Origin of perseveration in the trade-off between reward and complexity

Samuel J. Gershman
Department of Psychology and Center for Brain Science
Harvard University

Address for correspondence:
Samuel Gershman
52 Oxford St., Room 295.05 Cambridge, MA 02138
Phone: 773-607-9817
E-mail: gershman@fas.harvard.edu

Abstract

When humans and other animals make repeated choices, they tend to repeat previously chosen actions independently of their reward history. This paper locates the origin of perseveration in a trade-off between two computational goals: maximizing rewards and minimizing the complexity of the action policy. We develop an information-theoretic formalization of policy complexity and show how optimizing the trade-off leads to perseveration. Analysis of two data sets reveals that people attain close to optimal trade-offs. Parameter estimation and model comparison supports the claim that perseveration quantitatively agrees with the theoretically predicted functional form.

Keywords: decision making; information theory; reinforcement learning

Introduction

In his pioneering work on animal intelligence, Thorndike (1911) formulated several "laws" of learning. The most famous of these, the *law of effect*, stated that an action yielding a reward will become more likely to be repeated in the future. The lesser-known *law of exercise* stated that simply producing an action will make it more likely to be repeated in the future (and concomitantly, an action will become less likely to be repeated if it's not produced). The law of exercise implies a form of perseveration: with sufficient frequency of repetition, an action will be selected even if it no longer yields the highest reward among the possible choices. The law of exercise thus captures a key signature of habit, whereby repetition causes behavior to become "autonomous" from the agent's goals (Dickinson, 1985; Miller et al., 2019; Wood and Rünger, 2016).

Although these laws do not exhaustively determine action selection, they are supported by many studies of humans and other animals. For example, Lau and Glimcher (2005) found that monkeys performing a two-alternative forced choice task were influenced by the recent history of both rewards and choices, a finding that also extends to human subjects (e.g., Seymour et al., 2012). Under time pressure, people will frequently repeat previous actions despite intending to choose an alternative action (Betsch et al., 2004). Reaction times are also facilitated for response repetitions in serial choice reaction tasks (e.g., Bertelson, 1965; Rabbitt and Vyas, 1974). In everyday life, past actions predicts future actions (e.g., product choices in the supermarket; Riefer et al., 2017), even after controlling for other predictors such as conscious intentions and beliefs about social norms (Ouellette and Wood, 1998).

While perseveration has been ubiquitously documented, a basic puzzle is *why* it occurs at all. If the goal is to maximize reward, an agent's actions should be entirely predictable from its reward history; in other words, the law of exercise should be completely dominated by the law of effect. If anything, the need to explore actions in order to gain information about their consequences should induce a tendency *against* repeating past actions (Riefer et al., 2017; Schulz and Gershman, 2019).

A common theme in the psychology of habit is the idea that perseveration is somehow less effortful (Wood and Rünger, 2016). Some reinforcement learning models have conceptualized effort in terms of computational complexity; habits arise from action selection based on a look-up table of cached reward expectations, which demands less effort compared to goal-directed action selection based on planning with an internal model of the task (Daw, 2018). Consistent with this conceptualization, taxing cognitive resources (for example, by having subjects perform a secondary task or increasing the difficulty of planning) results in greater reliance on habit (Gershman et al., 2014; Kool et al., 2018b; Otto et al., 2013). The problem with this view of habit, as pointed out by Miller et al. (2019), is that it does not exactly correspond to Thorndike's law of exercise: caching reward expectations in a look-up table does not by itself produce a bias to repeat actions. Rather, this form of caching can be viewed as implementing Thorndike's law of effect. Miller et al. (2019) propose an alternative model that explicitly formalizes the law of exercise, whereby taking an action increases its habit strength independently from reward. While this model succeeds as a descriptive account of habitual action selection, it does not provide a computational rationale for perseveration. Thinking about this rationale in terms of computational complexity seems unpromising, since it's not obvious why looking up a cached habit value would be less cognitively expensive than looking up a cached reward expectation, and the latter is obviously more useful from the perspective of reward maximization.

A different approach to this question rests upon a distinction between *computational* and *statistical* complexity. Whereas computational (or time) complexity measures how much thinking is required to perform a task, statistical (or sample) complexity measures how much learning is required. Effort, in this case, corresponds to the difficulty of learning. Is it possible that habits are less statistically complex? In fact, Filipowicz et al. (2020) have shown that learning cached reward expectations is not necessarily more statistically complex than learning an internal model for planning (they did not directly address the perseverative notion of habit).

In this paper, we explore a different computational rationale for perseveration, based on the notion of *policy* complexity (McNamee et al., 2016; Parush et al., 2011; Still and Precup, 2012; Tishby and Polani, 2011). In the language of reinforcement learning theory, a policy $\pi(a|s)$ is a probabilistic mapping from states to actions (Sutton and Barto, 2018), where a state corresponds to the information about the environment that is needed for reward prediction. To implement a policy computationally, we would need to describe it in some programming language, and the description length of that program (e.g., in bits or nats) imposes a demand on memory resources. Intuitively, if a policy can be "compressed" to a short description length, it will be easier to remember, much in the same way that the benefits of compression have been studied in memory for digit verbal and visual stimuli (Brady et al., 2009; Mathy and Feldman, 2012; Nassar et al., 2018). As we will formalize later, it turns out that perseveration arises naturally from the imperative to reduce policy complexity. Perseveration is, in essence, a form of policy compression.

Policy complexity is conceptually different from computational complexity; one could have a policy with low policy complexity and high computational complexity, or vice versa. For example, finding the shortest path between two distant cities might be require an expensive optimization (high computational complexity), but the optimal path itself might be very simple, like staying on one highway for most of the trip (low policy complexity). In contrast, finding the shortest path between two locations in the same city might be cheap (low computational complexity), but the optimal path might be tortuous (high policy complexity), as anyone who has tried to get somewhere in Boston by car knows well.

The question addressed here is how people negotiate the trade-off between reward and policy complexity. The mathematical toolbox for answering this question comes from the branch of information theory known as *rate distortion theory* (Berger, 1971). The next section reviews the elementary concepts as they apply to policy optimization. Rate distortion theory allows us to derive the optimal trade-off function, which reveals that perseveration will occur for any resource-bounded agent. Since both reward and policy complexity are experimentally measurable, we can evaluate the degree to which choice data conform to the optimal trade-off function. Furthermore, by fitting parametrized policies to the choice data, we can evaluate how well the data match the theoretically predicted form of perseveration.

Methods

All code and data for reproducing the analyses described below is available at `https://github.com/sjgershm/reward-complexity`.

Theoretical framework

Rate distortion theory addresses the interface between information theory and statistical decision theory. Here we will adopt somewhat non-standard terminology, following Parush et al. (2011), in order to draw a clearer connection with the issues raised in the Introduction. We will assume that an agent either learns or has direct access to a *value function* $Q(s, a)$ that defines the expected reward in state s after taking action a. Each state is visited with probability $P(s)$, and an action is chosen according to a policy $\pi(a|s)$. In the language of information theory, we can think of the state distribution as a *source* and the policy as a *noisy channel*, mapping messages (states) to codewords (the internal representation), which are in turn mapped to output signals (actions). The average codeword length (or *rate*) necessary to encode a policy is equal to the mutual information between states and actions:

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

where $P(a) = \sum_s P(s)\pi(a|s)$ is the marginal probability of choosing action a. The maximum possible rate for a channel is called the *channel capacity*, which we denote by C.

The communication channel formulation is useful because it lets us see why compression makes sense. Real-world environments involve an astronomical number of states and actions, so a resource-limited system can't afford to represent all of them in a giant look-up table. This goes against the conventional wisdom that look-up tables are computationally cheap (e.g., Kool et al., 2018a); although they require little thinking (low computational complexity), they require a large number of bits (high policy complexity). Previous applications of rate distortion theory to psychology have used this insight to explain the factors influencing confusability in memory, on the assumption that items cannot be stored in a look-up table due to resource constraints (Sims et al., 2012; Sims, 2016).

Exactly how much to compress depends on the amount of reward that can be achieved for a given information rate. Let us denote the average reward by:

$$V^{\pi} = \sum_{s} P(s) \sum_{a} \pi(a|s) Q(s, a). \tag{2}$$

We can now formulate the optimization problem:

$$\pi^* = \operatorname*{argmax}_{\pi} V^{\pi} \tag{3}$$

$$\text{subject to } I^{\pi}(S; A) = C. \tag{4}$$

We have left implicit two other necessary constraints (action probabilities must be non-negative and sum to 1). This constrained optimization problem can be rewritten in a Lagrangian form:

$$\pi^* = \operatorname*{argmax}_{\pi} \beta V^{\pi} - I^{\pi}(S; A) + \sum_{s} \lambda(s) \left(\sum_{a} \pi(a|s) - 1 \right) \tag{5}$$

with Lagrange multipliers β and $\lambda(s)$. The optimal policy π^* has the following form (Parush et al., 2011; Still and Precup, 2012; Tishby and Polani, 2011):

$$\pi^*(a|s) \propto \exp\left[\beta Q(s, a) + \log P^*(a)\right]. \tag{6}$$

A number of recognizable components now come into view. The optimal policy is a softmax function, used ubiquitously in the reinforcement learning literature for both simulating artificial agents and modeling biological agents. The Lagrange multiplier β plays the role of the "inverse temperature" parameter, which regulates the exploration-exploitation trade-off via the amount of stochasticity in the policy (Sutton and Barto, 2018). When β is close to 0, the policy will be near-uniform, and as β increases, the policy will become increasingly concentrated on the action with maximum value. However, note that the derivation of the optimal policy makes no reference to exploration (see Still and Precup, 2012). Rather, β reflects the resource constraint—more precisely, the partial derivative of the rate with respect to the value:

$$\beta = \frac{dV^\pi}{dI^\pi(S;A)}. \tag{7}$$

This means that the inverse temperature for the optimal policy will be higher (choice stochasticity lower) when varying the rate has a greater effect on reward (see also Alexandre et al., 2019; Ortega and Braun, 2013).

Another important property of Eq. 6 is the $\log P^*(a)$. This implies that frequently chosen actions should bias the policy (i.e., perseveration), in accordance with Thorndike's law of exercise. We will empirically evaluate the specific functional form of perseveration implemented by Eq. 6, as described below.

The perseveration term implicitly depends on the optimal policy, since

$$P^*(a) = \sum_s P(s)\pi^*(a|s). \tag{8}$$

Thus, to find the optimal policy, we can use a variation of the classic *Blahut-Arimoto algorithm* (Arimoto, 1972; Blahut, 1972), alternating between updating the policy according to Eq. 6 and updating the marginal action distribution according to Eq. 8. by performing this optimization for a range of β values, we can construct a *reward-complexity curve* that characterizes the optimal policy for a given resource constraint.

Data sets

We evaluated the predictions of the theory developed in the previous section using two data sets. The first data set, reported in Collins (2018), consists of 91 subjects performing a reinforcement learning task in which the set size (the number of distinct stimuli, corresponding to states) varied across blocks (Figure 1A). On each trial, subjects saw a single stimulus, chose an action and received reward feedback. Each stimulus was associated with a single rewarded action. The experiment consisted of a learning and test phase (with no reward feedback), but here we only analyze the learning phase data. Each subject completed 14 blocks, half with set size 3 and half with set size 6. Each stimulus appeared 12-14 times in a block. No stimulus was repeated across blocks.

The second data set, reported in Steyvers et al. (2019), consists of 1000 subjects playing the task-switching game "Ebb and flow" on the Lumosity platform (Figure 1B). On each trial, subjects viewed moving leaves on a display and reported either the motion or pointing direction of the leaves. In this case, the state corresponds to a tuple (task, motion direction, pointing direction), defining 32 distinct states. Subjects played between 371 and 5227 trials, with a median of 2735

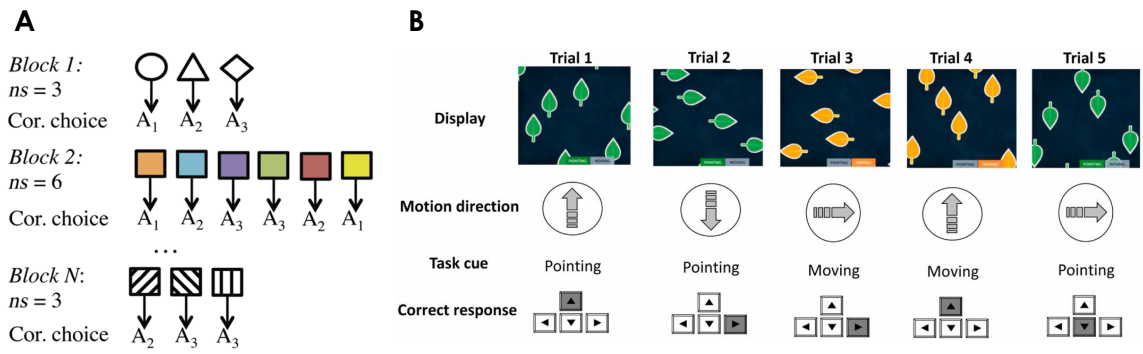

Figure 1: Schematic of experimental tasks. (A) In Collins (2018), subjects saw a single stimulus on each trial and then chose between 3 actions. Each stimulus corresponded to a state with a single rewarded action. The number of stimuli was varied across blocks. (B) In Steyvers et al. (2019), subjects viewed a display containing leaves moving coherently in one of 4 cardinal directions. The leaves also pointed in one of four cardinal directions. On some trials (indicated by orange leaf color) subjects made a motion direction judgment, and on other trials (indicated by green leaf color) subjects made a pointing direction judgment. Feedback was provided after each judgment.

trials (99% of subjects played over 1000 trials, so the task can be considered well-practiced for most subjects).

Mutual information estimation

To construct the empirical reward-complexity curve, one needs to estimate two quantities: average reward and the mutual information between states and actions. Estimation of average reward is straightforward, but estimation of mutual information is notoriously tricky (see Paninski, 2003). We used the Hutter estimator, which computes the posterior expected value of the mutual information under a Dirichlet prior (Hutter, 2002). We chose the Perks prior, which exhibits good performance when the joint distribution is sparse (Archer et al., 2013). The sparsity assumption is likely to hold true in the data sets analyzed here because there is a single rewarded action in each state. As shown in the Results, this produced empirical reward-complexity curves that mostly satisfied the theoretical bound.

Parameter estimation and model comparison

To quantitatively evaluate the theory, we fit models of the following form:

$$\pi(a|s) \propto \exp\left[\beta \hat{Q}(s,a) + \tau \log P(a)\right], \tag{9}$$

where $\hat{Q}(s,a)$ is an estimate of the expected reward and τ is a choice perseveration parameter. In model M1, we fit both β and τ as free parameters. In model M2, we fit only β, and forced τ to equal 1 (corresponding to the optimal policy in Eq. 6). Maximum likelihood parameter estimates were obtained using unconstrained optimization with 5 random initializations to avoid local maxima.

Model comparison was performed using a Bayesian random effects procedure (Rigoux et al., 2014). In brief, this procedure estimates the population-level frequency of each model, along with

the probability that an individual subject's data were generated by each model. We report the log model evidence favoring M2 over M1 for each subject, as well as the protected exceedance probability, which measures the probability that M2 is more likely in the population than M1, taking into account the probability of spurious differences due to randomness.

The procedure to obtain $\hat{Q}(s,a)$ and $P(a)$ for each trial was slightly different for the two data sets. For the Collins (2018) data set, the Q-values were initialized to 0, learning was modeled using a standard delta rule:

$$\Delta\hat{Q}(s,a) = \alpha[r - \hat{Q}(s,a)], \tag{10}$$

where α is a learning rate parameter (which was fitted to the data) and r is the received reward. The marginal action probability $P(a)$ was estimated using a 5-trial moving average (truncated at boundaries between blocks).

For the Steyvers et al. (2019) data set, we assumed that subjects had full knowledge of $Q(s,a)$ and simply hard-coded it into the policy. As with the Collins data set, the marginal action probability was estimated using a 5-trial moving average. Since there was no discrete block structure, no truncation of the moving average was applied.

Results

Figures 2 and 3 show the reward-complexity curves for the two data sets, with the empirical data superimposed. As predicted by the theory, reward generally increases monotonically with policy complexity, with values close to the optimal trade-off curve. Figure 2 also shows that policy complexity is higher for larger set sizes, resulting in lower average reward.

To quantify the agreement between theory and data, we used interpolation to identify the predicted average reward for each measured policy complexity value. These predictions were significantly correlated with the empirical average reward ($r = 0.84$ for the Collins data set, $r = 0.95$ for the Steyvers data set, both $p < 0.00001$). Despite this quantitative agreement, the data also indicate a salient deviation from the optimal reward-complexity curve: subjects with low policy complexity achieve lower average reward than would be predicted by the optimal policy. We quantified this by computing the correlation between the bias (how far a subject is from the theoretical curve) and policy complexity, finding a significant negative correlation for both data sets each measured policy complexity value ($r = -0.58$ for the Collins data set, $r = -0.8$ for the Steyvers data set, both $p < 0.00001$).

We next sought to evaluate the functional form of the policy described by Eq. 6. As described in the Methods, we fit two models to the choice data: M1, which fits the degree of choice perseveration as a free parameter to each subject separately, and M2, which forces the parameter to equal 1 (in accordance with the theory). The distributions of estimates for this parameter are shown in the left panels of Figure 4, revealing that they are concentrated around 1 (89% of the parameter estimates were between 0.5 and 1.5 for the Collins data set, and 96% in the Steyvers data set). This conclusion was further validated by random-effects Bayesian model selection, which strongly favored model M2 over M1 (protected exceedance probability greater than 0.99 for both data sets). M2 was favored for almost all subjects, as shown in the right panels of Figure 4. Taken together, these results show that the resource-constrained optimal policy provides a good quantitative model of perseveration in these data sets.

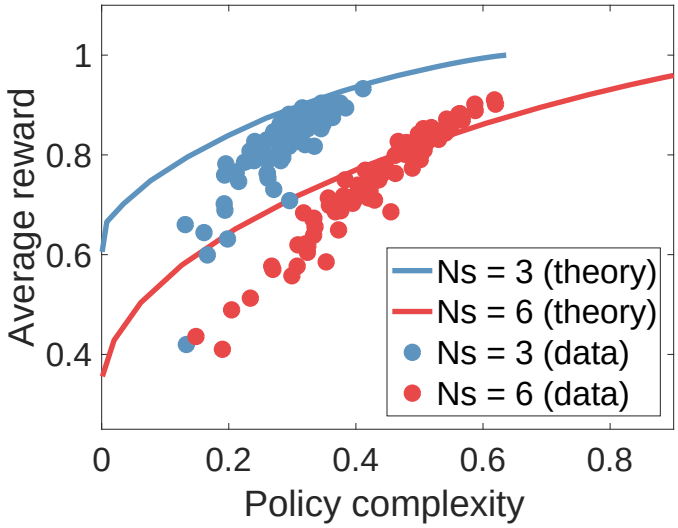

Figure 2: The reward-complexity trade-off, applied to data from Collins (2018). Each solid line shows the optimal trade-off function for a particular set size (Ns = 3 or 6). The circles show data from different blocks of trials, aggregated across subjects. Complexity is measured in nats.

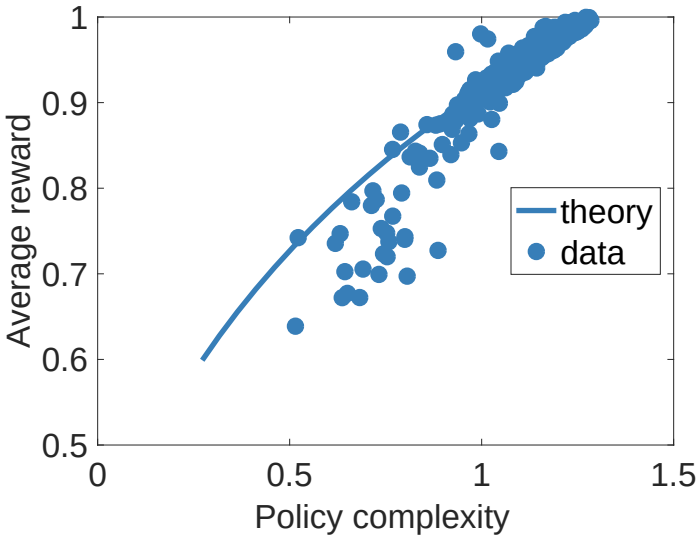

Figure 3: Reward-complexity trade-off, applied to data from Steyvers et al. (2019). The solid line shows the optimal trade-off function, and the circles correspond to individual subjects. Complexity is measured in nats.

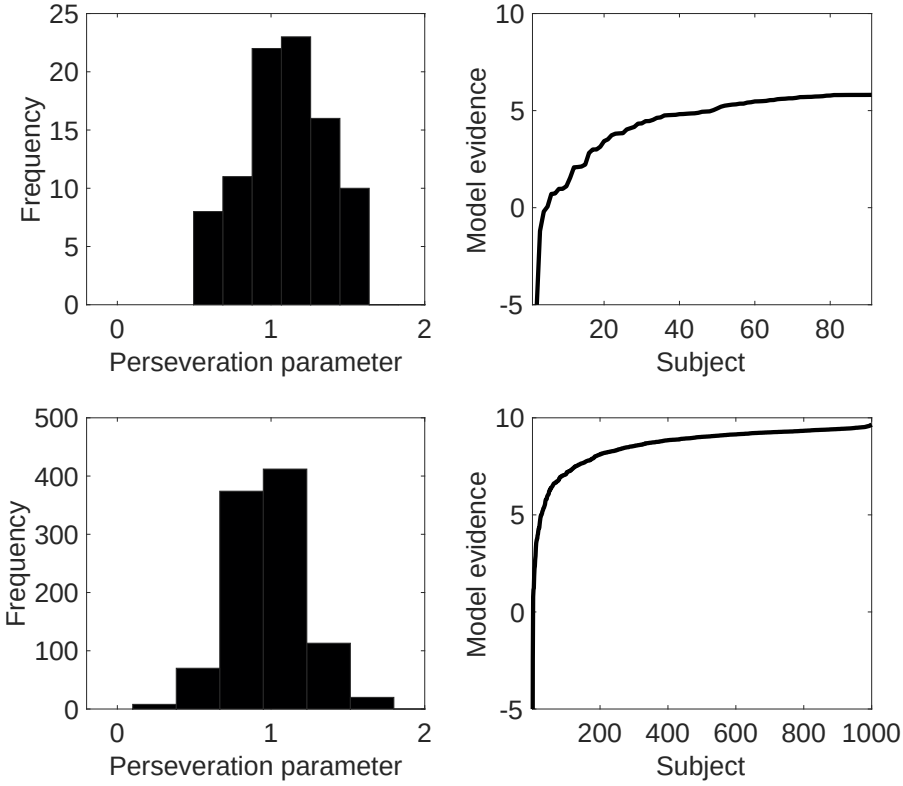

Figure 4: (Left) Histogram of maximum likelihood estimates for the choice perseveration parameter. (Right) Log posterior odds in favor of M2 (the optimal trade-off model) for each subject, sorted by increasing evidence. An evidence of 0 indicates equal support for M1 and M2. The top panels show the results for the data from Collins (2018); the bottom panels show the results for the data from Steyvers et al. (2019).

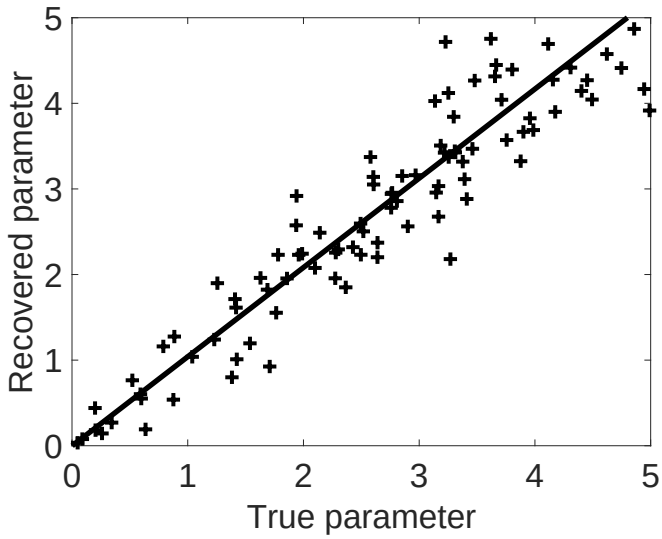

Figure 5: Parameter recovery from simulated data. Solid line shows the least-squares fit.

To ensure that our parameter estimation results are not a spurious consequence of the model structure (e.g., due to identifiability issues), we simulated data from Eq. 9 applied to the experimental design from the Steyvers study. The inverse temperature and choice perseveration parameters were sampled uniformly from the range $[0, 5]$. We then fit the model to these simulated data using the same procedure that we applied to the experimental data. Figure 5 shows a tight correlation between the true and recovered choice perseveration parameter estimates ($r = 0.94$), indicating that this parameter is indeed recoverable, bolstering our confidence in the analyses of parameter estimates for the experimental data.

Discussion

The idea that many aspects of cognition can be explained in terms of resource-bounded rationality has gained wide currency (Gershman et al., 2015; Lieder and Griffiths, 2019). The precise nature of the resource constraints and their implications is a matter of ongoing research. We contribute to this line of thought by applying rate distortion theory to a fundamental puzzle in psychology: why do humans and other animals perseverate? The answer provided here is that perseveration is a natural consequence of limitations on policy complexity. If the number of bits available to encode a policy is finite, then a resource-rational agent will exhibit perseveration. We showed empirically, using two data sets, that subjects or task conditions with different degrees of policy complexity yield predictable levels of reward attainment in accordance with the optimal reward-complexity trade-off. Our analyses also showed that the functional form of perseveration was quantitatively consistent with rate distortion theory. Nonetheless, there was a systematic deviation from the optimal trade-off function for subjects with low policy complexity.

In the Introduction, we highlighted a distinction between statistical complexity (the amount of data needed to learn a policy) and policy complexity (the number of bits needed to encode a

policy). However, these concepts are connected, because simpler policies are more easily learned. This follows from the general principle that *compression implies learning* (Blum and Langford, 2003), which can be formalized in a number of ways. For example, in the setting where $Q(s, a) = 1$ if $a \in \{0, 1\}$ is correct and 0 otherwise, the policy can be viewed as a binary classifier and the rewards can be viewed as labels (the standard supervised learning problem). Roughly speaking, if the number of bits required to describe the policy is much less than the number of samples, then we can guarantee accurate generalization to new samples (Blumer et al., 1987). The connection between compression and learning explains why the mutual information between states and actions can be used to measure both statistical complexity (see Filipowicz et al., 2020) and policy complexity.

The theoretical framework of rate distortion theory is highly abstract. We have made very few assumptions about the underlying cognitive mechanisms that produce a particular point on the reward-complexity curve. This contrasts with the modeling that was previously applied to the same data sets (Collins, 2018; Steyvers et al., 2019), which explored detailed mechanistic hypotheses. These different approaches have different advantages and disadvantages. Ultimately, we would like detailed mechanistic theories of cognition of the sort developed by Collins, Steyvers, and their colleagues. At the same time, the search for general principles can be usefully pursued at a more abstract level of the sort developed here. This has the advantage of allowing us to make general claims about the nature of cognition that transcend particular mechanistic implementations.

Rate distortion theory holds promise as a vehicle for general principles because it unifies two frameworks (information theory and statistical decision theory) that already by themselves have broad explanatory reach. Rate distortion theory has been successfully applied to many different cognitive phenomena, ranging from working memory (Sims et al., 2012; Sims, 2016) and absolute identification (Sims, 2018) to language (Zaslavsky et al., 2018) and motor control (Schach et al., 2018). A complete theory in these domains will eventually use mechanistic models to constrain the rate distortion analysis.

Acknowledgments

I am grateful to Anne Collins and Mark Steyvers for making their data available, and to Rahul Bhui and Chris Bates for helpful discussion. This research was supported by the Office of Naval Research (N00014-17-1-2984), the Center for Brains, Minds and Machines (funded by NSF STC award CCF-1231216), and a research fellowship from the Alfred P. Sloan Foundation.

References

Alexandre, Z., Oleg, S., and Giovanni, P. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, 123:5–18.

Archer, E., Park, I., and Pillow, J. (2013). Bayesian and quasi-Bayesian estimators for mutual information from discrete data. *Entropy*, 15:1738–1755.

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18:14–20.

Berger, T. (1971). *Rate Distortion Theory: A Mathematical Basis for Data Compression*. NJ: Prentice-Hall.

Bertelson, P. (1965). Serial choice reaction-time as a function of response versus signal-and-response repetition. *Nature*, 206:217–218.

Betsch, T., Haberstroh, S., Molter, B., and Glöckner, A. (2004). Oops, i did it again—relapse errors in routinized decision making. *Organizational Behavior and Human Decision Processes*, 93:62–74.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18:460–473.

Blum, A. and Langford, J. (2003). PAC-MDL bounds. In *Learning Theory and Kernel Machines*, pages 344–357. Springer.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1987). Occam's razor. *Information Processing Letters*, 24:377–380.

Brady, T., Konkle, T., and Alvarez, G. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, 138:487–502.

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

Daw, N. D. (2018). Are we of two minds? *Nature Neuroscience*, 21:1497.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 308:67–78.

Filipowicz, A. L. S., Levine, J., Piasini, E., Tavoni, G., Kable, J. W., and Gold, J. I. (2020). The complexity of model-free and model-based learning strategies. *bioRxiv*.

Gershman, S., Markman, A., and Otto, R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, 143:182–194.

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

Hutter, M. (2002). Distribution of mutual information. In *Advances in Neural Information Processing Systems*, pages 399–406.

Kool, W., Cushman, F. A., and Gershman, S. J. (2018a). Competition and cooperation between multiple reinforcement learning systems. In *Goal-directed Decision Making*, pages 153–178. Elsevier.

Kool, W., Gershman, S. J., and Cushman, F. A. (2018b). Planning complexity registers as a cost in metacontrol. *Journal of Cognitive Neuroscience*, 30:1391–1404.

Lau, B. and Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, 84:555–579.

Lieder, F. and Griffiths, T. L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, pages 1–85.

Mathy, F. and Feldman, J. (2012). Whats magic about magic numbers? chunking and data compression in short-term memory. *Cognition*, 122:346–362.

McNamee, D., Wolpert, D. M., and Lengyel, M. (2016). Efficient state-space modularization for planning: theory, behavioral and neural signatures. In *Advances in Neural Information Processing Systems*, pages 4511–4519.

Miller, K., Shenhav, A., and Ludvig, E. (2019). Habits without values. *Psychological Review*, 126:292–311.

Nassar, M., Helmers, J., and Frank, M. (2018). Chunking as a rational strategy for lossy data

compression in visual working memory. *Psychological Review*, 125:486–511.

Ortega, P. A. and Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469:20120683.

Otto, A. R., Gershman, S. J., Markman, A. B., and Daw, N. D. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, 24:751–761.

Ouellette, J. and Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin*, 124:54–74.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253.

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

Rabbitt, P. and Vyas, S. (1974). Interference between binary classification judgments and some repetition effects in a serial choice reaction time task. *Journal of Experimental Psychology*, 103:1181–1190.

Riefer, P. S., Prior, R., Blair, N., Pavey, G., and Love, B. C. (2017). Coherency-maximizing exploration in the supermarket. *Nature Human Behaviour*, 1:0017.

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *NeuroImage*, 84:971–985.

Schach, S., Gottwald, S., and Braun, D. A. (2018). Quantifying motor task performance by bounded rational decision theory. *Frontiers in Neuroscience*, 12.

Schulz, E. and Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, 55:7–14.

Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., and Dolan, R. (2012). Serotonin selectively modulates reward value in human decision-making. *Journal of Neuroscience*, 32:5833–5842.

Sims, C., Jacobs, R., and Knill, D. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, 119:807–830.

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

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, 360:652–656.

Steyvers, M., Hawkins, G. E., Karayanidis, F., and Brown, S. D. (2019). A large-scale analysis of task switching practice effects across the lifespan. *Proceedings of the National Academy of Sciences*, 116:17735–17740.

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

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Thorndike, E. L. (1911). Animal intelligence: Experimental studies.

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

Wood, W. and Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology*, 67:289–314.

Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115:7937–7942.