

Figure 3: Reaction times. The reaction time distributions for each participant are plotted as a function of the manipulated levels of conflict and volatility. Error bars represent bootstrapped 95% CIs.

Conclusions

Using a high-powered within-subject design, we replicated and expanded our previous work to show that different environmental signals modulate different aspects of the accumulation-of-evidence process during decision making. Future work will explore how pupil responses, as a proxy for noradrenergic activity, track with estimates of environmental volatility as a possible mechanism for the dynamic modulation of decision threshold.

Acknowledgments

K.B. was funded by the Behavioral Brain Research Training Program at the Center for the Neural Basis of Cognition at Carnegie Mellon University (National Institutes of Health grant T32GM081760). This project was partially funded by National Science Foundation Career Award 1351748.

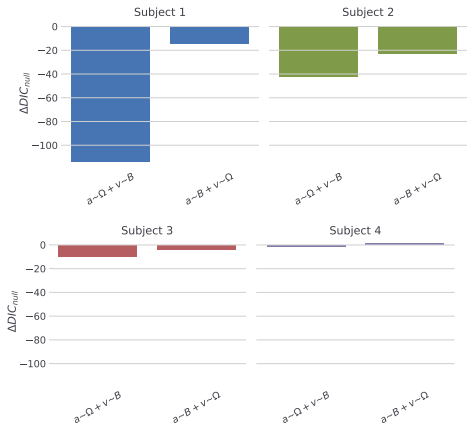

Figure 4: Deviance Information Criterion (DIC) scores for the hypothesized two-parameter model and the reversal of the hypothesized relationship. Model fit results are calculated relative to an intercept-only model. The hypothesized model fits the data best in three out of four replicates.

References

Bond, K., Dunovan, K., & Verstynen, T. (2018, 01). Value-conflict and volatility influence distinct decision-making processes.. doi: 10.32470/CCN.2018.1068-0

Dunovan, K., & Verstynen, T. (2016). Believer-skeptic meets actor-critic: Rethinking the role of basal ganglia pathways during decision-making and reinforcement learning. *Frontiers in neuroscience*, *10*, 106.

Dunovan, K., & Verstynen, T. D. (2017). Errors in action timing and inhibition facilitate learning by tuning distinct mechanisms in the underlying decision process.

Gauthier, I., & Tarr, M. J. (1997). Becoming a greeble expert: Exploring mechanisms for face recognition. *Vision research*, *37*(12), 1673–1682.

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic bulletin & review*, *24*(4), 1234–1251.

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(6), 679–692.

Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity reveals a novel dissociation between action and confidence. *Neuron*, *96*(2), 348–354.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). Hddm: hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, *7*, 14.

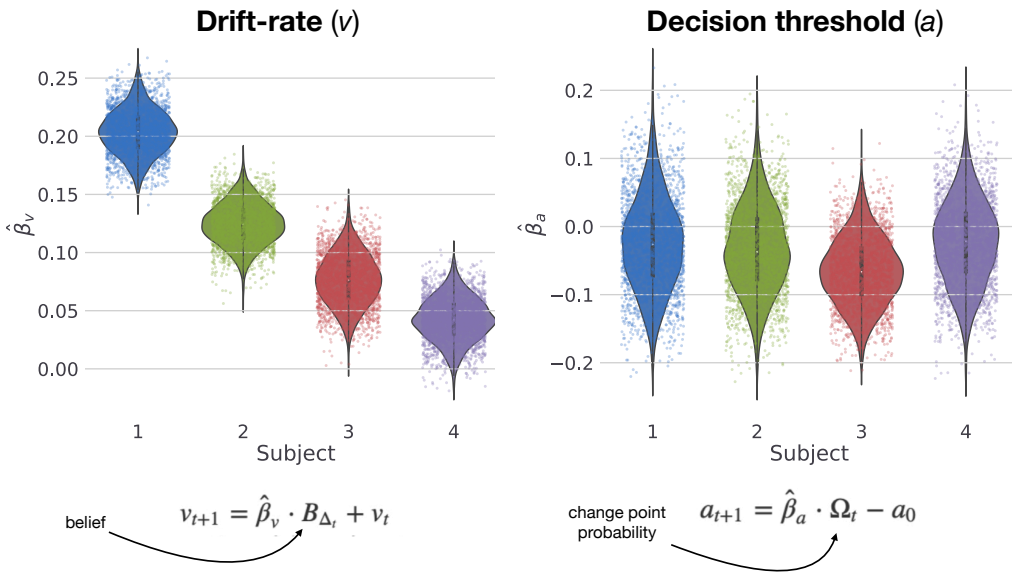

Figure 5: The posterior probability distributions of drift rate and decision threshold under varying conditions of conflict and volatility, plotted for each replicate.