Value-conflict and volatility influence distinct decision-making processes

Krista M. Bond
kbond@andrew.cmu.edu

Kyle Dunovan
kdunovan@andrew.cmu.edu

Timothy Verstynen
timothyv@andrew.cmu.edu

Department of Psychology and Center for the Neural Basis of Cognition
Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA

Abstract

Humans are capable of quickly adapting their decisions using multiple sources of environmental uncertainty. Drawing inspiration from the underlying neural substrates of adaptive decision-making, we propose a dynamic cognitive model in which feedback signals regarding the relative state-action value (i.e., conflict in the probability of reward between two choices) and the reliability of reward contingencies (i.e., likelihood that target value has changed) uniquely target the rate of evidence accumulation (*v*) and the amount of evidence needed to gate a decision (*a*), respectively. We experimentally vetted this model using a variant of the two-armed bandit task (*N* = 20), in which the level of value-conflict and the volatility of reward contingencies were independently manipulated between conditions. Model simulations and behavioral responses were fit to a hierarchical drift diffusion model and both showed similar patterns of changes in *v* and *a* across conditions, providing *prima facie* evidence that distinct estimates of environmental uncertainty target distinct components of decision processes.

Keywords: volatility; conflict; decision-making; uncertainty

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 propose 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 different 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 hypothesize that value-conflict will decrease the rate of evidence accumulation and that volatility in action-value associations will increase the amount of evidence needed to make a decision. The predictions of this model are vetted against behavioral observations from a sample of human participants ($N = 20$).

Methods

Cognitive Model

An adaptive variant of the drift diffusion model was used to simulate the results of our hypothetical learning model (see (Dunovan & Verstynen, 2017)). 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 affects 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).

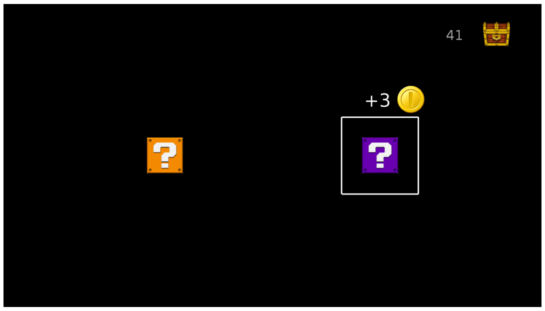

Figure 1: The behavioral task. Participants chose one of two ”mystery boxes” and received probabilistic reward. The total number of points earned was displayed next to a treasure box shown on the upper right portion of the screen.

Task

Participants Twenty participants were recruited from the Paid Psychology Subject Pool and the local community. They were paid $20 after completing all four conditions. This experiment was approved by the Institutional Review Board at Carnegie Mellon University.

Stimuli and procedure All participants completed four sessions of 600 trials each, according to a within-subjects design. To eliminate the effect of timing and its correlates on reward learning (Byrne, Hughes, Rossell, Johnson, & Murray, 2017; Murray et al., 2009), participants completed these conditions across days in counterbalanced order. On each trial, participants chose one of two mystery boxes that had the possibility of returning a set of coins (Figure 1). Probabilistic reward feedback was given in the form of points drawn from the normal distribution $N(\mu = 3, \sigma = 1)$. These points were displayed above the selected treasure box for 0.9 s. To prevent stereotyped responses, the inter-trial interval was sampled from the uniform distribution $U(0.25\text{ s}, 0.75\text{ s})$. Participants were instructed to obtain as many coins as possible by selecting one of the two boxes on each trial.

When participants responded in > 1 s, they received a message saying, ’Too slow! Choose quickly.’ When participants responded in $< .1$ s, they received a message saying, ’Too fast! Slow down. You can continue in 5 seconds,’ and they were required to wait for 5 seconds before continuing the experiment. In both of these cases, participants did not receive any reward feedback or earn any points, and the trial was repeated so that each participant performed 600 valid trials.

The high-value target was switched at each change point. The position of the high value target was pseudo-randomized on each trial to prevent prepotent response selections. Participants began with 600 points and lost one point for each incorrect decision.

Results

Behaviorally, we found that high conflict conditions decreased accuracy relative to low conflict conditions, and high volatil-

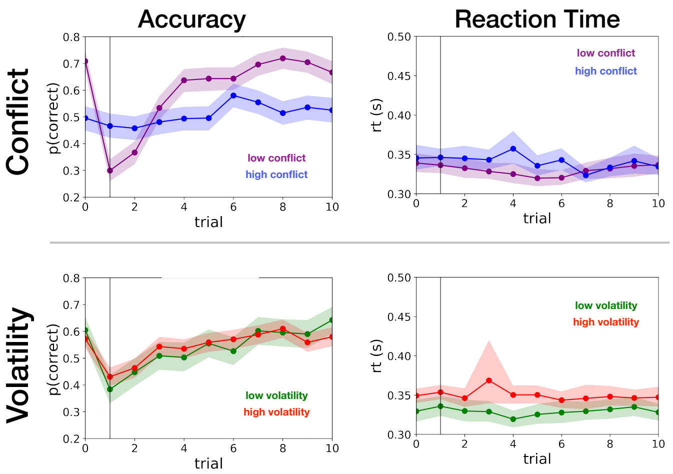

Figure 2: Reaction time and accuracy. Plots are aligned to the beginning of an epoch, where trial 0 is the trial before the start of the epoch and trial 1 marks the change point. Conflict affects response accuracy and volatility affects reaction time.

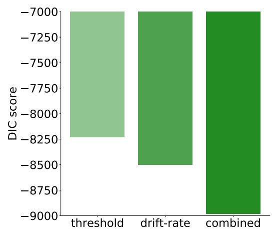

Figure 3: Deviance Information Criterion (DIC) scores for the threshold alone model, the drift-rate alone model, and the combined model. The combined model best accounts for the data.

ity conditions increased reaction time relative to low volatility conditions (Figure 2).

The RT distributions generated from the cognitive model and from human participants were then fit to a drift diffusion model (Wiecki, Sofer, & Frank, 2013). For the behavioral fits, we first wanted to confirm our assumption that both drift rate (v) and decision threshold (a) adapt across conditions. For these fits, we left either a single parameter or a pair of parameters free to be fit across conditions. Consistent with our hypothesis, we found strong evidence that the model which included both drift-rate and decision threshold adaptation best accounted for the data (Figure 3; DIC score difference for drift-alone and combined model = 479 points, DIC score difference between threshold-alone and combined model = 751 points).

Using the posterior probability distributions of drift-rate from both the simulations and the behavioral data, we found that for both datasets the drift-rate was lower in the high conflict condition than the low conflict condition (Figure 4; observed p(low conflict drift-rate $>$ high conflict drift-rate) = 1). By comparing

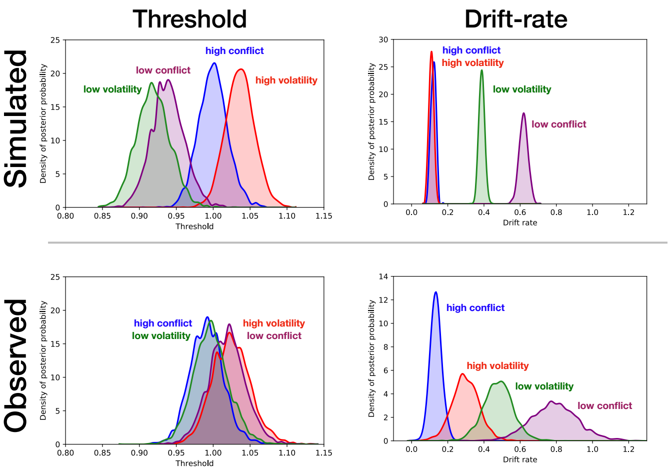

Figure 4: The posterior probability distributions of drift rate and decision threshold under conditions of conflict and volatility.

model and behavioral fits we also found qualitative similarities in the effect of volatility on the decision threshold; the decision threshold increased with volatility (observed p(high volatility threshold $>$ low volatility threshold) = 0.77)).

Conclusions

This combined modeling and behavioral study shows initial confirmatory evidence that different estimates of environmental uncertainty target distinct decision processes during learning, with value-conflict decreasing the speed of evidence accumulation and feedback volatility increasing the amount of evidence needed to make a decision.

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.

References

Byrne, J. E., Hughes, M. E., Rossell, S. L., Johnson, S. L., & Murray, G. (2017). Time of day differences in neural reward functioning in healthy young men. *Journal of Neuroscience*, *37*(37), 8895–8900.

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.

Murray, G., Nicholas, C. L., Kleiman, J., Dwyer, R., Carrington, M. J., Allen, N. B., & Trinder, J. (2009). Natures clocks and human mood: The circadian system modulates reward motivation. *Emotion*, *9*(5), 705.

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.

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.