Systems/Circuits

Policy Complexity Suppresses Dopamine Responses

Samuel J. Gershman[1] and Armin Lak[2]
[1]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138 and [2]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom

Limits on information processing capacity impose limits on task performance. We show that male and female mice achieve performance on a perceptual decision task that is near-optimal given their capacity limits, as measured by policy complexity (the mutual information between states and actions). This behavioral profile could be achieved by reinforcement learning with a penalty on high complexity policies, realized through modulation of dopaminergic learning signals. In support of this hypothesis, we find that policy complexity suppresses midbrain dopamine responses to reward outcomes. Furthermore, neural and behavioral reward sensitivity were positively correlated across sessions. Our results suggest that policy compression shapes basic mechanisms of reinforcement learning in the brain.

Key words: **dopamine; information theory; policy compression; reward prediction error**

Significance Statement

Decision-making relies on memory to store information about which actions to produce in which situations. This memory has limited capacity, which means that some information will be lost. The signatures of this information loss can be found in patterns of behavioral bias and randomness. However, relatively little is known about the neural mechanisms which ensure that actions achieve the highest possible reward given the limited capacity of decision memory. In this paper, we show that the neuromodulator dopamine is sensitive to the costs of memory, as predicted by a computational model of capacity-limited learning.

Introduction

Task performance is bounded by sensory and memory bottlenecks that limit the flow of information from task states to actions (Tishby and Polani, 2010; Gershman, 2020; Lai and Gershman, 2021). This implies that there is not a single performance optimum, but rather a spectrum of optima indexed by information capacity. This idea can be formalized by viewing an agent's policy (the probabilistic mapping from states to actions) as a communication channel characterized by the mutual information between states and actions, also known as *policy complexity* (Fig. 1). The channel's capacity is an upper bound on policy complexity, which in turn determines an upper bound on achievable task performance.

Agents should *compress* their policies to stay within the capacity limit. Optimally compressed policies have several signatures: they are stochastic, biased towards high frequency actions, and sensitive to the distribution/number of states. These signatures have been documented experimentally in humans (Lai and Gershman, 2024), and have been argued to account for a wide range of well-established behavioral phenomena, such as perseveration (Gershman, 2020) and undermatching (Bari and Gershman, 2023).

Despite the abundance of behavioral evidence for policy compression, we still do not understand how it is implemented neurally. One possibility, motivated by reinforcement learning models of policy compression (Lai and Gershman, 2021, 2024), is that the reward prediction error signals used for policy updating register a policy complexity penalty, thereby driving policies towards a balance between reward maximization and policy compression. Specifically, the error δ should take the following form:

$$\delta = r - V - \log \frac{P(a \mid s)}{P(a)}, \tag{1}$$

where r is the reward outcome, V is expected reward, and the last term is the policy cost (the probability of choosing action a in state s relative to the marginal probability of choosing a across all states), whose expectation is equal to policy complexity, $I(s; a)$. Note that since policy complexity is the mutual information between two variables, it is always non-negative, though on individual trials the policy cost can be negative.

Received Sept. 15, 2024; revised Dec. 17, 2024; accepted Dec. 23, 2024.
Author contributions: S.J.G. analyzed data; S.J.G. wrote the first draft of the paper; S.J.G. and A.L. edited the paper; A.L. designed research; A.L. performed research.
Shuze Liu and Bilal Bari provided helpful comments on an earlier draft. SG is supported by NIH grant U19 NS113201-01 and Air Force Office of Scientific Research grant FA9550-20-1-0413. AL is supported by grant 213465 from the Wellcome Trust.
The authors declare no competing financial interests.
Correspondence should be addressed to Samuel J. Gershman at gershman@fas.harvard.edu.
https://doi.org/10.1523/JNEUROSCI.1756-24.2024

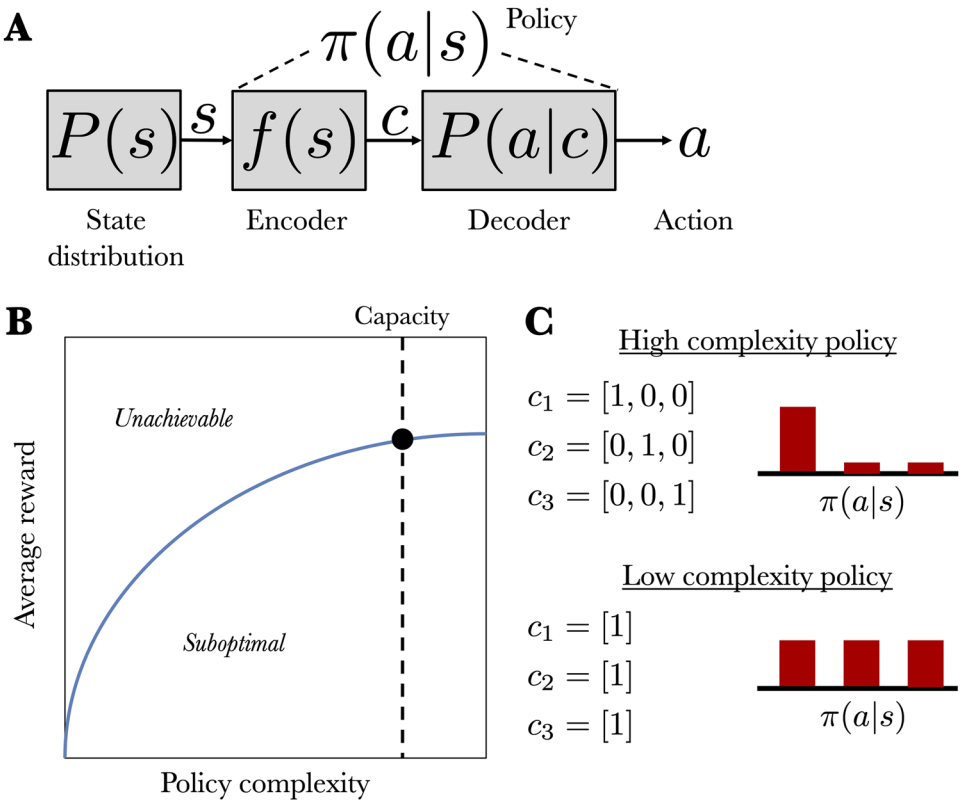

Figure 1. Policy compression framework. ***A***, State s is sampled from the state distribution and then compressed by an encoder into codeword c. At the time of action selection, the codeword is probabilistically decoded into an action a. The complete mapping from states to actions is the policy, $\pi(a \mid s)$. ***B***, The blue curve shows the optimal average reward achievable for each level of policy complexity (the mutual information between states and actions). A hypothetical capacity limit for an agent is shown as the dashed line; its intersection with the blue curve represents that agent's maximum achievable average reward. All points above the blue line are unachievable, and all points below it are suboptimal. ***C***, Two example policies, distinguished by their complexity. The high complexity policy has a capacity of 3 bits and yields a low entropy distribution over actions. In contrast, the low complexity policy has a capacity of 1 bit and yields a high entropy distribution over actions.

Since phasic dopamine signals classically conform to a reward prediction error signal (Schultz et al., 1997; Eshel et al., 2015; Gershman et al., 2024), we hypothesize that they will be suppressed by policy complexity. We tested this hypothesis using dopamine neuron recordings from mice during a perceptual decision task (Lak et al., 2020). The different stimuli in the task can be treated as distinct states (9 in total), allowing us to examine whether mice exhibit behavioral and neural signatures of policy compression.

Materials and Methods

Experimental procedure. The data analyzed in this paper were originally reported in Lak et al. (2020). We briefly summarize the data collection methods, referring readers to that paper for further details. The data came from 5 mice of either sex (55 sessions total) performing a perceptual decision task (Fig. 2) while the activity of dopamine neurons in the ventral tegmental area were monitored using fiber photometry of GCaMP signals. The fiber photometry used multiple excitation wavelengths (465 and 405 nm) modulated at different carrier frequencies (214 and 530 Hz) to allow for ratiometric measurements.

On each trial, mice were presented with a sinusoidal grating on either the left or right side of the monitor, and had to report the side using a wheel following an auditory Go cue (Fig. 2*A*,*B*). Task difficulty was controlled by the contrast of the grating. In addition, the reward magnitude for correct actions was asymmetric across blocks of trials (Fig. 2*C*).

Neural data analysis. Light collection, filtering, and demodulation were performed as previously described (Lerner et al., 2015) using the Doric photometry setup and Doric Neuroscience Studio Software (Doric Lenses Inc.). For each behavioral session, least-squares linear fit was applied to the 405 nm control signal, and the $\Delta F/F$ time series was then calculated as ((465 nm signal–fitted 405 nm signal)/fitted 405 nm signal).

Our analysis focused on the dopamine response at the time of reward feedback. Due to the relatively slow dynamics of the calcium signal, we averaged the signal between 300 and 800 ms following the outcome delivery, which encompasses the peak response. We normalized the response by subtracting the calcium signal averaged over a 200 ms window centered on the beginning of the trial. To aggregate across animals and sessions, we z-scored the responses within each session. We then fit a linear regression model to the responses with 4 regressors: an intercept, policy cost, action value (the average reward for the chosen action conditional on the current stimulus), and the outcome (water amount). The regression model aggregated data across all animals (i.e., used a fixed-effects structure), due to the small sample size; similar results were obtained using a linear mixed-effects model with random slopes and intercepts grouped by animal.

For visualization of the policy cost effect, we calculated partial residuals (Larsen and McCleary, 1972), the differences between observed and predicted responses with the policy cost term removed; plotting this against policy cost isolates the cost regressor's contribution after adjusting for the other regressors.

Policy compression model. Policy cost is defined as $\log P(a \mid s)/P(a)$, where $P(a \mid s)$ is the conditional probability of action a given state s (here taken to be the stimulus), and $P(a)$ is the marginal probability of action a. The probability distributions were estimated separately for each session. Policy complexity is defined as the average policy cost within a session. An animal's capacity is an upper bound on policy

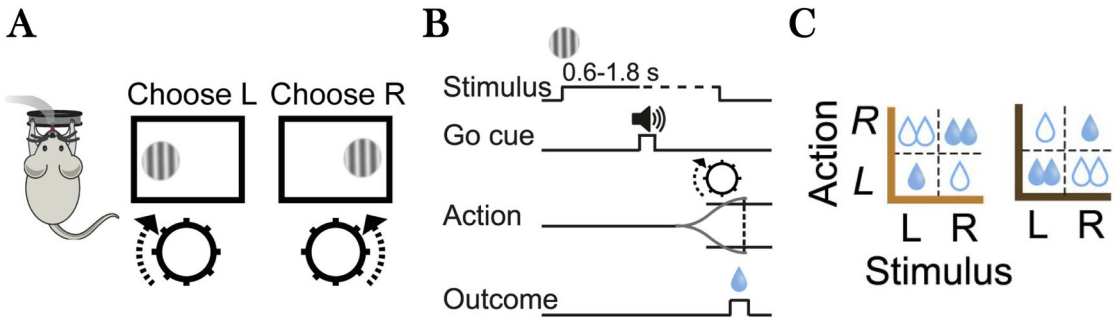

Figure 2. Task schematic. ***A***, Experimental interface. Mice reported the location (left or right) of a variable-contrast sinusoidal grating by turning a wheel. ***B***, Sequence of events on a single trial. Mice were required to await an auditory Go cue before responding, after which they received water reward for a correct action. ***C***, Reward structure. On different blocks, the correct action for one stimulus delivered twice as much reward as the correct action for the other stimulus (indicated by number of drops). The more-rewarded side switched in blocks of 50–350 trials. Unshaded drops indicate reward omission. Note that one and two unshaded drops both indicate reward omission; the number of unshaded drops indicates hypothetical expected reward.

complexity. The optimal capacity-limited policy is given by $P(a \mid s) \propto \exp[\beta Q(s, a) + \log P(a)]$, where $Q(s, a)$ is the average reward for taking action a in state s. The inverse temperature β is implicitly set based on the capacity limit and the task structure (see Lai and Gershman, 2021, for more details). We treated β as a free parameter, which we fit to the behavioral data using maximum likelihood estimation. To capture a small amount of state uncertainty, we smoothed the values across neighboring contrast levels.

Results

We first checked for behavioral signatures of policy compression. We used the Blahut-Arimoto algorithm to calculate the optimal reward-complexity frontier (Fig. 3*A*). This algorithm alternates between computing the optimal conditional distribution $P(a \mid s)$ and the optimal marginal distribution $P(a)$ until convergence. Each point on the frontier represents the maximum achievable average reward for a particular capacity limit. Points above the curve are unachievable, and points below the curve are suboptimal. We found that mouse behavior on this task was close to the optimal frontier, with a median deviation of 3.1%.

To gain an intuition for what different levels of policy complexity mean behaviorally, we can focus on two aspects of behavior: perseveration and stochasticity. Animals with low policy complexity are perseverative, choosing actions with high marginal probability. In other words, animals will tend to continue choosing an action that may no longer be relevant on the current trial. Animals with low policy complexity are also more stochastic in their action choices. Together, perseveration and stochasticity reduce the average reward for low-complexity policies.

We next tested whether the functional form of the optimal policy (see Materials and Methods) fit the choice data well. The optimal policy has only a single free parameter (the inverse temperature) which controls the balance between reward maximization and policy compression. When this parameter is small, the psychometric function should be shifted in the direction of high frequency actions, which we estimated using the session-specific bias (Fig. 3*B*). The optimal policy fit the data well, exhibiting a pronounced shift in the psychometric function depending on the session-specific bias, as measured by the average probability of choosing left or right within a session (Fig. 3*C*). Removing the bias term from the policy increased the Bayesian Information Criterion (ΔBIC = 378), thus supporting its inclusion.

Having established the behavioral plausibility of policy compression, we turned to an analysis of dopamine responses at the time of outcome. Linear regression with outcome, value, and policy cost regressors (see Materials and Methods) revealed significant effects for all three (Fig. 4*A*). Consistent with a reward prediction error, the outcome effect was positive ($t = 62.838$, $p < 0.0001$), and the value effect was negative ($t = -15.723$, $p < 0.0001$). Critically, the cost effect was negative ($t = -15.476$, $p < 0.0001$; Fig. 4*B*), consistent with the hypothesis that dopamine signals drive reinforcement learning away from high complexity policies. For comparison, a regression model predicting dopamine responses at the time of stimulus onset showed positive effects for both value ($t = 19.71$, $p < 0.0001$) and policy cost ($t = 6.29$, $p < 0.0001$), consistent with a cost-sensitive reward prediction error signal.

Removing the cost term from the outcome response model increased the Bayesian Information Criterion (ΔBIC = 228). Note that cost and outcome are correlated ($r = 0.53$), so the model comparison result is important for supporting our claim that the cost term is explaining substantial additional variance beyond its shared variance with the outcomes.

We conducted several other model comparisons to supplement this analysis. First, we confirmed that removing the outcome term (ΔBIC = 3503) and the action value term (ΔBIC = 235) produced inferior models, supporting the complete set of 3 terms we included in the original model specification. Second, we compared this model to one with a surprisal cost, $-\log P(a)$. The expectation of the surprisal cost is the action entropy. This model captures the proposal (e.g., Botvinick, 2007; Dreisbach and Fischer, 2012; Cavanagh et al., 2014) that high action entropy—a measure of response conflict—is costly (but see Rens et al., 2023), and this cost could be registered in the dopamine signal. We found that a model with surprisal instead of policy cost was disfavored (ΔBIC = 72), but that a model with both surprisal and policy cost was slightly favored (ΔBIC = −38). This suggests that response conflict may factor into dopamine responses beyond what is accounted for by policy cost. Finally, we compared the model against one with "stay" (action repetition) and "win-stay" (interaction between action repetition and last outcome) terms, based on the previous observation that dopamine responses show a win-stay effect (Lak et al., 2020). This model was also disfavored (ΔBIC = 237).

Finally, we tested whether the neural and behavioral results align with each other. According to some models (Lai and Gershman, 2021, 2024), the inverse temperature controls both the reward-compression tradeoff and the degree of reward sensitivity in the reward prediction error. This implies that the outcome coefficient in the neural regression should correlate

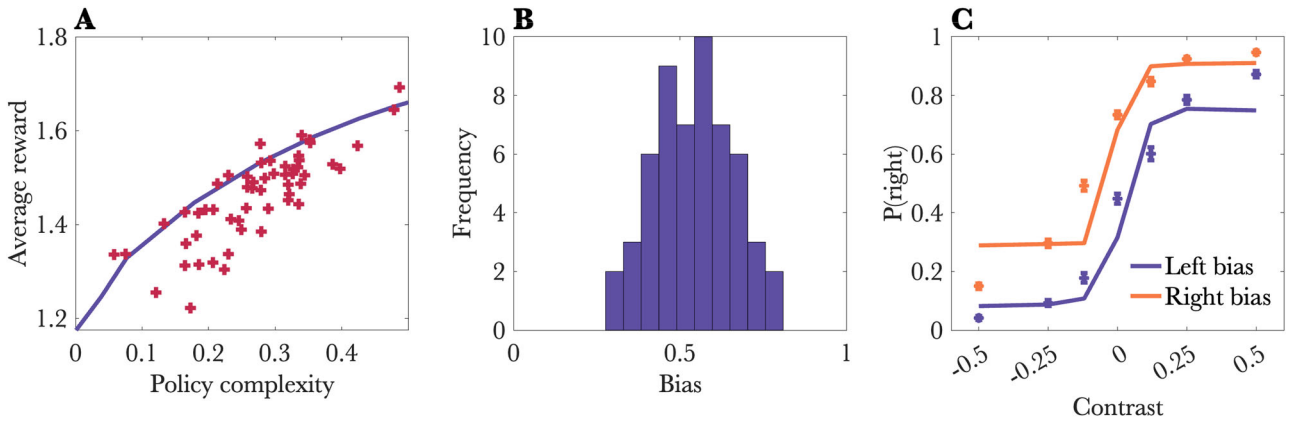

Figure 3. Behavioral results. ***A***, Task performance was close to the optimal reward-complexity frontier. Each cross represents a single session. Note that a few points are above the curve due to noise in estimation of policy complexity. ***B***, Histogram of bias (marginal probability of choosing "right") across sessions. ***C***, Probability of choosing "right" conditional on stimulus contrast and the session-specific bias. For visualization purposes, session-specific bias was dichotomized based on whether the P(right) in a session was less than 0.5 (a "Left bias" session) or greater than 0.5 (a "Right bias" session). Negative contrast values represent stimuli presented on the left; positive contrast values represent stimuli presented on the right. Solid lines show the model fit. All error bars show standard errors of the mean.

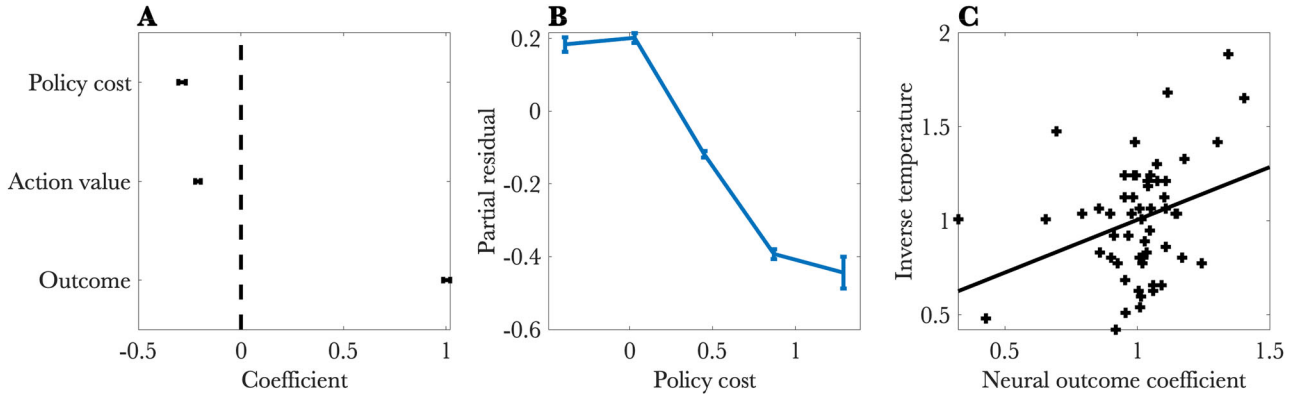

Figure 4. Neural results. ***A***, Regression coefficients for a linear model predicting the dopamine response at the time of reward feedback. ***B***, Partial residual plot for the policy cost regressor. ***C***, Behaviorally estimated inverse temperature plotted against the coefficient for the outcome regressor. Each cross represents a single session. All error bars show standard errors of the mean.

with the inverse temperature fit to behavior, consistent with the experimental data ($r = 0.33$, $p < 0.02$; Fig. 4*C*).

All of the neural analyses reported above were applied to baseline-corrected photometry signals, where the stimulus-evoked response was subtracted from the signal around the beginning of the trial (see Materials and Methods). This was done to mitigate lingering effects of transients at trial onset. Nonetheless, our results are robust to this baselining procedure, as shown in Figure 5, where we have not applied any baseline correction.

Discussion

Our study provides behavioral and neural evidence for policy compression in mice performing a perceptual decision task. Behaviorally, mice approximate the optimal reward-compression frontier, producing patterns of bias quantitatively consistent with the capacity-limited optimal policy. Neurally, dopamine responses to reward outcomes were suppressed by policy complexity, consistent with reinforcement learning models of policy compression (Lai and Gershman, 2021, 2024), and in contrast to models that locate capacity limits outside of the brain's error-driven reinforcement learning system (Collins et al., 2017). We note, however, that several studies have shown contributions of a capacity-limited working memory system to the reward expectations used for error-drive reinforcement learning (Collins, 2018; Collins and Frank, 2018), and hence the error signals themselves reflect the capacity limit. While such models are conceptually different from the one proposed here, they share the general idea that prediction error signals are shaped by capacity limits.

The idea that an information bottleneck constrains dopamine is buttressed by prior work. Schütt et al. (2024) showed that the population of dopamine neurons forms an efficient code for reward, with tuning curves that maximize information rate subject to a constraint on firing rate. At slower timescales, dopamine may also itself control bottlenecks by modulating sensitivity to sensory and reward signals (FitzGerald et al., 2015; Mikhael et al., 2021; Bari and Gershman, 2023), and by calibrating cognitive effort (Westbrook and Braver, 2016). In this study, we only examined the fast timescale component (phasic responses to reward).

We still lack a biologically plausible circuit model that synthesizes all of these observations. Our data suggest that any such circuit model should include projections from policy-sensitive regions to midbrain dopamine neurons. Searching for such projections will need to start with the identification of regions computing policy cost.

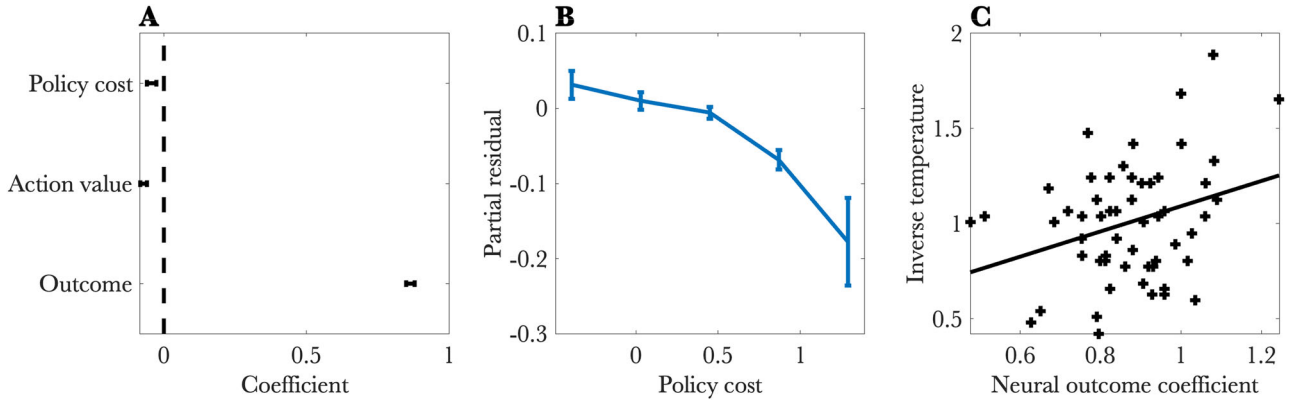

Figure 5. Neural results without baseline correction. Same analyses as shown in Figure 4, but without baseline correction. ***A***, regression coefficients for a linear model predicting the dopamine response at the time of reward feedback. ***B***, partial residual plot for the policy cost regressor. ***C***, behaviorally estimated inverse temperature plotted against the coefficient for the outcome regressor. Each cross represents a single session. All error bars show standard errors of the mean.

It has been proposed that the prefrontal cortex is organized into a hierarchy of cognitive control signals which guide action selection (Koechlin and Summerfield, 2007). Each of these control signals is derived from an information-theoretical analysis of the policy. In particular, Koechlin and Summerfield propose that premotor cortex tracks the action entropy $H[a] = \mathbb{E}[-\log P(a)]$, while lateral prefrontal cortex tracks the conditional entropy $H[a \mid s] = \mathbb{E}[-\log P(a \mid s)]$, with the state incorporating increasingly more information about stimuli, context, and past events along the anterior-posterior axis of lateral prefrontal cortex. The difference between the entropy and conditional entropy yields the policy complexity: $H[a] - H[a \mid s] = I(s; a)$. Thus, the policy cost (whose expectation is the policy complexity) could conceivably be computed based on differences in activity between lateral prefrontal and premotor areas.

Several lines of evidence indicate that prefrontal areas provide input to prediction error computation by dopamine neurons. Starkweather et al. (2018) showed that chemogenetic inactivation of medial prefrontal cortex eliminates the sensitivity of dopamine neuron activity to state uncertainty. Using the same perceptual decision task studied here, Lak et al. (2020) showed that medial prefrontal cortex encodes confidence-dependent action value; optogenetically suppressing this area altered learning putatively by increasing the prediction error. If the value signals conveyed by medial prefrontal cortex reflect a policy cost, then their transmission to dopamine neurons would enable the cost-sensitive error computation hypothesized here. One possibility is that medial prefrontal cortex integrates the control signals from the lateral prefrontal and premotor areas to compute cost-sensitive action values—a speculation broadly consistent with the role of medial prefrontal areas, particularly the anterior cingulate, in the integration of action costs and benefits (Walton et al., 2002; Rudebeck et al., 2006; Holroyd and McClure, 2015).

The data we analyzed in this paper came from a task that was not designed to specifically test the predictions of the policy compression framework. Recently, such tasks have been designed for studies of human decision making (Lai and Gershman, 2024). For example, Lai and Gershman (2024) studied how policy complexity varied with state and action probabilities, reward structure, and response deadlines. A promising direction for future research will be to develop these tasks for animal studies in conjunction with neural measurements of the underlying computational variables.

Code and Data Availability

All code and data for reproducing the analyses reported in this paper are available at https://github.com/sjgershm/dopamine-complexity.

References

Bari BA, Gershman SJ (2023) Undermatching is a consequence of policy compression. J Neurosci 43:447–457.
Botvinick MM (2007) Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. Cogn Affect Behav Neurosci 7:356–366.
Cavanagh JF, Masters SE, Bath K, Frank MJ (2014) Conflict acts as an implicit cost in reinforcement learning. Nat Commun 5:5394.
Collins AG (2018) The tortoise and the hare: interactions between reinforcement learning and working memory. J Cogn Neurosci 30:1422–1432.
Collins AG, Ciullo B, Frank MJ, Badre D (2017) Working memory load strengthens reward prediction errors. J Neurosci 37:4332–4342.
Collins AG, Frank MJ (2018) Within-and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proc Natl Acad Sci 115:2502–2507.
Dreisbach G, Fischer R (2012) Conflicts as aversive signals. Brain Cogn 78: 94–98.
Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N (2015) Arithmetic and local circuitry underlying dopamine prediction errors. Nature 525:243–246.
FitzGerald TH, Dolan RJ, Friston K (2015) Dopamine, reward learning, and active inference. Front Comput Neurosci 9:166836.
Gershman SJ (2020) Origin of perseveration in the trade-off between reward and complexity. Cognition 204:104394.
Gershman SJ, Assad JA, Datta SR, Linderman SW, Sabatini BL, Uchida N, Wilbrecht L (2024) Explaining dopamine through prediction errors and beyond. Nat Neurosci 27:1–11.
Holroyd C, McClure S (2015) Hierarchical control over effortful behavior by rodent medial frontal cortex: a computational model. Psychol Rev 122:54–83.
Koechlin E, Summerfield C (2007) An information theoretical approach to prefrontal executive function. Trends Cogn Sci 11:229–235.
Lai L, Gershman SJ (2021) Policy compression: an information bottleneck in action selection. In: *Psychology of learning and motivation*, Vol. 74, pp 195–232. Elsevier.
Lai L, Gershman SJ (2024) Human decision making balances reward maximization and policy compression. PLOS Comput Biol 20:e1012057.
Lak A, Okun M, Moss MM, Gurnani H, Farrell K, Wells MJ, Reddy CB, Kepecs A, Harris KD, Carandini M (2020) Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. Neuron 105: 700–711.
Larsen WA, McCleary SJ (1972) The use of partial residual plots in regression analysis. Technometrics 14:781–790.
Lerner TN, Shilyansky C, Davidson TJ, Evans KE, Beier KT, Zalocusky KA, Crow AK, Malenka RC, Luo L, Tomer R et al. (2015) Intact-brain analyses

reveal distinct information carried by SNc dopamine subcircuits. Cell 162: 635–647.
Mikhael JG, Lai L, Gershman SJ (2021) Rational inattention and tonic dopamine. PLoS Comput Biol 17:e1008659.
Rens N, Lancia GL, Eluchans M, Schwartenbeck P, Cunnington R, Pezzulo G (2023) Evidence for entropy maximisation in human free choice behaviour. Cognition 232:105328.
Rudebeck PH, Walton ME, Smyth AN, Bannerman DM, Rushworth MF (2006) Separate neural pathways process different decision costs. Nat Neurosci 9:1161–1168.
Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.
Schütt HH, Kim D, Ma WJ (2024) Reward prediction error neurons implement an efficient code for reward. Nat Neurosci 27:1–7.
Starkweather CK, Gershman SJ, Uchida N (2018) The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. Neuron 98:616–629.
Tishby N, Polani D (2010) Information theory of decisions and actions. In: *Perception-action cycle: models, architectures, and hardware*, pp 601–636. New York, NY: Springer.
Walton ME, Bannerman DM, Rushworth MF (2002) The role of rat medial frontal cortex in effort-based decision making. J Neurosci 22:10996–11003.
Westbrook A, Braver TS (2016) Dopamine does double duty in motivating cognitive effort. Neuron 89:695–710.