An information-theoretic perspective on speed-accuracy trade-offs and set-size effects

Shuze Liu (shuzeliu@fas.harvard.edu)
PhD Program in Neuroscience, Harvard University, 52 Oxford Street,
Cambridge, MA, USA

Lucy Lai (lucylai@g.harvard.edu)
PhD Program in Neuroscience, Harvard University, 52 Oxford Street,
Cambridge, MA, USA

Samuel J. Gershman (gershman@fas.harvard.edu)
Department of Psychology and Center for Brain Science, Harvard University, 52 Oxford Street,
Cambridge, MA, USA

Bilal A. Bari (bbari@mgh.harvard.edu)
Department of Psychiatry, Massachusetts General Hospital, 55 Fruit Street,
Boston, MA, USA

Abstract

Policies, the mappings from states to actions, require memory. The amount of memory is dictated by the mutual information between states and actions, or the *policy complexity*. High-complexity policies preserve state information and generally lead to greater reward compared to low-complexity policies, which discard state information and require less memory. Under our theory, high-complexity policies incur a time cost: they take longer to decode than low-complexity policies. This naturally gives rise to a speed-accuracy trade-off, in which acting quickly necessitates inaccuracy (via low-complexity policies) and acting accurately necessitates acting slowly (via high-complexity policies). Furthermore, the relationship between policy complexity and decoding speed accounts for set-size effects: response times grow as a function of set size because larger sets require higher policy complexity. Across two human experiments, we tested these predictions by manipulating intertrial intervals, environmental regularities, and state set sizes. In all cases, we found that humans are sensitive to time costs when modulating policy complexity. Altogether, our theory suggests that policy complexity constraints may underlie some speed-accuracy tradeoff and set-size effects.

Keywords: policy compression; speed-accuracy trade-off; set-size effects; reinforcement learning; resource rationality

Introduction

The brain has evolved to function under myriad cognitive resource constraints. Here, we focus on channel capacity, an upper bound on the amount of information that can be transmitted across a noisy channel. We model an agent that learns a policy, $\pi(a|s)$, a probabilistic mapping from states s to actions a. For a resource-rational agent, we formalize the cognitive cost as the mutual information between states and actions, $I(S;A)$, or *policy complexity*, and assume policies are subject to a capacity constraint, C, or an upper bound on policy complexity (Parush, Tishby, & Bergman, 2011; Sims, 2016; Gershman, 2020; Lai & Gershman, 2021). Shannon's noisy channel theorem states that the minimum expected number of bits to transmit a signal across a noisy information channel without error is equal to the mutual information. Higher policy complexity therefore demands more memory. We define the optimal policy, $\pi^* = \arg\max_\pi V^\pi$ subject to $I(S;A) \leq C$, where V^π is the expected reward under policy π. This can be solved using Lagrange multipliers, leading to the solution $\pi^*(a|s) \propto \exp(\beta Q(s,a) + \log P^*(a))$ where $Q(s,a)$ is the expected reward for taking action a in state s and $P^*(a) = \sum_s \pi^*(a|s)p(s)$ is the optimal marginal action distribution.

The optimal policy takes the form of the familiar softmax distribution, common in the reinforcement learning literature. Here, the Lagrange multiplier, β, plays the role of the inverse temperature parameter. Moreover, β is a function of the policy complexity: $\beta^{-1} = \frac{dV^\pi}{dI^\pi(S;A)}$. It is large at high policy complexity and small at low policy complexity. By varying β and calculating the optimal policy, we can trace out the reward-complexity frontier, which delimits the maximal trial-averaged reward obtainable for a given policy complexity (Figure 1A). In general, high-complexity policies yield more reward per trial than low-complexity policies. Moreover, low-complexity policies are dominated by the $\log P^*(a)$ term, a form of perseveration (state-independent actions) (Lai & Gershman, 2021).

Our formulation up to this point has ignored time costs. In order to understand why an agent would choose a low-complexity policy, let us assume states are represented as codewords through entropy coding, the canonical example of which is the Huffman code (Huffman, 1952). The Huffman code corresponds to a binary tree in which leaf nodes correspond to decoded states, where more complex state descriptions necessitate more leaf nodes, and therefore more bits. If we assume bits are inspected at a constant rate, then more complex policies take longer to read out to reveal the decoded action (Hick, 1952). Policies of high complexity necessitate more bits, and reading out these policies should take longer, necessitating longer response times (RTs). Moreover, given that bits are inspected at a constant rate, response times should be a linear function of policy complexity / description length, with some offset to reflect motor delay (Figure 1B).

To see how our theory predicts a speed-accuracy trade-off, let us assume subjects attempt to maximize *time-averaged* reward (rewards divided by time) (Balci et al., 2011; Drugowitsch, DeAngelis, Angelaki, & Pouget, 2015). This yields the relationship in Figure 1C, where we varied the intertrial interval (ITI). To maximize time-averaged reward, humans should decrease policy complexity when ITIs are short; although these policies result in less trial-averaged reward, they increase time-averaged reward because they allow agents to perform more actions due to smaller decoding time cost. Moreover, because the optimal policy includes a perseverative term ($\log P^*(a)$), the contribution of perseveration should be magnified at low policy complexity (low ITIs) because of the smaller β term. Regarding set-size effects, our theory predicts that response times should grow as a function of set size because larger sets require higher policy complexity (i.e., the policy must encode more states) to maximize time-averaged reward, which in turn demands longer decoding time.

Methods

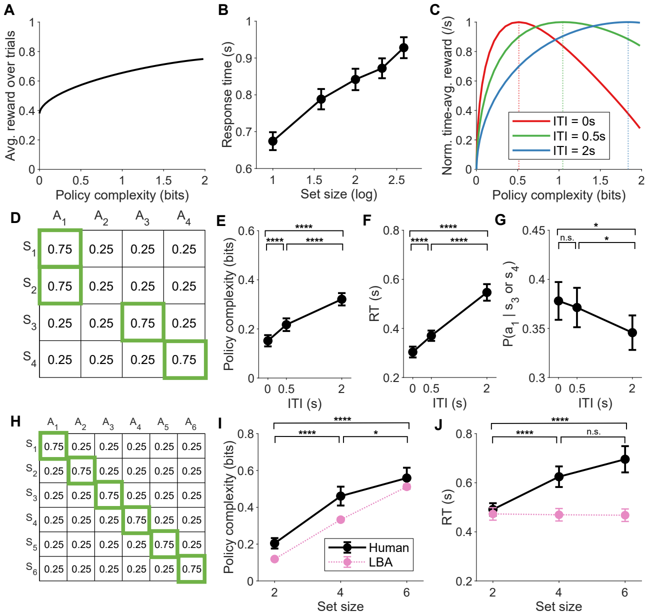
We used instrumental learning tasks, where participants pressed keyboard keys in response to images presented on a computer monitor.

In Experiment 1 ($N = 198$), on each trial, participants saw one of four possible images (states) and pressed one of four possible keys (actions). We varied the ITI (0s, 0.5s, or 2s) to modulate the optimal policy complexity (Figure 1C). Furthermore, the experiment was designed so two states shared the same optimal action a_1 (Figure 1D), to test the prediction that the marginal action distribution, $P^*(a)$, should affect low-complexity policies more. We first trained participants on all three ITI conditions (1 min each) to encourage them to learn the correct state-action mapping. They then completed three 3-minute test blocks, one for each ITI (order randomized).

Experiment 2 ($N = 99$) consisted of three separate conditions, each with a different number of available states (set size 2, 4, or 6; order randomized). The action set size was always 6. Each state had a unique optimal action (Figure 1H). Like Experiment 1, in each set-size condition, subjects were trained on the three ITIs (subjects saw 48 training trials/state to control for learning across set sizes). They then completed a 3-minute test block with ITI=2s.

After each trial in each experiment, subjects received reward feedback for 0.3s (green outline reporting reward or grey outline reporting no reward). Hence the total time of a trial was (RT + feedback time + ITI).

Results



	A_1	A_2	A_3	A_4
S_1	0.75	0.25	0.25	0.25
S_2	0.75	0.25	0.25	0.25
S_3	0.25	0.25	0.75	0.25
S_4	0.25	0.25	0.25	0.75

	A_1	A_2	A_3	A_4	A_5	A_6
S_1	0.75	0.25	0.25	0.25	0.25	0.25
S_2	0.25	0.75	0.25	0.25	0.25	0.25
S_3	0.25	0.25	0.75	0.25	0.25	0.25
S_4	0.25	0.25	0.25	0.75	0.25	0.25
S_5	0.25	0.25	0.25	0.25	0.75	0.25
S_6	0.25	0.25	0.25	0.25	0.25	0.75

Figure 1: A) Reward-complexity frontier. B) RT is linear in log encoded state information. Data from (Collins et al., 2014). C) Time-averaged reward over policy complexity, assuming linear RT-to-policy-complexity relationship. D) Experiment 1 reward structure. E-F) Policy complexity and RT vs. ITI. G) Probability of choosing the shared action in states where that action is suboptimal. H) Experiment 3 reward structure. I-J) Policy complexity and RT vs. ITI, overlaid with fitted LBA predictions.

Experiment 1: speed-accuracy trade-off

In Experiment 1, we manipulated ITIs to test whether participants adjust policy complexity to maximize time-averaged reward. Under longer ITIs, we predicted 1) higher policy complexity and 2) slower RTs, as this combination maximizes time-averaged reward. Because the marginal action distribution, $P(a)$, contributes less under higher policy complexity, we predicted 3) decreased perseveration with longer ITIs. To gain intuition, in the extreme case of policy complexity of 0 where subjects do not encode the stimuli at all, they should always pick action a_1, since this maximizes reward. We analyzed the policies for stimuli s_3 and s_4 to identify the effect of the marginal action distribution; for these stimuli, under high policy complexity, a_1 should be chosen infrequently since it is not the reward-maximizing option. However, as policy complexity decreases and the marginal action distribution has greater influence on the policy, a_1 should be chosen more often, being overall the best single action at low complexity.

Consistent with our predictions, participants achieved near-maximal trial-averaged reward as a function of policy complexity. As a function of ITI, they adopted more complex policies (Figure 1E) and slower RTs (Figure 1F). Furthermore, the influence of the marginal action distribution decreased with longer ITIs (Figure 1G). We validated our proposed linear relationship between policy complexity and RT by fitting linear mixed effects (LME) models, which yielded significant effects for policy complexity.

Experiment 2: set-size effects

In Experiment 2, we varied the state set size, since larger set sizes demand higher policy complexity to maximize reward. As predicted, participants increased policy complexity for larger set sizes, leading to longer RTs (Figure 1I-J).

To compare our theoretical framework with evidence accumulation models that can simultaneously predict choice and RT across multiple alternatives, we fit a linear ballistic accumulator model (LBA) to our data (Brown & Heathcote, 2008). The LBA included two mean drift rate parameters: one for the optimal action, and one for all suboptimal actions. While the best-fit LBA could capture the increase in policy complexity as a function of set size, it could not predict the relationship between RT and set size (Figure 1I-J).

Discussion

Across two experiments, we found that humans are sensitive to the time costs of decoding policies, and that this single relationship predicted behavior in domains as varied as speed-accuracy trade-offs and set-size effects. One novel contribution of our framework is the *linear* relationship between RT and policy complexity. The idea that retrieving a policy from memory incurs a time cost offers an alternative to sequential sampling models for describing RTs (Forstmann, Ratcliff, & Wagenmakers, 2016; McDougle & Collins, 2021). Second, our framework provides a normative interpretation of goal-directed $Q(s,a)$ and habitual $P(a)$ components of action selection, despite their different units (reward vs. frequency). The poor LBA fits are likely due to their insensitivity to past action frequencies in each set size, highlighting the contribution of both components to behavior.

Finally, by simultaneously accounting for multiple cognitive constraints (here, time and memory costs), we have unified several disparate findings. We suggest that concurrently accounting for time and memory costs may provide a normative basis for other seemingly disparate findings in psychology.

Acknowledgments

This work was supported by Grant R25MH094612 from the National Institute of Mental Health, Bethesda, MD (B.A.B.), Grant DRL-2024462 from the National Science Foundation, and the Harvard Brain Science Initiative Bipolar Disorder Seed Grant.

References

Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., & Cohen, J. D. (2011). Acquisition of decision making criteria: reward rate ultimately beats accuracy. *Attention, Perception, & Psychophysics*, *73*, 640–657.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153–178.

Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *Journal of Neuroscience*, *34*(41), 13747–13756.

Drugowitsch, J., DeAngelis, G. C., Angelaki, D. E., & Pouget, A. (2015). Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making. *eLife*, *4*.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual review of psychology*, *67*, 641–666.

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, 104394.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of experimental psychology*, *4*(1), 11–26.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, *40*(9), 1098–1101.

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In *Psychology of learning and motivation* (Vol. 74, pp. 195–232). Elsevier.

McDougle, S. D., & Collins, A. G. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic bulletin & review*, *28*, 20-39.

Parush, N., Tishby, N., & Bergman, H. (2011). Dopaminergic balance between reward maximization and policy complexity. *Frontiers in Systems Neuroscience*, *5*, 22.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181-198.