Time and memory costs jointly determine a speed-accuracy trade-off and set-size effects

Shuze Liu[1], Lucy Lai[1], Samuel J. Gershman[2,3], and Bilal A. Bari[4,5]

[1]PhD Program in Neuroscience, Harvard University, Cambridge, MA, USA

[2]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA, USA

[3]Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA, USA

[4]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA

[5]McLean Hospital, Harvard Medical School, Belmont, MA, USA

Author Note

Shuze Liu https://orcid.org/0000-0002-3946-5003

The online task, data, and analysis code for all experiments are available at https://github.com/LSZ2001/policycompression_timememorycosts.

We have no conflicts of interest to disclose.

Shuze Liu served as lead for data curation, formal analysis, investigation, project administration, visualization, and writing–original draft. Lucy Lai contributed to data curation, formal analysis, investigation, and writing–review & editing. Bilal Abdul Bari and Samuel Joseph Gershman served as lead for supervision and writing–review & editing, while contributing to writing–original draft equally. Samuel Joseph Gershman served as lead for funding acquisition. Shuze Liu, Lucy Lai, Bilal Abdul Bari, and Samuel Joseph Gershman all contributed to conceptualization and methodology equally.

Correspondence concerning this article should be addressed to Shuze Liu, PhD Program in Neuroscience, Harvard University, 52 Oxford St., Cambridge, MA 02138, United States. Email: shuzeliu@fas.harvard.edu

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 require less memory by discarding state information and exploiting environmental regularities. Under this 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 the number of possible states because larger state sets encourage higher policy complexity. Across three experiments, we tested these predictions by manipulating inter-trial intervals, environmental regularities, and state set sizes. In all cases, we found that humans are sensitive to both time and memory costs when modulating policy complexity. Altogether, our theory suggests that policy complexity constraints may underlie some speed-accuracy trade-offs and set-size effects.

Keywords: resource rationality, decision making, information theory, reinforcement learning, policy compression

Public significance statement: This study suggests that people make decisions to balance rewards and cognitive costs. Understanding this trade-off helps explain why complex decisions take more time or cognitive effort, and why we sometimes make mistakes when under time and memory pressure.

Time and memory costs jointly determine a speed-accuracy trade-off and set-size effects

All computational systems—the human brain included—are subject to physical constraints that limit the ability to store and transmit information. Decision making taxes these limited cognitive resources, bounding achievable performance. The framework of *resource rationality* formalizes this idea, treating decision making as a constrained optimization problem that considers not only performance but also the costs associated with making decisions (Bhui et al., 2021; Lieder & Griffiths, 2020). In biological systems, these costs are often formalized as time costs and memory costs, thought to be key factors underlying cognitive resource limitations (Bhui et al., 2021; Callaway et al., 2023; Lieder & Griffiths, 2020; Vul et al., 2014). Importantly, time and memory costs are typically studied in isolation, although it seems plausible that both should interact to influence behavior.

Time costs are typically studied in tasks where speed and accuracy trade off against one another (Garrett, 1922; Woodworth, 1899). In the domain of decision making, speed-accuracy trade-offs have been widely observed across numerous perceptual and memory-based decision tasks (Balci et al., 2011; Bogacz et al., 2010; Heitz, 2014; Hick, 1952). People can be made to trace out a speed-accuracy function through explicit instruction or with task designs that sharply favor particular strategies (Heitz, 2014; Wickelgren, 1977; Wu et al., 2023).

Early attempts at conceptualizing the computational logic underlying speed-accuracy trade-offs suggested that they arise from the increased information processing required for accurate responses, resulting in longer response times (RTs; Wickelgren, 1977). More recent work has argued that speed-accuracy trade-offs may arise from an imperative to maximize time-averaged reward (that is, reward per unit time; Simen et al., 2009). Evidence in favor of this normative principle has been observed in both perceptual decision making and cognitive control domains (Balci et al., 2011; Bogacz et al., 2010; Drugowitsch et al., 2015; Otto & Daw, 2019).

Despite this rich literature, gaps remain in our understanding of speed-accuracy trade-offs. There is limited work on how humans navigate this trade-off in multi-alternative, value-based task settings, a domain commonly studied in the reinforcement learning literature. This setting has relevance for more ecologically-meaningful behaviors (Pirrone et al., 2014). While there have been efforts to extend sequential sampling—a commonly used perceptual modeling framework—to characterize maximization of time-averaged reward in value-based decisions (Tajima et al., 2016, 2019), the resulting models primarily address tasks in which the decision-maker deliberates between stimuli, such that each stimulus elicits noisy evidence for one unique action. These models do not naturally extend to naturalistic value-based settings, in which humans must choose one of sometimes many actions in response to an environment state.

None of the models discussed above address the additional influence of memory costs on decision

making. A separate literature has demonstrated that decision making degrades as memory costs (i.e., the amount of information needed to implement the optimal policy) increase (Collins, 2018; Collins & Frank, 2012; Lai & Gershman, 2024). Models with limited memory capacity have been developed to explain this and related findings, but these models do not typically address the speed-accuracy trade-off.

Central to the present paper is the idea that time and memory costs are deeply intertwined: information stored in memory must be "decoded" into action, and this decoding process takes longer when more information is stored (Lai & Gershman, 2021). Memory incurs a time cost, and therefore decision-makers actually face a speed-accuracy-memory trade-off. Our goal is to understand this trade-off theoretically and explore it empirically.

In this paper, we develop a normative framework that jointly considers how time and memory costs influence decisions, and test its predictions in three instrumental learning experiments. We will show that across experimental conditions, human participants flexibly adjust their choice and RT profiles as predicted by the framework.

The policy compression framework

This section describes our theoretical framework formally. We first define how to optimize decision making under memory constraints. We then introduce time costs and link them to the memory constraints.

Memory-constrained policy optimization

The nervous system must contend with numerous constraints, including computational costs (Bossaerts et al., 2019), interference costs (Musslick et al., 2016), and metabolic costs (Gailliot & Baumeister, 2007), among other costs (Shenhav et al., 2017). Here, we will focus on how channel capacity, an upper bound on the amount of information that can be transmitted across a noisy channel (Miller, 1956; Shannon, 1948), affects decision making—both in terms of decisions as well as how quickly those decisions are made (Figure 1A,B).

For a resource-rational agent, we formalize memory usage as the mutual information between states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$, which we call the *policy complexity*:

$$I^{\pi}(S;A) = \sum_{s} P(s) \sum_{a} \pi(a|s) \log \frac{\pi(a|s)}{P(a)} \tag{1}$$

where $\pi(a|s)$ is the policy, a probabilistic mapping from states to actions, and $P(a) = \sum_s P(s)\pi(a|s)$ is the marginal probability of choosing action a. High complexity policies are ones that preserve state information (e.g., deterministic mappings from states to actions) whereas low complexity policies discard state information (e.g., random actions). In general, we assume that policies are subject to a capacity constraint, C, an upper bound on policy complexity. 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. Therefore, if the optimal policy requires more memory than the agent possesses, then the agent must *compress* its policy, or render it less state-dependent. We define the optimal policy, π^*, as:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, V^\pi, \text{ subject to } I^\pi(S;A) \leq C \tag{2}$$

where $V^\pi = \sum_s P(s) \sum_a \pi(a|s)\ Q(s,a)$ is the trial-averaged reward (i.e., reward per trial) under policy $\pi(a|s)$, and $Q(s,a)$ is the trial-averaged reward for taking action a in state s.

We can express the above constrained optimization problem in the following unconstrained Lagrange form:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \beta V^\pi - I^\pi(S;A) + \lambda(s)\left(\sum_a \pi(a|s) - 1\right) \tag{3}$$

where $\beta \geq 0, \lambda(s) \geq 0\ \forall s \in S$ are Lagrange multipliers.[1] Solving this equation yields the following optimal policy:

$$\pi^*(a|s) \propto \exp[\beta Q(s,a) + \log P^*(a)] \tag{4}$$

where $P^*(a) = \sum_s \pi^*(a|s)\ p(s)$ 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. Note that although β typically takes on the role of balancing exploration/exploitation in reinforcement learning, we made no such appeals in deriving this policy. Moreover, β is a function of the policy complexity:

$$\beta^{-1} = \frac{dV^\pi}{dI^\pi(S;A)} \tag{5}$$

At high policy complexity, when $\frac{dV^\pi}{dI^\pi(S;A)}$ is shallow, the optimal β is large and the policy is dominated by Q-values, which renders it state-dependent. At low policy complexity, the optimal β is close to 0, and Q-values have minimal impact on the policy. Moreover, low-complexity policies are dominated by the $\log P^*(a)$ term, a form of perseveration (state-independent actions). In general, high-complexity policies yield more reward per trial than low-complexity policies. 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 1C).

So far, we have treated memory as a constraint on policies, but it may also incur an additional subjective cost (henceforth "memory cost"). We hypothesize that this cost may manifest as perceived task difficulty (i.e., higher policy complexity is perceived as more cognitively demanding), possibly pushing policy complexity even lower than what would be predicted by the policy compression framework.

[1] $\lambda(s)$ terms ensure proper normalization: $\sum_a \pi(a|s) = 1$.

Time costs

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 state descriptions 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 RTs. We have previously observed a significant correlation between RT and policy complexity (Bari & Gershman, 2023; Lai & Gershman, 2021, 2024). 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 1D,E).

To see how the above theory predicts a speed-accuracy trade-off, let us assume agents attempt to maximize time-averaged reward (Balci et al., 2011; Drugowitsch et al., 2015). To make this concrete, let us take a simple premise where the agent perceives a state, selects an action after a response time, t_{RT}, and waits through an inter-trial interval (ITI) for t_{ITI} seconds before the next trial. The time-averaged reward under these conditions takes the following form:

$$V^{\pi}_{\text{time}}(I(S;A)) = \frac{V^{\pi}(I(S;A))}{t_{\text{RT}}(I(S;A)) + t_{\text{ITI}}} \tag{6}$$

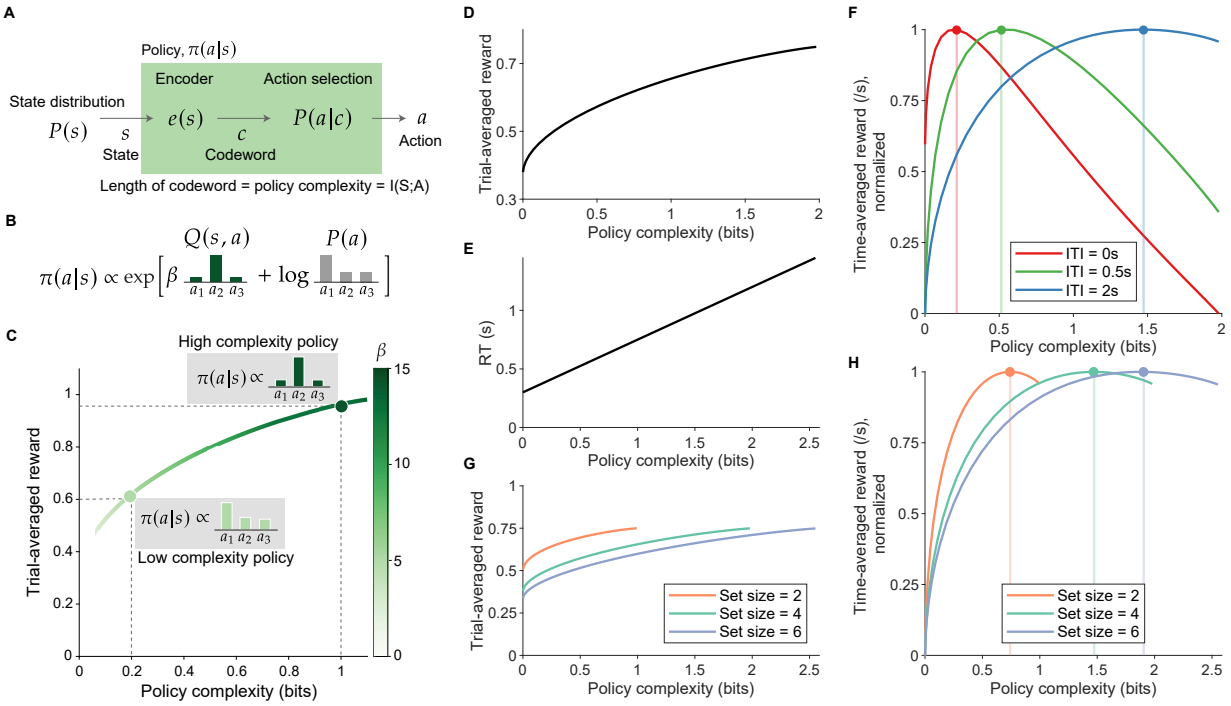
where $V^{\pi}_{\text{time}}(I(S;A))$, is the time-averaged reward. $V^{\pi}(I(S;A))$ is a function of policy complexity through the derivation of the optimal policy (Figure 1D) and $t_{\text{RT}}(I(S;A))$ is a function of policy complexity through the assumption of a linear relationship between RT and policy complexity (Figure 1E).

To see how the theory predicts a speed-accuracy trade-off, we can visualize the relationship between time-averaged reward and policy complexity in Figure 1F, where we varied the 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. This is an interpretation of the speed-accuracy trade-off founded in the notion of resource rationality. Regarding set-size effects, in which "set size" refers to the number of possible states/stimuli, the 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 (Figure 1G,H).

Acting faster can yield greater time-averaged reward under some task conditions (short ITIs or small stimulus set sizes), but necessitates a commitment to greater errors through lower policy complexity. Under other conditions (long ITIs or large stimulus set sizes), one can have the guarantee of fewer errors through higher policy complexity, though with the requirement of acting slower. A host of other predictions fall out of this single relationship, which we will elaborate on in the Experiment sections below.

Figure 1

Interplay of time and memory resources under policy compression. ***(A)*** *The policy as a communication channel. A state distribution $P(s)$ generates states s that are encoded into memory via an encoder, $e(s)$, yielding a codeword c. The codeword is then mapped onto an action a according to $P(a|c)$. Together, encoding and action selection produce the policy $\pi(a|s)$ that maps states to actions.* ***(B)*** *The optimal policy includes a state-dependent term, $Q(s,a)$, and a state-independent term, $\log P(a)$. The $\log P(a)$ term biases choices towards actions that are frequently chosen across all states. The β parameter determines the relative contribution of $Q(s,a)$ and $\log P(a)$, controlling the state-dependence of the policy. We highlight distributions for an example state.* ***(C)*** *A limit on the channel capacity results in a trade-off between reward and compression. The β parameter increases monotonically with policy complexity. We highlight two example optimal policies at different policy complexity levels. The optimal policies trace out the reward-complexity frontier, which delimits achievable performance for a given policy complexity.* ***(D)*** *Reward-complexity frontier for Experiment 1.* ***(E)*** *Proposed linear relationship between RT and policy complexity.* ***(F)*** *For Experiment 1, time-averaged reward as a function of policy complexity for each ITI under the linear RT-to-policy complexity relationship in (E); the optimal policy complexity for each condition is highlighted (vertical lines).* ***(G)*** *Reward-complexity frontiers for Experiment 3.* ***(H)*** *For Experiment 3, time-averaged reward as a function of policy complexity for each set-size condition under the linear RT-to-policy complexity relationship in (E). Panels A-C adapted from (Lai & Gershman, 2024)*



Experiment 1

We aimed to test whether humans incorporate both time and memory costs when making decisions. In Experiment 1, we manipulated ITIs to test whether humans adjust policy complexity to maximize time-averaged reward. We made the following predictions related to time costs. Under longer ITIs, we predict 1) higher policy complexity, which, given the framework's proposed relationship between policy complexity and RT, leads to 2) slower RTs, because more complex policies take longer to decode. According to the framework, this combination of policy complexity and RT maximizes time-averaged reward under longer ITIs (Figure 1F). Next, since higher policy complexity dictates a more deterministic mapping from states to actions, we predict 3) decreased action stochasticity (the conditional entropy of actions conditioned on state, $H(A|S)$) with longer ITIs. As detailed in the Introduction, the optimal policy is proportional to the exponentiated action values, weighted by an inverse temperature parameter β, and the marginal action distribution. At higher policy complexity, β increases, which decreases the influence of the marginal action distribution $P(a)$ on the policy. Assuming $P(a)$ is estimated and updated on a trial-by-trial basis, we predict 4) decreased perseveration (the probability of repeating the same action) with longer ITIs. Finally, we predict 5) decreased time-averaged reward; although higher policy complexity results in increased trial-averaged reward, this is offset by the longer time spent in the ITI.

We made two further predictions related to memory costs. If increased memory utilization is costly (Zenon et al., 2019), then we predict 6) longer ITIs should be associated with higher perceived difficulty, since maximizing time-averaged reward under longer ITIs necessitates greater policy complexity. Such difficulty measurements were available to us, as participants had ranked all experimental blocks by their perceived difficulty. Furthermore, we predict that participants will show 7) a systematic leftward bias in policy complexity, in which their empirical policy complexity is lower than what is optimal. This is a non-trivial prediction, since one would not expect this simply from maximizing time-averaged reward, as implementing a policy of slightly less or greater complexity results in similar time-averaged reward and there should therefore be no bias.

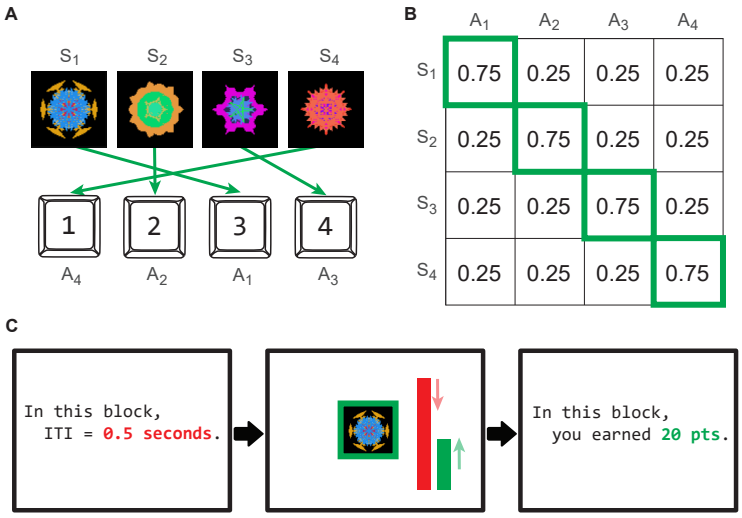
Materials and Methods

Participants

One-hundred participants (37 female, 62 male, 1 non-binary, 1 prefer not to say) were recruited. We selected the sample size based on the lowest estimated effect size (Cohen's $d = 0.312$) among dependent variables, according to estimates from a separate group of $N = 48$ pilot participants (data excluded from final analysis). All analyses were preregistered at https://aspredicted.org/blind.php?x=VF2_NH6. We excluded 3 participants for having an average RT for any block exceeding 5 seconds, leaving data from 97

Figure 2

Experiment 1 setup. ***(A)*** *The four possible states (images) and the corresponding optimal actions (key presses). The mapping between images and keys was randomized across participants.* ***(B)*** *Experiment 1 reward structure $Q(s, a)$. The optimal action for each state is indicated by a green border.* ***(C)*** *On every trial, the participant observes an image (state) and responds by pressing a key (action). Reward feedback is provided as a green border around the image if the action was rewarded or a gray border if action was not rewarded. Participants are able to track remaining time (red bar) and cumulative reward (green bar) for the block. After the block ends, participants receive feedback on the total reward gained in the block. Participants are informed of the block's ITI before starting.*



participants (35 female, 60 male, 1 non-binary, 1 prefer not to say) for subsequent analyses. Participants gave informed consent, and the Harvard University Committee on the Use of Human Subjects approved the experiment.

Procedure

Each participant completed three blocks of trials with ITIs of 0s, 0.5s, and 2s respectively. The block order was randomized across participants. Participants were informed of the ITI of each block. Participants were informed that they would receive a bonus proportional to their performance for each block (i.e., relative to the maximum reward attainable for each block).

There were four possible states (images) and four available actions, which are shared across blocks. Each stimulus was assigned a unique optimal action (Figure 2B). Participants were informed that the mapping from stimulus to action was held fixed across all blocks. This was done to minimize the learning of action values within blocks.

On each trial, participants were presented with one image (state) and responded by pressing one of several possible keyboard keys (actions; Figure 2A). Stimulus presentation was counterbalanced within runs of 8 trials, where the stimulus presentation order was randomized within each run and each of the four images appeared exactly twice per run. We did this to ensure a uniform state distribution $P(s)$, allowing us to better estimate policy complexity. See Supplementary Figure 1 for evidence that participants did not exploit this regularity. Reward delivery was binary and probabilistic: each state was associated with one optimal action (Figure 2B). After making a response, participants were given immediate feedback for 0.3s—either a green border around the image to indicate reward or a gray border to indicate no reward. We did not use punishment feedback. A fixation cross then appeared throughout the ITI. Each block lasted until 3 minutes elapsed and the current run of trials finished. Participants could track the remaining time and reward earned during the block, which were displayed as red and green bars, respectively. At the end of each block, they were provided with feedback on the total reward they earned in that block (Figure 2C).

Participants completed three 1-minute training blocks, one for each ITI condition, to familiarize themselves with the task and learn the mapping from stimulus to response. These data were not analyzed. Participants then completed the three 3-minute blocks where ITI was varied, as mentioned above. After completing the whole experiment, participants ranked the perceived difficulty for each block. Participants additionally completed the Barratt Impulsiveness Scale, which we did not analyze for this manuscript.

Statistical analysis

Due to the directional nature of the framework's predictions, all statistical tests were one-sided paired t-tests except for the test of optimal minus empirical policy complexity, which was a one-sided Wilcoxon signed-rank test due to strong non-normality. The one-sided test directions were preregistered. In the main text, we report all pairwise comparisons, their effect sizes, and the 95% confidence intervals (CI) of the effect sizes in Supplementary Tables 1 to 2.

We fit a linear mixed-effects (LME) model to determine the participant-specific relationship between average RT and policy complexity for a block. The fixed effects were the intercept and policy complexity and random effects were intercept and policy complexity (independent from each other), grouped by participant. We obtained parameter estimates using maximum likelihood estimation with the "fitlme" function in MATLAB R2023a.

Transparency and Openness

Data and code for this and subsequent experiments are available at https://github.com/LSZ2001/policycompression_timememorycosts.

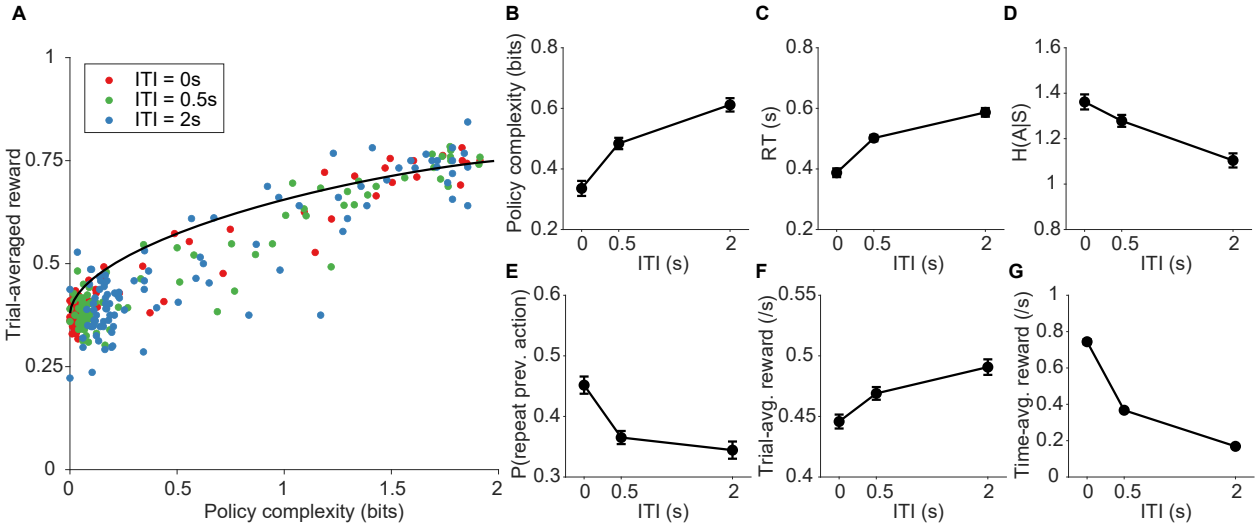
Results

Consistent with prior results, participants achieve near maximal trial-averaged reward as a function of policy complexity (Figure 3A; Bari & Gershman, 2023; Gershman, 2020; Gershman & Lai, 2021; Lai & Gershman, 2024). In line with our predictions that humans modulate policy complexity to maximize time-averaged reward, participants used policies of higher complexity in longer ITI blocks compared to shorter ITI blocks, $t(96) = -6.40$, $p < 10^{-8}$ (Figure 3B). Participants also adopted slower RTs in longer ITI blocks, $t(96) = -7.44$, $p < 10^{-10}$ (Figure 3C), consistent with the notion that higher complexity policies are slower to execute. Action stochasticity similarly decreased in longer ITI blocks as policies became concentrated on one action for each state, $t(96) = 4.39$, $p < 10^{-4}$ (Figure 3D). Perseveration, as predicted, decreased as a function of ITI, $t(96) = 4.09$, $p < 10^{-4}$ (Figure 3E). This overall led to a reduction in time-averaged reward as a function of ITI, $t(96) = 25.6$, $p < 10^{-44}$ (Figure 3F,G). All of these findings are consistent with the idea that humans are sensitive to time costs when adjusting policy complexity.

We next tested the hypothesis that humans are sensitive to memory costs when adjusting policy complexity. One readout of this is perceived task difficulty—if implementing a high complexity policy is

Figure 3

Experiment 1 sensitivity to time costs. ***(A)*** *Trial-averaged reward of participants across ITI conditions (color), and the theoretical upper bound (reward-complexity frontier; black) at each policy complexity level. Some data points lie above the optimal frontier due to the stochastic nature of reward delivery.* ***(B-G)*** *Mean±SEM of participant policy complexity (B), average RT (C), action stochasticity (D), perseveration (E), trial-averaged reward (F), and time-averaged reward (G) across ITI conditions. All SEM errorbars were within-participant (Cousineau et al., 2005).*



costly, then participants should perceive it as more cognitively demanding. This was indeed the case, as participants ranked the ITI = 0s condition the easiest—which demands the lowest policy complexity—and ranked the ITI = 2s condition the most difficult—which demands the highest policy complexity, $t(96) = -5.11$, $p < 10^{-6}$ (Figure 4A). Furthermore, we computed the Spearman correlation for each participant between their perceived difficulty rating and empirical policy complexity across ITI conditions and confirmed they were positively correlated at the single-participant level, $t(96) = 12.8$, $p < 10^{-22}$, Cohen's $d = 1.29$, 95% CI $[1.02, 1.56]$. Note that one would have predicted the opposite if the motor cost of the task conditions dominated, since shorter ITI conditions demand a higher frequency of button presses to maximize reward.

Finally, since we confirmed that higher policy complexity is costlier, we tested our prediction that participants should exhibit a leftward bias in policy complexity. First, we validated the proposed linear relationship between policy complexity and RT by fitting LME models to predict average RT as a function of policy complexity. The fitted model yielded significant effects for the intercept (fixed effects 0.301 ± 0.0251, $t(289) = 12.0$, $p < 10^{-26}$; random effects $SD = 0.161$) and policy complexity (fixed effects 0.445 ± 0.0400, $t(289) = 11.1$, $p < 10^{-23}$; random effects $SD = 0.0870$). Visually, empirical RTs and the LME-predicted RTs correlated well with one another (Figure 4B) and most participants shared linear time-cost functions that largely differed by intercept (Figure 4C). We next used the fitted participant-specific linear time-cost functions to estimate the optimal policy complexity for each ITI condition. For each participant, we compared empirical policy complexity to optimal and confirmed a leftward policy complexity bias for each ITI condition: ITI = 0s, $z = -2.48$, $p = 0.00665$, Cliff's $\delta = -0.468$, 95% CI $[-0.630, -0.307]$; ITI = 0.5s, $z = -2.14$, $p = 0.0160$, Cliff's $\delta = -0.401$, 95% CI $[-0.568, -0.233]$; ITI = 2s, $z = -8.20$, $p < 10^{-15}$, Cliff's $\delta = -0.674$, 95% CI $[-0.794, -0.555]$ (Figure 4D-F).

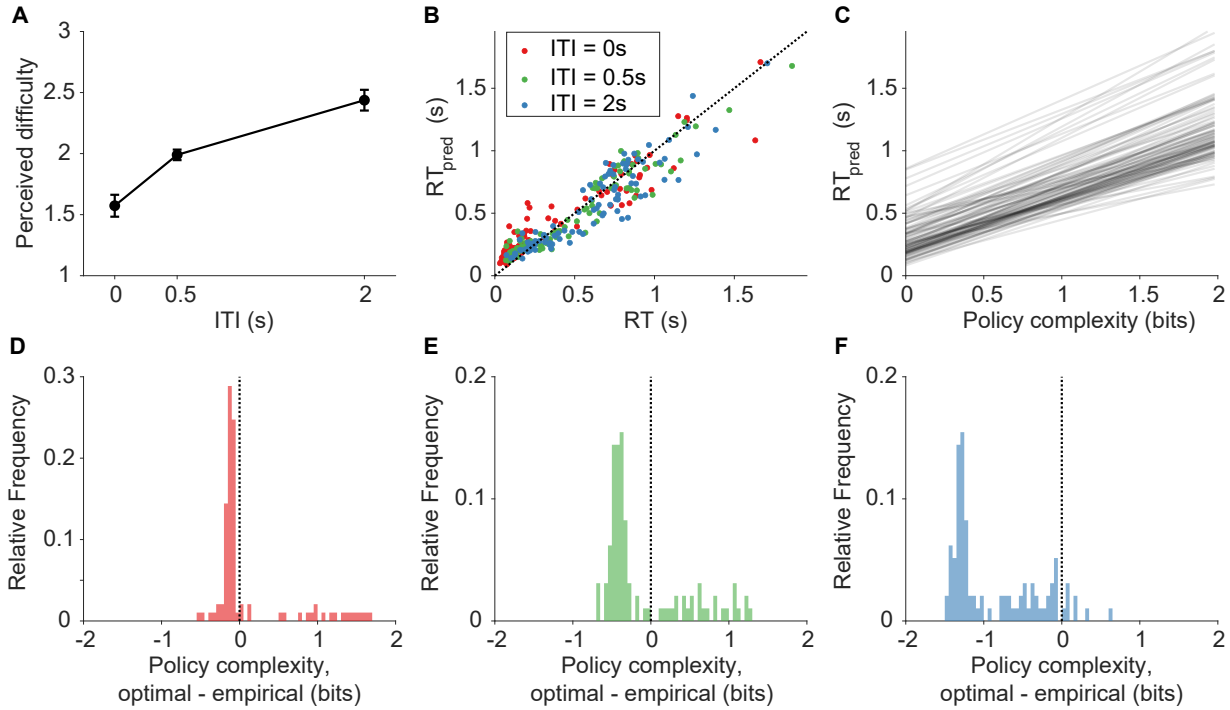
We further used these results to address an alternative explanation of our findings: perhaps a small subgroup of participants adjusted policy complexity but most of our findings can be attributed to a large subgroup of disengaged participants who consistently responded randomly across all ITI conditions. To test this explanation, we used the leftward policy complexity bias for the ITI = 2s condition (Figure 4F) and partitioned participants into two subgroups—a "low-complexity" and a "high-complexity" group. Consistent with our prior findings, this "low-complexity" subgroup significantly modulated policy complexity and RT across ITI conditions (Supplementary Figure 2).

Discussion

We found that participants in Experiment 1 were sensitive to ITI conditions, modulating their policy complexity and RT in the direction of time-averaged reward maximization. The experimental data

Figure 4

Experiment 1 linear mixed-effects modeling results and sensitivity to memory costs. (A) *Mean±SEM of perceived difficulty rankings (1 denotes easiest block; 3 denotes hardest) for each ITI condition.* ***(B)*** *Model-predicted RT and empirical RT, for each participant in each ITI condition.* ***(C)*** *Model-predicted linear relationship between average RT and policy complexity for each participant.* ***(D-F)*** *Difference in policy complexity between optimal (participant's predicted optimal policy complexity) and that participant's empirical complexity, for each ITI condition (left to right: 0s, 0.5s, 2s).*



supported all seven predictions of the policy compression framework, demonstrating that humans are sensitive to both time and memory costs when making decisions.

Experiment 2

In Experiment 2, we tested an additional prediction of policy compression: humans should exploit environmental regularities (e.g., multiple states sharing the same optimal action) when compressing their policies (endogenized by the $P^*(a)$ term in the optimal policy). We designed Experiment 2 to test this unique prediction and to replicate findings from Experiment 1. We introduced environmental regularity by having two states (s_1 and s_2) share the same optimal action a_1 (Figure 5A). This has the effect that the optimal marginal action distribution, $P^*(a)$ is non-uniform and favors that action. We hypothesized that

the effect of the optimal marginal action distribution would be greatest at low policy complexity, since marginal actions influence the policy more strongly at low complexity (Equation 4).

Materials and Methods

Participants

Two-hundred participants (113 female, 83 male, 4 prefer not to say) were recruited. All participants did not participate in Experiment 1. We selected the sample size based on the lowest estimated effect size (Cohen's $d = 0.245$) among dependent variables of interest, according to analyses of a separate group of $N = 50$ pilot participants (data excluded from final analysis). All analyses were preregistered at https://aspredicted.org/blind.php?x=VF2_NH6. The inclusion criterion was identical to Experiment 1. A total of 198 participants (112 female, 82 male, 4 prefer not to say) met this inclusion criteria. Participants gave informed consent, and the Harvard University Committee on the Use of Human Subjects approved the experiment.

Procedure

Task procedures were identical to those in Experiment 1, except that in Experiment 2, two of the four states (s_1 and s_2) shared the same optimal response a_1. The specific images and key presses were randomized across participants.

Statistical analysis

Statistical testing and errorbar visualization procedures were identical to those in Experiment 1. We report all pairwise comparisons, their effect sizes, and the 95% CIs of the effect sizes in Supplementary Tables 3 to 4. LME modeling procedures were identical to those in Experiment 1.

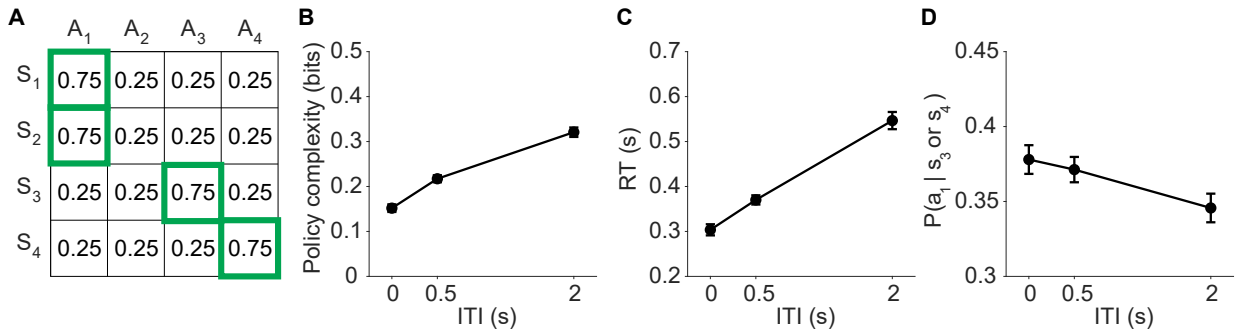
Results

Our findings were largely identical to what we found in Experiment 1. Participants achieved near-maximal trial-averaged reward as a function of policy complexity (Supplementary Figure 3B). Policy complexity increased as a function of ITI, $t(197) = -9.74$, $p < 10^{-18}$ (Figure 5B). RTs similarly slowed as a function of ITI, $t(197) = -7.95$, $p < 10^{-13}$ (Figure 5C). Action stochasticity ($t(197) = 1.95$, $p = 0.0264$), perseveration ($t(197) = 7.61$, $p < 10^{-12}$), and time-averaged reward ($t(197) = 34.8$, $p < 10^{-85}$) each decreased as a function of ITI. However, the 95% CI for the effect size of action stochasticity included 0 (Cohen's $d = 0.123$, 95% CI $[-0.002, 0.248]$).

According to the policy compression framework, participants should exploit the fact that states s_1 and s_2 share the same optimal action a_1 (Figure 5A), increasingly choosing a_1 as ITI decreases, because this favors lower policy complexity. To gain an intuition, in the extreme case where policy complexity equals to zero, participants ignore the stimuli entirely and they should always pick action a_1 since this

Figure 5

Experiment 2 behavioral results. ***(A)*** *Reward probability for each state-action pair.* ***(B-D)*** *Mean±SEM of participant policy complexity (B), average RT (C), and mean probability of choosing action a_1 in states s_3 and s_4 (D), across ITI conditions.*



maximizes reward. We looked at the policies for stimuli s_3 and s_4 to identify the effect of the marginal action distribution. For stimuli s_3 and s_4, under high policy complexity, a_1 should be chosen infrequently since this 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. Consistent with our prediction, the probability of choosing a_1 decreased monotonically as a function of ITI, $t(197) = 1.88$, $p = 0.0305$ (Figure 5D). However, the effect size was fairly small, with the 95% CI including 0 (Cohen's $d = 0.124$, 95% CI $[-0.006, 0.256]$). One limitation of this preregistered analysis is it only compares the most extreme datapoints, at ITI = 0s and 2s. We therefore performed a posthoc LME regression analysis utilizing data from all 3 conditions. We included intercept and ITI condition as fixed effects and independent random effects. The resulting fixed effect for ITI was significant (coefficient estimate -0.0164 ± 0.00766, $t(592) = -2.14$, $p = 0.0330$, random effects $SD = 0.00134$), providing additional evidence for a monotonically decreasing probability of choosing a_1 with increased ITI.

We further replicated Experiment 1's findings related to memory costs. Participants ranked the ITI = 0s condition the easiest and the ITI = 2s condition the hardest, $t(197) = -12.0$, $p < 10^{-24}$, and this trend held at the single-participant level, $t(197) = 23.3$, $p < 10^{-57}$, Cohen's $d = 1.66$, 95% CI $[1.44, 1.87]$. The LME had significant effects for the intercept (fixed effects 0.214 ± 0.0151, $t(592) = 14.2$, $p < 10^{-38}$; random effects $SD = 0.0336$) and policy complexity (fixed effects 1.02 ± 0.0115, $t(592) = 8.89$, $p < 10^{-17}$; random effects $SD = 0.617$). We used the same procedure as Experiment 1 to calculate participant-specific optimal policy complexity, and we again found a leftward bias in the difference between empirical and optimal policy complexity for the ITI = 2s condition, $z = -6.21$, $p < 10^{-9}$, Cliff's $\delta = -0.444$, 95% CI

$[-0.546, -0.343]$ (Supplementary Figure 3F). We did not find a leftward bias for the ITI = 0s and 0.5s conditions (ITI = 0s, $z = 11.7$, $p = 1.00$, Cliff's $\delta = 0.829$, 95% CI $[0.756, 0.902]$; ITI = 0.5s, $z = 7.57$, $p = 1.00$, Cliff's $\delta = 0.427$, 95% CI $[0.332, 0.523]$; Supplementary Figure 3F), likely because the optimal policy complexity of most participants for these ITI conditions (ITI = 0s, mean±SEM = 0.00345 ± 0.00144 bits; ITI = 0.5s, 0.0616 ± 0.00756 bits) was already very close to 0—the lowest possible policy complexity level. In contrast, in Experiment 1, the optimal policy complexity level for ITI = 0s was 0.227 ± 0.0153 bits, higher than the corresponding optimal policy complexity for Experiment 2. We conducted the same subgroup partition as in Experiment 1 and found that the "low-complexity" subgroup still significantly modulated policy complexity and RT as a function of ITI (see Supplementary Figure 2).

Discussion

In Experiment 2, we replicated findings from Experiment 1, which demonstrates their robustness. Most importantly, the data support an important prediction of the framework: subjects exploit environmental regularities by incorporating them into the marginal action distribution, $P(a)$, and that this effect is most pronounced at low policy complexity. This finding contributes to recent work demonstrating that participants exploit environmental regularities at low policy complexity (Lai & Gershman, 2024). We did, however, observe a systematic deviation from the predictions of the theory. Quantitatively, policy compression predicts that as policy complexity approaches 0, agents should deterministically choose the shared action a_1 for states s_3 and s_4; our participants systematically had higher entropy policies (i.e., their policies were more stochastic), leading to smaller effect sizes than predicted by the framework. Overall, this systematic deviation merits future investigation.

Experiment 3

We have so far demonstrated that humans modulate policy complexity in response to ITI manipulations to maximize time-averaged reward. However, manipulating ITI is not the only task condition that should modulate policy complexity and RTs. The relationship between policy complexity and decoding speed predicts set-size effects, a seemingly disparate domain: response times should grow as a function of set size because larger sets require higher policy complexity. As the set size grows and more stimuli must be encoded by the policy, the optimal policy complexity also grows to maximize time-averaged reward (Figure 1G,H).

In Experiment 3, we manipulated stimulus set size while keeping ITI fixed. We made the following predictions related to time costs: larger set sizes should be associated with 1) higher policy complexity, 2) slower RTs, 3) decreased perserveration, and 4) decreased time-averaged reward. We also made similar predictions related to memory costs: 5) greater set sizes should be associated with higher perceived

difficulty and 6) we should observe a systematic leftward bias in empirical policy complexity relative to optimal.

Materials and Methods

Participants

One hundred and one participants (54 female, 44 male, 2 non-binary, 1 prefer not to say) were recruited. All participants did not participate in either Experiment 1 or 2. We selected the sample size based on the lowest estimated effect size (Cohen's $d = 0.459$) among dependent variables of interest, according to analyses of a separate group of $N = 48$ pilot participants (data excluded from final analysis). All analyses were preregistered at https://aspredicted.org/ZSW_HFY. The inclusion criterion was identical to Experiments 1 and 2. A total of 99 participants (53 female, 43 male, 2 non-binary, 1 prefer not to say) met this inclusion criteria and were therefore included. Participants gave informed consent, and the Harvard University Committee on the Use of Human Subjects approved the experiment.

Procedure

The three test blocks had stimuli set sizes of 2, 4, and 6 stimuli respectively, and their order was randomized across participants. Each set size used unique images in order to make each set-size manipulation as independent as possible. The action set size was fixed at 6 across all set-size conditions. We used ITI = 2s for each block. Participants were informed of the ITI and set size of each block.

Each stimulus was associated with a unique optimal action. Stimuli were randomized and presented in counterbalanced runs of 8, 8, and 10 trials for set sizes 2, 4, and 6 respectively (each stimulus therefore appeared 4 times, 2 times, and 2 times respectively within each run).

For each set-size condition, participants first completed three training blocks with ITI = 0s, 0.5s, and 2s, similar to Experiments 1 and 2. We did this to encourage learning and minimize the length of training. To ensure similar learning across set-size conditions, we presented each stimulus 48 times during training (24 for ITI = 0s, 16 for ITI = 0.5s, and 8 for ITI = 2s) rather than training for a fixed time duration. Participants were told that the mapping from stimuli to actions remained fixed between training and test blocks. After completing the three training blocks, participants proceeded to the 3-minute test block of the same set-size condition. The structure, visual display, and duration of blocks were identical to Experiments 1 and 2.

After completing the whole experiment, participants ranked the perceived difficulty for each block. Participants additionally completed the Barratt Impulsiveness Scale, which we did not analyze for this manuscript.

Statistical analysis

Statistical testing and errorbar visualization procedures were identical to those in Experiment 1 and 2, except that tests were carried out between set-size conditions instead of ITI conditions. In the main text, we report comparisons between set size = 2 vs 6. We report all pairwise comparisons, their effect sizes, and the 95% CIs of the effect sizes in Supplementary Tables 5 to 6. LME modeling procedures were identical to those in Experiment 1 and 2.

Results

We found support for our predictions that humans are sensitive to time costs in response to set-size manipulations. Participants achieved near-maximal trial-averaged reward as a function of policy complexity (Figure 6A). Consistent with our predictions, policy complexity increased as a function of set size, $t(98) = -10.5$, $p < 10^{-17}$ (Figure 6B) and RTs slowed, $t(98) = -3.97$, $p < 10^{-4}$ (Figure 6C). Perseveration decreased as a function of set size, $t(98) = 4.00$, $p < 10^{-4}$ (Figure 6D). Finally, time-averaged reward decreased as a function of set size, $t(98) = 5.28$, $p < 10^{-6}$.

We additionally found support for our predictions regarding memory costs. Participants ranked the set size 2 condition as the easiest and the set size 6 condition as the hardest, $t(98) = -7.48$, $p < 10^{-10}$ (Figure 6E), and this trend held at the single-participant level, $t(98) = 17.8$, $p < 10^{-32}$, Cohen's $d = 1.77$, 95% CI $[1.46, 2.09]$. We fitted the LME and identified significant effects for the intercept (fixed effects 0.354 ± 0.0287, $t(295) = 12.4$, $p < 10^{-27}$; random effects $SD = 0.0381$) and policy complexity (fixed effects 0.568 ± 0.0977, $t(295) = 5.82$, $p < 10^{-7}$; random effects $SD = 0.568$). We used the same procedure as Experiment 1 to calculate participant-specific optimal policy complexity, and we again found a leftward bias in the difference between empirical and optimal policy complexity for all set-size conditions: set size 2, $z = -7.86$, $p < 10^{-14}$, Cliff's $\delta = -0.774$, 95% CI $[-0.865, -0.684]$; set size 4, $z = -7.71$, $p < 10^{-14}$, Cliff's $\delta = -0.759$, 95% CI $[-0.848, -0.670]$; set size 6, $z = -8.02$, $p < 10^{-15}$, Cliff's $\delta = -0.777$, 95% CI $[-0.867, -0.687]$ (Figure 6F).

Comparison with evidence-accumulation models

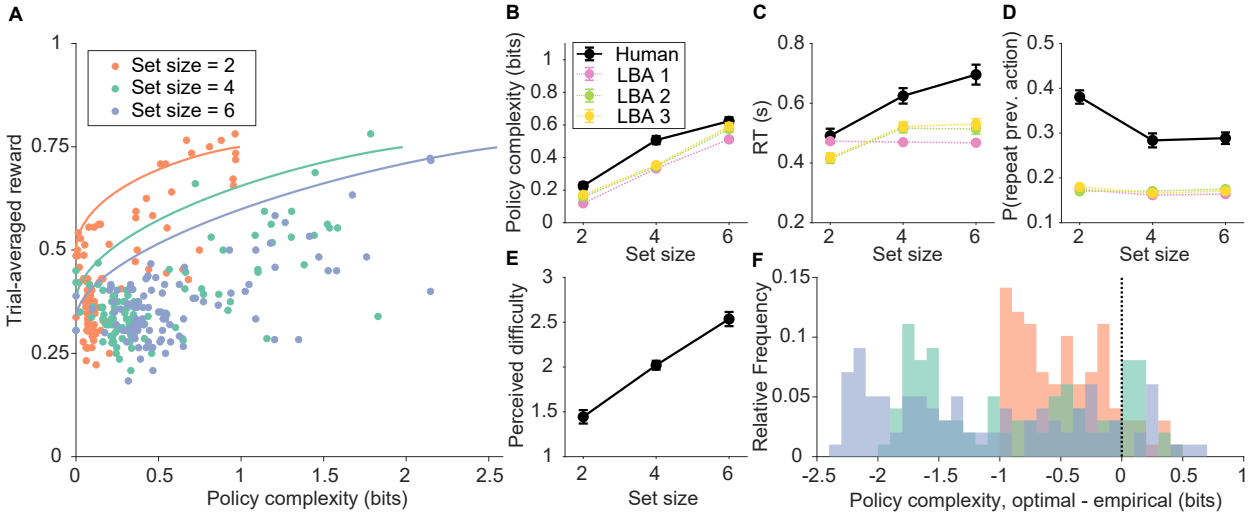
To check whether standard parameterizations of past models could already account for our experiment results, we sought to compare policy compression with other frameworks capable of generating choice and RT data in multi-alternative choice settings. We focused on evidence accumulation models as they have traditionally been employed in the domain. Typically applied to perceptual decisions, evidence accumulation models assume that observers integrate sequential noisy observations to a decision bound to make a choice, with the RT determined by the time it took to accumulate evidence (Forstmann et al., 2016). These models have been extended to value-based instrumental learning tasks by assuming agents

sequentially retrieve (noisy) Q-values from memory (Fontanesi et al., 2019; McDougle & Collins, 2021; Miletić et al., 2020; Pedersen et al., 2017; Tajima et al., 2019).

We focus specifically on the linear ballistic accumulator (LBA) class of models, as they are naturally well-suited for multi-alternative choice settings like our experiments (Brown & Heathcote, 2008; Donkin et al., 2009, 2011; McDougle & Collins, 2021). The simplest LBA assumes the following: each available action corresponds to one independent accumulator. On each trial, the start point of each accumulator is sampled uniformly between $[0, A]$. The drift rate k_i of each accumulator i is sampled from a Gaussian distribution $\mathcal{N}(v_i, s^2)$, where v_i is the mean drift rate of the accumulator and s^2 is the variance. As is standard, v_i takes one of two values: $v_i = v_{\text{correct}}$ for the correct action and $v_{\text{incorrect}}$ otherwise where $v_{\text{incorrect}} < v_{\text{correct}}$ (Donkin et al., 2011). We fixed $s = 0.1$ to ensure parameter identifiability (Donkin et al., 2009; McDougle & Collins, 2021). On each trial, each accumulator ballistically accumulates evidence and action i is taken when the first accumulator reaches a decision bound b. The trial's response time is

Figure 6

Experiment 3 behavioral and modeling results. ***(A)*** *Trial-averaged reward across set-size conditions (color). The solid line denotes the theoretical upper bound (reward-complexity frontier) for each set size. Note that some datapoints lie above the upper bound due to the stochastic nature of reward delivery.* ***(B-E)*** *Mean±SEM for empirical (black) policy complexity (B), RT (C), perserveration (D), and perceived difficulty (E), across set-size conditions. Using the three fitted LBA models (colors; in increasing number of free parameters), we simulated behavior for each task condition and computed the same behavioral statistics.* ***(F)*** *Difference in policy complexity between optimal (participant's LME-predicted optimal policy complexity) and empirical empirical complexity, for each set-size condition.*



the time taken for this first accululator to reach the bound plus some nondecision time t_0: $\text{RT} = t_0 + \frac{b-A}{k_i}$. The simplest LBA model contains 5 free parameters: $A, b, v_{\text{correct}}, v_{\text{incorrect}}, t_0$. This model formulation allows us to express the likelihood of observing a trial's choice and RT for a given set of parameter values. Therefore, the LBA can be fitted jointly on the choice and RT data.

We fit three variants of this LBA model to the Experiment 3 data, each with increasing degrees of freedom. Model 1 (LBA 1) is the LBA we have just described, with five free parameters shared across all three set-size conditions. Model 2 (LBA 2) fits $(v_{\text{correct}}, v_{\text{incorrect}})$ independently for each set size. Model 3 (LBA 3) fits all five parameters separately for each set size. After fitting these three models, we simulated choice and RT data over the same task conditions seen by participants. The LBA models each simulated data with increasing policy complexity as a function of set size, just like human participants (Figure 6B). However, they could not capture the increase in RT (Figure 6C) or decrease in perseveration as a function of set size (Figure 6D). As expected, these LBAs could generate a speed-accuracy trade-off in Experiment 1, but not perseveration (Supplementary Figure 4). We therefore reject this parameterization of LBAs since they cannot readily explain our findings.

Discussion

In Experiment 3, we have demonstrated the applicability of policy compression across a different form of task manipulation—stimulus set size, as opposed to ITI studied in Experiments 1 and 2. Importantly, both ITI and set-size manipulations induce human behavioral changes predicted by the framework. In addition, Experiment 3 shows that the framework is well-suited for a variety of value-based tasks, in which the environmental state does not feature multiple stimuli each favoring a different action. Such task setups induce behavioral patterns that existing LBA models could not easily accommodate, which highlights the generalizability of the policy compression framework. We do not exclude the possibility that other LBA parametrizations could explain our data; indeed, the policy compression framework offers a normative motivation for alternative parametrizations, as we discuss later.

General Discussion

Here, we developed a theoretical decision making framework that jointly considers both time and memory costs. Across three human instrumental learning tasks, we tested its predictions and validated its explanatory breadth in seemingly disparate domains. Overall, we found that humans are sensitive to the time cost of decoding policies when maximizing time-averaged reward.

The policy compression framework contributes to our understanding of value-based decision making in several ways. First, by considering both time and memory costs, the framework unifies several well-studied behavioral and cognitive variables that appear seemingly disparate. It links choice and RT

data for value-based decisions under a single coherent, normative framework, a longstanding goal of cognitive science, and separate from the approach of combining sequential sampling models with reinforcement learning models to generate both choices and RTs (Fontanesi et al., 2019; McDougle & Collins, 2021; Miletić et al., 2020; Pedersen et al., 2017; Tajima et al., 2019). This points to an area of future work for the policy compression framework, which does not naturally generate RT distributions. In this framework, policy complexity is the minimum average description length of the codewords used to decode actions, and therefore only maps onto average RT—a point statistic. In contrast, sequential sampling models are naturally suited for fitting and generating full RT distributions.

While descriptive models, such as LBAs, can generate a speed-accuracy trade-off (i.e., by lowering the decision bound, actions become faster but less accurate) and are capable of generating policies of higher complexity as a function of ITI and set sizes, we demonstrated here that a factorial combination of "vanilla" LBA parameters is insufficient to explain the host of predictions made by policy compression. For example, they were unable to account for perseverative effects. This is because the LBAs did not include a mechanism for "remembering" prior actions, and using action history to bias the tendency of future actions. One can imagine incorporating action history by specifying a prior over Q-values that favors previously chosen actions, by converting past action frequency (in units of probability) to Q-values (in units of reward). However, we did not include this or similar mechanisms in our analysis as such a decision is largely ad hoc in the absence of proper justification. The policy compression framework, however, can be interpreted as providing this justification for future process model development (e.g., normalized drift rates, adjusting accumulator start points as a function of action history) to better fit empirical behavior.

Policy compression offers normative insight into a broad set of seemingly disparate task manipulations—ITIs, environmental regularities, and stimulus set sizes. The framework made clear predictions that could readily be factored into the maximization of time-averaged reward (Equation 6). Previous normative frameworks are more limited in their explanatory breadth in this regard. For example, there is a normative account for how speed and accuracy should trade off as a function of ITI manipulations in drift diffusion models (Forstmann et al., 2016), but this framework does not readily generalize to tasks where there are more than two possible actions. The race model proposed by Tajima et al. (2019) accommodates multialternative settings, but their framework is limited to settings in which agents must choose one of multiple stimuli displayed simultaneously (i.e., each stimulus provides noisy evidence for its unique corresponding action). In this case, the number of actions is constrained to be the number of simultaneous stimuli. In our tasks where participants observed a single state and could make one of up to 6 actions, the normative insights provided by these past models—which would likely need to assume that a single stimulus elicits Q-value retrieval for all actions simultaneously—become less clear.

The policy compression framework, in contrast, provides a single explanation for both the speed-accuracy trade-off effects and set-size effects we observed: these arise due to the time cost of decoding policies.

One of the key features of policy compression is the inclusion of a state-independent term, $P^*(a)$, in modulating behavior, which we have argued provides a normative basis for perseveration (Gershman, 2020). Importantly, at low policy complexity, the influence of $P^*(a)$ is greatest, because this is when policies are highly state-independent. According to the framework, one role of $P^*(a)$ is to exploit environmental regularities when they exist and allow agents to maximize reward for no increase in the memory cost (since it does not require encoding state information). In Experiments 2 and 3, we validated this nontrivial prediction. Further, assuming participants estimate $P^*(a)$ on a trial-by-trial basis (e.g., via an iterative update process (Bari et al., 2024; Gershman & Lai, 2021; Lai & Gershman, 2024)), then there should be a greater tendency to repeat actions at low complexity. Across manipulations of ITIs, environmental regularities, and set sizes, this is what we observed.

A distinctive hypothesis of the policy compression framework is a *linear* relationship between RT and policy complexity. This was supported by LME fits across all experiments (see also Lai & Gershman, 2021, for additional evidence). In another experiment featuring 5 different set sizes, and therefore 5 independent measurements of policy complexity and RT for each participant, we again identified a linear relationship, suggesting that our identification of a linear fit was not a consequence of only having 3 datapoints per participant (Supplementary Figure 5). The success of the LME fits hints at the possibility that a discrete, bit-by-bit action decoding process, as opposed to sequential sampling of noisy evidence from memory, may better explain RTs in value-based decision making paradigms. Distinguishing this action decoding account from sequential sampling accounts appears to be a valuable area of inquiry. Generally speaking, it is unlikely that a Huffman code is exactly how the brain transmits information, since it likely requires unreasonably high precision. In support of this idea, the relationship between set size and RT in working memory tasks flattens for large set sizes (Longstreth, 1988; Seibel, 1963). However, given the relatively sparse state space of our experiments, we likely operated within the linear regime, explaining the quality of our RT-policy complexity fits.

In addition to time costs, our results also demonstrate that humans are sensitive to memory costs. Participants reported lower perceived difficulty in task conditions where they used low-complexity policies, across both ITI and set-size manipulations. This suggests that there is an intrinsic costliness to information gain. Consistent with this view, we observed a systematic leftward bias in empirical policy complexity, compared to optimal. These memory costs have previously been suggested to be a function of the amount of information required to update a prior distribution (i.e., via Kullback-Leibler divergence between prior and posterior distributions; Zenon et al., 2019). If we take the prior as $P^*(a)$ and the

posterior as $\pi(a|s)$, then this measure of divergence can explain the memory costs we have observed here. How this memory cost is instantiated in biological hardware remains a subject of future research.

Although the LME fits pointed to a linear RT to policy complexity relationship *within* participant, the same relationship *across* participants plateaued at high complexity (Supplementary Figure 3C). This could be due to a tendency for participants with steeper RT-policy complexity relationships to use low-complexity policies, whereas those with flatter relationships preferred to use high-complexity policies. Such behavior could be viewed as rational—if high-complexity policies cost too much time, it can make sense to employ low-complexity policies instead. This could be a valuable area of future work, since such relationships may explain a tendency towards impulsive behaviors.

Constraints on generality

We recruited participants from Amazon Mechanical Turk, using English as the instruction language. We acknowledge the possibility of behavioral differences induced by online versus in-person task presentation formats, as well as cross-cultural differences. However, we have no evidence suggesting that such variations would change our results significantly. Future work should be carried out to assess the robustness of our results to population group changes.

Conclusion

In summary, by considering both time and memory as concurrent resources consumed by decisions, we have developed a normative framework that specifies the relationship between habitual and goal-directed components of behavior, as well as their manifestation in choice and RT profiles. We have shown that the framework can predict speed-accuracy trade-offs and set-size effects, which demonstrates the potential of resource-rational analysis—interpreting decisions as optimizing a balance between reward and resource expenditure—in explaining human decisions. We believe future work that jointly considers multiple cognitive costs promises to have broad explanatory breadth of human behavior.

Acknowledgments

We are grateful to Jan Drugowitsch for helpful discussions. This work was supported by a Harvard Brain Science Initiative Bipolar Disorder Seed Grant, grant R25MH094612 from the National Institute of Mental Health, the National Science Foundation under Grant No. DRL-2024462, the Harvey Fellowship, and the Kempner Institute for the Study of Natural and Artificial Intelligence.

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.

Bari, B. A., & Gershman, S. J. (2023). Undermatching is a consequence of policy compression. *Journal of Neuroscience*, *43*(3), 447–457.

Bari, B. A., Krystal, A. D., Pizzagalli, D. A., & Gershman, S. J. (2024). Computationally-informed insights into anhedonia and treatment by κ-opioid receptor antagonism. *medRxiv*, 2024–04.

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21.

Bogacz, R., Hu, P. T., Holmes, P. J., & Cohen, J. D. (2010). Do humans produce the speed–accuracy trade-off that maximizes reward rate? *Quarterly journal of experimental psychology*, *63*(5), 863–891.

Bossaerts, P., Yadav, N., & Murawski, C. (2019). Uncertainty and computational complexity. *Philosophical Transactions of the Royal Society B*, *374*(1766), 20180138.

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

Callaway, F., Griffiths, T. L., Norman, K. A., & Zhang, Q. (2023). Optimal metacognitive control of memory recall. *Psychological Review*.

Collins, A. G. (2018). The tortoise and the hare: Interactions between reinforcement learning and working memory. *Journal of Cognitive Neuroscience*, *30*, 1422–1432.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*, 1024–1035.

Cousineau, D., et al. (2005). Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutorials in quantitative methods for psychology*, *1*(1), 42–45.

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*(2), 140–151.

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16*, 1129–1135.

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 (T. Behrens, Ed.). *eLife*, *4*.

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic bulletin & review*, *26*(4), 1099–1121.

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.

Gailliot, M. T., & Baumeister, R. F. (2007). The physiology of willpower: Linking blood glucose to self-control. *Personality and social psychology review*, *11*(4), 303–327.

Garrett, H. E. (1922). *A study of the relation of accuracy to speed* (Vol. 8). Columbia university.

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

Gershman, S. J., & Lai, L. (2021). The reward-complexity trade-off in schizophrenia. *Computational Psychiatry*, *5*(1).

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in neuroscience*, *8*, 86875.

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* (pp. 195–232, Vol. 74). Elsevier.

Lai, L., & Gershman, S. J. (2024). Human decision making balances reward maximization and policy compression. *PLOS Computational Biology*, *20*(4), 1–32. https://doi.org/10.1371/journal.pcbi.1012057

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Longstreth, L. E. (1988). Hick's law: Its limit is 3 bits. *Bulletin of the Psychonomic Society*, *26*(1), 8–10.

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.

Miletić, S., Boag, R. J., & Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia*, *136*, 107261.

Miller, G. A. (1956). The magic number seven plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*, 91–97.

Musslick, S., Dey, B., Ozcimder, K., Patwary, M. M. A., Willke, T. L., & Cohen, J. D. (2016). Parallel processing capability versus efficiency of representation in neural networks. *Network*, *8*(7).

Otto, A. R., & Daw, N. D. (2019). The opportunity cost of time modulates cognitive effort. *Neuropsychologia*, *123*, 92–105.

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

Pirrone, A., Stafford, T., & Marshall, J. A. (2014). When natural selection should optimize speed-accuracy trade-offs. *Frontiers in neuroscience*, *8*, 80741.

Seibel, R. (1963). Discrimination reaction time for a 1,023-alternative task. *Journal of experimental psychology*, *66*(3), 215.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, *40*, 99–124.

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: Empirical tests of theoretical predictions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1865.

Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019). Optimal policy for multi-alternative decisions. *Nature neuroscience*, *22*(9), 1503–1511.

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature communications*, *7*(1), 12400.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85.

Woodworth, R. S. (1899). Accuracy of voluntary movement. *The Psychological Review: Monograph Supplements*, *3*(3), i.

Wu, S., Éltető, N., Dasgupta, I., & Schulz, E. (2023). Chunking as a rational solution to the speed–accuracy trade-off in a serial reaction time task. *Scientific reports*, *13*(1), 7680.

Zenon, A., Solopchuk, O., & Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, *123*, 5–18.

Supplementary Materials

This supplementary materials file contains the full set of statistical test results and effect sizes, as well as more comprehensive supplementary figures.

Statistical Test Tables

Supplementary Table 1

Experiment 1 statistical test p-values. *Red colors denotes statistically nonsignificant tests under the* $p < 0.05$ *cutoff.*

	ITI = 0s vs 0.5s	**ITI = 0.5s vs 2s**	**ITI = 0s vs 2s**
Policy complexity (bits)	8.92×10^{-5}	8.22×10^{-5}	2.82×10^{-9}
Average RT (s)	1.49×10^{-7}	3.64×10^{-5}	2.15×10^{-11}
Action stochasticity (H(A|S); bits)	0.0528	1.78×10^{-4}	1.48×10^{-5}
Perserveration (P(repeat previous action))	3.66×10^{-5}	0.161	4.56×10^{-5}
Time-averaged reward (reward/s)	7.86×10^{-35}	2.09×10^{-49}	5.15×10^{-45}
Perceived difficulty	1.77×10^{-4}	1.10×10^{-5}	8.36×10^{-7}
Policy complexity (bits), low complexity subgroup	1.49×10^{-12}	0.0359	2.05×10^{-4}
Policy complexity (bits), high complexity subgroup	8.22×10^{-4}	3.76×10^{-4}	5.22×10^{-8}
Average RT (s), low complexity subgroup	1.28×10^{-6}	8.78×10^{-4}	2.13×10^{-9}
Average RT (s), high complexity subgroup	0.00140	0.00604	4.74×10^{-5}

Supplementary Table 2

Experiment 1 Cohen's d effect sizes and their 95% confidence intervals. *Red color denotes effect directions opposite to the framework's prediction, or CIs that include 0.*

	ITI = 0s vs 0.5s	**ITI = 0.5s vs 2s**	**ITI = 0s vs 2s**
Policy complexity (bits)	−0.244 [−0.377, −0.117]	−0.202 [−0.311, −0.0979]	−0.452 [−0.614, −0.303]
Average RT (s)	−0.302 [−0.425, −0.189]	−0.238 [−0.361, −0.122]	−0.550 [−0.726, −0.391]
Action stochasticity (H(A|S); bits)	0.129 [−0.0282, 0.291]	0.268 [0.122, 0.422]	0.402 [0.216, 0.620]
Perserveration (P(repeat previous action))	0.337 [0.172, 0.512]	0.0842 [−0.0839, 0.255]	0.406 [0.205, 0.620]
Time-averaged reward (reward/s)	2.34 [1.96, 2.79]	3.18 [2.72, 3.73]	3.71 [3.16, 4.37]
Perceived difficulty	−0.599 [−0.943, −0.273]	−0.673 [−1.00, −0.367]	−1.00 [−1.44, −0.599]
Policy complexity (bits), low complexity subgroup	−0.203 [−0.268, −0.148]	−0.267 [−0.573, 0.0258]	−0.549 [−0.877, −0.249]
Policy complexity (bits), high complexity subgroup	−0.423 [−0.715, −0.164]	−0.415 [−0.682, −0.178]	−0.851 [−1.21, −0.549]
Average RT (s), low complexity subgroup	−0.398 [−0.580, −0.237]	−0.279 [−0.466, −0.106]	−0.690 [−0.947, −0.467]
Average RT (s), high complexity subgroup	−0.334 [−0.575, −0.117]	−0.322 [−0.596, −0.0712]	−0.664 [−1.04, −0.339]

Supplementary Table 3

Experiment 2 statistical test p-values. *Red color denotes statistically nonsignificant tests under the* $p < 0.05$ *cutoff.*

	ITI = 0s vs 0.5s	**ITI = 0.5s vs 2s**	**ITI = 0s vs 2s**
Policy complexity (bits)	1.99×10^{-7}	1.11×10^{-9}	7.78×10^{-19}
Average RT (s)	9.43×10^{-8}	1.09×10^{-9}	7.22×10^{-14}
Action stochasticity (H(A|S); bits)	0.0305	0.374	0.0264
Perserveration (P(repeat previous action))	1.23×10^{-4}	3.74×10^{-9}	5.43×10^{-13}
Time-averaged reward (reward/s)	3.70×10^{-65}	1.15×10^{-100}	1.77×10^{-86}
Perceived difficulty	1.17×10^{-12}	1.02×10^{-20}	1.55×10^{-25}
$P(a_1 \mid s_3 \text{ or } s_4)$	0.331	0.0478	0.0305
Policy complexity (bits), low complexity subgroup	3.17×10^{-13}	6.47×10^{-21}	1.13×10^{-26}
Policy complexity (bits), high complexity subgroup	4.18×10^{-4}	0.763	0.0150
Average RT (s), low complexity subgroup	1.17×10^{-7}	2.21×10^{-9}	1.08×10^{-12}
Average RT (s), high complexity subgroup	0.0376	0.0671	0.00762

Supplementary Table 4

Experiment 2 Cohen's d effect sizes and their 95% confidence intervals. *Red color denotes effect directions opposite to the framework's prediction, or CIs that include 0.*

	ITI = 0s vs 0.5s	**ITI = 0.5s vs 2s**	**ITI = 0s vs 2s**
Policy complexity (bits)	-0.190 $[-0.265, -0.117]$	-0.288 $[-0.385, -0.195]$	-0.497 $[-0.613, -0.388]$
Average RT (s)	-0.219 $[-0.303, -0.137]$	-0.444 $[-0.594, -0.300]$	-0.603 $[-0.769, -0.446]$
Action stochasticity (H(A|S); bits)	0.0983 $[-0.00472, 0.203]$	0.0190 $[-0.0982, 0.136]$	0.123 $[-0.00168, 0.248]$
Perserveration (P(repeat previous action))	0.160 $[0.0750, 0.248]$	0.285 $[0.190, 0.385]$	0.456 $[0.332, 0.586]$
Time-averaged reward (reward/s)	2.19 $[1.92, 2.47]$	3.68 $[3.30, 4.11]$	3.46 $[3.08, 3.87]$
Perceived difficulty	-0.880 $[-1.13, -0.638]$	-1.10 $[-1.35, -0.873]$	-1.62 $[-1.95, -1.32]$
$P(a_1 \mid s_3 \text{ or } s_4)$	0.0242 $[-0.0850, 0.134]$	0.0961 $[-0.0172, 0.211]$	0.124 $[-0.00600, 0.256]$
Policy complexity (bits), low complexity subgroup	-0.701 $[-0.901, -0.514]$	-1.04 $[-1.28, -0.829]$	-1.28 $[-1.53, -1.45]$
Policy complexity (bits), high complexity subgroup	-0.577 $[-0.960, -0.245]$	0.135 $[-0.250, 0.532]$	-0.414 $[-0.826, -0.0385]$
Average RT (s), low complexity subgroup	-0.365 $[-0.508, -0.228]$	-0.575 $[-0.775, -0.386]$	-0.750 $[-0.968, -0.544]$
Average RT (s), high complexity subgroup	-0.242 $[-0.534, 0.0281]$	-0.117 $[-0.283, 0.0391]$	-0.337 $[-0.639, -0.0652]$

Supplementary Table 5

Experiment 3 statistical test p-values. *Red color denotes statistically nonsignificant tests under the* $p < 0.05$ *cutoff.*

	Set size 2 vs 4	**Set size 4 vs 6**	**Set size 2 vs 6**
Policy complexity (bits)	6.64×10^{-11}	0.00552	4.66×10^{-18}
Average RT (s)	2.02×10^{-4}	0.0991	6.82×10^{-5}
Perserveration (P(repeat previous action))	3.54×10^{-4}	0.580	6.14×10^{-5}
Time-averaged reward (reward/s)	1.22×10^{-6}	0.0585	3.92×10^{-7}
Perceived difficulty	4.39×10^{-8}	3.66×10^{-6}	1.58×10^{-11}
Policy complexity (bits), low complexity subgroup	1.15×10^{-6}	0.476	1.09×10^{-10}
Policy complexity (bits), high complexity subgroup	6.71×10^{-7}	1.19×10^{-4}	2.71×10^{-10}
Average RT (s), low complexity subgroup	0.105	0.918	0.461
Average RT (s), high complexity subgroup	2.27×10^{-4}	0.0334	6.41×10^{-6}

Supplementary Table 6

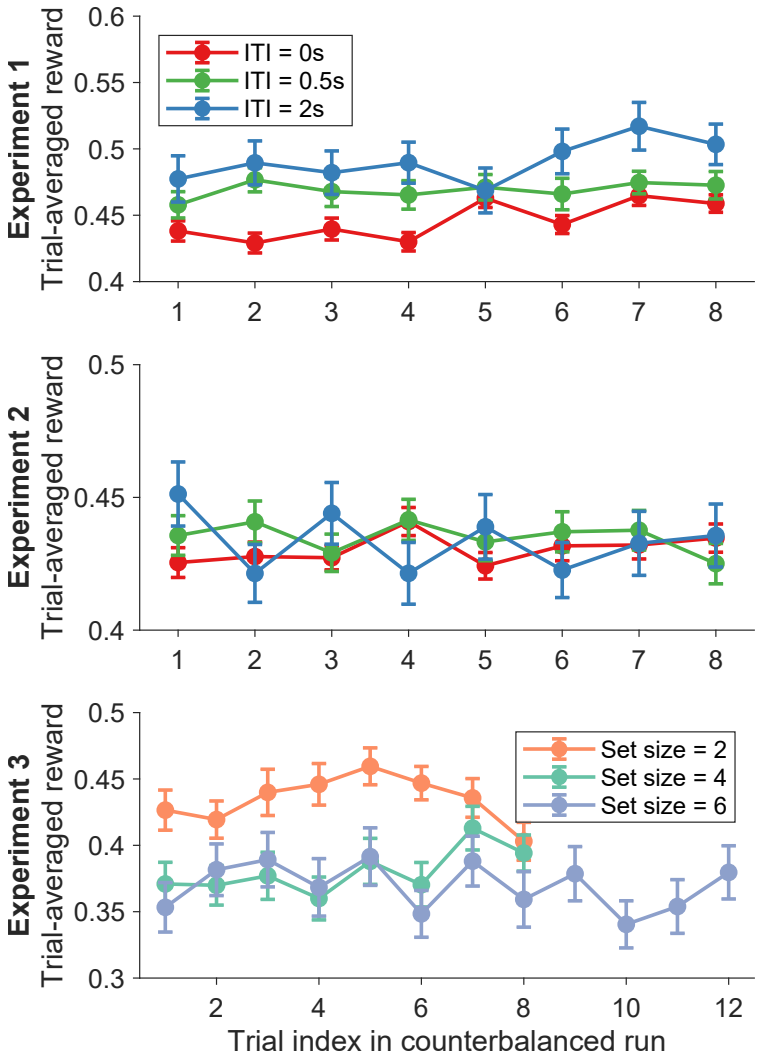
Experiment 3 Cohen's d effect sizes and their 95% confidence intervals. *Red color denotes effect directions opposite to the framework's prediction, or CIs that include 0.*

	Set size 2 vs 4	**Set size 4 vs 6**	**Set size 2 vs 6**
Policy complexity (bits)	−0.757 [−1.00, −0.531]	−0.253 [−0.456, −0.0582]	−1.05 [−1.31, −0.813]
Average RT (s)	−0.383 [−0.604, −0.172]	−0.148 [−0.378, 0.0790]	−0.487 [−0.748, −0.239]
Perserveration (P(repeat previous action))	0.396 [0.169, 0.635]	−0.0202 [−0.220, 0.179]	0.390 [0.193, 0.598]
Time-averaged reward (reward/s)	0.606 [0.358, 0.871]	0.176 [−0.0454, 0.402]	0.754 [0.460, 1.07]
Perceived difficulty	−0.903 [−1.25, −0.579]	−0.778 [−1.14, −0.443]	−1.413 [−1.86, −1.01]
Policy complexity (bits), low complexity subgroup	−0.783 [−1.15, −0.456]	−0.00823 [−0.286, 0.269]	−1.23 [−1.65, −0.860]
Policy complexity (bits), high complexity subgroup	−0.731 [−1.07, −0.445]	−0.581 [−0.928, −0.278]	−1.20 [−1.65, −0.834]
Average RT (s), low complexity subgroup	−0.175 [−0.461, 0.103]	0.164 [−0.0695, 0.4046]	−0.0146 [−0.321, 0.291]
Average RT (s), high complexity subgroup	−0.718 [−1.17, −0.325]	−0.462 [−0.997, 0.0372]	−1.15 [−1.74, −0.650]

Supplementary Figures

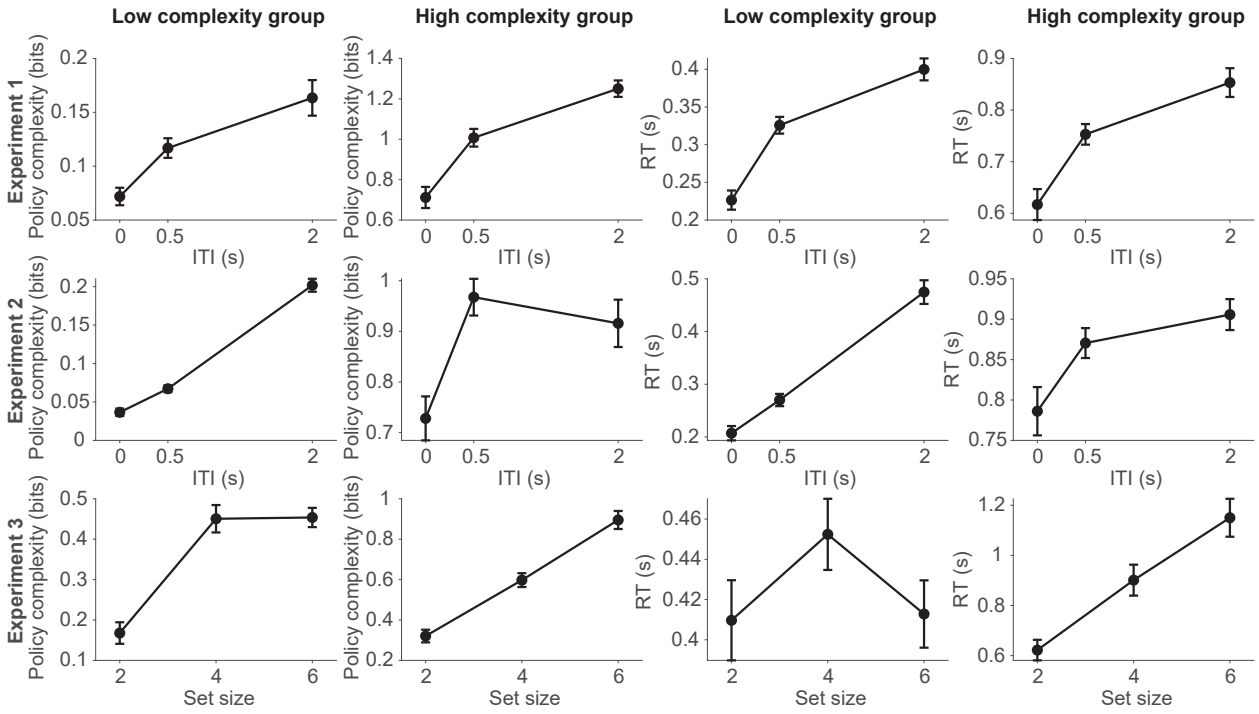
Supplementary Figure 1

Trial-averaged reward as a function of trials within each run. *Mean±SEM errorbars are scaled for within-participant visualization (Cousineau et al., 2005).*



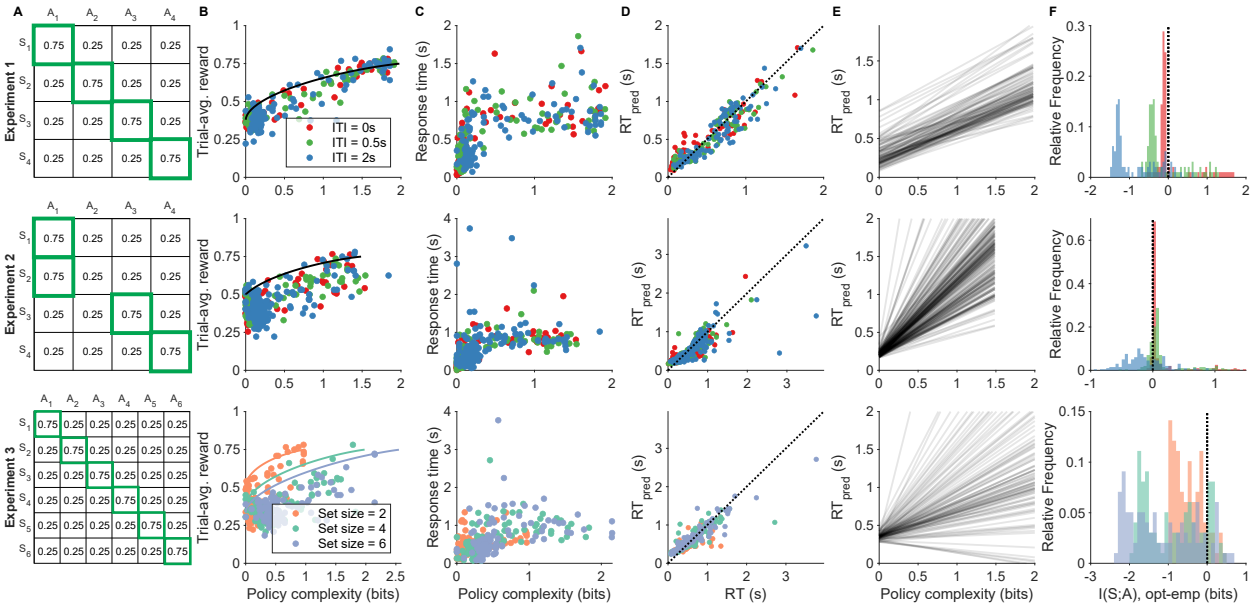
Supplementary Figure 2

Subgroup analysis. *For each Experiment, we divided participants into a "low complexity" and "high complexity" subgroup.* ***Row 1:*** *Mean±SEM of policy complexity and average RT, for Experiment 1 participants in the low and high-complexity subgroups.* ***Row 2:*** *Experiment 2 results.* ***Row 3:*** *Experiment 3 results.*



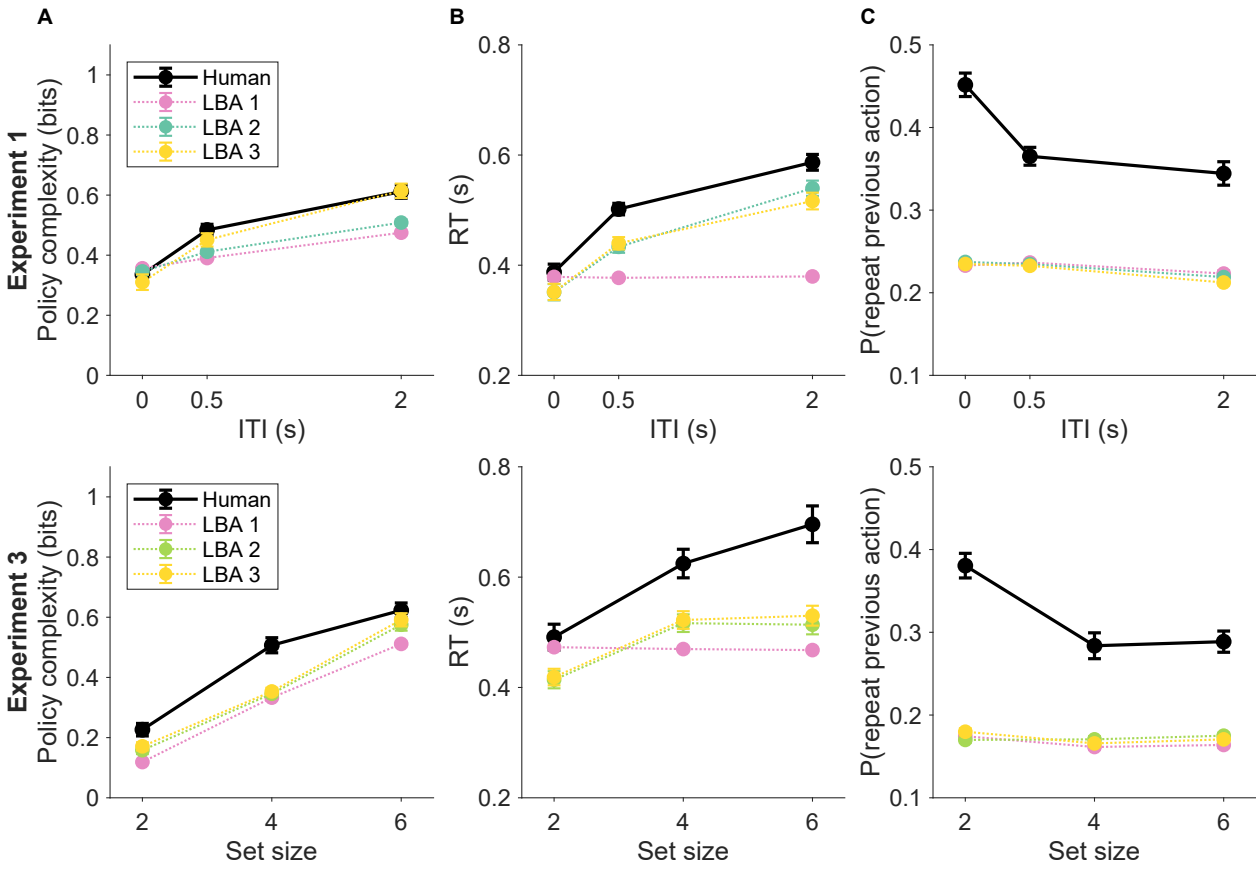
Supplementary Figure 3

Reward-complexity relationships and LME model fits for all experiments. Row 1: Experiment 1 results. ***(A)*** *Reward structure.* ***(B)*** *Reward-complexity frontier, overlaid with empirical policy complexity and trial-averaged reward (colors denote task condition).* ***(C)*** *Empirical average RT to policy complexity relationship.* ***(D)*** *LME model-predicted RT and empirical RT, for each participant in each condition.* ***(E)*** *LME model-fitted linear relationship between RT and policy complexity for each participant.* ***(F)*** *Histogram of policy complexity difference between a participant's LME-predicted optimal policy complexity and that participant's empirical policy complexity. Panels A, B, D, E, F are replicated from the main Figures.* ***Row 2: Experiment 2 results.*** *Panels A and F are replicated from the main Figures.* ***Row 3: Experiment 3 results.*** *Panels B and F are replicated from the main figures.*



Supplementary Figure 4

LBA model predictions. Row 1: Experiment 1 predictions. (A-C) *Empirical (black) and LBA-predicted (color) mean±SEM policy complexity (A), RT (B), and perseveration (C) of participants, as a function of ITI.* ***Row 2: Experiment 3 predictions.*** *These 3 panels are replicates of Figure 6B-D to facilitate comparison. Note that the definition of LBA 2 is different across Experiments 1 and 3 (denoted by a different color). In Experiment 1, LBA 2 allows the bound height b to vary across ITI conditions. in Experiment 3, LBA 2 allows the mean drift rate ($v_{correct}, v_{incorrect}$) to differ across set sizes.*



Supplementary Figure 5

Experiment 3 pilot behavioral and modeling results. ***(A,B)*** *Mean±SEM of policy complexity (A) and RT (B) across set-size conditions.* ***(C)*** *LME model-predicted RT and empirical RT for each participant in each set-size condition (color).* ***(D-F)*** *LME model-predicted linear relationship between RT and policy complexity (black line) for three example participants, overlaid with empirical policy complexity and RT (mean±SEM across trials in each set-size condition, hence not scaled for within-participant visualization).*