Précis of *Policy Compression: Acting with Limited Cognitive Resources*

Lucy Lai

Introduction

One of the most striking features of cognition is the ability to learn about and navigate a complex, information-rich world. From deciding what to eat to learning a new skill, humans and other animals excel at consolidating and transforming vast amounts of environmental input to guide behavior. However, this remarkable ability is also fundamentally constrained by limits on cognitive resources. The concept of *bounded rationality*—doing the best with what one has—has long been used as a framework for understanding how cognitive resource constraints shape decision making. This perspective highlights how resource limitations can lead to deviations from the idealized, optimal behavior of an unbounded agent (Simon, 1957; Rubinstein, 1998; Gigerenzer and Selten, 2002; Kahneman, 2003). Bounded rationality posits that agents employ heuristics and simplified decision rules to manage the computational complexities of real-world problems.

The related concept of *resource rationality* refines bounded rationality by formalizing it in computational terms, linking cognitive limitations to the physical constraints of the systems executing the decisions (Lieder and Griffiths, 2019; Bhui et al., 2021). All machines, including the brain, face physical constraints that limit their capacity to store and transmit information. These constraints necessitate trade-offs between maximizing reward and minimizing computational cost, which has important implications for the system's behavior. For example, agents may sacrifice some potential rewards to simplify decision making and reduce cognitive load. The resource rational perspective proposes that apparent deviations from rationality are not failures of the system, but rather reflect the optimal use of limited resources.

Despite these valuable insights, existing theories often lack precise, quantitative models that explain how agents learn to balance cognitive costs and rewards in real-time, especially in the domain of action selection. Furthermore, while the time costs of decision making (i.e., the number of mental operations required to implement decisions) have been extensively studied (Daw et al., 2005; Gershman et al., 2014; Kool et al., 2018; Callaway et al., 2022), their representational costs (i.e., the amount of memory storage required) remain relatively underexplored. We are left with lingering questions: **How do agents simplify decision making through interactions with their**

environment? How does the trade-off between reward and cognitive cost shape both behavior and the neural mechanisms that support it?

This dissertation addresses these questions by developing a theoretical framework that explains how biological agents optimize behavior within the constraints of limited cognitive resources. Specifically, I formalize the concept of **policy compression**—the simplification of action policies to reduce their representational costs—and develop an online algorithm that learns to dynamically balance rewards and costs through interactions with the environment. **Through computational modeling, behavioral experiments, and lesion studies, this dissertation investigates the mechanisms and implications of policy compression, explaining how agents achieve complex behavior within the limits of their cognitive resources.**

The dissertation is organized as follows: Chapter 1 introduces the theoretical framework of policy compression, formalizing the trade-off between cognitive cost and reward. Chapter 2 grounds this theory in empirical evidence, reinterpreting a variety of behavioral and neural phenomena through the lens of cognitive resource constraints. Chapters 3 and 4 present new experimental evidence showing how humans actively exploit environmental structure to simplify decision making and reduce cognitive load. Chapter 5 proposes that the brain uses policy compression as a cost-efficient strategy to balance robustness and flexibility in adaptive behavior. Finally, Chapter 6 concludes by exploring future directions and the broader implications of policy compression.

Chapter 1: Theoretical Foundations of Policy Compression

> The work in Chapters 1 and 2 was published in *Psychology of Learning and Motivation* (Lai and Gershman, 2021).

Action selection demands memory. When you play an instrument, drive to work, or prepare a meal, your brain is retrieving stored information about *policies*, the mappings from states of the world to actions (Sutton and Barto, 2018). In the language of information theory, policies can be described as communication channels that transmit information about states of the world to guide action selection (Figure 1A). The *complexity* of a policy reflects the amount of memory required to store it, which can be formally quantified in *bits* of information. Since the brain's memory capacity is finite, policies must be "compressed" as much as possible, discarding redundant information and reducing precision where it's not needed. Intuitively, if a policy can be compressed, it will be easier to remember.

To provide an intuitive example, imagine you are preparing a meal for your family. In this case, states correspond to family members, actions correspond to dishes, and policies are probabilistic mappings from family members to dishes (Figure 1B). If everyone in your family is happy to eat the same dish, you can ignore the state entirely and just take the same action (prepare the same dish) repeatedly. Such a policy is "compressed" (low complexity) in the sense that it consumes

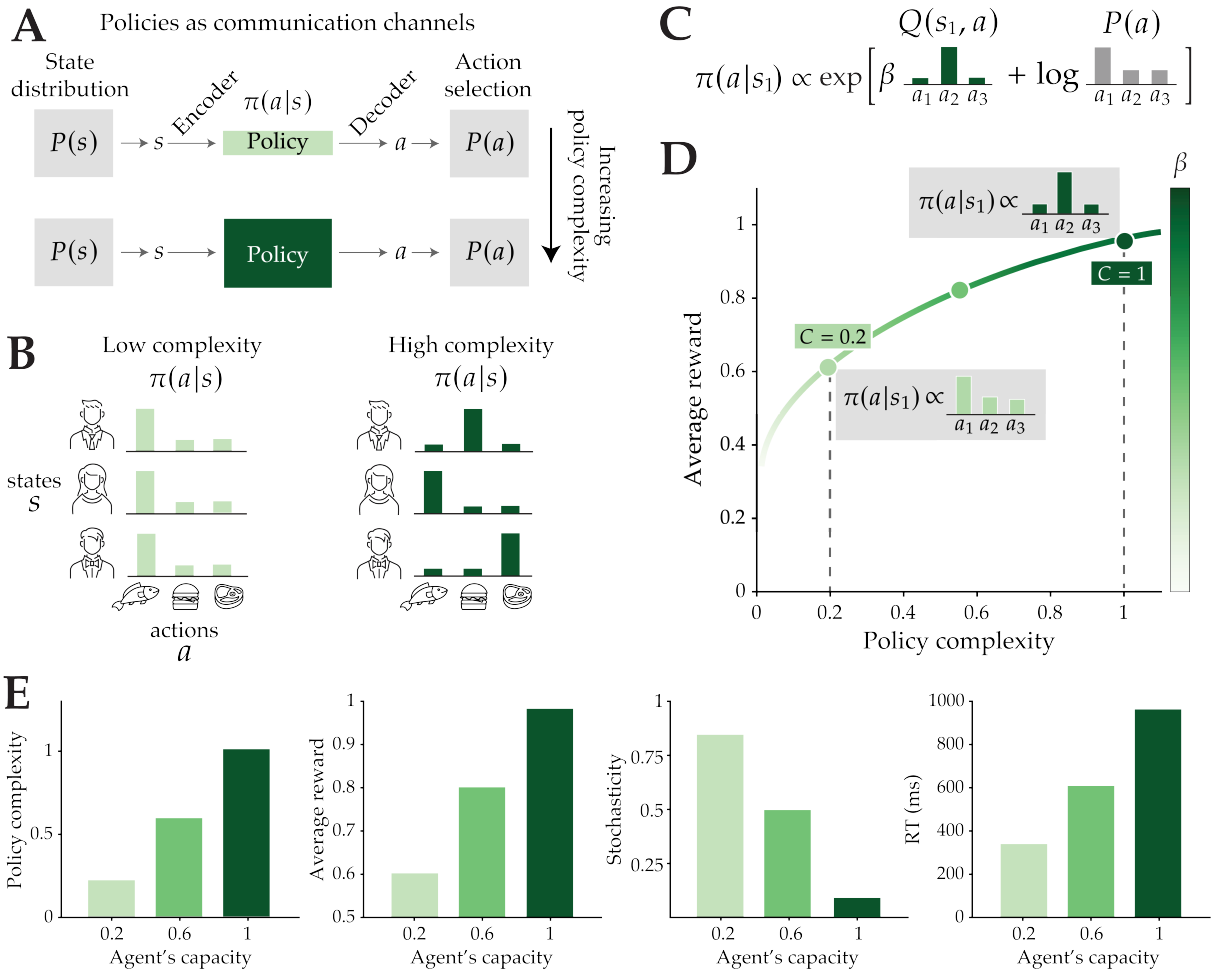

Figure 1: **Policy compression. (A)** Policies can be described as communication channels that map states to actions. The channel's storage capacity is the policy complexity. **(B)** Cooking dinner for your family represented as a low complexity policy (cooking everyone the same dish) and high complexity policy (cooking everyone different dishes). **(C)** The optimal policy combines information about the most rewarding action(s) in each state $Q(s, a)$ with the action(s) that is chosen most frequently overall $P(a)$. The trade-off parameter, β, determines the relative contribution of $Q(s, a)$ and $P(a)$. Example distributions depict action selection in one state. **(D)** A limit on the agent's capacity, C, results in a trade-off between reward and complexity. Low complexity policies (low β) are biased towards $P(a)$. High complexity policies (high β), are biased towards $Q(s, a)$. **(E)** An agent's capacity limit affects various aspects of behavior, including policy complexity, earned reward, stochasticity of actions, and response times (RT).

fewer bits of memory compared to one in which you need to remember separate dishes for each family member (high complexity).

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 preferred dishes between family members. These capacity constraints create a trade-off between maximizing reward (e.g., satisfying family members' preferences) and minimizing the cost of representing action policies in memory (Figure 1D). By framing action selection as a capacity-limited communication channel, we can formalize this trade-off and derive an optimal policy; i.e., a behavioral strategy that maximizes rewards subject to constraints on memory resources (Figure 1C):

$$\pi^*(a|s) \propto \exp[\beta Q(s,a) + \log P^*(a)]. \tag{1}$$

This equation illustrates how capacity-limited agents balance selecting the most rewarding action in a given state, $Q(s,a)$, with relying on their history of frequently chosen actions, $P(a)$. The resulting behavior integrates information from both the current state and past actions. The trade-off parameter, β, indirectly represents the agent's capacity limit, which constrains the maximum policy complexity the agent can achieve. This term also determines the relative influence of $Q(s,a)$ and $P(a)$, which shapes a range of behavioral outcomes. Agents with higher capacity can earn more reward, behave less randomly, and take longer to execute decisions, implications we will explore in more detail in the next chapter (Figure 1E).

In the remainder of the chapter, I develop a tractable reinforcement learning algorithm that learns optimal policies through repeated interactions with the environment. This novel algorithm incrementally adjusts the policy using reward feedback while incorporating a penalty for complexity. In summary, **policy compression marries reinforcement learning with information theory by modeling policies as capacity-limited channels that balance the demands of memory and reward.** In subsequent chapters, I demonstrate how this model explains a wide range of behavioral and neural phenomena, reframing them as manifestations of policy compression.

Chapter 2: The Peculiarities of Policy Compression

Why do we sometimes choose randomly, even when we know the best option? Why do old habits persist, even when circumstances change? Why does it take longer to make decisions in some situations than others? These seemingly unrelated questions about human behavior can all be understood through the lens of policy compression. In this chapter, **I use illustrative simulations and experimental reanalysis to unify a variety of behavioral and neural phenomena—including stochasticity, perseveration, response times, chunking, and navigation—under the framework of policy compression.** While each individual phenomenon may be explained by alternative the-

ories, I make the case that they can be understood collectively as reflections of a single underlying principle. Here, I highlight three of these phenomena and discuss intriguing applications of this framework to illustrate its broad relevance.

Stochasticity. Imagine reading a restaurant menu, unsure which dish to order. Even though the waiter recommends a popular dish, you might still choose another at random. This randomness, or stochasticity, in decision making has often been attributed to exploratory behavior or unexplained noise in how we evaluate options (Schulz and Gershman, 2019). Policy compression offers an alternative explanation: randomness is not a flaw but a feature of capacity-limited decision making. When cognitive resources are scarce, introducing randomness allows us to reduce the complexity of our decision policies while still approximating good outcomes. Empirical data confirm the theory's prediction that increasing cognitive load leads to more random choices (Collins et al., 2014). Randomness is thus a rational adaptation to cognitive resource constraints.

Perseveration. Our tendency to stick with old habits can similarly be reframed as a rational strategy under limited cognitive capacity. It is more efficient to rely on past choices rather than calculating the best course of action for every situation—so why *not* order the same dish you always do? In classical tasks like reversal learning, where reward contingencies abruptly change, the inability to quickly adapt behavior reflects people's reliance on a simpler, low complexity policy. Simulations reveal that agents with reduced capacity exhibit stronger perseverative tendencies, consistent with empirical data (Hassett and Hampton, 2017; Collins, 2018; Gershman, 2020). These findings help explain perseveration not as a failure of flexibility but as an optimal solution given limited cognitive resources.

Response times. The more menu options to chose from, the longer it takes to pick one! Hick's Law, a classic finding in cognitive psychology, states that response times increase logarithmically with the number of possible options (Hick, 1952; Hyman, 1953). But why should this be the case? Policy compression provides an elegant computational explanation: the more complex a policy, the longer it takes to implement (Proctor and Schneider, 2018). Selecting an action involves processing information proportional to the complexity of the policy, analogous to how running a longer piece of code requires more time. Policy compression predicts that people with higher cognitive capacity will learn more complex policies and exhibit slower response times, a finding supported by reanalysis of human behavior in a simple choice task (Collins, 2018).

Applications in psychiatry, neuroscience, and machine learning. The implications of policy compression extend beyond everyday behaviors. In psychiatry, compression offers a new lens for understanding cognitive deficits in a range of disorders. For example, patients with schizophrenia show reduced policy complexity and increased perseveration, consistent with the behavior of capacity-limited agents (Gershman and Lai, 2021). By linking symptoms to underlying resource constraints, policy compression provides a foundation for targeted interventions and diagnostics.

Policy compression also generates testable hypotheses in neuroscience. For instance, the reinforcement learning algorithm described in Chapter 1 penalizes agents for high complexity policies,

and this penalty influences the reward prediction error (RPE), a signal thought to be mediated by phasic dopamine (Schultz et al., 1997). This leads to an interesting and testable prediction: phasic dopamine levels should vary systematically with policy complexity.

Beyond cognitive science, the principles of policy compression have implications for designing machine learning algorithms. By promoting generalization and avoiding overfitting, compression conserves memory resources while enabling machines to learn more efficiently. Compression also facilitates multitask learning by allowing shared representations across tasks, though this comes at the cost of increased vulnerability to interference—a fundamental trade-off in resource-limited systems. A deeper understanding of policy compression can provide valuable insights for both biological and artificial systems that face physical constraints.

Chapter 3: Human Decision Making Balances Reward Maximization and Policy Compression

> The work in this chapter was published in *PLOS Computational Biology* (Lai and Gershman, 2024).

While policy compression has been successfully applied to explain diverse behavioral phenomena, including perseveration and undermatching (Lai and Gershman, 2021; Gershman, 2020; Gershman and Lai, 2021; Bari and Gershman, 2023), these findings have largely relied on *post hoc* analyses of previously published datasets. In this chapter, **I design novel experiments that directly test the unique predictions of policy compression,** allowing us to explore new hypotheses about learning under cognitive constraints.

Policy compression predicts that the structure of relationships between states, actions, and rewards influences how agents simplify their policies. To test this, I designed three tasks that manipulated the distribution of states and actions to encourage policy compression. Here, I highlight one task as a representative example: participants learned to make specific key presses (actions) in response to visual stimuli (states) to earn rewards (Figure 2A). In the "Test" condition, one action (e.g.,"J") consistently delivered rewards across all states, enabling participants to compress their policies by ignoring state-specific information. In other words, participants could repeatedly choose the same rewarding action (e.g.,"J") rather than tailoring their responses to each state, reducing the task's demand on memory.

Policy compression predicted three key behavioral outcomes (Figure 2B-C): (1) participants adopted simpler, lower-complexity policies in the "Test" condition compared to a "Control" condition where such structure was absent; (2) compression enabled participants to earn more reward in the "Test" condition; and (3) participants exhibited a choice bias, favoring the shared action over other equally optimal actions. Crucially, this bias is not predicted by traditional reinforcement learning models, which would assume equal selection among all optimal actions. These findings suggest that people adapt their degree of policy compression based on environmental structure.

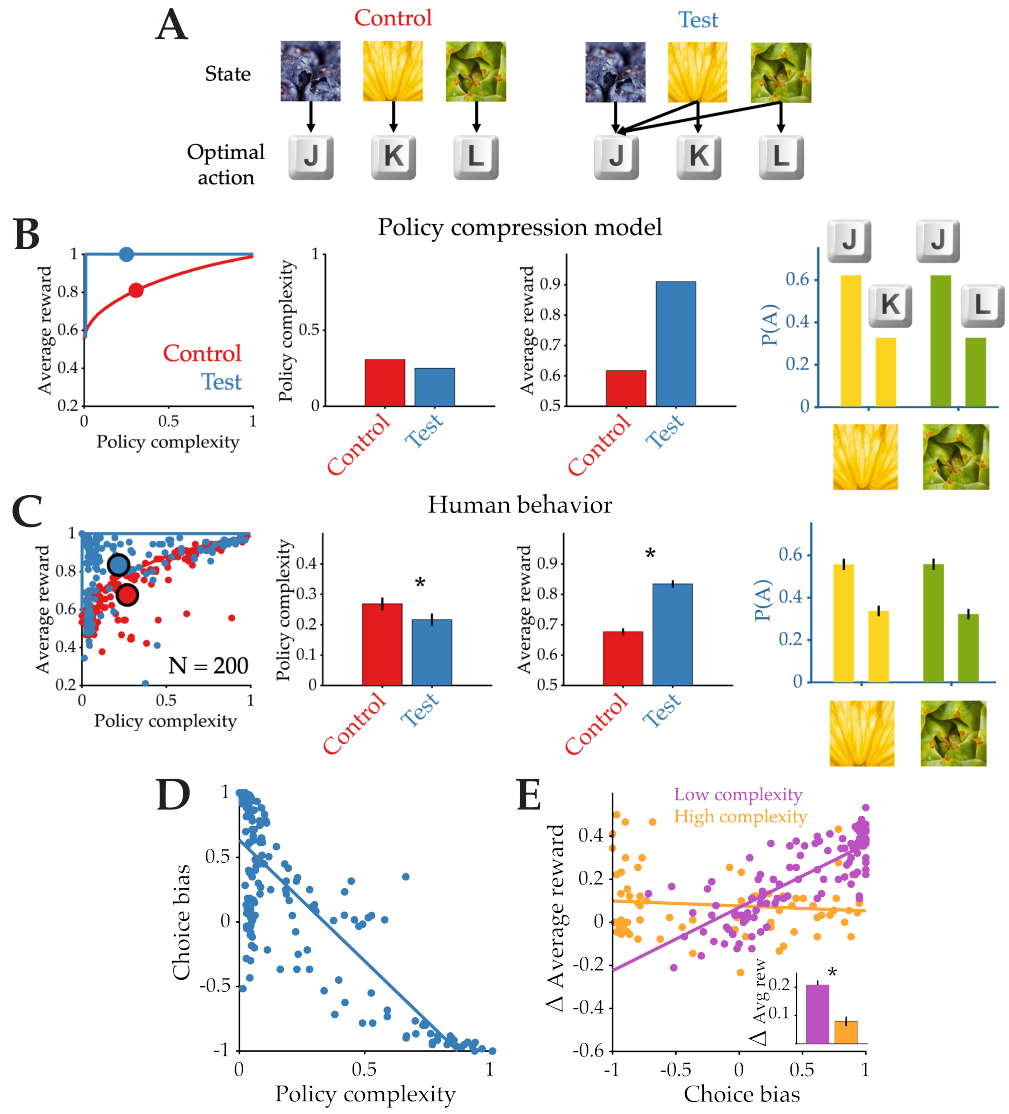

Figure 2: **Human decision making behavior aligns with key predictions of policy compression.** **(A)** A simple instrumental learning task designed to encourage compression, where some states share an optimal action. **(B)** Policy compression predicts distinct behavioral signatures and choice biases. **(C)** Human behavior closely matches model predictions.**(D)** Low policy complexity is associated with stronger choice bias. **(E)** Individuals with lower policy complexity leverage their choice biases to earn more reward.

Across all tasks, participants consistently favored simpler policies, leveraging redundancies in state-action mappings to reduce memory demands. As a result, choices were systematically biased towards actions that were chosen most frequently across states. Choice bias increased under higher memory load and among participants with lower policy complexity (Figure 2D), who benefitted more from their bias by earning more rewards for less effort (Figure 2E). Finally, participants further compressed their policies and displayed greater choice bias under time pressure, confirming the prediction (from Chapter 2) that actions are selected through the time-sensitive decoding of compressed representations (Lai and Gershman, 2021). Importantly, these results cannot be explained by models that lack capacity constraints on policy complexity, including those that incorporate working memory contributions to reinforcement learning (Collins and Frank, 2012). Taken together, these findings demonstrate that people exploit environmental structure to simplify their policies and provide strong experimental support for the policy compression framework.

Understanding human behavior under cognitive constraints has practical implications for designing decision environments that promote better choices, especially in high-stakes contexts. For example, choice architecture strategies, such as default options, leverage perseverative biases to influence decision making. Automatic enrollment in retirement plans (Beshears et al., 2007) and default renewable energy options (Pichert and Katsikopoulos, 2008) show how defaults can positively impact decisions for individuals who lack the time or cognitive capacity to fully evaluate alternatives. By incorporating quantitative models like policy compression, decision environments can be strategically engineered to align with behavioral tendencies, shaping choice behavior more effectively (Dan and Loewenstein, 2019).

Chapter 4: Action Chunking as Conditional Policy Compression

The work in this chapter is currently under review at *Cognition* (Lai et al., 2022).

The policy compression framework introduced in Chapter 1 does not yet address how agents might simplify behavior by taking advantage of temporal structure in their environment—an important limitation, given that natural environments often exhibit temporal continuity of states and actions. In this chapter, I revisit the concept of action chunking and reframe it as a natural consequence of policy compression (an idea I briefly introduce in Chapter 2). In doing so, I derive novel predictions and test them through new experiments, highlighting how agents exploit temporal regularities to conserve cognitive resources.

Many skills in our everyday lives are learned by sequencing actions toward a desired goal. These action sequences can become "chunks," where individual actions are grouped together and executed as a single unit, improving their efficiency (Miller, 1956). Action chunking is a hallmark of skill learning and habitual behavior that has been extensively studied (Terrace, 1991; Verwey, 1996, 1999; Sakai et al., 2003; Miyapuram et al., 2006; Bo and Seidler, 2009; Banca et al., 2023). However, a puzzle remains as to *why* and *under what conditions* chunking occurs. Existing mod-

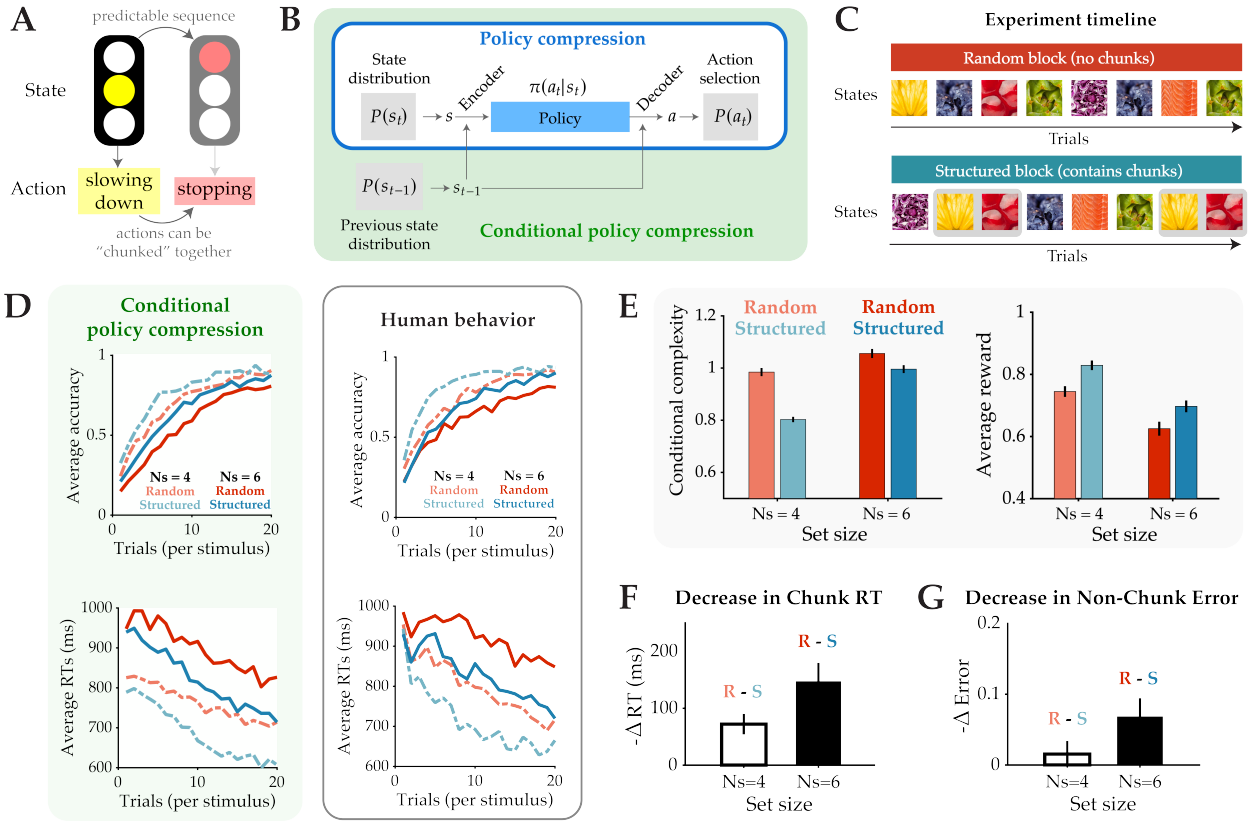

Figure 3: **Action chunking as policy compression. (A)** Temporal structure enables the current action (stopping the car) to be informed by the previous state (the yellow light) and action (slowing down). Because the current state (red light) provides redundant information, the agent can ignore it, reducing memory demands. In other words, the agent does not need to explicitly remember that red means "stop." **(B)** Conditional policy compression reduces the representational cost of policies by leveraging additional sources of information, such as temporal dependencies between states. **(C)** Participants learned to associate key presses (actions) with visual stimuli (states). Some experiment blocks contained structured temporal dependencies, or state "chunks" (grey boxes), to encourage action chunking. **(D)** Temporal structure leads to faster learning, higher accuracy, and shorter response times (RT). Human behavior aligns with model predictions. **(E)** Action chunking reduces conditional policy complexity, enabling participants to earn more reward with less cognitive effort. **(F)** Chunking increases under a higher memory load (Ns = 6), evidenced by a greater decrease in action chunk RT. **(G)** Chunking reduces error in not-chunked (unpredicted) states under high memory load (Ns = 6).

els suggest that chunking reduces the time cost of actions by reducing their latency (Dezfouli and Balleine, 2012). However, I propose an alternative explanation based on policy compression: **action chunking reduces the *memory* required to store and execute action sequences** (Lai and Gershman, 2021).

To illustrate this, imagine you are driving through an intersection. You know that a yellow traffic light is always followed by a red light, so upon seeing yellow, you can initiate an action sequence of slowing down and stopping the car—all without needing to remember the specific action for the red light (Figure 3A). By reducing demands on memory, you also act faster, which explains the time benefits of chunking. Notably, viewing chunking as saving memory resources makes unique predictions: action chunking should (1) increase with memory load, where the pressure to compress representations is greater, and (2) free cognitive resources for processing other information.

To formalize these ideas, I extended the model in Chapter 1 to develop **conditional policy compression**, which posits that **agents reduce cognitive costs by using additional sources of information—such as temporal structure—to simplify policies** (Figure 3B). By using information from prior states to guide action selection, agents can ignore the current state, thereby reducing the memory required for action selection. As a result, actions are selected faster *and* more accurately, as agents can efficiently anticipate and execute future actions based on previous states, reinforcing the link between policy complexity and response time.

To test this theory, I designed a serial reaction time task where participants learned to associate key presses (actions) with visual stimuli (states). Critically, I manipulated the predictability of state sequences, creating conditions that encouraged action chunking (Figure 3C). Results from the experiments confirmed four key model predictions: (1) Temporal structure led to faster learning, higher accuracy, and shorter response times, replicating previous findings in action chunking (Figure 3D). (2) Chunking reduced *conditional policy complexity*—the memory required for action selection—enabling participants to earn *more* rewards with *less* cognitive effort (Figure 3E). (3) Chunking increased under higher memory load: participants showed greater reductions in response time when the number of states was larger, consistent with the greater pressure to compress policies (Figure 3F). (4) Finally, chunking freed working memory resources, improving accuracy even on non-chunked (unpredicted) stimuli (Figure 3G).

In summary, action chunking emerges as a behavioral strategy for managing limited memory resources by leveraging temporal structure in the environment. By applying a novel theoretical framework to a classic phenomenon, this work highlights the adaptive value of chunking for resource-limited agents.

Chapter 5: Policy Regularization Enables Robustness and Flexibility in Motor Sequence Learning

Having demonstrated that human behavior aligns with the principles of policy compression, I now turn to the neural mechanisms that enable cost-efficient, adaptive behavior. Compression enhances generalization and robustness by allowing rewarding actions to be reused across states or contexts. However, this strategy has limitations in dynamic environments, where flexibility is essential for adapting to changing demands. This raises a fundamental question: **how do capacity-limited neural systems balance robustness with the flexibility to adapt to new environmental contexts?**

To explore this question, we must first reimagine policy compression as a form of *regularization* that encourages the adoption of simpler policies. The optimal "control" policy introduced in Chapter 1 (Eq. 1) is regularized by a "default" policy, meaning it is biased toward reusing behaviors that are rewarded across contexts to minimize representational cost (Schulman et al., 2015; Levine, 2018; Abdolmaleki et al., 2018) (Figure 4A). When necessary, the control policy can override the default to adapt to new environmental demands, though it incurs a cost proportional to its deviation from default behavior. For example, a pianist learning a new piece might rely on familiar finger patterns mastered through practice (their default policy), but also remains flexible to learn new sequences. This balance between robustness and flexibility ensures adaptive behavior while conserving cognitive resources.

Given their well-studied role in learning and action selection, I hypothesized that cortico-striatal circuits in the brain implement policy regularization to achieve this balance (Figure 4B) (Joel et al., 2002; O'Doherty et al., 2004; Niv, 2009). Within the striatum, the dorsolateral striatum (DLS) and dorsomedial striatum (DMS) serve distinct roles and compete for behavioral control (Turner et al., 2022): the DLS supports robust, habitual responding, while the DMS enables flexible, goal-directed behaviors (Killcross and Coutureau, 2003; Daw et al., 2005; Yin and Knowlton, 2006). I argue that this division of labor reflects a strategy to maximize reward while reducing cognitive costs: the DLS stores the default policy, while the DMS remains flexible to learn from environmental feedback. Additionally, cortical regions such as motor cortex (MC) and prefrontal cortex (PFC) provide the state and action representations upon which these policies are learned (Haddon and Killcross, 2006; Corbit and Balleine, 2003; Killcross and Coutureau, 2003; Nguyen et al., 2023; Mizes et al., 2024).

To test these ideas, my collaborators and I employed a motor sequence learning task paired with lesion experiments to examine how these brain regions contribute to adaptive behavior (Figure 4C). Rats were trained on two tasks: one required them to respond to visual cues with lever presses (CUE), while the other involved reproducing a consistent sequence of lever presses from memory (AUTO) (Mizes et al., 2023; Mizes, 2023; Mizes et al., 2024). The CUE task assessed animals' ability to flexibly respond to a dynamic environment, while the AUTO task evaluated their

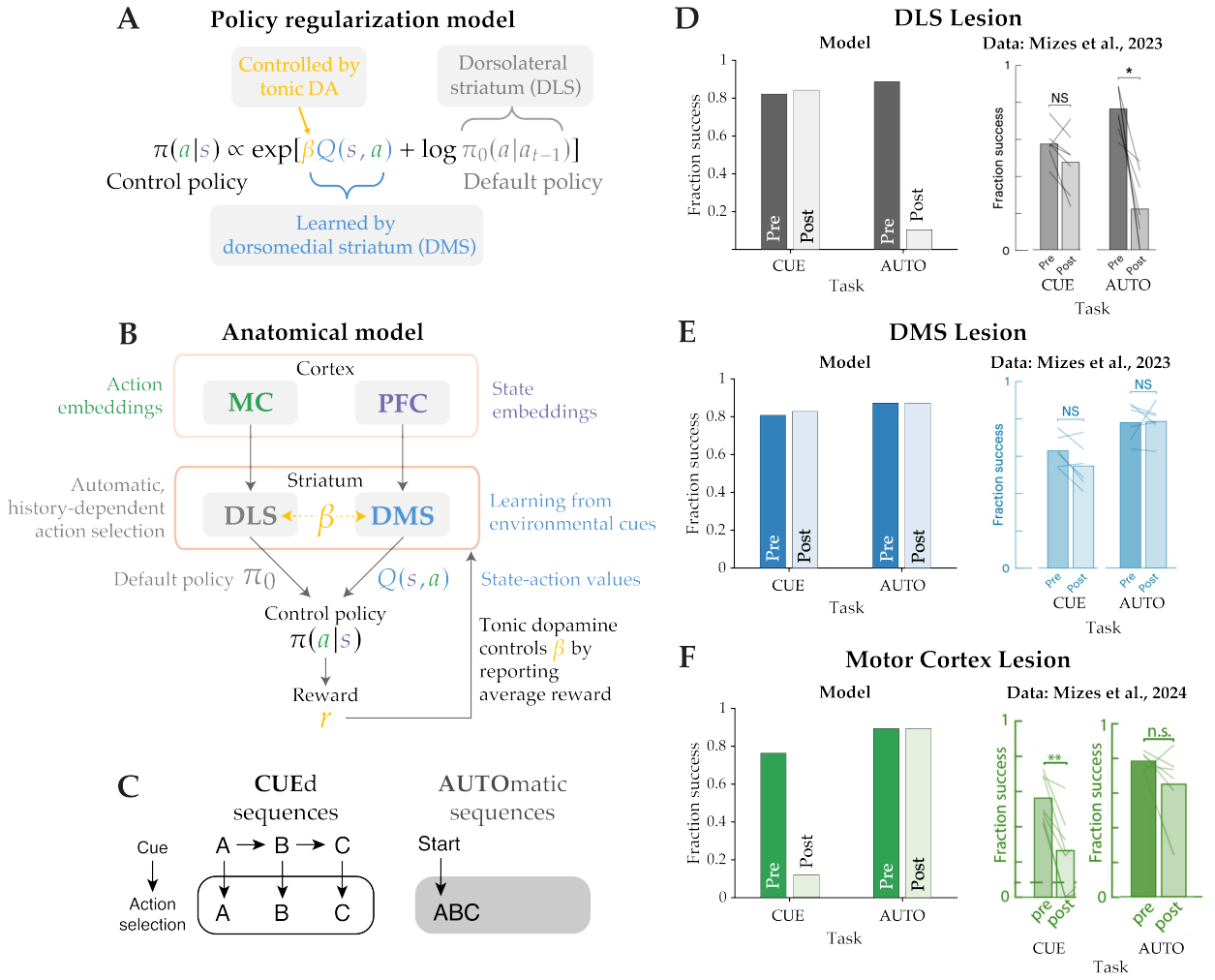

Figure 4: **Policy regularization in cortical-striatal circuits. (A)** The optimal control policy balances state-action values, learned by the dorsomedial striatum (DMS), with a default policy, learned by the dorsolateral striatum (DLS). **(B)** Proposed computational roles of cortico-striatal regions. **(C)** The "piano" task: rats responded to a sequence of visual cues with lever presses (CUEd) or reproduced a consistent lever sequence from memory (AUTOmatic). **(D-F)** Model simulations replicate key findings from lesion studies, revealing the distinct roles of the DLS, DMS, and MC in cued and automatic motor sequence execution.

execution of automatized, default behaviors—much like how a pianist can switch between reading new sheet music and playing a riff from memory.

Through model-simulated "lesions" to different brain regions, I reproduced key experimental findings (Killcross and Coutureau, 2003; Ostlund et al., 2009; Mizes et al., 2023; Bhatia et al., prep): (1) DLS lesions impaired automatic behaviors but preserved flexibility in cue-based tasks (Figure 4D); (2) DMS lesions impaired the learning of novel associations but spared learned behaviors (Figure 4E); and (3) MC lesions disrupted flexible responding to cues without affecting automatic behaviors (Figure 4F). These results align with the hypothesized computational roles of these regions within the framework of policy regularization.

Our model generated several novel predictions for future experiments: First, DMS lesions should impair learning of new state-action contingencies while preserving existing skills. Second, due to the competitive interaction between the two regions, DLS lesions should enhance learning of cued behaviors, while DMS lesions should accelerate habitual, automatic responding. Third, MC and PFC lesions should specifically disrupt skill acquisition through their inputs to DLS and DMS, respectively. Finally, increasing cognitive load should strengthen reliance on default behaviors, highlighting the adaptive value of automatization for managing limited cognitive resources.

I have detailed a normative theory of the functional organization of cortico-striatal circuits, proposing that the brain uses policy compression as a cost-efficient strategy to balance robustness and flexibility. While previous models focused on the neural dynamics of adaptive control (Mizes et al., 2023, 2024), our approach takes a different perspective by asking *why* the brain might be organized this way in the first place. By combining policy compression with existing mechanistic models, we can develop a more complete understanding of how the brain enables efficient, adaptive behavior within biological constraints.

Chapter 6: Conclusion

In this dissertation, I have explored the mechanisms underlying policy compression and its implications for cognitive resource allocation in learning and decision making. By combining computational modeling, behavioral experiments and lesion studies, I have demonstrated how a single theoretical framework can explain diverse aspects of action selection—from decision making to motor skill learning.

Looking ahead, the theory of policy compression opens numerous avenues future research. Some key questions include: (1) How do different types of cognitive cost—such as the cost of time, memory, and data—interact to shape behavior? (2) Can these cost measures help quantify and differentiate dimensions of psychiatric disorders, leading to more precise diagnostic tools and targeted interventions? (3) What roles do dopamine and other neural mechanisms play in implementing policy compression? (4) How does policy compression scale in high-dimensional and ecologically valid decision-making environments, where cognitive demands are more complex

and dynamic? By answering these questions, we can deepen our understanding of how resource constraints shape behavior and provide new tools for addressing challenges in neuroscience, psychiatry, and AI.

Beyond its theoretical contributions, policy compression offers practical applications. Understanding how humans compress and optimize behavior could inspire the design of more user-friendly technologies that better account for human cognitive limitations. It could also guide the development of more efficient machine learning algorithms by incorporating principles of human-like resource management. Finally, this framework offers new approaches for quantifying individual differences in cognitive processing, with potential applications in both clinical assessment and personalized interventions. The future of policy compression is expansive.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. (2018). Maximum a posteriori policy optimisation. *arXiv*.

Banca, P., Ruiz, M. H., Gonzalez-Zalba, M. F., Biria, M., Marzuki, A. A., Piercy, T., Sule, A., Fineberg, N. A., and Robbins, T. W. (2023). Action-sequence learning, habits and automaticity in obsessive-compulsive disorder. *Elife*, 12.

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

Beshears, J., Choi, J. J., Laibson, D., and Madrian, B. C. (2007). The importance of default options for retirement saving outcomes: Evidence from the USA. In Kay, S. J. and Sinha, T., editors, *Lessons from Pension Reform in the Americas*, page 59–87. Oxford Academic.

Bhatia, C., Hardcastle, K., and Ölveczky, B. P. (in prep). The role of dorsalmedial striatum in learning flexible motor sequences.

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

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.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., and Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nat. Hum. Behav.*, 6(8):1112–1125.

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.

Corbit, L. H. and Balleine, B. W. (2003). The role of prelimbic cortex in instrumental conditioning. *Behav. Brain Res.*, 146(1-2):145–157.

Dan, O. and Loewenstein, Y. (2019). From choice architecture to choice engineering. *Nat. Commun.*, 10(1):2808.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, 8(12):1704–1711.

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

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

Gershman, S. J. and Lai, L. (2021). The reward-complexity trade-off in schizophrenia. *Computational Psychiatry*, 5(1):38–53.

Gershman, S. J., Markman, A. B., and Otto, A. R. (2014). Retrospective revaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.*, 143(1):182–194.

Gigerenzer, G. and Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*. MIT Press.

Haddon, J. E. and Killcross, S. (2006). Prefrontal cortex lesions disrupt the contextual control of response conflict. *J. Neurosci.*, 26(11):2933–2940.

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.

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

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.*, 15(4-6):535–547.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93:1449–1475.

Killcross, S. and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb. Cortex*, 13(4):400–408.

Kool, W., Gershman, S. J., and Cushman, F. A. (2018). Planning complexity registers as a cost in metacontrol. *J. Cogn. Neurosci.*, 30(10):1391–1404.

Lai, L. and Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In *Psychology of Learning and Motivation*, volume 74, pages 195–232.

Lai, L. and Gershman, S. J. (2024). Human decision making balances reward maximization and policy compression. *PLoS Computational Biology*, 20(4):e1012057.

Lai, L., Huang, A. Z., and Gershman, S. J. (2022). Action chunking as conditional policy compression. *PsyArXiv*.

Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv*.

Lieder, F. and Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.*, 43:e1.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

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.

Mizes, K. G. C. (2023). *Distinct neural substrates for flexible and automatic motor sequence execution*. PhD thesis, Harvard University.

Mizes, K. G. C., Lindsey, J., Escola, G. S., and Ölveczky, B. P. (2023). Dissociating the contributions of sensorimotor striatum to automatic and visually guided motor sequences. *Nat. Neurosci.*, 26(10):1791–1804.

Mizes, K. G. C., Lindsey, J., Escola, G. S., and Ölveczky, B. P. (2024). The role of motor cortex in motor sequence execution depends on demands for flexibility. *Nat. Neurosci.*, 27(12):2466–2475.

Nguyen, Q. N., Michon, K. J., and Lee, T. G. (2023). Dissociable causal roles of dorsolateral prefrontal cortex and primary motor cortex as a function of motor skill expertise. *bioRxiv*, page 2023.10.20.563280.

Niv, Y. (2009). Reinforcement learning in the brain. *J. Math. Psychol.*, 53(3):139–154.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454.

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. *J. Neurosci.*, 29(25):8280–8287.

Pichert, D. and Katsikopoulos, K. V. (2008). Green defaults: Information presentation and pro-environmental behaviour. *J. Environ. Psychol.*, 28(1):63–73.

Proctor, R. W. and Schneider, D. W. (2018). Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology*, 71(6):1281–1299.

Rubinstein, A. (1998). *Modeling bounded rationality*. MIT press.

Sakai, K., Kitaguchi, K., and Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Exp. Brain Res.*, 152(2):229–242.

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. (2015). Trust region policy optimization. *arXiv*.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

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

Simon, H. A. (1957). *Models of Man*. Wiley.

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

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

Turner, K. M., Svegborn, A., Langguth, M., McKenzie, C., and Robbins, T. W. (2022). Opposing roles of the dorsolateral and dorsomedial striatum in the acquisition of skilled action sequencing in rats. *J. Neurosci.*, 42(10):2039–2051.

Verwey, W. B. (1996). Buffer loading and chunking in sequential keypressing. *Journal of Experimental Psychology: Human Perception and Performance*, 22(3):544–562.

Verwey, W. B. (1999). Evidence for a multistage model of practice in a sequential movement task. *Journal of Experimental Psychology: Human Perception and Performance*, 25(6):1693–1708.

Yin, H. H. and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nat. Rev. Neurosci.*, 7(6):464–476.