Computationally Rational Reinforcement Learning: Modeling the Influence of Policy and Representation Complexity

Zeming Fang (fangz5@rpi.edu) & Chris R. Sims (simsc3@rpi.edu)
Department of Cognitive Science, Rensselaer Polytechnic Institute
110 8th St, Troy, NY 12180 USA

Abstract

In recent years, several models of human reinforcement learning have been proposed that balance rationality (maximizing expected utility) against cognitive costs. Lai and Gershman (2021) proposed a model in which the cognitive cost was assumed to be the policy complexity, defined in terms of information theory as the mutual information between the sensory input and behavioral response. Here, using evidence from a published data set (Collins & Frank, 2012), we show that this model fails to account for the "set size effect" in learning: humans' learning efficiency decreases when the number of the presented stimuli increases. We therefore propose an alternative computational model, in which cognitive cost constitutes not only the policy complexity, but also the representation complexity—the amount of information conveyed from sensory inputs to internal representations. We quantify information processing cost as the combination of representation complexity and policy complexity. The resulting model captures the set size effect in an instrumental learning paradigm.

Keywords: Computationally rational; Reinforcement Learning; Information theory; Set size effect

Introduction

Human working memory is known to be capacity limited. A well-established consequence of this is the *set size effect*–namely, humans' memory performance systematically decreases as the number of items to be stored in memory increases (Ma, Husain, and Bays (2014)). Much existing work has sought to quantify what is meant by working memory capacity and explain the set size effect. One example is the work by Sims (2016), who formalized working memory capacity as a limited pool of information quantity that enables a cognitive function (e.g., store a stimulus) or process (e.g. making decisions). The information resource can be subdivided into portions, and more items to be stored implies less resource allocated to encode each item, resulting in lower recall precision per item. Up until now, however, most research on working memory has not examined how these limits might impact other cognitive systems.

Collins and Frank (2012) and Collins, Brown, Gold, Waltz, and Frank (2014) studied how working memory limits impact humans' reinforcement learning (RL). They reported an analog of the set size effect in an instrumental learning paradigm, and showed that a standard RL model (M^{RL} model in this article) cannot capture this phenomenon.

Gershman and Lai (2020) reexamined Collins et al. (2014), and proposed a computationally rational (Gershman, Horvitz, & Tenenbaum, 2015) account of humans' suboptimal learning performance. The mathematical framework they used is known as *rate distortion theory* (Berger, 1971). This framework provides the tools for predicting the highest achievable performance under a given information capacity constraint, and hence is directly applicable to explaining human learning performance under a limited pool of (information-theoretic) resource. They considered the capacity constraint as *policy complexity* (Tishby & Polani, 2011; Still & Precup, 2012; Lerch & Sims, 2018), which measures the rate of information extracted from states and transmitted to actions. They concluded that in general human participants optimized this reward-policy complexity trade-off, and humans' suboptimal performance can be understood as a compromise to limited policy complexity.

While interpreting humans' suboptimal performance, Gershman and Lai (2020) did not explicitly address how the set size effect emerges in human learning. This article seeks to fill this gap. Intuitively, one expects that when learning in larger set size conditions, the overall cognitive cost is higher, and hence humans will rationally trade task performance against rising cognitive costs. By analyzing the data set in Collins and Frank (2012), we show that policy complexity does not suffice as an explanation for human behavior: the policy complexity to reach optimal performance does not necessarily increase with the set size. This observation also violates humans' experience that the larger set size task is more difficult. This implies that the policy complexity is not sufficient for cognitive cost, other complementary constitutions are needed. We then considered another information notion, *representation complexity* (Tishby & Polani, 2011; Genewein, Leibfried, Grau-Moya, & Braun, 2015; Zenon, Solopchuk, & Pezzulo, 2019), measuring information transmitted about environmental state to an agent's internal representation.

Directly measuring internal representations is a notoriously difficult problem because they are latent constructs. In this article, we resort to a model-based analysis to understand how the set size impacts human learning performance. We compare three classes of models: a standard RL model as a benchmark, two RL models with policy complexity adopted from Lai and Gershman (2021) to show the failure of policy complexity, and another two that explains cognitive cost as the summation of both representation complexity and policy complexity to interpret the set size effect.

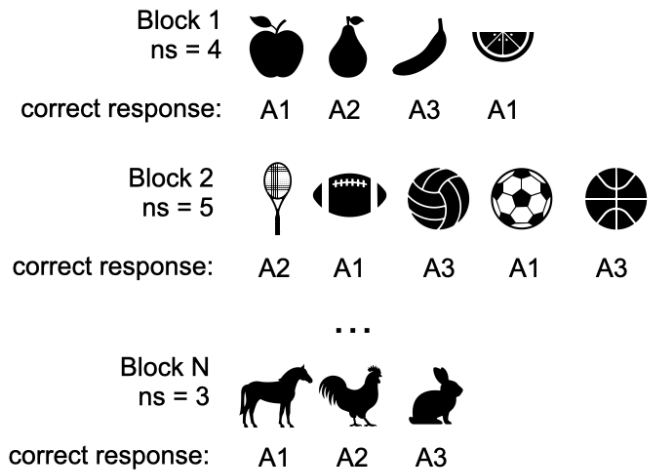

Figure 1: Schematic of experimental task studied by Collins and Frank (2012). On each trial, subjects were shown one single stimulus and were instructed to choose one of three actions. Each stimulus corresponded to one correct action and the number of stimuli varied across blocks. Note that the stimuli shown here are for illustrative purposes and are not the actual stimuli used in the experiment.

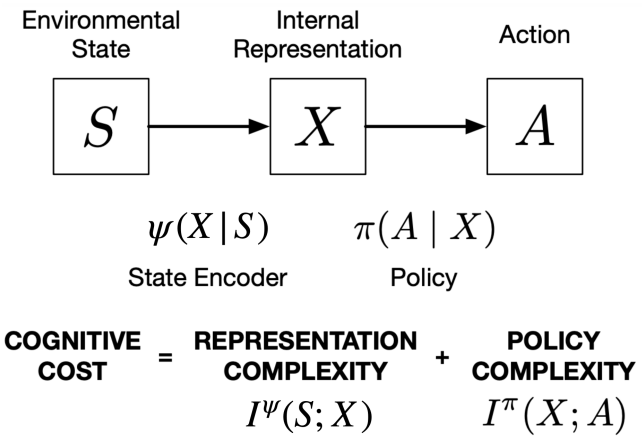

Figure 2: Schematic of cognition process. The biological sensory signal of the input stimuli S are encoded to internal mental representation X, and based on which human make decisions A.

Methods

Data set

We tested a series of models on the data set reported in Collins and Frank (2012). This data set consists of 78 subjects' learning performance in a multi-armed bandit task. On each trial, subjects were shown one single visual stimulus (drawn from categories such as sports, fruits, etc.) and were instructed to quickly choose a key among three alternatives. Each response was followed with a binary outcome, either 1 (reward) or 0 (no reward). For each stimulus, the reward was deterministically associated with only one of the three responses. All stimuli were repeated 9-15 times within a block, and did not appear across blocks. The set size *ns* (the number of different stimuli within a block) systematically varied across blocks, ranging from 2 to 6 (Figure 1). Each subject completed 19 blocks, six in which with $ns = 2$, four with $ns = 3$, three blocks each with $ns = 4$, 5, or 6. See Collins and Frank (2012) for complete details.

Computational rationale of policy complexity

In the standard RL scenario, decision-making involves two variables: *environmental state* S and the *action* A. In the language of information theory, we can think of this cognitive process as an information channel: a policy $\pi(a|s)$ that maps the environmental states S onto a probability distribution over actions A. According to information theory, the average computational demands necessary to convey information over this 'policy channel' is equal to the *mutual information* between state and action, the general equation of which is:

$$I(Y;Z) = \sum_i p_y(y_i) \sum_j p_{z|y}(z_j|y_i) \log \frac{p_{z|y}(z_j|y_i)}{p_z(z_j)} \qquad (1)$$

where Y means the sender and Z, the receiver. The calculation of mutual information requires us to know the marginal distribution of both variables, p_y and p_z, as well as the channel statistics, $p_{z|y}$.

Gershman and Lai (2020) and Lai and Gershman (2021) considered the mutual information $I^{\pi}(S;A)$ as policy complexity. This is correct with the implicit assumptions that humans have full access to the environmental state (Tishby & Polani, 2011) and that they do not rely on internal representations of stimuli. Under these two assumptions, human decision-making is much like the stimulus-response (S-R) mapping in classic behaviorism.

Computational rationale of representation complexity

Instead of considering humans' decision-making as an S-R process, we introduce a third construct: the encoded *internal representation* of the state, X. Humans may now respond A to the given representation X. We may now consider the whole decision process as a cascade information channel (Figure 2): a state encoder $\psi(x|s)$ that maps the environmental states S onto a probability distribution over internal representations X, followed by a policy $\pi(a|x)$ that maps the mental states X to a distribution over actions A. The mutual information $I^{\Psi}(S;X)$ is considered as representation complexity and the policy complexity is now $I^{\pi}(X;A)$.

The advantage of introducing the representation is to allow the emergence of abstractions, which is thought of as a hallmark of intelligence (Kemp, Perfors, & Tenenbaum, 2007;

Gershman & Niv, 2010). When the environment is very complicated with an unaffordable information cost, an adaptive agent can cluster environmental states with a similar policy to lower the information cost during the state encoding stage (Genewein et al., 2015). However, the goal of this article is to identify what constitutes the cognitive cost, and the formation of adaptive representations is beyond our focus. In the present paper, we implemented a simple fixed state encoder ψ (see Models section for details).

Models

RL baseline: $\mathbf{M^{RL}}$ We use the RL baseline from Collins and Frank (2012). The computational goal of the RL baseline model is to find a policy that maximizes the expected total reward over all trials within a block,

$$\max_{\pi} E[r_t|p_s,\pi] \tag{2}$$

where p_s represents the prior knowledge about the state distribution and it is a uniform distribution in this experiment, in keeping with the experiment design where stimuli are uniformly sampled. r_t is the reward subjects received at trial t.

To achieve this, the model learns a state-action value $Q(s,a)$ and a policy $\pi(a|s)$ to guide action selection. Both the Q function and the policy are updated after each trial t. The update of the Q function follows:

$$Q^t(s_t,a_t) = Q^{t-1}(s_t,a_t) + \alpha_q[r_t - Q^{t-1}(s_t,a_t)] \tag{3}$$

where α_q is the learning rate. The s_t and a_t are the observed current state and action. We use the superscript t to note temporally changing variables.

To balance exploration and exploitation in the RL baseline model, the policy is formalized as the output of the softmax function of the most recent Q value,

$$\pi^t(s,a) = \frac{\exp[\beta Q^t(s,a)]}{\sum_j \exp[\beta Q^t(s,a_j)]} \tag{4}$$

where $\beta \geq 0$ is the inverse temperature parameter that controls the degree of stochasticity in the policy (Sutton & Barto, 2018).

The only parameters for the RL baseline are the learning rate α_q and the inverse temperature β for the policy. To apply the model to behavioral data, we fit both parameters via maximum likelihood estimation.

Policy complexity: $\mathbf{M}^{\pi}_{(1)}$ For the model that considers policy complexity, the computational goal is to maximize expected utility while ensuring that the policy complexity does not exceed a fixed capacity limit:

$$\max_{\pi} E[r_t|p_s,p_a,\pi] \quad s.t.\ I^{\pi}(S;A) \leq C \tag{5}$$

where p_a is the marginal action distribution and C denotes the channel capacity–the maximum available cognitive resource. Equation 12 can be rewritten in a Lagrangian form:

$$\max_{\pi} \beta E[r_t|p_s,p_a,\pi] - I^{\pi}(S;A) \tag{6}$$

where $\beta \geq 0$ regulates the tradeoff between external reward and policy complexity. When $\beta \to \infty$, the agent can be consider fully rational; when $\beta \to 0$, the agent sticks with its prior policy p_a.

To solve equation 6, we use the gradient-based process model developed in Lai and Gershman (2021). For more details, see (Lai & Gershman, 2021, appendix). This is an "actor-critic" model using the "policy gradient" algorithm (Sutton & Barto, 2018) to incrementally update the parameterized policy π_θ (the parameters of which are θ) and value function V_w (the parameters of which are w). In the original paper, all parameters are initialized as 0. while in this article we initialized the value parameters w as 1 as it provided a better fit to the data.

In each timestep t, the model first estimates the value of the current state s_t and the current policy of the state:

$$\hat{V}_w(s_t) = w^{t-1} \cdot \mathbb{I}(s_t) \tag{7}$$

and

$$\hat{\pi}_\theta(a|s_t) = \exp(\beta\theta^{t-1} \cdot \mathbb{I}(s_t) + \log p_a(a)) \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function that returns an one-hot encoding of the input. Note that $\hat{\pi}_\theta(a|s_t)$ is a distribution over action a.

The model the update the critic using :

$$w^t = w^{t-1} + \alpha_w \mathbb{I}(s_t)\delta \tag{9}$$

where α_w means the learning rate of the parameters of the value function, and $\delta = \beta r_t - \log \frac{\hat{\pi}_\theta(a_t|s_t)}{p_a^{t-1}(a_t)} - \hat{V}_w(s_t)$ is the prediction error.

The update of the critic is divided into two sub-steps. The first step is to update the policy:

$$\theta^t_{a_t} = \theta^{t-1}_{a_t} + \alpha_\theta \mathbb{I}(s_t)\beta[1 - \hat{\pi}_\theta(a_t|s_t)] \tag{10}$$

where α_θ is the learning rate of the policy parameter and $\theta^t_{a_t}$ means the parameters for action a_t. The update of the policy is followed by the update of the marginal action distribution:

$$p_a^t(a) = p_a^{t-1}(a) + \alpha_a[\hat{\pi}_\theta(a|s_t) - p_a^{t-1}(a)] \tag{11}$$

To use this model, we need to fit four hyperparameters $\{\alpha_w, \alpha_\theta, \alpha_a, \beta\}$. Note this model is a general case of the RL baseline. When $\alpha_a = 0$ and p_a is an uniform distribution, this model collapses to a policy gradient variant of RL baseline.

Policy complexity: $\mathbf{M}^{\pi}_{(2)}$ As shown in the results section below, the basic policy complexity model ($\mathbf{M}^{\pi}_{(1)}$) does not predict a set size effect in human learning. The limitation stems from assuming a single tradeoff parameter (β) for all set sizes. To avoid this limitation we can fit a specific β to set size *ns*.

Thus, $\mathbf{M}^{\pi}_{(2)}$ is exactly the same as $\mathbf{M}^{\pi}_{(1)}$ except it has eight hyperparameters $\{\alpha_w, \alpha_\theta, \alpha_a, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6\}$. The subscript of the β represents the set size the β is fit to.

Representation + policy complexity: $\mathrm{M}^{\psi+\pi}_{(1)}$ This model considers the cognitive cost as summation of both representation complexity and policy complexity, as illustrated in Figure 2. The resulting objective is

$$\max_{\pi} E[r_t|p_s,p_x,p_a,\psi,\pi] \quad s.t.\ I^{\psi}(S;X)+I^{\pi}(X;A)\leq C \quad (12)$$

and the corresponding Lagrangian form is,

$$\max_{\pi} \beta_{ns}E[r_t|p_s,p_x,p_a,\psi,\pi]-I^{\psi}(S;X)-I^{\pi}(X;A) \quad (13)$$

where p_x means the prior belief about the internal representations and is assumed as a uniform distribution. Representations X are generated probabilistically according to the state encoder $\psi(x|s)$. For example, "apple" and "orange" may evoke very distinct sensory representations, but are mapped to one latent representation because they both have the same optimal response (and hence are functionally, if not perceptually, equivalent). In this paper, we implemented a simple (non-adaptive) model for the state encoder ψ inspired by the ε-greedy policy in RL (Sutton & Barto, 2018). An environmental state s has $1-\varepsilon$ probability to be recognized as s and has $\frac{\varepsilon}{|S|-1}$ probability be recognized as any of stimuli other than s. Increasing ε increases the "noise" in the state encoder, and hence reduces its information-theoretic channel capacity. The motivation behind this design is that we need a noisy categorical distribution (environmental state s is recorded as a categorical variable in the data) that may collapse to a one-hot encoding (the indicator function $\mathbb{I}(\cdot)$ in equation 8, assuming humans participants had full access to the environmental state) if humans are really optimal state encoders. If the fitted ε is 0, we may conclude humans develop perfect representations for the external stimuli in this simple experiment paradigm.

To implement a gradient-based RL model with a state encoder ψ, we only need to change the indicator function of the state indicator function $\mathbb{I}(s_t)$ to $\psi(x|s_t)$. The nine hyperparameters of this model are $\{\alpha_w,\alpha_\theta,\alpha_a,\varepsilon,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6\}$. When $\varepsilon=0$, $\mathrm{M}^{\psi+\pi}_{(1)}$ collapses to $\mathrm{M}^{\pi}_{(1)}$.

Representation + policy complexity: $\mathrm{M}^{\psi+\pi}_{(2)}$ The previous model utilized a gradient-based optimization procedure to achieve the learning objective. We also tested a gradient-free normative model based on Tishby and Polani (2011) and Genewein et al. (2015). The model is built upon RL baseline M^{RL} with a same critic formulation and update rule.

The actor component of the model is conditional on the internal representation x and action a. Since we have no access to the latent representation in the observed data, we can only infer the representation-action value function $Q_{bel}(x,a)$ following (Genewein et al., 2015),

$$Q^t_{bel}(x,a)=\sum_x p(s|x)Q^t(s,a) \quad (14)$$

where $p(s|x)=p_s(s)\psi(x|s)/p_x(x)$ is the Bayesian posterior over s given x and ψ follows the same design with $\mathrm{M}^{\psi+\pi}_{(1)}$.

With the representation-action function $Q_{bel}(x,a)$, we can formulate the optimal update of the actor as,

$$\pi^t(a|x)=\frac{\exp[\beta_{ns}Q^t_{bel}(x,a)+\log p_a^{t-1}(a)]}{\sum_j \exp[\beta_{ns}Q^t_{bel}(x,a_j)+\log p_a^{t-1}(a_j)]} \quad (15)$$

This is the optimal policy update for a given value function (Tishby & Polani, 2011). In contrast to the gradient-based update of the previous model, this model would be expected to learn more quickly.

The update of marginal policy p_a follows equation 11. The hyperparameters of $\mathrm{M}^{\psi+\pi}_{(2)}$ are $\{\alpha_q,\alpha_a,\varepsilon,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6\}$.

Optimal policy As a benchmark for evaluating our models, we also determined the optimal policy for a learning agent. To achieve the optimal solution, we can simply use the RL baseline model with $\alpha_q=1$ (high learning rate) and $\frac{1}{\beta}=0$ (no exploration). This is a consequence of the particular task environment, as there is exactly one action that is deterministically rewarded for each stimulus.

Results

Model fits and the set size effect

Figure 3 compares human and model learning curves. As expected, the RL baseline (M^{RL}) does not reproduce the set size effect. More surprisingly, a model incorporating policy complexity ($\mathrm{M}^{\pi}_{(1)}$) also fails to account for this effect. This model utilizes a fixed utility–complexity tradeoff parameter (β) for all set sizes. Model $\mathrm{M}^{\pi}_{(2)}$ fits separate parameters for each set size, but offers no explanation as to why this parameter should differ according to set size. The models that incorporate both policy complexity and representation complexity were able to demonstrate the set size effect.

Table 1: Models' goodness-of-fit.

-	NLL	SSE
M^{RL}	28135.358	0.463
$\mathrm{M}^{\pi}_{(1)}$	26299.911	0.256
$\mathrm{M}^{\pi}_{(2)}$	25889.932	0.083
$\mathrm{M}^{\psi+\pi}_{(1)}$	**25784.780**	**0.078**
$\mathrm{M}^{\psi+\pi}_{(2)}$	26347.952	0.089

Table 1 summarizes the negative log-likelihood (NLL) and sum-of-squared-error (SSE) for all models. NLL evaluates how well the model accounts for the experimental data, and SSE measures the degree of similarity between the model's predictive learning curves and that of humans. In terms of these two criteria, $\mathrm{M}^{\psi+\pi}_{(1)}$ accounts best for subjects' behaviors. However, $\mathrm{M}^{\psi+\pi}_{(1)}$ fails to capture one observation in human data: human follows a nearly optimal learning curve in set size 2 and 3. This phenomenon is only captured by $\mathrm{M}^{\psi+\pi}_{(2)}$.

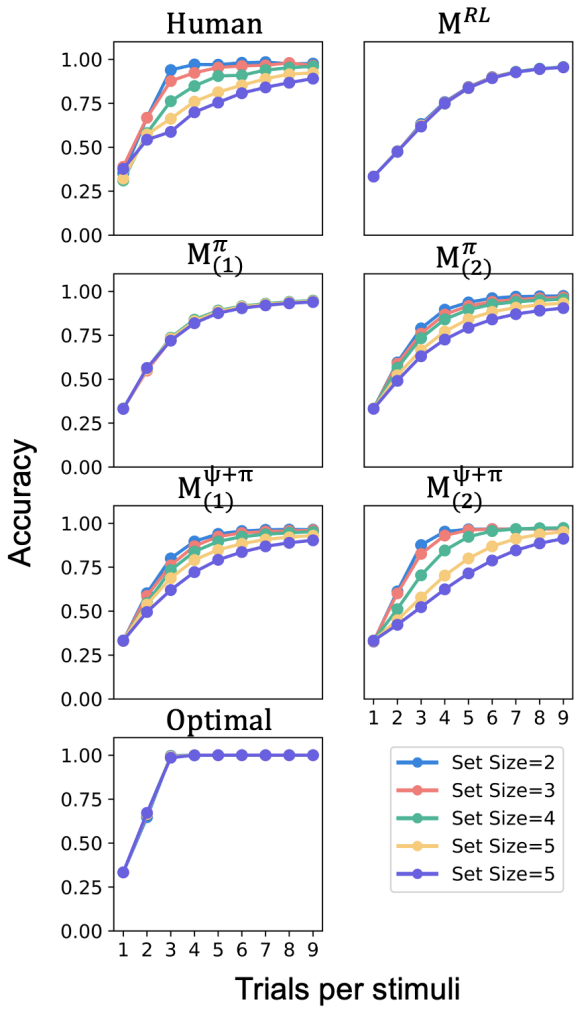

Figure 3: Model results. Learning curves generated using the fit parameters for each set size. Accuracy indicates proportion of responses that were rewarded.

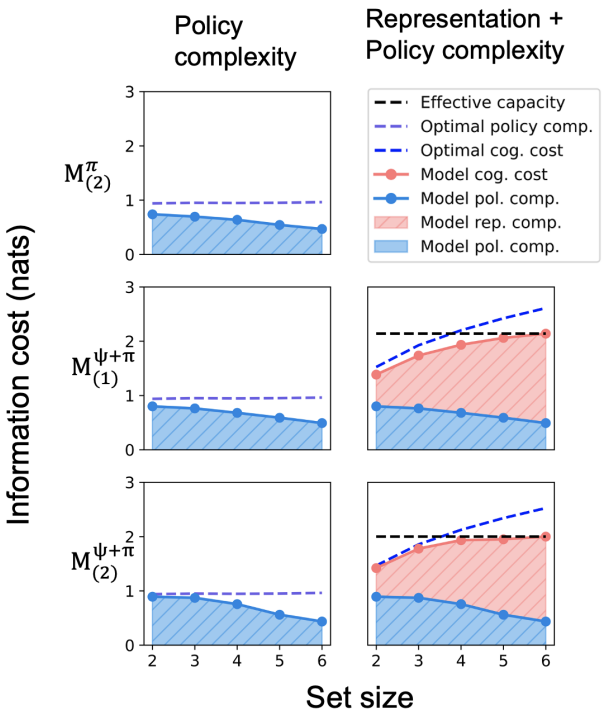

Figure 4: Cognitive cost estimation. The left column shows the complexity of both models' policy (blue shaded region) and the optimal policy complexity (purple dashed line). The right shows the total cognitive cost (red solid line), constituted of policy complexity (blue shaded region) and representation complexity (red shaded region) for both models. The working memory capacity (black dashed line) equals the maximum of models' cognitive cost. The blue dashed line indicates the total cognitive cost for the optimal policy. All quantities are measured in nats

Policy complexity does not account for the set size effect

The key assumption of our information-theoretic approach is that humans behave rationally subject to a fixed cognitive resource (the information constraint), and consequently their suboptimal task performance is explainable via this constraint. In this sense, if we estimate the cognitive cost of the optimal policy, the amount of information (measured in nats) to encode this policy may increase monotonically with the set size, whereas the cognitive cost for the empirical human policy should saturate to a certain value. We may consider this asymptotic value as the effective constraint on cognitive cost for human participants.

Figure 4 shows the model-based cognitive cost estimation. Details for estimating policy complexity and representation complexity are given in the Methods section above. The left column shows the estimation of policy complexity revealing two salient features: 1) the optimal policy complexity $I^{opt}(X;A)$, calculated using equation 1, is almost constant over set sizes instead of monotonically increasing (purple dashed line), hence larger set sizes do not appear to be more cognitively demanding according to this model; 2) the empirical policy complexity monotonically decreases instead of saturating at a fixed channel capacity (blue shaded region). A problematic question therefore arises: if the tasks in all set sizes are equally complex from an information-theoretic perspective, why do human participants adopt simpler policies in a larger set size conditions?

Neither of these properties is readily explainable from the perspective of computational rationality. We, therefore, conclude that policy complexity alone does not adequately explain human cognitive costs in this experiment.

Representation complexity plus policy complexity captures the set size effect

The right column of figure 4 displays the estimation for both policy complexity and representation complexity (cognitive

cost). The number of nats required to encode the combined representation and optimal policy ($I^{\Psi}(S;X) + I^{opt}(X;A)$) increases monotonically with set size (blue dashed line). However, whereas the task demands for optimal performance grow monotonically with set size, the empirically estimated cognitive costs (red solid line) appear to grow much slower in $M_{(1)}^{\Psi+\pi}$ and reach an asymptote at ~ 1.998 nats in $M_{(2)}^{\Psi+\pi}$. Consistent with our expectation, both properties imply the existence of an upper limit on the cognitive capacity that is the sum of both representational complexity and policy complexity.

This formulation of cognitive cost captures and quantifies the subjective experience that increasing the set size increases the cognitive difficulty of the task. In addition, while the estimated policy complexity saturates (or tends to saturate), Figure 4 also shows that rising representation complexity (red shaded region) imposes extra constraints on policy complexity $I^{\pi}(X;A)$ as the set size increases, answering the question we asked in the last paragraph. According to this model, for set size $ns = 2,3$ conditions, human decision-makers are able to perform near-optimally because the total cognitive cost is below the available capacity. However, when $ns = 4,5,6$, as the state representation complexity grows, human decision makers must resort to an increasingly suboptimal policy to prevent total cognitive cost from exceeding a maximum limit.

Conclusions

In this article, we proposed a new model the optimizes the resource-rational computational goal. Comparing with a similar model published (Griffiths, Lieder, & Goodman, 2015), a large improvement has been made in predicting a deterministic reinforcement learning task. The empirical results indicated that the progress was made because of refining three modeling assumptions: (i) constructing the cost as the sum of representation and policy complexity, (ii) estimating the complexities using a wrong prior, and (iii) updating the model in terms of distribution.

Many suboptimal decisions can be explained as a trade-off between maximizing utility and minimizing costs or constraints imposed by limited cognitive resources (Sims, 2016; Lerch & Sims, 2018; Gershman, 2020). We contribute to this line of thought by arguing that there are two separate sources of cognitive demand in a reinforcement learning setting: representation complexity, and policy complexity. Through a model-based analysis, we showed that the total cognitive cost incorporating both of these constructs appears to saturate to an upper limit in human reinforcement learning. This tentatively suggests the existence of a fixed cognitive resource that can be allocated to a learning task.

Based on this conclusion, we made one further step to interpret how the set size leads to humans' suboptimal performance in the (Collins & Frank, 2012) experiment. Although a larger set size is not necessarily more complicated in terms of policy complexity, it does requires the human subjects to hold more representations of the world stimuli. Humans, thus, have to seek a simpler policy to balance the rising cognitive cost.

In future work, we seek to increase the quality of our model-based analysis by developing more accurate models that better describe humans' learning and decision-making under limited cognitive resources. We expect the following properties from a better model: First, instead of fitting the tradeoff parameters β_{ns} for each set size to describe humans' policy, we can model the principle humans may follow in balancing the reward and the resource. Also, our models assume that human sensory processing is fixed. However, substantial evidence supports that human sensory channel might be adaptive (see review Orhan, Sims, Jacobs, and Knill (2014)). Perhaps, human subjects start with a less resource-efficient sensory code but end up with more efficient coding, allowing humans to learn a more rewarding but complicated policy. This change might be observed only after extensive training because the update of the sensory channel should follow an extremely small learning rate due to it is hardwired in the human neural system.

References

Berger, T. (1971). The source coding game. *IEEE Transactions on Information Theory*, *17*(1), 71–76.

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

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*(7), 1024–1035.

Genewein, T., Leibfried, F., Grau-Moya, J., & Braun, D. A. (2015). Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, *2*, 27.

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

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gershman, S. J., & Lai, L. (2020). The reward-complexity trade-off in schizophrenia. *bioRxiv*.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, *20*(2), 251–256.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217–229.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Devel-*

opmental science, *10*(3), 307–321.

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. *Psychology of Learning and Motivation S*.

Lerch, R. A., & Sims, C. R. (2018). Policy generalization in capacity-limited reinforcement learning.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature neuroscience*, *17*(3), 347.

Orhan, A. E., Sims, C. R., Jacobs, R. A., & Knill, D. C. (2014). The adaptive nature of visual working memory. *Current directions in psychological science*, *23*(3), 164–170.

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

Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, *131*(3), 139–148.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tishby, N., & Polani, D. (2011). Information theory of decisions and actions. In *Perception-action cycle* (pp. 601–636). Springer.

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