Decisions about reward and effort for the learning and control of dynamical systems

Harrison Ritz
Cognitive, Linguistic, & Psychological Sciences
Brown University
Providence, RI 02912
hritz@brown.edu

Matthew R. Nassar
Carney Institute for Brain Science
Brown University
Providence, RI 02912
matthew_nassar@brown.edu

Michael J. Frank
Cognitive, Linguistic, & Psychological Sciences
Carney Institute for Brain Science
Brown University
Providence, RI 02912
michael_frank@brown.edu

Amitai Shenhav
Cognitive, Linguistic, & Psychological Sciences
Carney Institute for Brain Science
Brown University
Providence, RI 02912
amitai_shenhav@brown.edu

Abstract

We live in a dynamic world, controlling our thoughts and actions to optimize costs and benefits. While this form of continuous dynamic control has been well-characterized in the domain of motor control, it remains unclear how we learn and deploy analogous control over linear systems when making abstract planning decisions involving reward maximization and effort minimization. The current experiment presents a novel decision-making task in which participants learned how their actions would influence a simple dynamical system. We found that participants appeared to learn a model of this system, and used it to make choices that traded-off rewards and effort. We modeled participants decision-making under an optimal control framework, inferring the latent objective function used to make choices. We found that these objective functions captured key features of participants' cost-benefit trade-off. Our results offer a promising avenue for understanding dynamic control in a non-motoric domain, with potential implications for models of cognitive control.

Keywords: decision-making, optimal control, cost-benefit analysis, effort

Acknowledgements

Thanks to Allison Loynd for assistance with data collection.

1 Introduction

Behavioral and cognitive control often requires people to make continuous actions over continuous spaces (e.g. deciding how fast to run, or how much time to invest in a project), weighing the costs of different actions against the benefits of their outcomes [1]. These control decisions are shaped both by the dynamics of our environment and the causal influence of different actions. While these continuous behavioral dynamics have been explored in online motor control [2], little is known about how people implement state-space control in more abstract domains. The current experiment uses a novel task to understand how people a) learn a linear control model and b) use this model to earn reward and avoid effort. We used an optimal control framework to model people's decisions, finding that our inferred objective functions captured several salient features of participants' behavior.

2 Methods

2.1 Participants

Twenty-six individuals participated in Experiment 1 for course credit or pay (Mean(SD) age = 19(1.1); 13 females). All participants could additional earn a performance-based monetary bonus (see Task). All participants provided informed consent in compliance with our University's Institutional Review Board.

2.2 Task

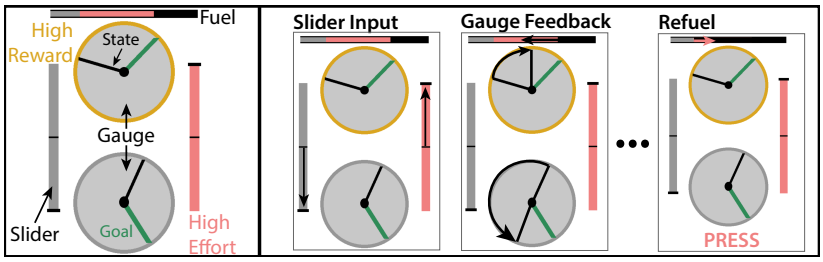

Figure 1: Left: Participants provide inputs on high/low effort sliders to reach goals on high/low reward gauges. The 'fuel bar' tracks the cumulative input magnitude. Right: Participants propose inputs, and then observe the outcomes. When their inputs drain the 'fuel' below a threshold, they must make repeated key presses to 'refuel'.

Participants performed a computerized task in which they made effort-demanding decisions to earn monetary rewards, based on their knowledge about the input dynamics to a simple linear system. Their goal was to align the needles on a set of gauges with a goal orientation (see Figure 1). Participants saw two gauges and two sliders on the screen (Figure 1). Each gauge had needles indicating participants' current state ('State Needle'), as well as a needle indicating their current goal ('Goal Needle'). Participants used their two sliders to move their state needle towards their goal needle. On each trial, participants moved each slider up or down from its zero-starting-point with their keyboard, submitted their response with the spacebar, and then saw the outcome of their inputs. Each slider moved the State Needle on both gauges, and the input-output mapping was linear with Gaussian noise.

When participants successfully moved their State Needle close to the goal (within 20 degrees), they received a reward and the location of that goal was re-sampled at least 90° away. Participants received a fixed reward for each goal, and a reward depending on how close the other gauge's needle was from its goal (exponential decay with 50% reward at 45° from the goal).

We manipulated the rewards for achieving goals and the costs of providing inputs to see how these participants weighed these in their decision-making. On each trial, one of the gauges had a gold outline, providing twice as much reward as the standard gauge. We randomly re-sampled which of the two gauges would have high rewards each time a goal was reached.

Participants' inputs required physical effort. However, instead of requiring effort exertion on every trial, inputs drained a token resource ('fuel') depending on their magnitude. When the fuel fell below a threshold, participants had to effortfully replenish this resource ('refuel') by repeatedly pressing a key (50-60 times). We manipulated effort by making the red slider use resources at twice the rate of the grey slider. Participants could always see their current fuel level and the fuel that would be required for their inputs.

Participants completed extensive instruction followed by 12 blocks of 20 trials. During each block, participants learned a new input-output mapping. This mapping was constrained such that each gauge was influenced much more by one slider (100-200°per max input) than the other slider (0-50°per max input). We randomly selected 24 trials at the end of the experiment, and paid participants $1 per 1000 points earned on those trials.

2.3 Computational Model

We analyzed participants behavior under an optimal control framework, inferring the latent objective function that gave rise to their choices. This model was made up of two major components: a learning process that inferred the input-output mapping, and a choice process that weighed the costs and benefits of different actions.

We modeled participants learning process using a simple gradient descent algorithm, the Windrow-Hoff learning rule [3], commonly used to train linear neural networks. This algorithm predicts how the state needle ($x = [gauge_1, gauge_2]^T$) will change based on a linear model of the system and its inputs, $x_{t+1} = Ax_t + \hat{B}_t u_t$. This model is a combination of the system's intrinsic dynamics (A, here set to identity), and the input dynamics (B, inferred through observations, and $u = [slider_1, slider_2]^T$). On each trial the algorithm observes a prediction error:

$$\delta_{t-1} = x_t - (x_{t-1} + \hat{B}_{t-1} u_{t-1}) \tag{1}$$

The algorithm adjusts $\hat{B}$ in the direction of the gradient that minimizes prediction errors. The rate of learning is controlled by α, and on trials where participant do not use a slider the corresponding input mapping decays by ζ.

$$\hat{B}_t = \hat{B}_{t-1} + \alpha \delta_{t-1} u_{t-1}^T \tag{2}$$

$$\hat{B}(:, u_{t-1} = 0)_t = \zeta \hat{B}(:, u_{t-1} = 0)_t \tag{3}$$

The choice algorithm uses this system model to generate an objective function, consisting of a linear combination of sub-objectives. This objective function is over actions, determining the relative value of candidate inputs ($u_t^{i,j}$). This algorithm was heavily influenced by the standard linear quadratic regulator algorithm commonly used in optimal control, but includes custom sub-objectives and affine costs to better account for participants' behavior.

The first sub-objective (Q) prefers inputs that minimize the expected distance from each goal ($g = [gauge_1, gauge_2]^T$), weighted by a preference for low vs high rewards. For notational simplicity we've fixed the high/low reward gauges, whereas they alternated over time in the task and model.

$$Q^{i,j} = [q_{highRew}, q_{lowRew}] \times abs\big(g_t - (x_t + \hat{B}_t u_t^{i,j})\big) \tag{4}$$

The second sub-objective (R) was to minimize the input magnitude, regularizing actions and minimizing the effort demands according to the relative preference for the high and low effort slider.

$$R^{i,j} = [r_{highEff}, r_{lowEff}] \times abs\big(u_t^{i,j}\big) \tag{5}$$

We included additional sub-objectives to account for participants' strategies and heuristics. First, the algorithm minimized the difference in input magnitude from trial-to-trial, accounting for perseverative responding to the previously-used slider.

$$S^{i,j} = [s, s] \times abs\big(|u_t^{i,j}| - |u_{t-1}|\big) \tag{6}$$

The algorithm also had a preference for actions that reduced the planning complexity of inputs and outputs. With respect to outputs, the algorithm preferred actions that only achieved one goal, and ones that moved towards both goals equally. For notational simplicity, the expected goal distance for each gauge is represented as $\epsilon_t^{m,n}$, based on Equation 4.

$$V^{i,j} = [v_{xor}, v_{and}] \times \left[abs\left(\frac{|\epsilon_t^m| - |\epsilon^n|}{|\epsilon^m| + |\epsilon^n|} \right),\ abs\left(0.5 - abs\left(\frac{|\epsilon^m| - |\epsilon^n|}{|\epsilon^m| + |\epsilon^n|} \right) \right) \right]^T$$

Finally, the algorithm also preferred strategies that reduced the planning complexity of inputs, using the same transformation as (7) on u^i and u^j. This term ($W^{i,j}$) preferred inputs that used one slider at a time, or used both sliders equally.

We summed across these objective functions in order to generate the net objective function ($O^{i,j}$). We passed this objective function through a softmax function, converting it into a probability that a given action would be selected given the set of parameters that defined the learning and choice processes. We used the algorithms above to calculate the choice probabilities at 21 equally-spaced inputs in each slider, linearly interpolating this likelihood at participants' choice. We

used this likelihood function to maximize the *a posteriori* probability of participants' choices under this model using a hierarchical expectation-maximization procedure with mean-field approximation [4]. In addition to the above-mentioned parameters, we also fit a lapse parameter that mixed the model likelihood with the chance rate for over input bins ($\frac{1}{21^2}$), improving the robustness of our parameter estimation.

3 Results

3.1 Behavioral analysis

Participants quickly increased their reward rate during the first few trials (during putative model learning), and maintaining this reward rate as the goal locations changed (Figure 2). This is a qualitative marker of model-based control, as a model-free controller should not demonstrate this generalization. These group-level results were evident in most of the individual participants.

Participants' slider inputs (see Figure 2 for example participants) showed a common tendency to use one slider at a time (inputs on the axis in Figure 2) or use both inputs with the same magnitude (inputs on the diagonal in Figure 2), reducing the dimensionality of their inputs. Gauge movement (Figure 2) demonstrated a similar but weaker trend (e.g., movements in the direction of one goal or the other, but not both), with several participant showing a clear bias towards the high-reward gauge.

Finally, we tested whether our manipulation of reward and effort influenced participants' choices. We found that participants reached the high-reward goal more than twice as often as the low-reward goal (sign-rank test; $z(25)$=4.32, p=1.58e-5). Participants were also more likely to use the low-effort slider ($z(25)$=2.50, p=.012), and gave smaller inputs to the high-effort slider when they did use it ($z(25)$=-2.30, p=.0031).

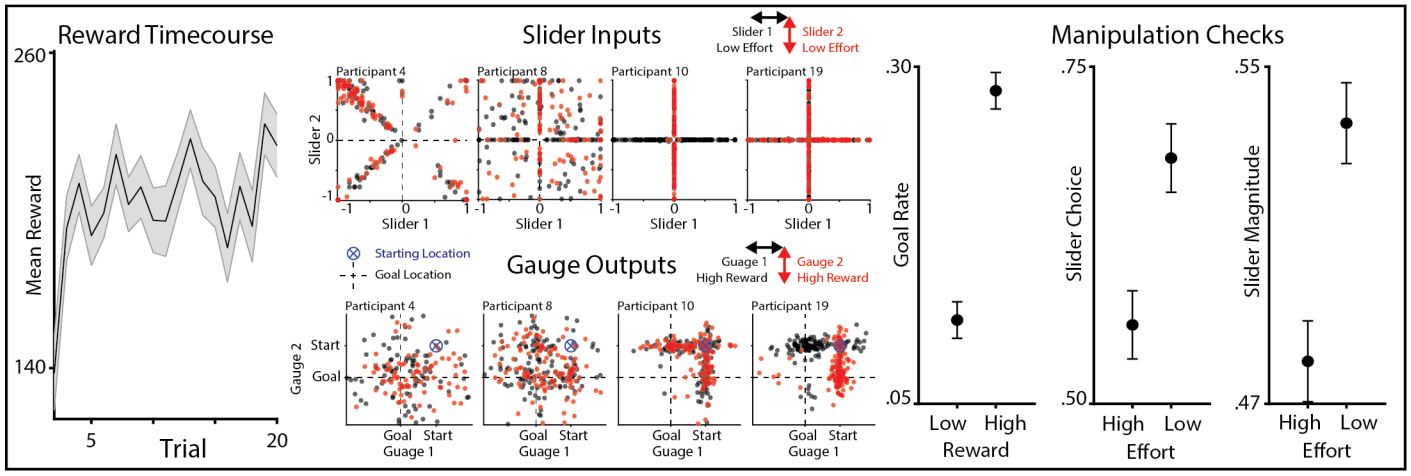

Figure 2: Left: participants' mean reward for each trial, averaged over blocks. Center, top: the distribution of four example participants' inputs, colored according to the effort condition. On-axis choices reflect the usage of a single slider (e.g., Participant 10), whereas off-axis choices reflect a combination of sliders. Center, bottom: the same participants' movement on the gauges, normalized to the proportion of travel from the starting location (blue X) towards the goal (dashed axis), and colored according to reward condition. Movements towards only one of the goals occur on the vertical or horizontal axes in this space (e.g., Participant 19). Right: Mean goal achievement, slider choice, and input magnitude on chosen sliders, separated by reward or effort condition.

3.2 Model-based analysis

We fit our model to participants' choices, and then used the best-fitting parameterization to simulate behavior on each trial (see Figure 3). Like participants, our model showed the fastest learning in the earliest trials of a block. However, the model did not learn as quickly as participant, and had lower asymptotic performance.

We also compared the distribution of slider inputs and gauge outputs between our model and participants (Figure 3). We compared two participants that had relatively poor model fits (Participants 4 & 8; ∼5th percentile) and two participants that had relatively good model fits (Participants 10 & 19; ∼90th percentile). Qualitatively, our model captures the coarse distribution of participant inputs. A salient contrast is that our model is restricted to a symmetrical input preference, whereas some participants show idiosyncratic response biases. We found that participants demonstrating

a 'single slider' strategy were better fit by our model, likely in part due to lower-variance behavior. Our model does a poorer job capturing the distribution of participants' gauge outputs, indicating a need to improve the learning and/or choice components of our model.

Consistent with our behavioral findings, we found that our output (Q) and input (R) parameters significantly differed between reward and effort conditions (sign-rank on participant-level parameters; reward: $z(25) = 3.20$, p=2.63e-5; effort: $z(25) = 3.19$, p=.0014). We found that these parameters effects were strongly correlated with our behavioral measures of reward and effort sensitivity (rank correlation; Q & Goal Rate: $r(24) = .91$; R & Slider Choice: $r(24) = .88$; R & Slider Magnitude: $r(24) = .74$).

Finally, we confirmed that the maximally parameterized model was the best fit to the data. We compared this model to reduced models without either Q, R, S, V, or W sub-objectives, using Bayesian model selection on the Laplacian-approximated likelihood (a complexity-penalized measure of model fit). Within this set of models, we observed a high probability that the maximal model best explained the most participants (PXP > 0.99; BOR = 3.18e-7), accounting for 43% of the variance in the median participant relative to a uniform policy (Pseudo-R^2 quartiles: [29%, 52%]).

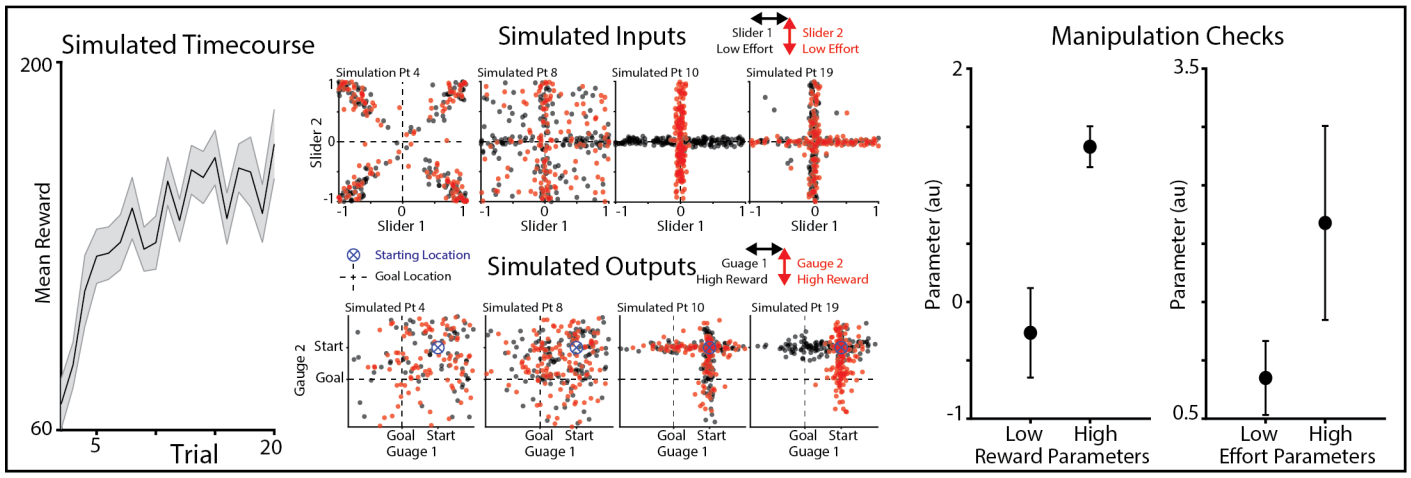

Figure 3: Model-simulated behavior, with the same organization as Figure 2.

4 Discussion

This experiment offers a preliminary look into how people make graded actions in continuous environments, outside of the motor control domain in which this modelling approach has traditionally been studied. Participants' control of this dynamical system was sensitive to the costs and benefits of their actions, and our optimal control analysis was able to identify several predictive features of their decision process. Although our model's slow learning demands a more complete account of participants' learning, potentially including better priors and/or uncertainty-dependent learning, we believe that optimal control offer a promising framework for modeling the dynamics of learning and control. In particular, this framework is general enough that it may help explain the cognitive control process that are thought to similarly depend on cost-benefit arbitration [1,5].

5 References

1. Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron, 79(2), 217-240.

2. Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. Nature neuroscience, 5(11), 1226.

3. Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits (No. TR-1553-1). Stanford Electronics Labs.

4. model-fitting code available at: https://github.com/sjgershm/mfit

5. O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural computation, 18(2), 283-328.