Policy compression: an information bottleneck in action selection

Lucy Lai[1,*], and Samuel J. Gershman[2]
[1]Program in Neuroscience, Harvard University
[2]Department of Psychology and Center for Brain Science, Harvard University
[*]Correspondence: lucylai@g.harvard.edu

Contents

Abstract

The brain has evolved to produce a diversity of behaviors under stringent computational resource constraints. Given this limited capacity, how do biological agents balance reward maximization against the costs of representing complex action policies? In this chapter, we examine behavioral evidence for this reward-complexity trade-off. First, we introduce a theoretical framework that formalizes the idea of *policy compression*, or the reduction in cognitive cost of representing action policies by making them simpler. We then describe how a wide range of behavioral phenomena, including stochasticity, perseveration, response time, state and action chunking, and navigation are brought together under this framework. Finally, we discuss how our model can be used to probe the neural underpinnings of policy compression and their dysfunction in psychiatric illness.

Keywords: action selection, reinforcement learning, rational behavior, resource-rationality

1 Introduction

All action demands memory. When you go shopping, drive to work, or prepare a meal, your brain is retrieving stored information about *policies*, the mappings from states of the world to actions. Like all mappings in the brain, policies are capacity-limited: a finite physical storage medium imposes an upper bound on the number of bits (the description length) that can be used to specify policies. The need to economize on description length means that policies should be compressed as much as possible, discarding redundant bits and reducing precision where it's not needed. We will shortly formalize policy compression, but first we provide some intuitions.

Imagine you are preparing a meal for your family. In this case, states correspond to family members, actions correspond to dishes, and policies are mappings from family members to dishes (Figure 1). If you're lucky, everyone in your family will eat the same thing. This means that you can ignore the state entirely and just take the same action (prepare the same dish) repeatedly. Such a policy is compressed in the sense that it consumes fewer bits of memory compared to one in which you need to remember separate dishes for each family member. This illustrates the concept of *redundancy reduction*: remembering separate dishes would be redundant because they would simply be copies of the same dish. Remembering a single dish for everyone eliminates this redundancy.

If you are a mentally taxed parent, you might need to compress your policies more aggressively. Your children won't be able to tell the difference between Greek and Italian olive oil, so there's no need to distinguish between dishes that differ in that one ingredient. Just choose randomly! A random policy requires fewer bits than a deterministic policy, because you no longer need to remember which action to take in a particular state. Similarly, it's not worth spending bits on a policy for the teenager who rarely shows up at dinner; you can safely compress your policy by choosing randomly. These examples illustrate the concept of *precision reduction*: compression can be achieved by forgetting distinctions that don't matter.

The idea of compression has played an important role in theories of short-term memory (Miller, 1956; Brady et al., 2009; Nassar et al., 2018; Mathy and Feldman, 2012; Norris and Kalm, 2020), but until recently it has been comparatively neglected in theories of action selection. Despite this neglect, we will show that many aspects of action selection (stochasticity, perseveration, response time, and chunking) can be viewed as forms of policy compression. To set the stage, we will begin by introducing a general information-theoretic framework for understanding compression, adapted from applications to memory and perception research (Gershman, 2021; Sims, 2016). This will allow us to derive the optimal policy under a given capacity constraint and deduce empirical predictions from this policy.

2 Action selection as a communication channel

At first glance, it is somewhat counter-intuitive to think of action selection in terms of communication; in what sense are actions communicating anything? To understand why this makes sense, let us first consider memory more broadly as a channel for transmitting information about the past for use in the future. In the same way, selecting an action requires transmitting information about states to guide future action.

In our running example, this means that while you are cooking, you must be able to remember *who* will be at dinner (state) in order to know what dish (action) to cook for each family member. Similarly, cooking a particular dish (action) provides information about the particular family member (state) you are serving. Note that the current state might include features of the past (e.g., stimuli, actions, rewards, etc.), and hence it is appropriate to think of the state as a kind of historical record or summary statistic.[1] (For example, a "feature" of your teenager is that they rarely show up at dinner.)

Unfortunately, our brains are *not* perfect at remembering all of the state information needed to guide action. A distracted or tired chef might misremember who wants to eat what dish, and may even confuse

[1] In reinforcement learning theory (Sutton and Barto, 2018), "state" has a technical meaning related to Markov decision processes: the state is a sufficient statistic for predicting future states and rewards. In other words, an agent can forget the past once it knows what state it's currently in.

preferred dishes between family members. In other words, there are capacity limits on memory that constrain the amount of information that can be faithfully transmitted. Viewing action selection as a communication channel allows us to formalize these capacity limits using the language of information theory. From this foundation we can analyze the structure of optimal policies under capacity limits, and derive practical algorithms for policy compression.

As diagrammed in Figure 1A, the channel input is generated by a distribution $P(s)$ over states (s). The channel encodes each state into a *codeword* c; this is the step at which compression occurs, as we discuss below. Conditional on the codeword, the channel selects an action a according to $P(a|c)$. Taken together, these two steps (encoding and action selection) produce the policy $\pi(a|s)$ mapping states to actions. An illustration of these steps using our running example is shown in Figure 1B.

Compression can be quantified in terms of the state's description length, the length of the codeword. Since we can always translate the codeword into binary strings of 0's and 1's (bits), we can compare the description lengths of codewords in units of bits. The channel capacity places a limit on the average description length of codewords. In particular, the minimum number of bits needed for error-free transmission of the state identity is given by the mutual information between states and actions (Shannon, 1948):

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

In the context of action selection, we will refer to the information rate as the *policy complexity* because policies that are more highly state-dependent (i.e., action probabilities vary to a greater degree across states) require on average more bits to encode. For instance, a policy in which you have to make different dishes for each family member is more complex than one in which you can make the same dish for everyone.

There are a number of complementary ways to understand policy complexity. Suppose you observe only the inputs (states) to the channel; how much do these observations help you predict the outputs (actions)? Policy complexity is mathematically equivalent to the reduction in uncertainty about the actions conditional on the states:

$$I(S;A) = H(A) - H(A|S), \tag{2}$$

where

$$H(A) = -\sum_a P(a) \log P(a) \tag{3}$$

is the entropy of the marginal action distribution, expressing uncertainty about the actions prior to observing the states, and

$$H(A|S) = -\sum_s P(s) \sum_a \pi(a|s) \log \pi(a|s) \tag{4}$$

is the conditional entropy, expressing the uncertainty about the actions after observing the states, averaged over the action distribution. Intuitively, knowing *who* a distracted (low capacity) chef is currently cooking for provides relatively little information about *what* they are going to serve.

We can also understand policy complexity as the degree of uncertainty reduction about the inputs conditional on the outputs:

$$I(S;A) = H(S) - H(S|A). \tag{5}$$

In other words, if I only observe an agent's actions, policy complexity measures how well I can infer the unobserved state driving those actions. Watching a distracted chef prepare dishes provides relatively little information about *whom* they are about to serve.

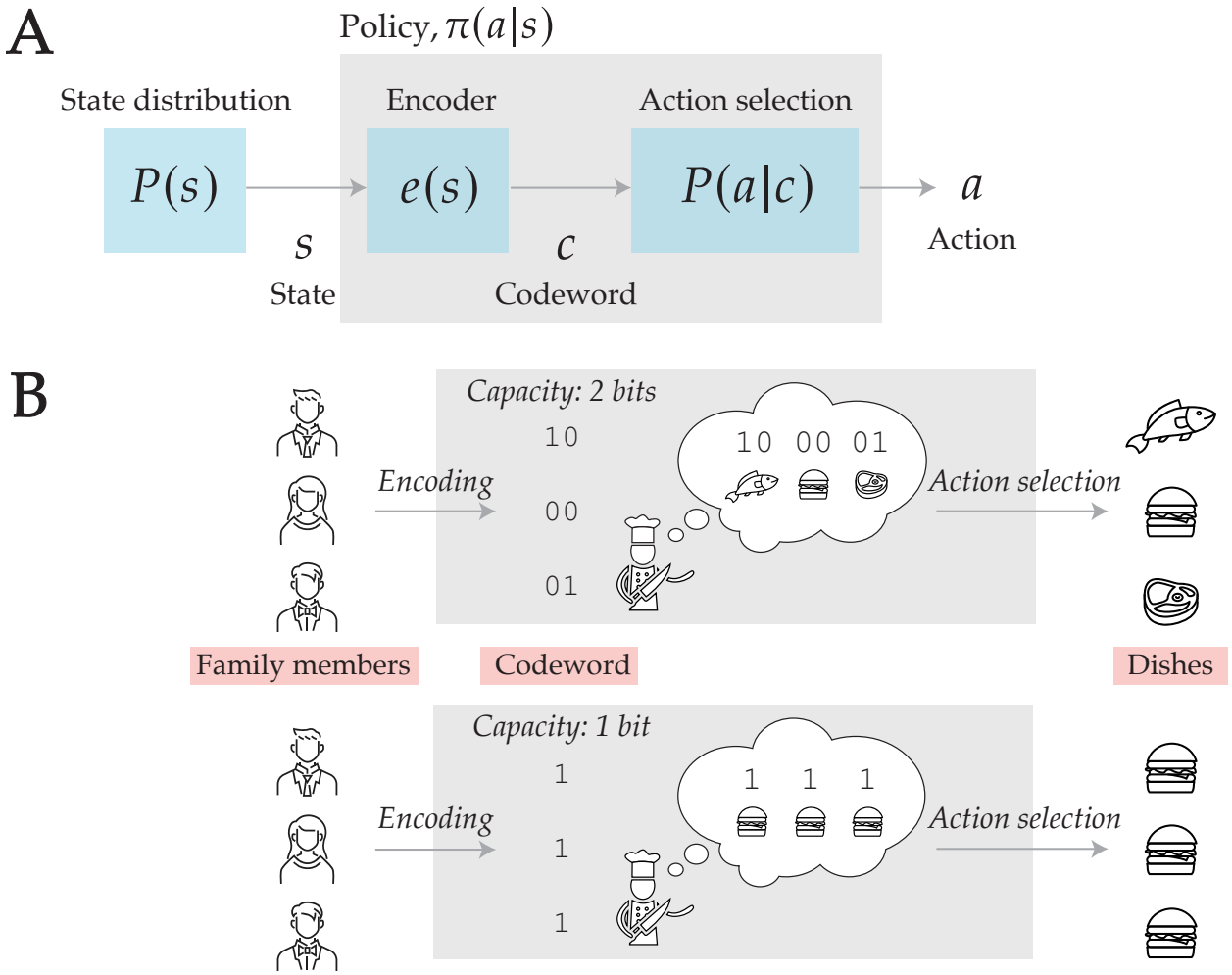

Figure 1: **The policy as a communication channel.** (A) We start with an distribution $P(s)$ over states (s). Each state is encoded into memory via an *encoder*, $e(s)$, yielding a codeword c. The codeword is then mapped onto an action a according to $P(a|c)$. Together, encoding and action selection produce the policy $\pi(a|s)$ mapping states to actions. (B) Imagine three family members (states) at dinner that a chef must account for. Each family member can be represented by a codeword, whose length is constrained by the chef's capacity limit (measured in bits). This corresponds to the amount of memory they have allocated to remembering relevant information about family members (e.g., whether they have a preferred dish, have dietary restrictions, etc.). The codeword representations are used to select the dishes (actions) that the chef will cook. (Top) In the case where the chef has 2 bits of memory at allocate, they can differentiate between family members' food preferences. (Bottom) However, when the chef can only allocate 1 bit of memory to the family member-dish mapping, they do not differentiate between food preferences and end up making the same dish for everyone.

3 Compression as a trade-off between reward and complexity

In this section, we consider the channel design problem: how should the brain optimize for accurate encoding and decoding with the goal of minimizing distortion (or equivalently, maximizing reward), given capacity limitations? First we need to clarify what it means to be optimal. If the goal is lossless (error-free) transmission, then the goal is to choose an encoder that achieves the Shannon bound, with average description length equal to the information rate. If there is no channel noise, the inputs can be unambiguously inferred from the outputs, and hence the conditional entropy $H(S|A)$ is 0. According to Eq. 5, the policy complexity is then equal to the source entropy, $H(S)$. The shortest average description length is thus also equal to the source entropy.[2]

The optimal error-free code under noiseless transmission can be achieved by a family of algorithms known as *entropy coding*, which assign codewords to each state s; the codeword's length (in bits) is equal to the state's *surprisal*, $-\log P(s)$. In fact, the expected surprisal is equal to the source entropy, thus satisfying the Shannon bound. The canonical algorithm for entropy coding is Huffman coding (Huffman, 1952), which constructs a binary tree whose leaf nodes correspond to input symbols (states). The binary code for each state can be thought of as a sequence of instructions for traversing the tree and terminating at a leaf node to reveal the encoded state. This decoding procedure has interesting implications for understanding response time, as we will see later.[3]

The basic problem with error-free transmission as a theory of action selection is that the brain is not, and cannot, be error-free. As was recognized long ago by Von Neumann (1958), the brain is a low precision communication system, corrupted by many sources of noise (see Faisal et al., 2008, for a contemporary review). Thus, a compression scheme like Huffman coding, which eliminates all redundancy, is *prima facie* implausible. In unreliable communication systems like the brain, redundant bits are needed to correct transmission errors (Bhui and Gershman, 2018; Tkačik et al., 2010). However, if capacity is too low, there will not be sufficient bits available to correct all errors. This raises the question: how should the brain allocate bits when there are not enough to go around?

Rate-distortion theory was developed to answer this question (Berger, 1971). The key idea is that bits should preferentially go to transmitting information that matters. Applied to action selection, the theory stipulates a distortion function $d(s, a)$ that measures the cost of outputting a when the state is s.[4] The optimization problem is to minimize the expected distortion $D = \mathbb{E}[d(s, a)]$ subject to a constraint on the information rate, or equivalently minimize the information rate subject to a constraint on the expected distortion. In the context of action selection, it is often more natural to work with the reward $Q(s, a)$, the mirror image of the distortion. The application of rate-distortion theory to action selection has been developed theoretically by a number of different authors (Tishby and Polani, 2011; Parush et al., 2011; Lerch and Sims, 2018; Fox et al., 2015; Still and Precup, 2012; Grau-Moya et al., 2018). Our summary of these ideas is condensed and simplified (for example, we do not address sequential decision problems), so the interested reader is referred to these papers for more technical details.

It can be shown that the highest achievable expected reward for a given capacity constraint is a monotonically increasing and concave function of policy complexity (see Figure 3). There is thus a trade-off between reward and policy complexity: greater compression of states (lower complexity) can only be achieved at the expense of reward. In subsequent sections, we will explore the empirical implications of this trade-off.

A number of these implications can be deduced from the functional form of the optimal policy:

$$\pi^*(a|s) = \frac{\exp[\beta Q(s, a) + \log P^*(a)]}{\sum_{a'} \exp[\beta Q(s, a') + \log P^*(a')]}. \tag{6}$$

[2]This is an informal statement of Shannon's source coding theorem (Shannon, 1948).

[3]See Brady et al. (2009) and Norris and Kalm (2020) for further examples of studies that use Huffman coding as a psychological model.

[4]Traditionally, the channel output is conceived as a reconstruction of the state, so that distortion refers intuitively to the difference between the input and reconstruction. Thus, the term "distortion" is somewhat confusing in the context of action selection, where we take it to mean the cost function.

This is the familiar softmax equation, ubiquitous in studies of reinforcement learning, psychophysics, and econometrics. The parameter β is commonly referred to as the *inverse temperature*, and controls the degree of stochasticity; as β increases, the policy concentrates on the action with highest reward. In rate-distortion theory, β has another interpretation; its inverse (the temperature) is the slope of the reward-complexity trade-off function:

$$\beta^{-1} = \frac{dV}{dI(S;A)}, \tag{7}$$

where $V = \mathbb{E}[Q(s,a)]$ is the expected reward. The second term inside the softmax,

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

captures a form of perseveration, a bias towards actions that are chosen frequently across all states. Low complexity policies compress the state and therefore cannot distinguish between policies for different states. As a consequence, the optimal policy will be close to the marginal distribution over actions, ignoring the state (see empirical examples from Figure 3).

How can we actually compute the optimal policy? Notice that the perseveration term depends on the optimal policy, so there is a circularity in the definition. The Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972) harnesses this circularity by iterating between computing $P(a)$ and $\pi(a|s)$, a process that will converge to the optimal values.

One drawback of the Blahut-Arimoto algorithm is that computing $P(a)$ requires a marginalization over the entire state space. If the state space is very large or continuous (as in many real-world environments), this marginalization might be computationally intractable. In the Appendix, we derive a tractable algorithm based on a reinforcement learning formulation (see also Gershman and Lai, 2020). This algorithm incrementally modifies the policy based on reward feedback, with the critical property that the agent is penalized for complex policies. While the algorithm will eventually converge to the optimal policy, its incremental nature means that it will spend a non-trivial amount of time away from the optimal trade-off curve. We will later show that this allows it to explain empirical deviations from the optimal curve.

4 Behavioral signatures of policy compression

In the following sections, we review behavioral phenomena that can be interpreted in terms of policy compression. For some of these phenomena, we present illustrative simulations using the process model presented in the Appendix. Our goal is synthetic rather than discriminative: although each individual phenomenon may be explained by alternative theories, we make the case that they can be understood collectively as reflections of a single underlying principle.

4.1 Stochasticity

Imagine that you are shopping for wine at the grocery store. You don't know much about wine, so you select one at random and later discover that you like it. However, when you return to the wine aisle the next week, you decide to randomly select a different wine instead of picking the one you chose last time. This example illustrates the stochasticity inherent in our daily choice behaviors.

Why are human actions stochastic even when faced with the same choice options? There are a number of different answers to this question (see Icard, 2019). One conventional answer in the framework of reinforcement learning is that stochasticity facilitates exploration (Schulz and Gershman, 2019). In order to identify the most rewarding action, an agent must sample multiple actions, and injecting randomness into the policy (e.g., via the softmax equation) is a simple way to accomplish this. However, the exploration perspective does not explain why action selection is apparently stochastic even in situations where the payoffs and probabilities are fully known (e.g., Mosteller and Nogee, 1951). In economics, a standard explanation of

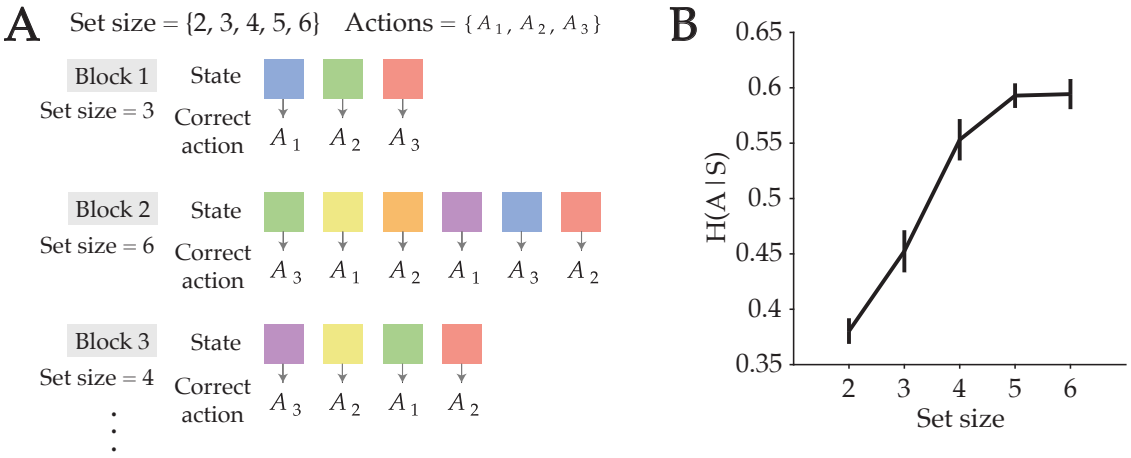

Figure 2: **Stochasticity as a function of set size.** (A) In the task developed by Collins and Frank (2012), subjects saw a single stimulus on each trial and chose between 3 actions. Each stimulus had a single rewarded action. The number of states, or set size, varied between 2 and 6 across blocks. (B) Conditional entropy as a function of set size in data from healthy controls in Collins et al. (2014). Error bars show standard error of the mean.

stochasticity is the hypothesis that actions are evaluated based on a random utility function that specifies a level of inherent randomness when sampling actions in proportion to their expected payoffs. But this just pushes the question back further: why is the utility function random?

The results from rate-distortion theory presented earlier provide a different foundation for stochastic action selection. Any capacity-limited agent must be stochastic if they are operating at the optimal reward-complexity frontier. Remarkably, this optimal stochasticity should take the form of a softmax policy, consistent with the way in which psychologists and economists have long modeled action selection.[5] In the economics literature, this insight was derived within the framework of "rational inattention" (Matějka and McKay, 2015), which is mathematically equivalent to the Lagrangian formulation of the rate-distortion objective function (see Appendix and Denti et al., 2019).

Rate-distortion theory not only rationalizes the softmax policy; it also links the temperature (commonly treated as a free parameter that is fit to data) as a function of the reward-complexity trade-off. Specifically, Eq. 7 shows that the temperature corresponds to the slope of the reward-complexity trade-off function. Because the trade-off function is concave, and hence the slope is monotonically decreasing in the policy complexity, the optimal policy is more stochastic for low complexity policies.

One implication is that increasing cognitive load, which should reduce policy complexity or force a fixed complexity to be distributed across more states, will cause action selection to be more stochastic. Collins and Frank (2012) developed an experimental paradigm, a kind of "contextual multi-armed bandit" task, that allows us to test this hypothesis (Figure 2A). In their experiment, subjects were shown a stimulus (the state variable in our terminology) and were tasked with selecting one of 3 actions. If they chose the correct action, they were rewarded. Critically, the number of states (the set size) was manipulated across blocks. Collins and Frank reported that performance degraded with set size. If an agent's policy complexity decreases with set size, or if the policy complexity is distributed across more states, then we should expect higher stochasticity for larger set sizes. Using the conditional entropy $H(A|S)$ as a measure of stochasticity, this prediction is confirmed (Figure 2B).[6]

[5]The softmax policy is known as the multinomial logit policy in economics; see McFadden (2001).

[6]This analysis used data from a follow-up study (Collins et al., 2014), which we have previously re-analyzed (Gershman and Lai, 2020).

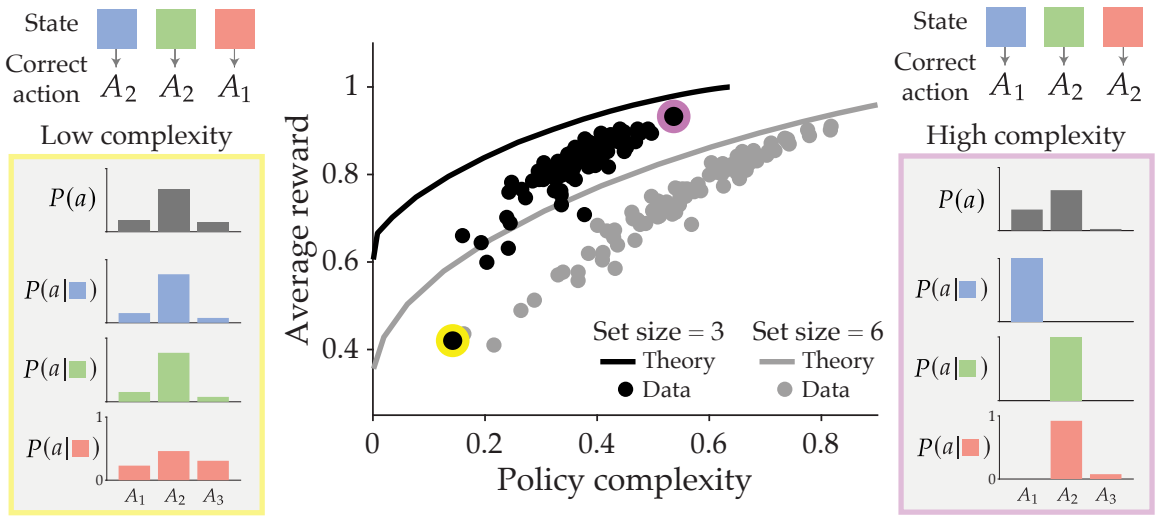

Figure 3: **The reward-complexity trade-off describes the optimal degree of policy perseveration.** Adapted from Gershman (2020). (Middle panel) Theoretically optimal reward-complexity trade-off curves (solid lines) for a particular set size (3 or 6). Each data point represents a subject's performance aggregated across blocks of the same set size. (Left panel) Marginal and conditional action probabilities from an example subject (set size = 3) with a low complexity policy. Notice the similarity between the marginal action probability $P(a)$ and the policies conditioned on different states. (Right panel) An example subject with a high complexity policy. Notice the deviation of the state-dependent policies from the marginal action probability.

4.2 Perseveration

The tendency to perseverate on past policies, even when inappropriate, is ubiquitous. It has been observed in studies of operant conditioning (Thorndike, 1911; Lau and Glimcher, 2005), perceptual decision making (Verplanck et al., 1952; Howarth and Bulmer, 1956; Fründ et al., 2014), and choice reaction time (Bertelson, 1965). Perseveration has often been viewed as a kind of irreducible nuisance factor, or as a reflection of habit formation (Dickinson, 1985; Miller et al., 2019). What remains unclear in these accounts is why perseveration should happen at all. Is there a computational logic underlying its existence?

The capacity-limited optimal policy provides one possible answer. Any capacity-limited agent operating at the optimal reward-complexity frontier will exhibit a bias to take actions that have been frequently chosen in the past. This bias comes from the $\log P^*(a)$ term in Eq. 6. The influence of this bias on the overall action policy will depend on the agent's capacity limits, reflected in the value of β.

Gershman (2020) argued that policy complexity can be interpreted as an inverse measure of perseveration, because it is higher to the extent that state-specific policies diverge from the marginal policy. When this divergence is low, it means that states exert a weak degree of control over actions, and hence there is a tendency to choose actions with the same probability across all states (see Figure 3, left panel). Given this interpretation of policy complexity, the empirical reward-complexity trade-off function tells us whether a particular degree of perseveration is optimal given a particular agent's capacity limit. Specifically, we can say that the agent is optimal if their empirical trade-off coincides with the optimal trade-off for a given policy complexity.

Gershman (2020) estimated the empirical trade-off function using data from a version of the contextual bandit task discussed above (Collins, 2018). Overall, subjects were close to the optimal trade-off curve, though subjects with low policy complexity exhibited a systematic deviation (Figure 3). This deviation was

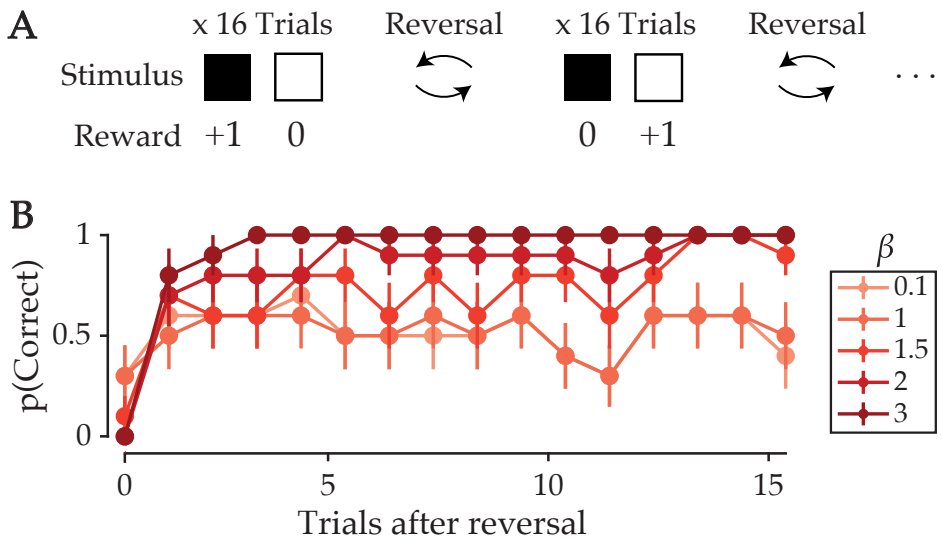

Figure 4: **Simulated reversal learning under varying capacity constraints.** (A) In a serial reversal learning task, one stimulus is rewarded while the other is not. After a set number of trials (or when a performance criterion is reached), the reward contingencies reverse. (B) Proportion of trials where the "correct" choice was made as a function of the number of trials after reversal. Performance is averaged across 10 reversals. Trial 0 indicates the first trial after the reward contingencies reverse; therefore, agents should still respond according to the contingencies in effect prior to reversal.

later explained by Gershman and Lai (2020) as a possible consequence of suboptimal learning: subjects with lower learning rates, as estimated using the actor-critic model presented in the Appendix, showed a greater deviation.

Gershman (2020) also investigated whether the pattern of perseveration in this data set followed the functional form in Eq. 6. While it is common for separate coefficients to be estimated for the value and perseveration terms that are entered into the softmax, the optimal policy suggests that only a single coefficient (β) is necessary. Consistent with this hypothesis, a model with a single coefficient outperformed a model with two coefficients.

Perseveration of policies can often be observed as behavioral inflexibility in response to a change in the environment. For example, consider the widely-used serial reversal learning task, in which selecting one of two stimuli is reinforced while the other is not (Figure 4A). After a subject has learned to consistently choose the reinforced stimulus, the reward contingencies reverse. In this scenario, perseveration would correspond to the inability to adapt choice behavior after reversal.

Hassett and Hampton (2017) showed that monkeys trained on this task were slower to adapt to reversals when working memory was taxed by manipulations of cognitive load. This result suggests that working memory is critical for behavioral flexibility as measured by reversal speed, a finding consistent with the idea that perseverative action biases are modulated by capacity limits on memory. This kind of perseveration has also been quantified in rule-switching tasks such as the Wisconsin card sort task (Berg, 1948) and the dimensional card sort task (Zelazo, 2006), where perseveration is not simply a repetition of earlier actions but repetition of an entire policy.

To explicitly demonstrate the relationship between perseveration and capacity limits on memory, we simulated the performance of five agents with varying capacity constraints on a generic serial reversal learning

task (Figure 4A).[7] If behavioral flexibility varies as a function of capacity constraint, we should expect to see faster rule reversals for high capacity agents, and "stickier" choice behavior for low capacity agents, indicating the perseverative influence of a previously-learned policy. Indeed, we see that high capacity agents are faster at learning the new reward mapping than low capacity agents, attaining 100% accuracy just a few trials after reversal has occured (Figure 4B).

Notice that the "state" in this task is not the stimulus (as it was in the previous example), but rather, the underlying reward contingencies or context. This again underscores the definition of perseveration as the degree of state-independence of the policy: low capacity agents will not be able to distinguish between different reward contexts and therefore will not perform reversals. At the lowest extreme (e.g., $\beta = 0.1$), agents do not even learn to select the rewarded stimulus due to the strong bias of the marginal action probability (which we assume be uniform across actions at the beginning of learning). This causes an agent to continue responding stochastically (e.g., choosing both stimuli with equal probability) throughout the entire task. As the capacity constraint increases, agents become better able to distinguish between reward contexts, allowing them to detect when the underying state has changed. When the stimulus that was once associated with reward is suddenly no longer rewarded (in Figure 4B, we take this to be Trial 0 after reversal), agents are aware that a state change has occured and reverse their action policy accordingly.

We have shown that the tendency to perseverate on past policies can be explained by limitations on cognitive capacity: when capacity is low, policies will be biased by the actions taken most often in recent history. When capacity is high, policies will be more sensitive to the current state. Our framework allows us to recast perseveration in the light of an optimal solution under a given capacity limit instead of treating it as a suboptimal behavioral nuisance.

4.3 Response time

In a seminal paper, Hick (1952) studied response times in a task where subjects made speeded responses to one of N possible targets. Hick found that mean response time was an approximately logarithmic function of the number of targets, a regularity now known as *Hick's law* (see Proctor and Schneider, 2018, for a recent review). Hick's law holds not only for target selection (commonly known as *choice reaction time* tasks), but also for the contextual multi-armed bandit task studied by Collins and colleagues (Collins and Frank, 2012; Collins et al., 2014; Collins, 2018), as shown in Figure 5: mean response time is well approximated by a linear function of log set size (see also McDougle and Collins, 2020, for more detailed analyses of response times in this task).

Hick used an information-theoretic analysis to derive his law. Recall from our discussion of entropy coding that the optimal description length for state s in a noiseless channel is $-\log P(s)$. If each state is equally likely (as in Hick's experiments), then $P(s) = 1/N$ and the optimal description length is $\log N$. In a Huffman code, this corresponds to the number of bits that need to be inspected to reveal the coded state. Thus, if we assume that bits are inspected at a constant rate, we arrive at Hick's law.

More generally, response time should be a linear function of the description length, which can be manipulated even when the number of states is held fixed (Hyman, 1953). Following this logic, we reasoned that people with lower capacity should have longer description lengths (higher policy complexity) and therefore longer response times. Consistent with this prediction, policy complexity was significantly correlated ($r = 0.35, p < 0.0001$) with mean response time across subjects in the data from Collins (2018).

A number of studies have found that the slope of the set size function decreases with practice (Hale, 1968; Mowbray and Rhoades, 1959; Wifall et al., 2016; Teichner and Krebs, 1974). One explanation for this finding is that optimal compression depends on knowing the state probabilities, which must be learned. Studies of compression effects in short-term memory have shown that these effects emerge over the course of training (Brady et al., 2009; Ngiam et al., 2019).

[7] In the rest of the simulations, we specify values of β to imply given capacity limits (see Eqs. 6 and 7). This allows us to evaluate behavior as a function of capacity under the assumption that β implicitly defines a point on the reward-complexity trade-off curve. Alternatively, the value of β can be learned from experience. We leave this possibility as an open question for future research.

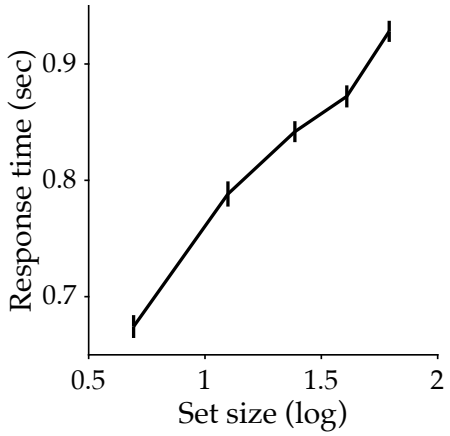

Figure 5: **Average response time as a function of log set size**. Data taken from healthy controls in Collins et al. (2014) performing the simulus-response task depicted in Figure 2A. Error bars show standard error of the mean.

As mentioned earlier, algorithms like Huffman coding seem to require implausibly high precision relative to the level of noise in the brain. This point is relevant to Hick's law, because some studies indicate that the set size function flattens out for very large set sizes, deviating from the law (Seibel, 1963; Longstreth, 1988). This would make sense if there is a limit to the number of bits that can be assigned to each codeword, which would constitute a lossy code if the limit is sufficiently low.

A final point about Hick's law, and set size effects more generally, is that these effects are only observed when a policy needs to be retrieved from memory (Proctor and Schneider, 2018). For example, response time increases with set size if subjects have to move a cursor to target locations indicated symbolically by letters, but not if the target location corresponds to the location of the stimulus (Dassonville et al., 1999). This observation is broadly consistent with our argument that the memory demand of policies acts as an information bottleneck in action selection.

4.4 Action chunking

Imagine you are learning to make coffee. You carefully measure out the beans, grind them in a coffee grinder, put the grounds into your coffee maker, and press a button. Over time, you notice that you get faster and faster at making coffee, and eventually, it requires little thought or attention—you could do it with your eyes closed. This example illustrates the idea that learned sequences of actions eventually become automatic and quick to execute.

Many skills in our everyday lives, such as making coffee, are learned by repeatedly sequencing actions in the service of a desired goal. The action sequence can become automatic such that once started, it must be brought to completion. These observations underlie the defining characteristics of action "chunks"—the reflexive association of a number of independently produced actions into rapidly executed action sequences (Dezfouli and Balleine, 2012; Lashley, 1951; Sakai et al., 2003; Botvinick, 2008).

Examples of action chunking are ubiquitous, especially in the sequence learning literature (Terrace, 1991; Verwey, 1999; Sakai et al., 2003; Miyapuram et al., 2006). Animals show evidence of action chunking in the formation of habitual action sequences (Graybiel, 1998a; Jin and Costa, 2010; Jin et al., 2014). In the machine learning literature, action chunking is closely related to "options" (Precup et al., 1998; Precup, 2000), or temporally extended action sequences that allow agents to plan more efficiently and accurately.

In an important paper on sequence learning, Sakai et al. (2003) showed that human subjects can learn a visuomotor sequence by spontaneously chunking elementary movements together, where each chunk acts as a single action unit. They showed that execution time decreased as a function of sequence repetition during

learning. To test for the formation of chunks, they shuffled the visuomotor sequences and found that the performance on a shuffled sequence was both faster and more accurate when the action chunks in the original sequence were preserved (within the shuffled arrangement) compared to when they were destroyed. These results are also consistent with evidence that people reuse learned action chunks, even when the chunks are suboptimal for the task at hand (Huys et al., 2015).

Action chunks are advantageous because they allow for the production of rapid action sequences without having to rely on the selection of individual elements. In this view, an action chunk can be treated as a single response such that selecting a familiar movement pattern only involves a single processing step. It has been suggested that the formation and expression of action chunks provides a mechanism for the execution of action repertoires that would otherwise be too biologically costly to implement (Graybiel, 1998b; Ramkumar et al., 2016). In this way, chunks reduce the amount of memory necessary to execute a sequence of actions by effectively compressing state information. Consistent with this idea, some studies have demonstrated a relationship between spatial working memory capacity and the learning of new action chunks (Bo and Seidler, 2009; Seidler et al., 2012).

Why might action chunks cost less from an information processing perspective? Recall that the complexity of an action policy depends on the degree to which it is state-dependent. In other words, we can think of policy complexity as quantifying the amount of memory that must be devoted to the state information when selecting actions. When the sequence of actions involved in making coffee becomes an action chunk, there is no longer a need to pay close attention to the "states" associated with each action (e.g., the particular brand of coffee beans being used, the grinder setting, and particular coffee maker). Furthermore, this sequence of actions would be the same regardless of whether you were making coffee at home or at your friend's home (assuming that you aren't particularly picky about your choice of coffee beans). Every morning, this familiar sequence is initiated by the first action of picking up your bag of coffee beans, and brought to completion without needing to process state information beyond noting where the coffee beans are stored.

In the following example, we show that multi-step action sequences are preferred when there is an imperative to reduce policy complexity. Specifically, as capacity is reduced, the preference for multi-step action chunks should increase, while the action execution time should decrease. This is because for low complexity agents, it is more cost- and time-effective to select an action chunk of length n than it is to select n independent actions.

Imagine a task environment in which every state cues a specific rewarded action (Figure 6A). States appear in sequence, and an agent's task is to select the correct action as quickly as possible.[8] Agents can select from five different actions: four "independent" ones, as well as one action "chunk" (Figure 6B).[9] Following previous work suggesting that an action chunk acts as a single action unit (Graybiel, 1998b; Sakai et al., 2003), we also assume that it takes an equal amount of time to execute an independent action as it does to execute an action chunk. If a chunk of length n is selected on trial t, the actions in the chunk set will be executed until $t + n - 1$ (instead of using the policy to select actions). In our process model (as described in the Appendix), selecting an action chunk will inherently cost less because the agent does not need to factor in policy costs from time $t + 1$ to $t + n - 1$. This effectively means that agents can ignore all incoming state information until they are finished executing the action sequence.

If the task environment consists of predictable temporal relationships across states (e.g., if a certain sequence of states occurs over and over again), agents will naturally begin to select the same sequence of actions in response to this state sequence. In our specific example, agents frequently observe the same three state sequence (red, green, blue) within the "Train" block of the task (Figure 6C). Later, in a "Test" block inspired by the task in Sakai et al. (2003), this three state sequence is destroyed by shuffling the order in which states appear. This manipulation provides a way to observe how the reuse of learned action chunks when they are no longer advantageous varies as a function of the agent's memory capacity.

[8] This is similar to many tasks used to study motor skill learning, for example, the Discrete Sequence Production task (Verwey, 1999) or the Serial Reaction Time Task (Nissen and Bullemer, 1987; Robertson, 2007).

[9] We include only one possible action chunk in this example for simplicity, though we recognize that any combination of actions of any length could also be considered an action chunk. We also do not address how these chunks are learned, and leave that as an open question for future research.

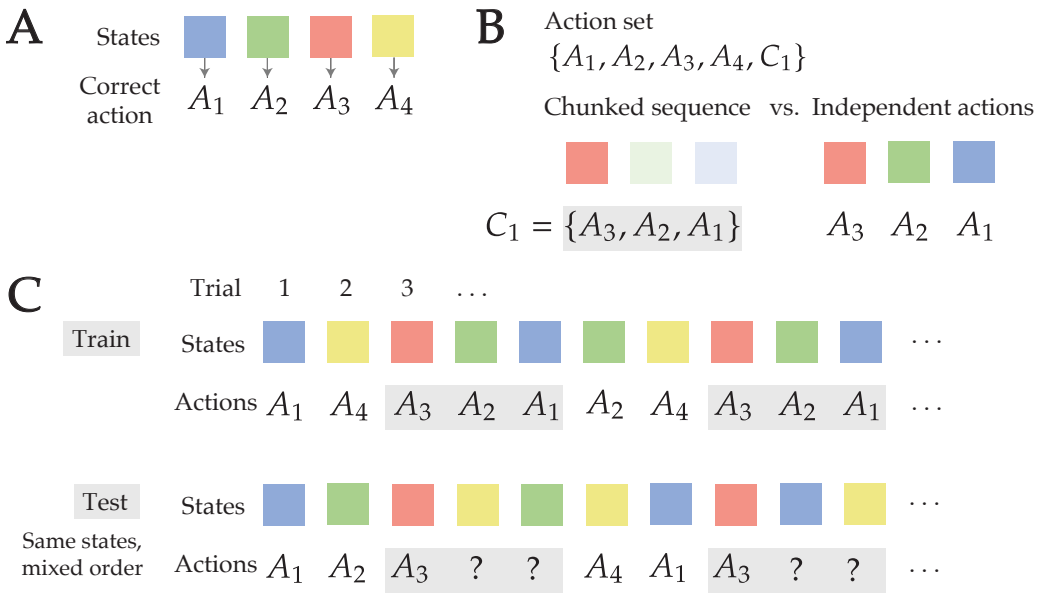

Figure 6: **A generic action chunking task.** (A) A stimulus-response task with four states and four corresponding actions. Selecting the correct action in each state leads to a reward of +1. (B) The valid action set. For simplicity, imagine that agents can either select independent actions (A_1, A_2, A_3, A_4), or select the action chunk C_1, which is composed of the action sequence: $\{A_3, A_2, A_1\}$. Notice that for the same three state sequence (red, green, blue), selecting three correct independent actions will lead to the same total reward (3) as if the action chunk was selected. However, in selecting the action chunk, agents no longer have to pay attention to the states following the chunk-initiating state (red). (C) Agents first learn the correct state-action pairings in a "Train" block. The task is designed such that a specific state sequence (red, green, blue) reoccurs often. Over the course of training, agents should recognize this state sequence and choose to select the action chunk C_1 in the red state instead of taking independent actions in each of the states occuring in sequence. In the "Test" block, agents are exposed to the same states as seen in the "Train" block, but in randomized order. Now that the reoccuring state sequence is eliminated, one can measure the degree of chunking learned in the "Train" block by observing the actions taken in the two states following the original chunk-initating state.

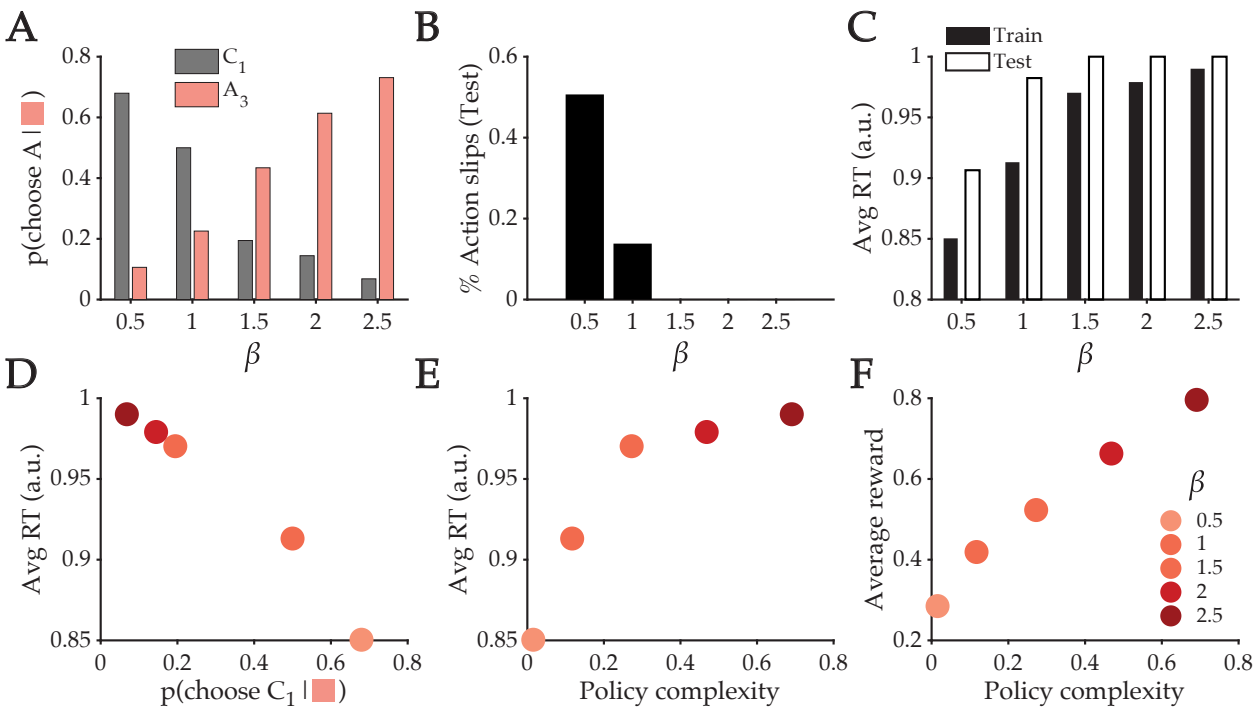

Figure 7: **Action chunking under resource constraints.** (A) Proportion of trials where A_3 or C_1 was chosen in the red state (see Figure 6). (B) The percentage of "action slips" in the "Test" block (quantified as choosing C_1 in the red state when it is no longer the optimal action) as a function of capacity, β. (C) Average response time (RT) per trial as a function of capacity. (D) Average RT as a function of the proportion of trials where C_1 was chosen in the red state. (E) Average RT as a function of the policy complexity. (F) The reward-complexity trade-off for different capacities.

Agents can either choose to select actions upon presentation of each state ("independent" selection of A_3, A_2, A_1), or can choose to select the action chunk that consists of three actions in sequence, $C_1 = \{A_3, A_2, A_1\}$. While these two action plans lead to the same total reward (3), selecting the action chunk frees agents from having to pay attention to the states following the chunk-initiating state (red). As noted previously, this reduces policy costs, since the agent can ignore some state information. Since there are fewer states to decode information from, the action sequence is also produced more quickly.

In Figure 7A, we show that the preference for selecting multi-step action chunks C_1 in the chunk-initiating red state decreases as a function of the capacity. This is consistent with the intuition that low capacity agents should prefer chunked action sequences because they compress information, while high capacity agents can afford to retain more state information (and therefore take state-specific actions). As stated earlier, one consequence of learning a preference for action chunks is that they are often reused even when this is suboptimal (Huys et al., 2015). To show this, we counted the number of "action slips" (cf. Norman, 1981), defined as the percentage of trials where the action chunk C_1 was taken in the chunk-initiating state when it was no longer optimal, and saw that it was higher at lower capacity constraints.

Additionally, we confirmed that the average response time per trial increased in the "Test" block across all capacity constraints (but especially for lower-capacity agents), reflecting the increase in processing time necessary when chunks are no longer being used. This means that the average response time overall should decrease as a function of the proportion of trials where the action chunk was selected (Figure 7D). In analog

to Figure 5, we also observe the RT increasing as a function of policy complexity (Figure 7E), consistent with the information-theoretic analysis of Hick's Law.

4.5 State chunking

Real-world environments contain a large number of states, and the brain's capacity limit means that it is typically not efficient to represent all states with the same fidelity. Especially in environments where states are correlated (e.g., if they lead to the same reward or policy), it can be advantageous and cost-effective to merge similar states into state "chunks."

State chunking confers several advantages. By generalizing across similar states, information learned in one state can be efficiently re-used and transferred across other states that share similar reward and transition structures (Abel et al., 2019; Lehnert and Littman, 2019; Tomov et al., 2020; Lehnert et al., 2020). Recent work has suggested that for long-term benefits, agents should focus on learning reward-predictive state abstractions (Lehnert and Littman, 2019; Lehnert et al., 2020), implying that if states share common futures, then they will become chunked together. However, there are also situations in which generalization may not be advantageous. For example, if the reward structure of an environment suddenly changes in a manner that is not supported by the learned state chunk, an agent will lose out on reward if it continues to follow the same policy.

In the following example, we will demonstrate that state chunking arises under the imperative to compress policies. Specifically, we will show that lower capacity agents are more likely to chunk states leading to similar reward, but are also less likely to adapt to changes in the reward environment. In contrast, higher capacity agents are able to flexibly and quickly adapt to changes in the environment in a state-specific manner.

Consider a task environment with three distinct states (blue, green, and red) and two actions (A_1 and A_2; Figure 8A). In the "Train" block, agents learn that taking action A_1 in the blue and green states and action A_2 in the red state leads to reward. Note that after training, the marginal probability $P(a)$ is biased towards A_1. In the "Retrain" block, agents learn a new reward mapping for the green state—now, taking action A_2 in the green state is rewarded. After retraining, the marginal probability $P(a)$ is now biased towards A_2. To quantify the amount of state chunking, we introduce a "Test" block where we measure the proportion of trials where an agent choses action A_2 in the blue state.

Figure 8B depicts the relative contribution of the state-action values and the marginal action probability to the final policy. Critically, action selection in the "Test" block will depend on the agent's given capacity limit as stipulated by β. If there is no need for compression (β is high), then the agent should respond to the blue state exactly as it did in the "Train" block. However, if the capacity constraint is low and there is a need to compress information (β is low), agents will end up relying on the marginal action probability, which allows them to ignore some state information. This means that the proportion of choosing action A_2 in the blue state in "Test" will be greater than in the "Train" block.

We will first examine learning in the "Retrain" block. Figure 8C shows the probability of choosing A_2 in the green state as a function of the capacity constraint (β). Note that agents with higher capacity constraints are better at adapting to the new reward structure in the "Retrain" block, choosing A_2 more often than in the "Train" block. Figure 8D shows that the change in proportion of trials where A_2 is taken is increasing as a function of capacity constraint, indicating a greater ability to adapt to the new reward structure at higher capacity constraints.

To understand how capacity constraints affect the chunking of the blue and green states, we can examine choice behavior in the "Test" block. As mentioned previously, if the blue and green states were chunked together in the "Train" block, we should see an increase in the probability of choosing A_2 in the blue state from "Train" to "Test" blocks. Figure 8E shows that the probability of choosing A_2 decreases as a function of the capacity constraint in the "Train" block, and remains roughly constant before decreasing in the "Test" block.

A nuanced result is shown in Figure 8F, where we show that the change in proportion of trials where A_2 is chosen is increasing and then decreasing as a function of capacity constraint. To understand this nuance, consider the lowest capacity agent ($\beta = 0.1$), who selects actions based on their marginal action

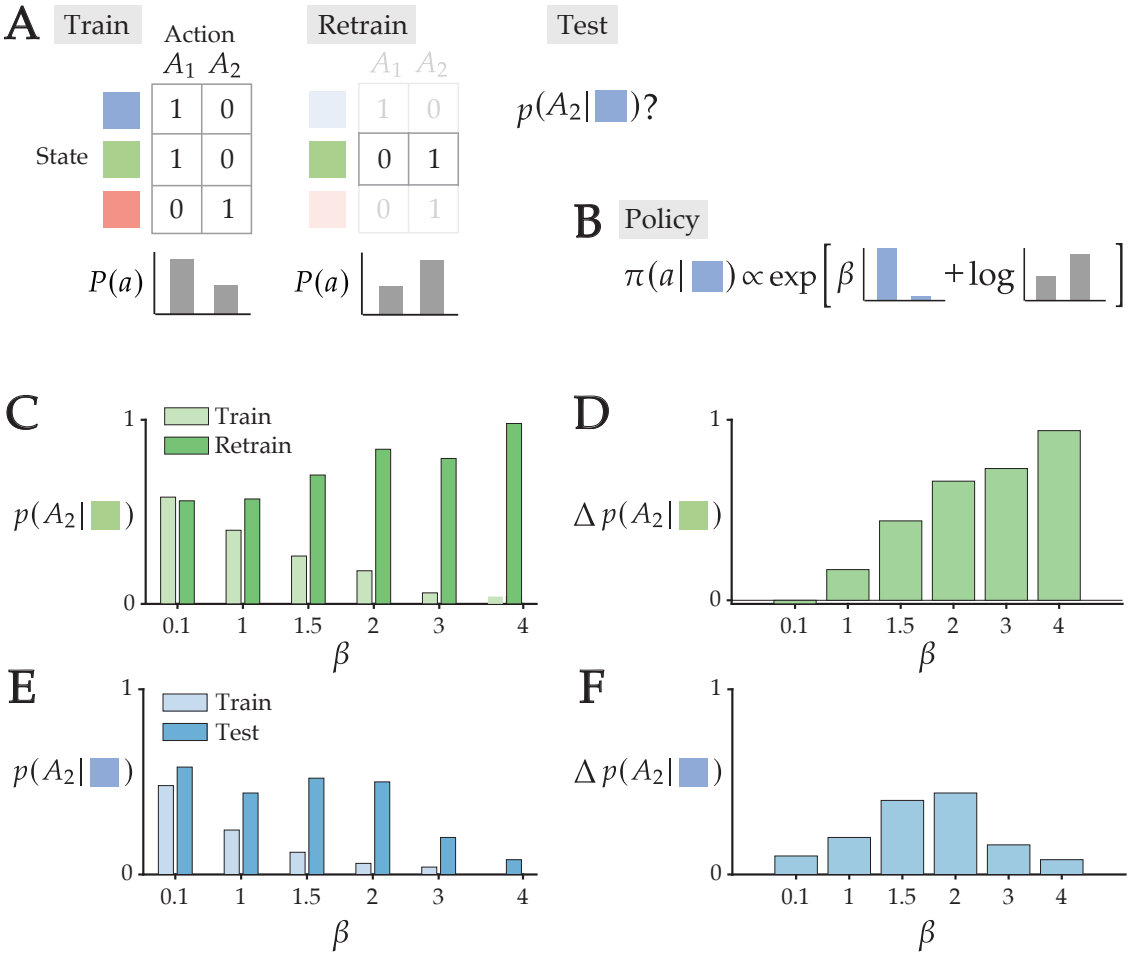

Figure 8: **State chunking under resource constraints.** (A) Consider a task with three distinct states (blue, green, and red) and two actions (A_1 and A_2). In the "Train" block, agents learn that taking action A_1 in the blue and green states and action A_2 in the red state leads to reward. After training, the marginal probability P(a) is biased towards A_1. In the "Retrain" block, agents learn a new reward mapping for the green state—now, taking action A_2 in the green state is rewarded. After retraining, the marginal probability P(a) is now biased towards A_2. To quantify the amount of state chunking, we introduce a "Test" block where we measure the proportion of trials where an agent choses action A_2 in the blue state. (B) A graphical depiction of the policy, decomposed into the state-action values and the marginal action probability. The relative influence of each component on action selection will depend on the value of β. (C) The proportion of choosing A_2 in the green state as a function of the capacity constraint for the "Train" and "Retrain" blocks. (D) The change in proportion of trials where A_2 is chosen in the green state as a function of capacity constraint (Δp = Retrain - Train). (E) The proportion of choosing A_2 in the blue state as a function of the capacity constraint for the "Train" and "Test" blocks. (F) The change in proportion of trials where A_2 is chosen in the blue state as a function of the capacity constraint (Δp = Test - Train).

probability $P(a)$, assumed to be uniform across actions at the beginning of learning. As a result, responding remains stochastic throughout both "Train" and "Test" blocks, so there is no significant change between the probability of choosing A_2 in the blue state across blocks. However, for intermediate capacity constraints, the policy is influenced by both the state-action values and the marginal action probability (Figure 8B). As capacity increases ($\beta = 1, 1.5, 2$), the marginal action probability $P(a)$ will become more peaked and influence the policy in the blue state, causing the agent to choose A_2 more frequently in "Test" relative to "Train." But as capacity continues increasing ($\beta = 3, 4$), the probability of choosing A_2 begins decreasing, because now the state-action values learned in "Train" alone should dominate the policy.

These results demonstrate that state chunking can arise out of a need to compress state information by merging states that lead to the same reward when there are limits on capacity. At one extreme, low capacity agents do not learn much at all, and their responding remains stochastic. At the other extreme, high capacity agents can perfectly learn and store state-dependent policies, and flexibly adapt to new reward structures. Between the two extremes, there is a dynamic range of behavior that varies depending on how much state-specific action information an agent can afford to store in memory.

4.6 Navigation

Navigation is a form of goal-directed behavior that requires an organism to plan complex sequences of actions towards a desired location. Maze tasks such as the Morris Water Maze, radial arm maze, and T-maze have been used in the animal learning literature to study navigation, and behavioral performance is typically measured by the amount of time it takes subjects to reach a goal location. However, this performance criterion ignores the complexity cost of a particular action trajectory, which is related to how *state-specific* a trajectory is.

To illustrate this, imagine that you have the choice of driving to two grocery stores, one that is closer and one that is farther away. While driving time alone might suggest that the closer store is preferred, it may not take into account the fact that driving to the closer store requires you to turn at a number of intersections (high policy complexity), while driving to the store that is farther away would only require staying on a highway for most of the trip (low policy complexity). The former route requires you to remember more about the specifics of your current location (or state) in order to successfully navigate to your final destination, while the latter does not. This example highlights the fact that the state-specificity of a navigational policy incurs a computational cost that must be taken into account by behaving agents.

In a recent study, Amir et al. (2020) used the reward-complexity trade-off to analyze the learning processes of mice navigating to a hidden platform in the Morris Water Maze task. Specifically, they quantified the trade-off between the value and complexity of an animal's swimming trajectories across four days of learning. The value of a swimming trajectory was related to its energy cost and was correlated with swimming time, while the complexity of a trajectory provided a measure of the computational cost needed to generate specific, goal-directed motor plans at any given location. Complexity was measured relative to the swimming trajectories of naive animals, which by definition had the lowest policy complexities. Trained animals exhibited swimming trajectories that were often shorter and more direct, which increased the value of the trajectory but also incurred a higher complexity cost. This increased cost reflects the fact that trained animals took into account their current location or state on a moment-by-moment basis in order to orient their swimming direction towards the goal platform.

Animals tended to optimize for value early in learning while reducing policy complexity later in learning (by finding less costly trajectories that maintained the same value). This corresponds to movement *along* the optimal reward-complexity curve over the course of training. To summarize the learning state of an animal at a given time, Amir et al. (2020) fit values of β to the data from each day and found that it increased over the four days of learning.[10] The trade-off between value and complexity effectively captured the behavioral

[10]In our simulations, we have chosen to assume a fixed value of β for simplicity. But as this study suggests, β is likely to evolve over the course of learning as agents move along the optimal reward-complexity frontier. Future work should investigate how β can be learned depending on the optimal policies and reward statistics unique to particular task environments.

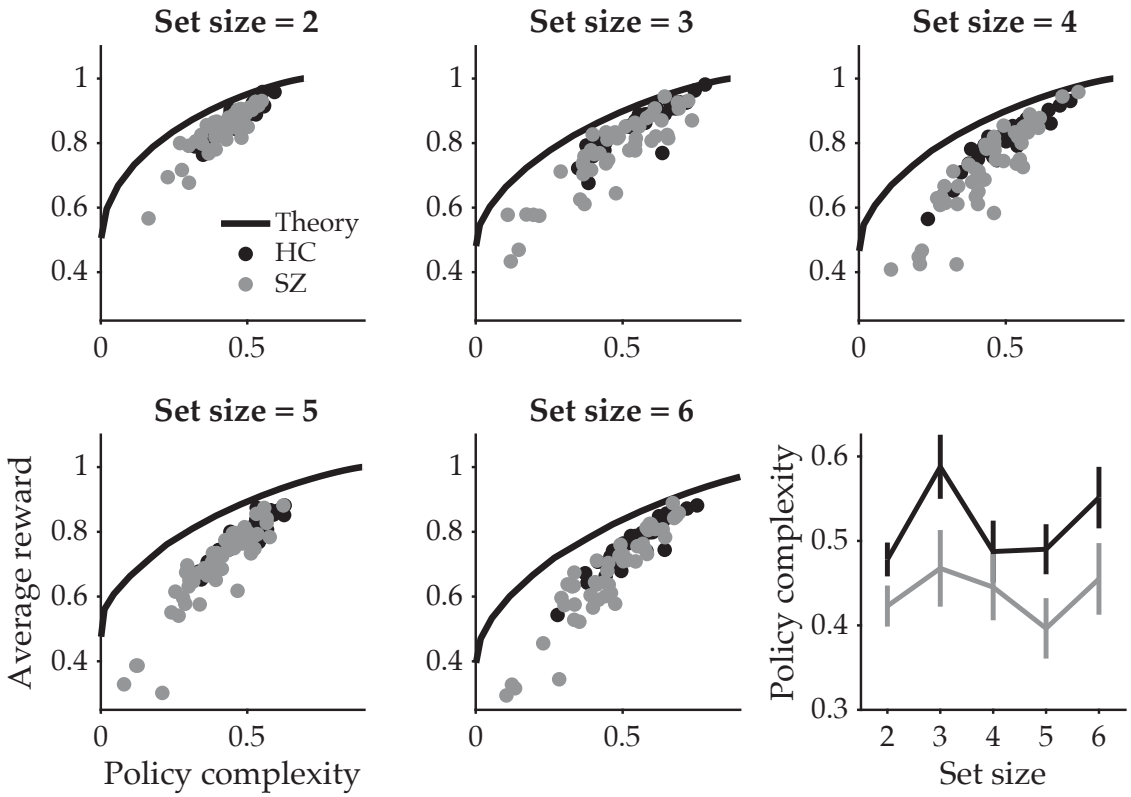

Figure 9: **The reward-complexity trade-off for schizophrenia patients and healthy controls.** Adapted from Gershman and Lai (2020), data from Collins et al. (2014). The optimal reward complexity curves (solid lines) for each given set size (2 to 6). Solid circles are the empirical reward-complexity values for each subject (HC = healthy controls; SZ = schizophrenia patients). (B) Policy complexity as a function of set size. Error bars show 95% confidence intervals.

dynamics of navigational learning: as mice gained more details about the location of the goal platform, they increased the precision of their motor commands and were able to quickly swim towards the platform from any starting location in the tank.

4.7 Psychiatry

What happens when humans fail to optimize the reward-complexity trade-off? As mentioned previously, the incremental nature of our proposed learning algorithm (see Appendix) means that it will spend a non-trivial amount of time away from the optimal reward-complexity trade-off curve. In fact, as long as learning is incomplete, agents will exhibit suboptimality. This suboptimality in learning offers one potential explanation for the maladaptive behaviors observed in many psychiatric conditions (e.g., schizophrenia, OCD, Tourette's, Parkinson's disease). While we recognize that there could be various explanations for the behavioral deficits observed in these conditions, we focus primarily on the influence of deficits in working memory capacity on behavior.

Working memory and cognitive effort deficits are a well-known characteristic of schizophrenia (Culbreth et al., 2018; Forbes et al., 2009). It therefore stands to reason that these patients should exhibit pronounced sub-optimality in their reward-complexity trade-off. To pursue this question, Gershman and Lai (2020) compared the performance of patients to healthy controls in the Collins contextual bandit task depicted in

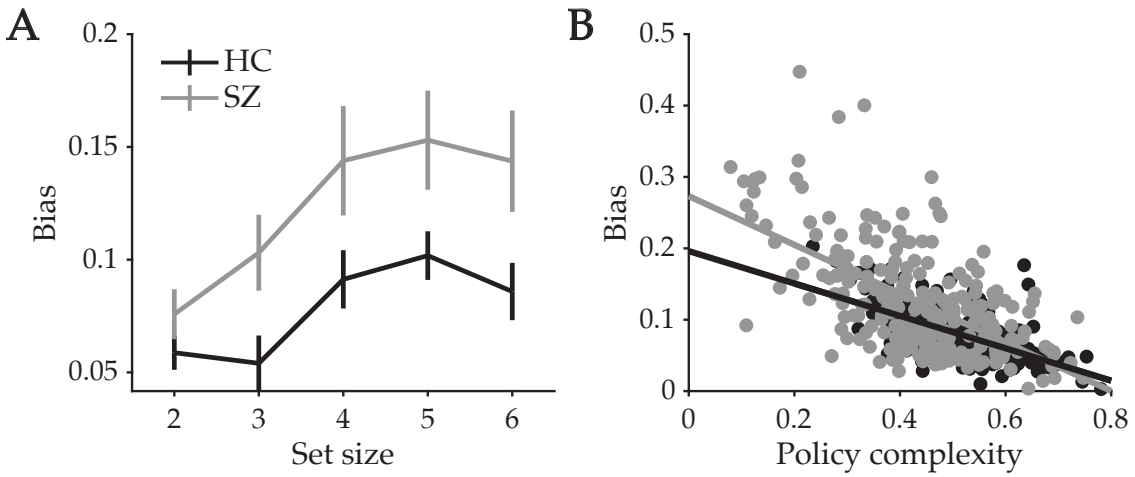

Figure 10: **Bias in schizophrenic patients and healthy controls.** Adapted from Gershman and Lai (2020), using data from Collins et al. (2014). (A) Bias, or the deviation of the empirical from the optimal reward-complexity curves, as a function of set size for the schizophrenia (SZ) and healthy control (HC) groups. Error bars show 95% confidence intervals. (B) Bias as a function of policy complexity.

Figure 2. Figure 9 shows the optimal and empirical reward-complexity trade-offs, as well as the average policy complexity across patients and healthy controls for tasks of varying set sizes, demonstrating that patients had systematically lower policy complexity but appeared to lie along the same empirical reward-complexity trade-off curve as healthy controls. Additionally, Gershman and Lai showed that empirical deviations from the optimal curve ("bias") in both populations increased as a function of set size and decreased as a function of policy complexity (Figure 10). This pattern of bias could be explained as a consequence of suboptimal learning, as reflected in the lower learning rates of schizophrenic patients as compared to healthy controls.

In probabalistic reversal learning tasks, schizophrenic patients have been shown to achieve fewer reversals than controls as well as decreased win-stay/lose-shift behavior. Patients also do not see errors as being informative of a context shift, suggesting some insensitivity to the underlying reward state (Culbreth et al., 2016; Schlagenhauf et al., 2014). One large study of probabalistic reversal learning in schizophrenia found that patients exhibit more suboptimal behavior as compared to healthy controls (Reddy et al., 2016). Specifically, schizophrenic patients had a higher proportion of lose-stay (selecting the same stimulus when it was previously unrewarded) and win-shift (selecting an alternate stimulus when the previous stimulus was rewarded) behaviors than did healthy controls.

Could it be that schizophrenia patients have overall lower capacity limits, which affect learning in cognitive tasks? To explore this possibility, we used the same reversal learning simulation in Figure 4 to show that the need to compress policies can lead to suboptimal behaviors in reversal strategy. In Figure 11, we show that as capacity is reduced, the proportion of lose-stay and win-shift behaviors increases, a result consistent with data from schizophrenic patients (Reddy et al., 2016).

The computational phenotyping of psychiatric illness is still in its infancy. However, the trade-off framework offered here provides a way to evaluate systematic suboptimalities in behavior resulting from deficits in memory capacity. While we have only focused here on the behavioral deficits observed in schizophrenia, these principles could be applied to other psychiatric conditions as well.

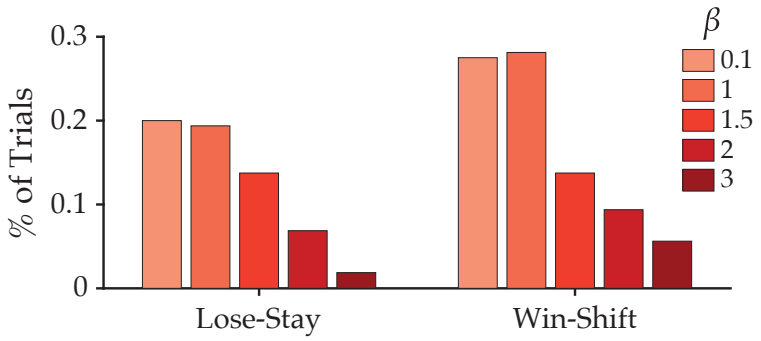

Figure 11: **Suboptimalities in reversal learning.** The proportion of trials where agents exhibited lose-stay (selecting the same stimulus when it was previously unrewarded) and win-shift (selecting an alternate stimulus when the previous stimulus was rewarded) behavior as a function of different capacity constraints. When the capacity limit is reduced, the proportion of both suboptimal strategies increases.

5 Neural signatures of policy compression

Now that we have seen all the ways in which action selection is shaped by constraints on memory capacity, it seems reasonable to ask: How do neurons in the brain learn the appropriate degree of policy compression? Some answers may be found in our cost-sensitive actor-critic model, which learns the appropriate policy complexity given a biological agent's resource constraints (see Appendix for more details). From an anatomical perspective, our model suggests a computational rationale for the massive compression factor in the mapping from cortex to striatum—if most incoming state information can be disregarded when selecting actions, the basal ganglia can afford to compress incoming cortical information (Bar-Gad et al., 2003).

From a mechanistic learning perspective, our model implements compression by incrementally modifying the policy based on reward feedback, with the critical assumption that an agent is penalized for policies that deviate from the marginal action distribution. This penalty enters into the reward prediction error (RPE), which in turn affects the policies that an agent will learn. This has direct biological implications: if learning is sensitive to the desire to compress policies, we should see the RPE signal, thought to be communicated by phasic dopamine, vary as a function of a subject's policy complexity.

While there have not yet been any experimental studies directly investigating the relationship between dopaminergic RPE and policy complexity, several studies from the action chunking literature have hinted at possible downstream effects within the striatum that reflect a neural signature of policy compression. Some have suggested that the "recoding" within the striatum seen during stimulus-response learning is responsible for chunking the representations of action sequences so that they can be implemented as single performance units (Graybiel, 1998b). Moreover, the learning of action chunks is often reflected in the emergence of "start/stop" activity within the striatum that brackets the beginning and end of learned sequences, as well as neural activity that is sustained throughout the execution of an entire sequence (Graybiel, 1998a; Smith and Graybiel, 2013; Jin and Costa, 2010; Jin et al., 2014). This reorganization of neural activity could be one implementation of compression: the "start/stop" activity indicates when a subject's attention must be given to the state (namely, at the start and termination of an action chunk), while the sustained activity reflects the fact that an action sequence currently in execution is impervious to the environmental state (and therefore, the neural activity is the same across states). This kind of neural activity would reflect a measure of policy complexity by signaling the moment-by-moment state-dependence of an animal's actions. Similar sensitivity for learned behavioral sequences has been observed in parts of the lateral and dorsomedial prefrontal cortices, with the latter shown to be necessary for the formation of action chunks (Shima et al., 2007; Ostlund et al., 2009).

Beyond the downstream effects of phasic dopamine, others have suggested that tonic dopamine levels

may act as a pseudo-temperature signal by directly modulating striatal excitability and thus tuning the trade-off between reward and policy complexity (Parush et al., 2011). If tonic dopamine is a neural correlate of the β parameter described in our framework, it should reflect a subject's capacity constraint and predict various behavioral measures of compression as detailed in this chapter. For example, Rutledge et al. (2009) found that perseveration increases in Parkinson's disease and decreases with dopamine therapy. This result is actually opposite what is predicted by the model in Parush et al. (2011), but nonetheless indicates that tonic dopamine plays a role in governing perseveration.

6 Compression and learning

In communication systems, compression is a solution to the problem of limited capacity—the overflow of data, so to speak. But in learning systems, compression is a solution to the problem of limited experience—the underflow of data. In order to generalize effectively from finite data, it is necessary to have an inductive bias favoring some generalizations over others. Otherwise, an unlimited number of generalizations are equally plausible given a finite data set. In statistical machine learning, inductive biases that provide generalization guarantees have been formalized under different assumptions and learning objectives. It turns out that a number of these generalization guarantees can be understood as statements that "compression implies learning" (Blum and Langford, 2003; Blumer et al., 1987).

To get a feel for why this is true, consider an agent that receives a stream of state-action-reward samples. The agent learns a policy from these "training" samples, and the question then is whether this policy will generalize effectively to new "test" samples. That is, will following the learned policy yield high reward? If the policy is complex enough, then it will favor actions that yield high reward on the training samples, but a policy that is too complex can also fit noise in the data, and hence *overfit*, yielding low reward on the test samples. Thus, policies that balance reward and complexity are more likely to generalize.

The link between compression and learning has been examined from a different angle by recent work on the nature of multitasking constraints (Musslick et al., 2017, 2020; Sagiv et al., 2018). The motivating puzzle is why, given billions of neurons, the brain suffers from an inability to simultaneously perform certain tasks at the same time? The basic answer is that even a relatively small amount of overlap in neural pathways can produce a catastrophic degree of cross-talk (Feng et al., 2014). This begs the question why the brain is evolved to be so susceptible to cross-talk. The answer proposed by Musslick and colleagues is that the upside of pathway overlap is efficiency of learning: shared representations facilitate generalization by reducing the number of separate parameters that need to be learned across tasks (or states). This recapitulates the point that compression (in this case, sharing of representations) is necessary for effective generalization. To mitigate the deleterious effects of cross-talk on task performance, the brain has an additional mechanism, cognitive control, which acts to selectively and strategically potentiate specific representations.

7 Conclusion

We have argued that the trade-off between reward and complexity is a fundamental optimality principle in action selection with broad empirical implications. Seemingly unrelated phenomena (perseveration, stochasticity, response time, state and action chunking, and navigation) are woven coherently together within this framework. A unifying theme is that memory limitations play an important role in governing action selection. In this sense, our framework intersects with recent work using rate-distortion theory to understand human memory (Bates et al., 2019; Bates and Jacobs, 2020; Sims et al., 2012; Sims, 2016; Nagy et al., 2020). This intersection suggests that we can continue to derive new insights into action selection by drawing parallels with other capacity-limited memory systems.

Acknowledgements

We are grateful to Dan McNamee for helpful comments. This research was supported by the Center for Brains, Minds, and Machines (funded by NSF STC award CCF-1231216) and a Graduate Research Fellowship from the NSF.

References

Abel, D., Arumugam, D., Asadi, K., Jinnai, Y., Littman, M. L., and Wong, L. L. S. (2019). State abstraction as compression in apprenticeship learning.

Amir, N., Suliman, R., Tal, M., Shifman, S., Tishby, N., and Nelken, I. (2020). Value-complexity tradeoff explains mouse navigational learning. *PLoS Comput. Biol.*, 16(12):e1008497.

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

Bar-Gad, I., Morris, G., and Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Prog. Neurobiol.*, 71(6):439–473.

Bates, C. J. and Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*, 127:891–917.

Bates, C. J., Lerch, R. A., Sims, C. R., and Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, 19(2):11–11.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *J. Gen. Psychol.*, 39:15–22.

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

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

Bhui, R. and Gershman, S. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, 125:985–1001.

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.

Bo, J. and Seidler, R. D. (2009). Visuospatial working memory capacity predicts the organization of acquired explicit motor sequences. *J. Neurophysiol.*, 101(6):3116–3125.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, 12:201–208.

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.

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

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

Culbreth, A. J., Gold, J. M., Cools, R., and Barch, D. M. (2016). Impaired activation in cognitive control regions predicts reversal learning in schizophrenia. *Schizophr. Bull.*, 42(2):484–493.

Culbreth, A. J., Moran, E. K., and Barch, D. M. (2018). Effort-based decision-making in schizophrenia. *Current Opinion in Behavioral Sciences*, 22:1–6.

Dassonville, P., Lewis, S. M., Foster, H. E., and Ashe, J. (1999). Choice and stimulus–response compatibility affect duration of response selection. *Cognitive Brain Research*, 7:235–240.

Denti, T., Marinacci, M., and Montrucchio, L. (2019). A note on rational inattention and rate distortion theory. *Decisions in Economics and Finance*, pages 1–15.

Dezfouli, A. and Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, 35:1036–1051.

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.

Faisal, A. A., Selen, L. P., and Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, 9:292–303.

Feng, S. F., Schwemmer, M., Gershman, S. J., and Cohen, J. D. (2014). Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience*, 14:129–146.

Forbes, N., Carrick, L., McIntosh, A., and Lawrie, S. (2009). Working memory in schizophrenia: a meta-analysis. *Psychological Medicine*, 39:889.

Fox, R., Pakman, A., and Tishby, N. (2015). Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*.

Fründ, I., Wichmann, F. A., and Macke, J. H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, 14:9–9.

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

Gershman, S. J. (2021). The rational analysis of memory. In Wagner, M. K. . A., editor, *Oxford Handbook of Human Memory*. Oxford University Press.

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

Grau-Moya, J., Leibfried, F., and Vrancx, P. (2018). Soft q-learning with mutual-information regularization. In *International Conference on Learning Representations*.

Graybiel, A. M. (1998a). The basal ganglia and chunking of action repertoires. *Neurobiol. Learn. Mem.*, 70(1-2):119–136.

Graybiel, A. M. (1998b). The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory*, 70:119–136.

Hale, D. (1968). The relation of correct and error responses in a serial choice reaction task. *Psychonomic Science*, 13:299–300.

Hassett, T. C. and Hampton, R. R. (2017). Change in the relative contributions of habit and working memory facilitates serial reversal learning expertise in rhesus monkeys. *Anim. Cogn.*, 20(3):485–497.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Dxperimental Psychology*, 4:11–26.

Howarth, C. and Bulmer, M. (1956). Non-random sequences in visual threshold experiments. *Quarterly Journal of Experimental Psychology*, 8:163–171.

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

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., and Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proc. Natl. Acad. Sci. U. S. A.*, 112(10):3098–3103.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, 45:188.

Icard, T. (2019). Why be random? *Mind*.

Jin, X. and Costa, R. M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, 466:457–462.

Jin, X., Tecuapetla, F., and Costa, R. M. (2014). Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nature Neuroscience*, 17:423–430.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014.

Lashley, K. S. (1951). The problem of serial order in behavior. In Jeffress, L., editor, *Cerebral Mechanisms in Behavior*, pages 112–136. Wiley.

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.

Lehnert, L. and Littman, M. L. (2019). Successor features combine elements of Model-Free and model-based reinforcement learning.

Lehnert, L., Littman, M. L., and Frank, M. J. (2020). Reward-predictive representations generalize across tasks in reinforcement learning. *PLoS Comput. Biol.*, 16(10):e1008317.

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

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

Matějka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105:272–98.

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

McDougle, S. D. and Collins, A. G. (2020). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, pages 1–20.

McFadden, D. (2001). Economic choices. *American Economic Review*, 91:351–378.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63:81–97.

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

Miyapuram, K. P., Bapi, R. S., Pammi, C. V. S., Ahmed, and Doya, K. (2006). Hierarchical chunking during learning of visuomotor sequences. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 249–253.

Mosteller, F. and Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, 59:371–404.

Mowbray, G. and Rhoades, M. (1959). On the reduction of choice reaction times with practice. *Quarterly Journal of Experimental Psychology*, 11:16–23.

Musslick, S., Saxe, A., Hoskin, A. N., Reichman, D., and Cohen, J. D. (2020). On the rational boundedness of cognitive control: Shared versus separated representations.

Musslick, S., Saxe, A. M., Özcimder, K., Dey, B., Henselman, G., and Cohen, J. D. (2017). Multitasking capability versus learning efficiency in neural network architectures. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, page 829– 834.

Nagy, D. G., Török, B., and Orbán, G. (2020). Optimal forgetting: Semantic compression of episodic memories. *PLOS Computational Biology*, 16:1–28.

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.

Ngiam, W. X., Brissenden, J., and Awh, E. (2019). "memory compression" effects in visual working memory are contingent on explicit long-term memory. *Journal of Experimental Psychology: General*, 148:1373–1385.

Nissen, M. J. and Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cogn. Psychol.*, 19(1):1–32.

Norman, D. (1981). Categorization of action slips. *Psychological Review*, 88:1–15.

Norris, D. and Kalm, K. (2020). Chunking and data compression in verbal short-term memory. *Cognition*, 208:104534.

Ostlund, S. B., Winterbauer, N. E., and Balleine, B. W. (2009). Evidence of action sequence chunking in goal-directed instrumental conditioning and its dependence on the dorsomedial prefrontal cortex. *Journal of Neuroscience*, 29:8280–8287.

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

Precup, D. (2000). *Temporal abstraction in reinforcement learning*. PhD thesis, University of Massachusetts Amherst.

Precup, D., Sutton, R. S., and Singh, S. (1998). Theoretical results on reinforcement learning with temporally abstract options.

Proctor, R. W. and Schneider, D. W. (2018). Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology*, 71:1281–1299.

Ramkumar, P., Acuna, D. E., Berniker, M., Grafton, S. T., Turner, R. S., and Kording, K. P. (2016). Chunking as the result of an efficiency computation trade-off. *Nature Communications*, 7:1–11.

Reddy, L. F., Waltz, J. A., Green, M. F., Wynn, J. K., and Horan, W. P. (2016). Probabilistic reversal learning in schizophrenia: Stability of deficits and potential causal mechanisms. *Schizophr. Bull.*, 42(4):942–951.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

Robertson, E. M. (2007). The serial reaction time task: Implicit motor skill learning? *J. Neurosci.*, 27(38):10073–10075.

Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. A., and Glimcher, P. W. (2009). Dopaminergic drugs modulate learning rates and perseveration in parkinson's patients in a dynamic foraging task. *Journal of Neuroscience*, 29:15104–15114.

Sagiv, Y., Musslick, S., Niv, Y., and Cohen, J. D. (2018). Efficiency of learning vs. processing: Towards a normative theory of multitasking. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, page 1004–1009.

Sakai, K., Kitaguchi, K., and Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Experimental Brain Research*, 152:229–242.

Schlagenhauf, F., Huys, Q. J. M., Deserno, L., Rapp, M. A., Beck, A., Heinze, H.-J., Dolan, R., and Heinz, A. (2014). Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *Neuroimage*, 89:171–180.

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

Seibel, R. (1963). Discrimination reaction time for a 1,023- alternative task. *Journal of Experimental Psychology*, 66:215–226.

Seidler, R. D., Bo, J., and Anguera, J. A. (2012). Neurocognitive contributions to motor skill learning: the role of working memory. *J. Mot. Behav.*, 44(6):445–453.

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

Shima, K., Isoda, M., Mushiake, H., and Tanji, J. (2007). Categorization of behavioural sequences in the prefrontal cortex. *Nature*, 445(7125):315–318.

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.

Smith, K. S. and Graybiel, A. M. (2013). A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron*, 79(2):361–374.

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.

Teichner, W. and Krebs, M. (1974). Laws of visual choice reaction time. *Psychological Review*, 81:75–98.

Terrace, H. S. (1991). Chunking during serial learning by a pigeon: I. basic evidence. *J. Exp. Psychol. Anim. Behav. Process.*, 17(1):81–93.

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. Macmillan Press.

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

Tkačik, G., Prentice, J. S., Balasubramanian, V., and Schneidman, E. (2010). Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences*, 107:14419–14424.

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., and Gershman, S. J. (2020). Discovery of hierarchical representations for efficient planning. *PLoS Computational Biology*, 16:e1007594.

Verplanck, W., Collier, G., and Cotton, J. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology*, 44:273–282.

Verwey, W. B. (1999). Evidence for a multistage model of practice in a sequential movement task. *J. Exp. Psychol. Hum. Percept. Perform.*, 25(6):1693–1708.

Von Neumann, J. (1958). *The Computer and the Brain*. Yale University Press.

Wifall, T., Hazeltine, E., and Mordkoff, J. T. (2016). The roles of stimulus and response uncertainty in forced-choice performance: an amendment to hick/hyman law. *Psychological Research*, 80:555–565.

Zelazo, P. D. (2006). The dimensional change card sort (DCCS): a method of assessing executive function in children. *Nat. Protoc.*, 1(1):297–301.

Appendix

In this section, we derive a generalization of the process model presented in Gershman and Lai (2020). Code used to reproduce the simulations is available at `http://github.com/lucylai96/plm/`.

The optimization problem facing an agent is to maximize expected reward subject to the constraint that the policy complexity (information rate) cannot exceed the agent's capacity C. Expected reward under policy π is defined as follows:

$$V^{\pi} = \sum_{s} P(s) \sum_{a} \pi(a|s) Q(s,a), \tag{9}$$

where $P(s)$ is the probability of state s and $Q(s,a)$ is the expected reward in state s after taking action a.

To solve the constrained optimization problem, we write it in Lagrangian form, with Lagrange multipliers β and $\lambda(s)$:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \beta V^{\pi} - I^{\pi}(S;A) + \sum_{s} \lambda(s) \left(\sum_{a} \pi(a|s) - 1 \right). \tag{10}$$

The optimal policy π^* has the form stated in Eq. 6. The question we address here is how to tractably find the optimal policy.

We can cast the optimization problem in a form amenable to reinforcement learning by rewriting the Lagrangian as follows (leaving the non-negativity and summation constraints on π implicit):

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[\beta r - \log \frac{\pi(a|s)}{P(a)} \right], \tag{11}$$

By taking the gradient of the objective function with respect to the policy parameters and using it to incrementally modify the policy, we obtain a "policy gradient" algorithm (Sutton and Barto, 2018) that will converge to the optimal policy. The algorithm takes the form of an "actor-critic" architecture consisting of a paramtrized policy (the actor) and a value estimator (the critic). Critically, this algorithm does not require marginalizing over the state space, as in the Blahut-Arimoto algorithm.

Following the functional form of the optimal policy (Eq. 6), we parametrize the "actor" component of the model according to:

$$\pi_{\theta}(a|s) \propto \exp\left[\beta \theta_a \cdot \phi(s) + \log P(a) \right], \tag{12}$$

where θ denotes the adjustable policy parameters and $\phi(s)$ denotes a set of state features. These features will vary across task domains. Taking the gradient of the objective function with respect to θ yields the following learning rule after taking action a in state s and receiving reward r:

$$\Delta\theta_a = \alpha_{\theta} \phi(s) \delta [1 - \pi_{\theta}(a|s)] \beta, \tag{13}$$

where α_{θ} is the actor learning rate and

$$\delta = \beta r - \log \frac{\pi(a|s)}{P(a)} - \hat{V}(s) \tag{14}$$

is the prediction error of the "critic" $\hat{V}(s)$, an estimator of the expected cost-sensitive reward which we parametrize as a linear function of state features $\phi(s)$:

$$\hat{V}(s) = \mathbf{w} \cdot \phi(s), \tag{15}$$

with adjustable parameters $\mathbf{w}$ updated according to:

$$\Delta\mathbf{w} = \alpha_V \phi(s) \delta \tag{16}$$

with critic learning rate α_V.[11]

Finally, we incrementally estimate the marginal action probabilities with an exponential moving average:

$$\Delta P(a) = \alpha_P[\pi(a|s) - P(a)] \tag{17}$$

where α_P is another learning rate parameter.

[11] In Gershman and Lai (2020), the actor learning rate (but not the critic learning rate) is scaled by $1/t$ to ensure that the actor eventually converges to the optimal policy by satisfying the Robbins-Munro conditions for stochastic approximation algorithms (Robbins and Monro, 1951). This also ensures that the actor learning rate will generally be slower than the critic learning rate, a typical theoretical requirement of these algorithms (Konda and Tsitsiklis, 2000). For simplicity, here we omit the $1/t$ scaling.