Cognition 264 (2025) 106201

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full length article

Action chunking as conditional policy compression

Lucy Lai [a,b,*,1], Ann Z.X. Huang [a,1], Samuel J. Gershman [c]

[a] Program in Neuroscience, Harvard University, United States of America
[b] Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Japan
[c] Department of Psychology and Center for Brain Science, Harvard University, United States of America

ARTICLE INFO

Dataset link: https://github.com/lucylai96/chunking/

Keywords:
Action selection
Chunking
Reinforcement learning
Resource rationality
Decision making
Information bottleneck

ABSTRACT

Many skills in our everyday lives are learned by sequencing actions towards a desired goal. The action sequence can become a "chunk" when individual actions are grouped together and executed as one unit, making them more efficient to store and execute. While chunking has been studied extensively across various domains, a puzzle remains as to why and under what conditions action chunking occurs. To tackle these questions, we develop a model of *conditional* policy compression—the reduction in cognitive cost by conditioning on an additional source of information—to explain the origin of chunking. We argue that chunking is a result of optimizing the trade-off between reward and conditional policy complexity. Chunking compresses policies when there is temporal structure in the environment that can be leveraged for action selection, reducing the amount of memory necessary to encode the policy. We experimentally confirm our model's predictions, showing that chunking reduces conditional policy complexity and reaction times. Chunking also increases with working memory load, consistent with the hypothesis that the degree of policy compression scales with the scarcity of cognitive resources. Finally, chunking also reduces overall working memory load, freeing cognitive resources for the benefit of other, not-chunked information.

1. Introduction

In his seminal 1956 paper, George Miller proposed that organizing multiple pieces of information into a single "chunk" could help circumvent the limited capacity of working memory (Miller, 1956). Chunking is thought to reduce cognitive load by representing information in a more efficient form. It has been studied extensively across various domains, including visual statistical learning (Lengyel et al., 2021; Orbán et al., 2008), visual and verbal short-term memory (Chen & Cowan, 2005; Franco & Destrebecqz, 2012; Gobet et al., 2001; Mathy et al., 2024; Nassar et al., 2018; Norris & Kalm, 2021; Orhan & Jacobs, 2013), serial order memory (Mathy & Feldman, 2012; Thalmann et al., 2019), language processing (Christiansen & Chater, 2016; Perruchet & Vinter, 1998), task set learning (Bouchacourt et al., 2020), skill learning (Du et al., 2022; Haith & Krakauer, 2018), and action sequence learning (Banca et al., 2023; Bo & Seidler, 2009; Miyapuram et al., 2006; Sakai et al., 2003; Terrace, 1991; Tosatto et al., 2022; Verwey, 1996, 1999).

While chunking encompasses a broad range of cognitive phenomena, we focus specifically on action chunking—the grouping of sequential motor responses into unified representations. In sequence learning tasks, such as the serial reaction time task, participants learn to execute a series of actions in response to cued stimuli that appear in a patterned order. The ability to group individual actions together into chunks enables faster execution times and higher accuracy (Dezfouli & Balleine, 2012; Sakai et al., 2003), a strategy also observed in animals during the acquisition of habitual behaviors (Graybiel, 1998; Jin & Costa, 2010; Jin et al., 2014). Notably, Dezfouli and Balleine (2012) modeled action chunking as an adaptive strategy for reducing decision time costs: if multi-step action chunks can be executed faster than selecting individual actions, chunking becomes advantageous by increasing overall reward rate.

These approaches emphasize the *computational* (i.e., time) costs of action selection while largely overlooking its *representational* (i.e., memory) demands. However, in other domains of cognitive science, chunking has long been studied as a strategy for compressing information in memory. Human learning and memory are believed to be constrained by limited cognitive resources (Baddeley, 1992; Cowan, 2001; Ma et al., 2014; Miller, 1956; Oberauer et al., 2016), and chunking is thought to support working memory efficiency by leveraging similarity-based compression (Chekaf et al., 2016; Kowialiewski et al., 2022; Mathy et al., 2024; Pothos, 2007). In language processing, cognitive resource

* Correspondence to: Department of Cognitive Science, University of California, San Diego, United States of America.
E-mail addresses: lai@ucsd.edu (L. Lai), annhuang@g.harvard.edu (A.Z.X. Huang), gershman@fas.harvard.edu (S.J. Gershman).
[1] Equal contribution.

https://doi.org/10.1016/j.cognition.2025.106201
Received 9 October 2024; Received in revised form 6 May 2025; Accepted 2 June 2025
Available online 1 July 2025

constraints are thought to shape the emergence of structured, learnable patterns (Christiansen & Chater, 2016). Relatedly, memory limitations have been shown to influence chunking in language acquisition (Frank et al., 2010; Goldwater et al., 2009). Computational models such as TRACX (French et al., 2011), PARSER (Perruchet & Vinter, 1998), and the competitive chunking model (Servan-Schreiber & Anderson, 1990) capture how learners form chunks by extracting structure from statistical regularities in sequences, while models based on the Minimum Description Length (MDL) principle have formalized chunking as a data compression process that reduces representational complexity (Mathy & Feldman, 2012; Robinet et al., 2011). Although these studies did not directly examine action selection, they suggest that action chunking may serve not only to reduce time costs but also to economize on limited memory storage.

In this study, we propose an alternative perspective that action chunking emerges as a consequence of bounded memory resources. Following previous work (Dezfouli & Balleine, 2012), we develop a formal model of action chunking within the reinforcement learning (RL) framework, a mathematical account of instrumental learning that describes how agents learn to associate states (stimuli) with actions (responses) in order to maximize reward (Sutton & Barto, 2018). The mapping from states to actions is called a *policy*, and storing a policy in memory places demands on cognitive resources: more complex policies require more memory to store and retrieve. We propose that chunking serves as an optimal strategy for reducing these representational costs.

Specifically, we show that action chunking emerges from an information-theoretic model that maximizes reward while minimizing the representational complexity of the policy. In the language of information theory, a policy can be thought of as a noisy communication channel that maps individual states to codewords (the internal representation), which are then decoded into actions. The channel's function is to discard redundant information about states that is not needed for effective action selection (Lai & Gershman, 2021). The average codeword length (the information rate required to encode the policy in memory) is equal to the mutual information between states and actions, $I(S;A)$, or the *policy complexity*. This complexity is bounded by an agent's *channel capacity*, a measure of its available storage space. Prior work has shown that the highest achievable expected reward is a monotonically increasing and concave function of policy complexity (Gershman, 2020; Lai & Gershman, 2021; Parush et al., 2011; Tishby & Polani, 2011), implying a trade-off between reward and policy compression: higher reward demands more complex policies, but resource limitations constrain complexity. Compression of policy representations thus comes at the cost of reward.

In our previous work (Lai & Gershman, 2021), we proposed that chunking can reduce policy complexity by collapsing states with similar optimal actions or by binding sequences of actions into a single unit in memory. For example, if two states yield the same optimal action, they can be grouped into a single memory representation (Fig. 1A; see also Lai and Gershman (2024)). Alternatively, if one state reliably follows another, their associated actions can be combined into an action chunk (Fig. 1B). In both cases, chunking reduces the memory cost of representing a policy. Furthermore, if policy complexity correlates with decision time, as we have proposed in Lai and Gershman (2021, 2024), then chunked actions (which are lower in complexity) should also be faster to execute.

In this paper, we empirically test the hypothesis that action chunking arises from *conditional* policy compression—the reduction in complexity when an agent leverages structured temporal information in the environment (Fig. 2A). We designed an RL task in which participants learned the correct action to take in different states (indicated by visual stimuli). Critically, we manipulated both the predictability of state sequences and the overall memory load (i.e., the number of states), building on previous work showing that cognitive load affects instrumental learning via "set size" effects (Collins, 2018; Collins et al., 2017; Collins & Frank, 2012, 2018). Based on our theoretical framework, we hypothesized that both temporal predictability and memory load would facilitate the formation of action chunks, which we measured as a decrease in errors and response time. Following previous work, we also predicted that chunking would free up memory resources and lead to improved performance for other, unchunked information (Kowialiewski et al., 2022; Mathy et al., 2024; Thalmann et al., 2019).

To evaluate our theory, we compare a *conditional* policy compression model—where agents maximize reward subject to a constraint on *conditional* policy complexity—to an *unconditional* policy compression model, Lai and Gershman (2021, 2024). Our empirical findings support the conditional model, providing a normative and mechanistic account of how structured temporal input and memory limitations together shape the emergence of action chunks.

2. Conditional policy compression as a model of action chunking

In this section, we adapt our policy compression framework (Lai & Gershman, 2021) to the problem of action chunking. We start from the assumption that capacity-limited agents will exploit structure in their environments to compress their policies, which can include redundancy in the reward-maximizing action in each state. This is because, if multiple states share the same optimal action, the agent does not need to pay as much attention to the state in order to select the best action (Fig. 1A and Lai and Gershman (2024)). However, the standard policy compression framework (Fig. 2A, blue rectangle) does not address how an agent might take advantage of *temporal structure* in their environment. This is an important problem, as natural environments have temporal continuity of both states and actions.

Here, we propose that action chunking takes advantage of the temporal structure of the environment to compress policies, by way of states being fully predicted by other states in time. For example, if two states have different optimal actions, but one state reliably predicts another in time (e.g., a yellow traffic light → red traffic light), then the actions associated with each state can also be chunked together as one action unit (e.g., slowing down → stopping the car). This effectively allows the agent to pay less attention to the deterministically predicted state in order to know what to do next (Fig. 1B).

As we developed in Gershman (2020) and Lai and Gershman (2021), policies can be thought of as communication channels that transmit information about the environmental state (source) to guide action selection (output). The policy *complexity* is a measure of the amount of mutual information (in bits) between states and actions. Action selection therefore requires an agent to first reconstruct the state identity (or source) in order to guide an appropriate behavioral response.

In this view, one can think about temporally-correlated state information as providing additional data for the reconstruction of the source. This "side" information is available to both the encoder and decoder (Fig. 2A). In this set up, the minimum number of bits (the minimum communication rate) needed for error-free transmission of the source identity is the *conditional mutual information* (Gray, 1972; Niu et al., 2023, Fig. 2B). In the context of action selection, we can refer to this information rate as the *conditional policy complexity*, defined as the mutual information between states and actions after conditioning on an additional information source, such as the state on the previous trial S_{t-1}:

$$I^{\pi}(S_t; A_t|S_{t-1}) = \sum_{s_{t-1}} P(s_{t-1}) \sum_{s_t} P(s_t|s_{t-1}) \times \sum_{a_t} \pi(a_t|s_t, s_{t-1}) \log \frac{\pi(a_t|s_t, s_{t-1})}{P(a_t|s_{t-1})} \quad (1)$$

In environments where $I(S_t; S_{t-1})$ is high, the current state s contains redundant information about the previous state s_{t-1}, which limits the capacity of S_t to carry unique information about A beyond what is already conveyed by S_{t-1}. We hypothesize that capacity-limited agents take advantage of this correlated side information and in effect,

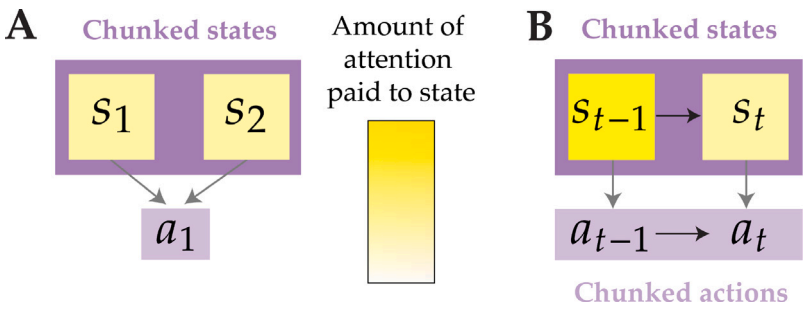

Fig. 1. Two ways to chunk. (A) If two states lead to the same action, they can be described by one codeword and become "chunked" together as one unit in memory. (B) If two states lead to two different actions, but one state is fully predicted by another state, the two-state sequence can be fully described by the first state, and the corresponding actions can be chunked together into an "action chunk". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

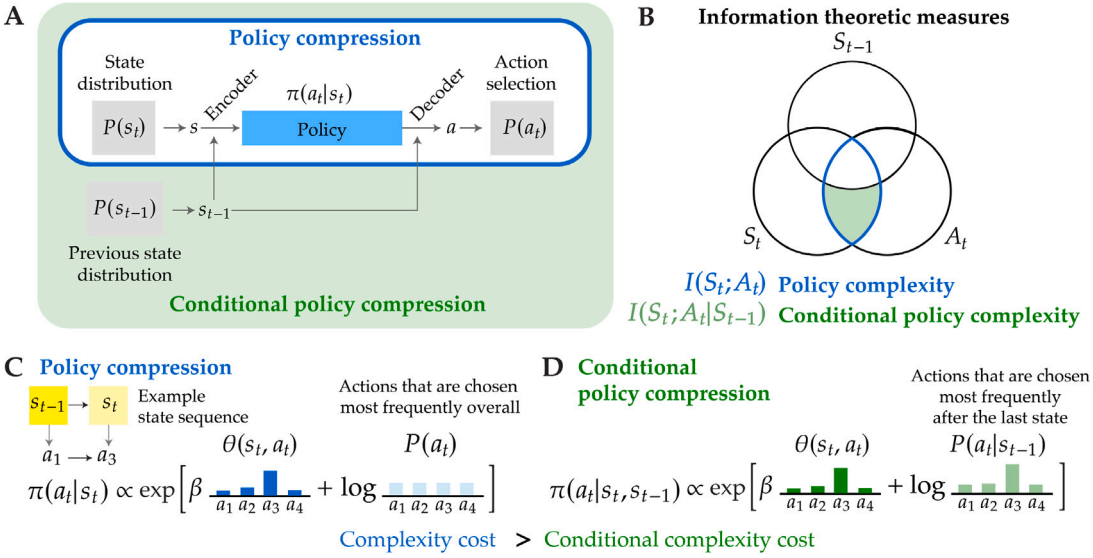

Fig. 2. Conditional policy compression. (A) The policy as a communication channel. A state distribution $P(s_t)$ generates states s that are encoded into memory via an encoder, yielding a codeword (not shown) whose length is bounded by the channel capacity (policy complexity). The codeword is then decoded and mapped onto an action a_t. Together, the encoding and action selection process produce the policy $\pi(a_t|s_t)$ that maps states to actions. Conditional policy compression introduces an additional information source, in this case the state on the previous trial s_{t-1}. This side information source is available to both encoder and decoder. (B) A Venn diagram visual of the relationship between multiple information theoretic quantities: the state distribution on the current trial (S_t), the action distribution on the current trial (A_t), and the state distribution on the previous trial (S_{t-1}). The policy complexity is the information shared by the state and action distribution, while the conditional policy complexity is the *unique* information between states and actions after accounting for the information provided by the previous state. (C) In the unconditional policy compression model, the optimal policy combines state–action values $\theta(s_t,a_t)$ with a marginal action probability term $P(a_t)$ that biases the policy towards actions that are chosen frequently across all states (in this example it is assumed to be uniform). The trade-off term, β, determines the relative contribution of $\theta(s_t,a_t)$ and $P(a_t)$, thereby controlling how state-dependent action selection is. Example distributions depict action selection in one state. (Inset) Example state sequence, where s_t reliably follows s_{t-1}, and therefore, the agent learns to choose a_3 after a_1. The unconditional model does not capture the temporal dependence of a_1 on s_{t_1} in the marginal distribution. (D) In the conditional policy compression model, the optimal policy combines $\theta(s_t,a_t)$ with a *conditional* marginal action probability term $P(a_t|s_{t-1})$ that biases the policy towards actions that are frequently chosen given a particular previous state. Given the example state sequence in (C), a_3 has high probability in the marginal term. Under the same value of β, the conditional complexity cost is less than the complexity cost, as the deviation between the action policy and the marginal distribution is smaller in the former. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

learn reward-maximizing policies subject to an upper bound on their *conditional policy complexity*:

$$\begin{aligned} &\underset{\pi}{\text{argmax}} && V^{\pi} \\ &\text{subject to} && I^{\pi}(S_t;A_t|S_{t-1}) \le C \end{aligned} \tag{2}$$

where C is the agent's capacity limit and $V^{\pi} = \mathbb{E}[r|\pi]$, the expected reward ($r$) conditional on policy π. This optimization problem can be solved in a Lagrangian form:

$$\pi^* = \underset{\pi}{\text{argmax}}\, \beta V^{\pi} - I^{\pi}(S_t;A_t|S_{t-1}) + \sum_{s_t} \lambda(s_t) \left(\sum_{a_t} \pi(a_t|s_t,s_{t-1}) - 1 \right), \tag{3}$$

with Lagrange multipliers $\beta \ge 0$ and $\lambda(s) \ge 0$ (the 3rd term ensures proper normalization, which we leave implicit in subsequent equations).

2.1. Learning action chunks

We can now adapt our cost-sensitive actor-critic learning model to the problem of action chunking. The optimization problem that the agent faces (expressed in terms of an expectation over states) is:

$$\pi^* = \underset{\pi}{\text{argmax}}\, \mathbb{E}\left[\beta r - \log \frac{\pi(a_t|s_t,s_{t-1})}{P(a_t|s_{t-1})}\right]. \tag{4}$$

To find the optimal policy π^*, the cost-sensitive agent must find the policy parameters θ^* that maximize expected reward relative to the conditional policy complexity *cost*. We use the term *cost* to refer to the trial-by-trial cost (i.e., $\log \frac{\pi(a_t|s_t,s_{t-1})}{P(a_t|s_{t-1})}$), and *complexity* to refer to its expectation (i.e., $I(S_t;A|S_{t-1})$).

For our task, we can employ a tabular parametrization with parameters θ for simplicity, although our model can also accommodate continuous state spaces (via function approximation), making it applicable

to a wider range of environments:

$$\pi_\theta(a_t = a|s_t = s, s_{t-1}) \propto \exp[\beta\theta(s,a) + \log P(a_t = a|s_{t-1})]. \quad (5)$$

This parametrization was chosen because it corresponds to the optimal functional form of the policy under the Lagrangian specified above. The parameter β determines the relative contribution of θ and $P(a|s_{t-1})$, thereby controlling how dependent action selection is on the *previous state*. As described in previous work (Lai & Gershman, 2021), β also implicitly indexes an agent's capacity constraint, with higher values indicating higher capacities and therefore, a higher dependence of actions on the current state.

After taking action a_t in state s_t and observing reward r, the policy parameters are updated via the policy gradient (i.e., the gradient of the Lagrangian with respect to the policy parameters):

$$\Delta\theta(s,a) = \begin{cases} \alpha_\theta \delta[1 - \pi_\theta(a_t|s_t = s, s_{t-1})]\beta & \text{if } a_t = a \\ -\alpha_\theta \delta \pi_\theta(a_t|s_t = s, s_{t-1})\beta & \text{if } a_t \neq a \end{cases} \quad (6)$$

where α_θ is the actor learning rate and

$$\delta = \beta r - \log \frac{\pi_\theta(a_t|s_t, s_{t-1})}{P(a_t|s_{t-1})} - \hat{V}(s_t), \quad (7)$$

is the prediction error of the critic $\hat{V}(s)$, which is updated according to:

$$\Delta\hat{V}(s_t) = \alpha_V \delta, \quad (8)$$

where α_V is the critic learning rate. We estimate the marginal action probabilities with an exponential moving average:

$$\Delta P(a_t|s_{t-1}) = \alpha_P[\pi_\theta(a_t|s_t, s_{t-1}) - P(a_t|s_{t-1})], \quad (9)$$

where α_P is a learning rate parameter and $P(a_t|s_{t-1})$ is the probability of taking action a_t given that the previous state was s_{t-1}.

Finally, the trade-off parameter β is adaptively optimized to increase the *conditional* policy complexity up to the agent's capacity constraint, C:

$$\Delta\beta = \alpha_\beta \left(C - \hat{I}\right), \quad (10)$$

where $\hat{I}$ is the agent's estimate of its own conditional policy complexity, updated with an exponential moving average:

$$\Delta\hat{I} = \alpha_I \left[\log \frac{\pi_\theta(a_t|s_t, s_{t-1})}{P(a_t|s_{t-1})} - \hat{I}\right], \quad (11)$$

with learning rate α_I. We will refer to this as the "conditional policy compression" model.

2.2. The relationship between conditional policy complexity and response times

To generate response times (RTs) from our models, we made two linking assumptions. The first comes from the algorithmic relationship between policy complexity and decoding time. RTs partly reflect how long it takes to decode a policy, which in turn depends linearly on the code length under standard algorithms like Huffman decoding (Lai & Gershman, 2021, 2024). We therefore assumed that RT is linearly related to the policy cost (the code length for the policy on a given trial), which indexes the cost of taking a specific action a in the current state s by quantifying the deviation of the policy $\pi(a_t|s_t, s_{t-1})$ from the conditional action probability $P(a_t|s_{t-1})$: $\log \frac{\pi_\theta(a_t|s_t,s_{t-1})}{P(a_t|s_{t-1})}$. Policy complexity is the expectation of the policy cost. Under this assumption, RTs will be slower on trials with greater deviation between the current policy and the conditional action distribution.

The second assumption comes from learning effects: empirical RTs are known to decrease over time, since greater "action uncertainty" at the onset of learning produces slower initial RTs (Hick, 1952; McDougle & Collins, 2020; Proctor & Schneider, 2018). As a policy is learned, it becomes less stochastic as participants discover the correct action to take in each state. We assumed that RT is monotonically related to the entropy of the policy on a given trial:

$$H = -\sum_{a_t} \pi_\theta(a_t|s_t, s_{t-1}) \log \pi_\theta(a_t|s_t, s_{t-1}). \quad (12)$$

To summarize, conditional policy cost measures how much influence the previous state has on action selection in the current state, and policy entropy measures how variable the policy is at the current timestep. Combining these two quantities, we can specify a simple linear regression model relating policy cost and entropy to response time (in milliseconds; see also Ballard and McClure (2019)):

$$\log \mathrm{RT} = \log \left[t_0 + b_1 \left(\log \frac{\pi_\theta(a_t|s_t)}{P(a_t)}\right) + b_2 H\right] + \epsilon, \quad (13)$$

where t_0 is non-decision time and $\epsilon \sim N(0, \sigma^2)$ is Gaussian random noise.

3. Experimental methods

3.1. Instrumental learning task

We developed an instrumental learning task that allows us to test the predictions of our theoretical framework. A key feature of this task is the presence of structured transitions between states on some blocks. We predicted that participants would exploit these structured transitions to compress their policies via chunking.

Participants had to learn the correct key press response (action) associated with each image (state) through trial-and-error (Fig. 3A). Specifically, participants were presented with one image at a time, and had to learn which key to press to obtain a deterministic reward. On each trial, the image was presented for a maximum duration of 2.5 s and the trial ended as soon as a key press was made. If no key was pressed after 2.5 s, the next trial began automatically. Feedback was presented after each trial as either an orange border (correct) or no border (incorrect) around the image, indicating a reward value of +1 or 0, respectively (Fig. 3B). Each stimulus was presented 20 times. Participants were instructed to respond as accurately and fast as possible to obtain a performance-based monetary bonus proportional to the amount of average reward they earned in the task.

Each state was uniquely associated with one rewarded action. Despite the one-to-one correspondence between states and correct actions, we tried to encourage independent learning of actions across states by instructing the participants that finding the correct action for one state was not informative about the correct action for another state, so that multiple states could share the same correct action. Each block of the experiment was either a "Random" block or a "Structured" block. In Random blocks, the sequence order of states was completely random, meaning that the transition probability from one state to another was uniform across the state space. In Structured blocks, sequences of two states where one state was deterministically followed by another (e.g., s_1 is always followed by s_3) consistently reappeared in the block, and the rest of the states were randomly distributed in the sequence (Fig. 3C). This meant that the states that were part of the deterministic state sequence always appeared together in the block, and never independently. We designed the Structured block in this way assuming that participants might exploit this predictable temporal structure and learn corresponding action "chunks" (e.g., $a_3 \rightarrow a_1$) that reduce their cognitive load. Participants were not given any block-specific instructions that distinguished Random from Structured blocks, though different stimuli were used each block to eliminate any confound of stimulus familiarity (i.e., each block was learned *de novo*).

To test our model's prediction that a higher cognitive load should increase action chunking, we also manipulated the number of distinct states in a block (i.e., the set size). Each participant completed the four block types for two different set sizes, Ns = 4 and Ns = 6 (Ns: number of states). The number of actions also matched the number of

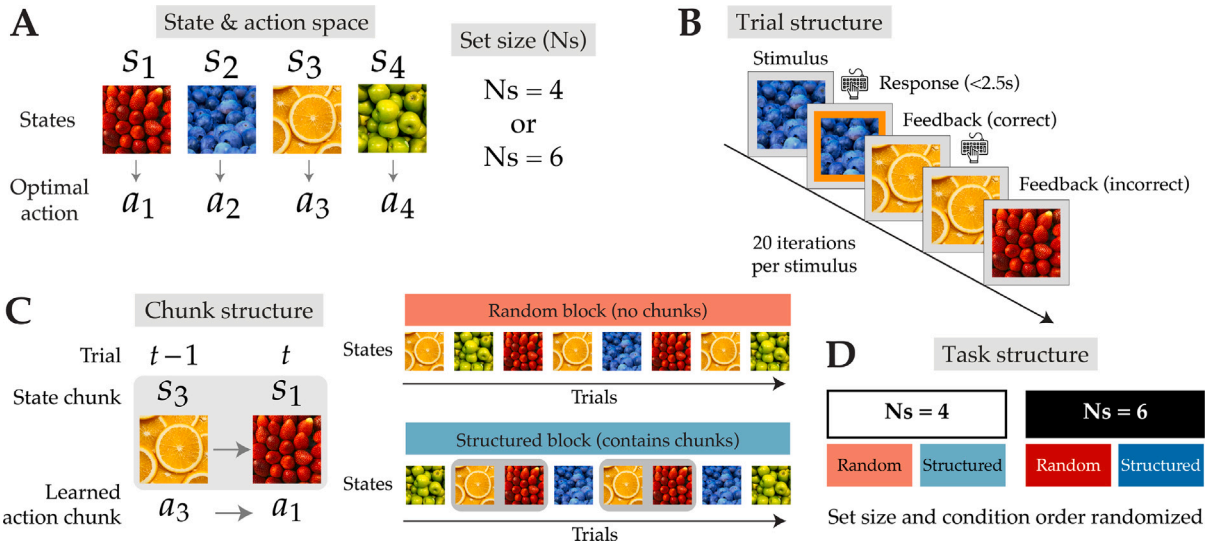

Fig. 3. Instrumental learning task. (A) The state and action space for the Ns = 4 task. Each state (image) has a unique correct action (key press). The state and action space for the Ns = 6 task was the same except that there were 6 distinct states and actions. (B) An example trial structure. Participants press a key and receive feedback in the form of an orange border around the image when the correct action was selected, and no border when the incorrect action was selected. Each stimulus is repeated for 20 iterations. (C) An example of a state "chunk" that may appear in the sequence of trials. In this case, the strawberry image always follows the orange image, encouraging participants to learn a 2-action chunk, $a_3 \rightarrow a_1$. In Random blocks (red), all states are presented in a random sequence, while in Structured blocks (blue), the state chunk is embedded throughout the state sequence, and the non-chunk states are randomly intermixed in between. (D) Each participant completed both block types (Random and Structured) for both set size conditions (Ns = 4 and Ns = 6). Set size and block condition were randomized across subjects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

states (4 actions for Ns = 4, 6 actions for Ns = 6), since there was a deterministic mapping from states to actions. Distinct sets of stimuli were also used across set sizes. Each participant completed both block conditions (Random and Structured) for one set size before moving onto the next set size (Ns = 4 and Ns = 6). The set size and block conditions were randomized across participants (3D). By comparing the task performance and RTs between the Random and Structured blocks, we can assess whether people are forming action chunks when there is temporal structure in their environment.

3.2. Participants

Eighty-one participants completed our study on Amazon Mechanical Turk and received monetary compensation. All subjects gave electronic written consent before beginning the study. Participants were paid a base pay of $4 and a performance bonus of up to $8 for completing the task. On average, the overall pay was $10.50. We excluded 5 participants for nonsensical key-pressing behavior (i.e., pressing the same key over and over again with a response time < 200 ms) or a lack of responses for more than 20 consecutive trials (i.e., the participant let the experiment run without engaging in the task). This left us with 76 participants for data analysis.

3.3. Model fitting

We compared the conditional policy compression model with the unconditional policy compression model. The models only differ in what they are optimizing, either the policy complexity $I(S_t; A_t)$, or the conditional policy complexity $I(S_t, A_t|S_{t-1})$, but otherwise share the same learning rules (in effect, the unconditional policy compression model replaces $\pi(a_t|s_t, s_{t-1})$ and $P(a_t|s_{t-1})$ in Eqs. (2)–(11) with $\pi(a_t|s_t)$ and $P(a_t)$, respectively; see also 2C and D). The two models were equivalent in their number of free parameters, which are summarized in Table 1.

We used maximum likelihood estimation to jointly fit the choice and response time data for each subject individually. Parameter constraints were defined according to Table 1. We fixed the non-decision time t_0 to 300 ms and σ to 0.5 for all models, following previous work (Lai & Gershman, 2024). We also fixed the following learning rates to avoid degeneracy in the parameter space: α_p, α_I, and $\alpha_\rho = 0.01$ (which are the learning rates for the default policy, policy complexity, and average reward, respectively). We note that fixing these parameters does not typically change the qualitative predictions of each model. We chose to focus on these qualitative predictions because standard quantitative model comparison metrics did not strongly discriminate between the models for these data.

Table 1
Fitted parameters and their bounds.

Parameter	Range
C	[0, 2]
β_0	[1, 10]
α_θ	[0, 1]
α_V	[0, 1]
α_β	[0, 1]
b_1	[0, 500]
b_2	[0, 500]
Total # Parameters	7

3.4. Calculating empirical policy complexity

Empirical policy complexity was estimated from each subject's behavior per task condition. We did this using a custom-written MATLAB package, which is available at https://github.com/lucylai96/chunking.

4. The benefits of chunking

To reiterate our main hypothesis, we argue that chunking and its benefits (reduced error and response time) arise as a result of the pressure to reduce conditional policy complexity when working memory is limited. In the Introduction, we defined policy complexity as a measure of the amount of memory that must be devoted to state information during action selection. As discussed earlier, one way of reducing complexity is by choosing the same action regardless of state, or in other words, by bringing the action distribution closer to the marginal action probability $P(a_t)$ via the β parameter in the policy

Table 2
Information measures for Random and Structured sequences. For simplicity, we represent states as numbers and assume a perfect agent that produces the correct action for each state. The first 20 states in the sequence are shown, but the information measures were calculated on a sequence of 80 states, with each state appearing 20 times (as in our experiment). State chunks are **bolded** in the Structured examples. Note that for both Random and Structured sequences, the unconditional policy complexity $I(S_t; A_t)$ is the same. However, when state chunks are present, $I(S_t; S_{t-1})$ increases for the sequence, which decreases $I(S_t; A_t|S_{t-1})$.

Block type	Sequence	$I(S_t; A_t)$	$I(S_t; S_{t-1})$	$I(S_t; S_{t-1}\|A_t)$	$I(S_t; A_t\|S_{t-1})$
Random	[4, 1, 1, 2, 2, 4, 2, 2, 4, 4, 3, 3, 3, 4, 1, 1, 3, 1, 3, 2...]	1.301	0.104	0.083	1.28
Random	[3, 2, 1, 4, 3, 2, 2, 1, 3, 2, 4, 3, 4, 1, 4, 1, 2, 3, 4, 1...]	1.302	0.159	0.087	1.229
Structured	[4, 3, 3, 4, **2, 1,** 4, 3, 3, **2, 1,** 4, 4, **2, 1,** 3, 3, **2, 1,** 3...]	1.300	0.633	0.120	0.788
Structured	[**2, 1,** 3, **2, 1,** 4, 3, 3, **2, 1,** 3, 4, **2, 1,** 3, 4, **2, 1,** 4, 3...]	1.301	0.626	0.119	0.793

(Fig. 2C). Chunks are another particular way of ignoring state information during action selection, by bringing the action distribution closer to the *conditional* action probability $P(a_t|s_{t-1})$. The conditional policy compression model implements the benefits of action chunking by allowing the agent to leverage information from the previous state and ignore some incoming information about the current state for action selection. As a result, actions are selected faster *and* more accurately, because the agent can simply "look up" its next action based on the previous state. This is an important departure from previous applications of policy compression, because we are asserting that the behaviorally-relevant unit of cognitive load (given redundancy in the environment) is the conditional policy complexity.

Viewed another way, the benefits of chunking are enabled by the fact that temporal predictability from $s_{t-1} \rightarrow s_t$ increases both $I(S_t; S_{t-1})$ and $I(S_t; S_{t-1}|A_t)$ and thus decreases the conditional policy complexity $I(S_t; A_t|S_{t-1})$, which can be seen in the following decomposition derived from the chain rule of mutual information:

$$I(S_t; A_t|S_{t-1}) = I(S_t; A_t) - I(S_t; S_{t-1}) + I(S_t; S_{t-1}|A_t). \quad (14)$$

The higher $I(S_t; S_{t-1})$ is (the more redundancy between S_t and S_{t-1}), the greater the benefits of chunking should be. The quantity $I(S_t; S_{t-1}|A_t)$ measures how much information the previous state S_{t-1} provides about the current state S_t after conditioning on the agent's current action A_t. For perfect action selection in a Random environment, $I(S_t; S_{t-1}|A_t) \approx I(S_t; S_{t-1})$[2] and the two terms cancel out, while in Structured environments, $I(S_t; S_{t-1}) > I(S_t; S_{t-1}|A_t) > 0$. Knowing the current action may account for some of the relationship between S_t and S_{t-1} when an agent is in a state chunk ($I(S_t; S_{t-1}) > I(S_t; S_{t-1}|A_t)$); however, even after conditioning on the action, the previous state may still provide additional information about the current state ($I(S_t; S_{t-1}|A_t) > 0$).

To illustrate the relationships from Eq. (14) more concretely, Table 2 shows the corresponding information measures for four example sequences, two Random and two Structured (from the Ns = 4 condition). Note that while the unconditional policy complexity $I(S_t; A_t)$ is similar across all sequences (assuming a perfect agent that always produces the correct action for each state), the conditional policy complexity $I(S_t; A_t|S_{t-1})$ is significantly reduced for the Structured sequences.

In the following sections, we test the conditional policy compression model against the unconditional policy compression model to confirm four hypotheses: (1) chunking increases accuracy and reduces response time when there is structure in the environment, (2) people seek to maximize reward while reducing conditional policy complexity, (3) chunking increases under a higher cognitive load, and (4) chunking frees cognitive resources for the benefit of other, not-chunked information.

5. Task produces behavioral features of action chunking

We first showed that our task was able to produce behavioral evidence of action chunking similar to previous studies. Previous work using the serial reaction time task (which our task is a simplified version of) has reported faster learning, higher accuracy, and shorter response times in conditions where there exists a predictable sequence of states (Desmurget & Turner, 2010; Matsuzaka et al., 2007; Sakai et al., 2003).

Fig. 4A shows how participants' accuracy evolves as a function of the number of trials per stimulus. Participants learned faster in Structured blocks than in Random blocks, and in the Ns = 4 (dotted lines) versus Ns = 6 (solid lines) set size conditions. Participants achieve higher average accuracy in Structured blocks than in Random blocks [mixed-effects ANOVA: $F(1300) = 16.93$, $p < 0.001$] and in the Ns = 4 versus Ns = 6 set size condition [$F(1300) = 32.56$, $p < 0.001$]. There was no interaction between block type and set size [$F(1300) = 0.19$, $p = 0.660$].

The difference in performance between Random and Structured blocks is mirrored by the conditional policy compression model (Fig. 4B), but not by the unconditional policy compression model (Fig. 4C). This highlights an important qualitative difference between the models in how well they capture participants' use of structured temporal information. However, as expected, both models are able to reproduce the performance difference between set sizes, as sensitivity to memory load is a key feature of policy compression.

Next, we analyzed how participants' RTs evolved over the course of learning (Fig. 5A). Participants' average RT was shorter in Structured blocks than in Random blocks [mixed-effects ANOVA: $F(1300) = 28.41$, $p < 0.001$], and for the Ns = 4 versus Ns = 6 condition [$F(1300) = 59.63$, $p < 0.001$]. Again, we see that the difference in RTs between Random and Structured blocks is captured by conditional policy compression (Fig. 5B), but not by the unconditional policy compression (Fig. 4C), though both models do reflect the RT difference between set size conditions. This is because, as hypothesized, participants are leveraging information from the previous state to make action selection faster in the current state. Since the unconditional policy compression model does not have access to previous state information, it cannot account for this RT benefit in Structured blocks.

6. Chunking reduces conditional policy complexity

One assumption of the conditional policy compression model is that participants' *conditional* policy complexity, which quantifies the unique information shared by states and actions given the previous state, is upper-bounded by their capacity constraint. In contrast, the unconditional policy compression model assumes that the constraint applies instead to participants' (marginal) policy complexities. The two models make qualitatively different predictions when it comes to

[2] Theoretically, if there are truly no statistical regularities in the Random blocks, then the conditional and unconditional policy complexities should be approximately equal: $I(S_t; A_t|S_{t-1}) = I(S_t; A_t)$. The small discrepancies in Table 2 and Fig. 6 can be explained by the fact that Random blocks may still contain some unintended patterns or regularities because of our pseudorandom experimental design ("pseudo" because we had to ensure that every stimulus appeared 20 times). We verified this by generating random sequences of four or six states and calculating the mutual information between the current and previous state $I(S_t; S_{t-1})$ for different numbers of trials per state (stimuli). It seems that for the number of trials in our experiment, there will always be a small discrepancy between the conditional and unconditional complexities (i.e., $I(S_t; S_{t-1}) \approx 0.1$) in Random blocks. If we increase the number of trials per stimuli, this discrepancy decreases. The remaining discrepancies may have to do with variance in our mutual information estimator.

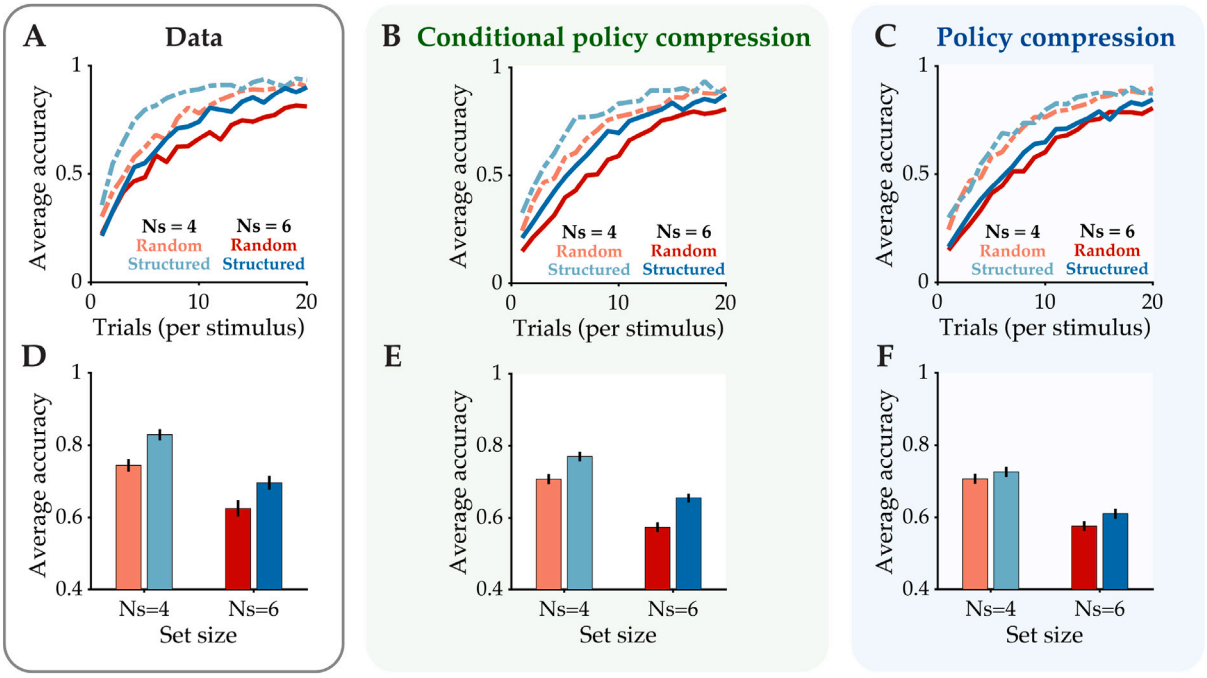

Fig. 4. Accuracy increases during learning. (A) Participants' average accuracy as a function of the number of trials (per stimulus). (B) Same as (A) for data simulated from the conditional policy compression model. (C) Same as (A) for data simulated from the policy compression model. (D) Participants' average accuracy as a function of set size. (E) Same as (D) for data simulated from the conditional policy compression model. (D) Same as (D) for data simulated from the policy compression model. All error bars indicate standard error.

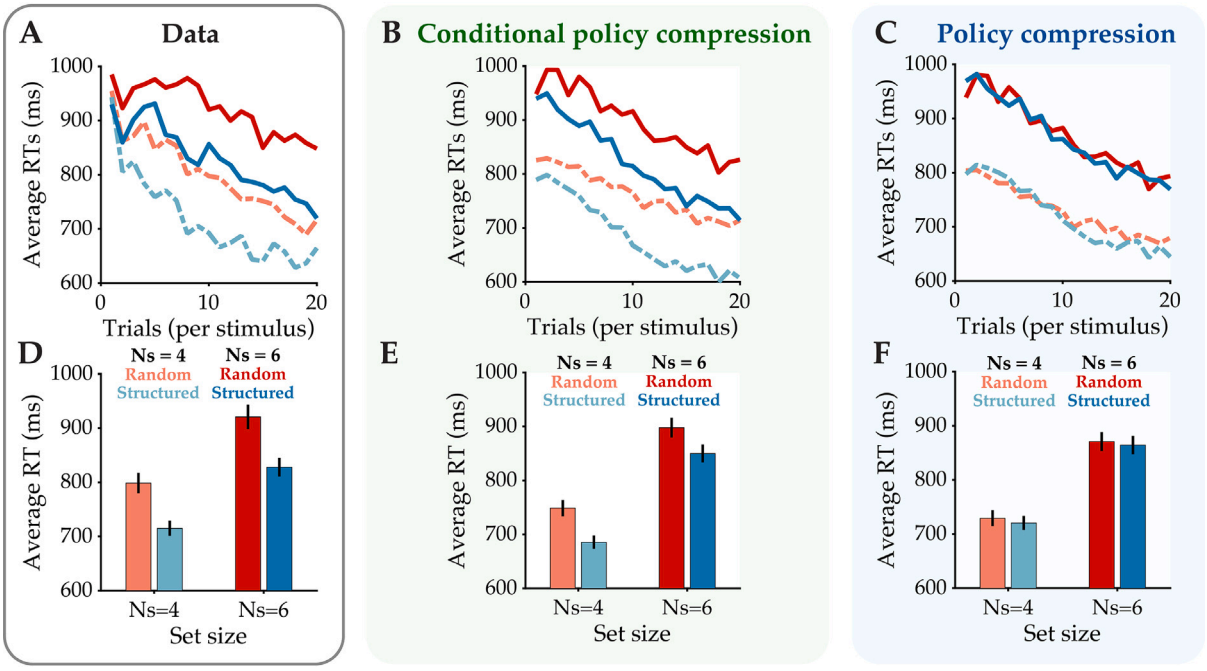

Fig. 5. Response time decreases during learning. (A) Participants' average response times (RT) as a function of the number of trials (per stimulus). (B) Same as (A) for data simulated from the conditional policy compression model. (C) Same as (A) for data simulated from the policy compression model. (D) Participants' average RTs as a function of set size. (E) Same as (D) for data simulated from the conditional policy compression model. (D) Same as (D) for data simulated from the policy compression model. All error bars indicate standard error.

empirical data: assuming that the task demands exceed participants' capacity limit, the unconditional policy compression model predicts that participants' policy complexities should approach a roughly similar value across block types and set size conditions (a result we observed in Gershman & Lai, 2021), while the conditional policy compression model predicts the same pattern for the conditional policy complexity.

To analyze these two possibilities, we computed the empirical conditional and unconditional policy complexities for each participant in each experimental block and set size condition (see Methods). Fig. 6A and B shows participants' average reward plotted as a function of policy complexity, broken down by set size and block type, while Fig. 6E and F show the same for conditional policy complexity. Several features of the data stand out: average policy complexity was higher in Structured than

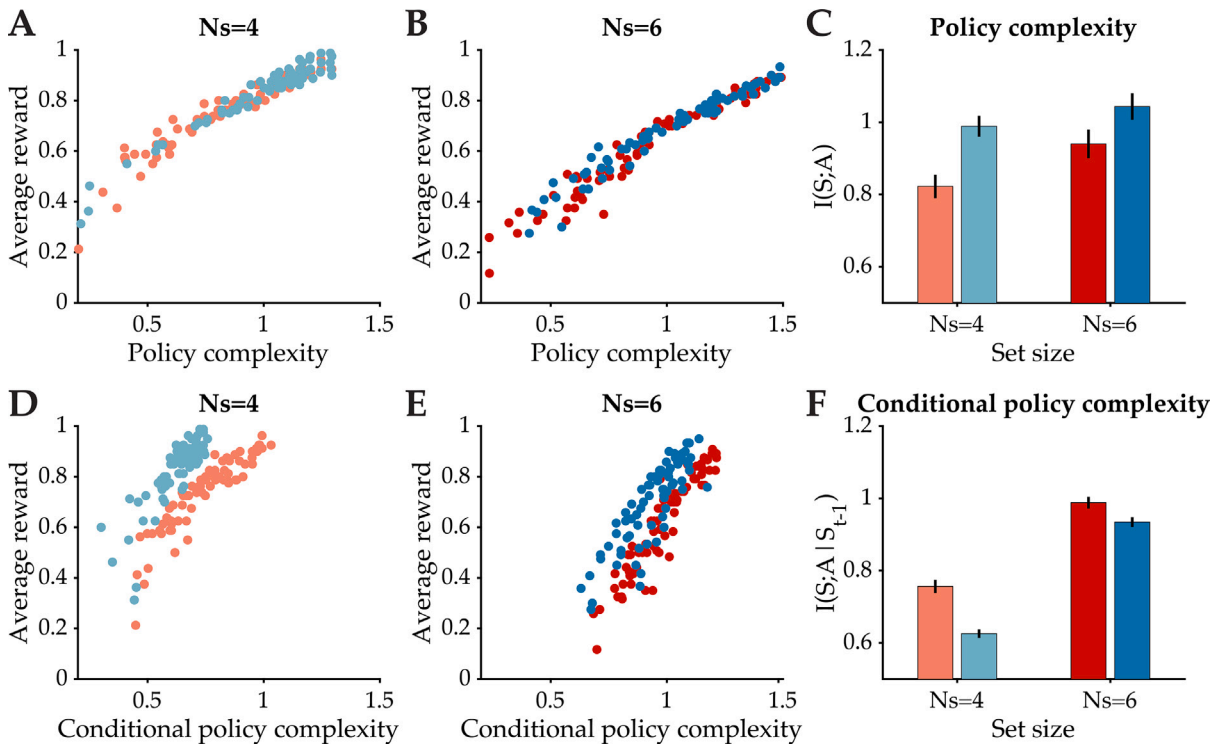

Fig. 6. Reward-complexity trade-offs. (A) Participants' average reward as a function of policy complexity for the Ns = 4 condition. Each data point on the plot represents a single subject's performance in one block condition (Structured or Random). (B) Participants' average reward as a function of policy complexity for the Ns = 6 condition. (C) Average policy complexity as a function of set size. (D) Participants' average reward as a function of conditional policy complexity for the Ns = 4 condition. (E) Participants' average reward as a function of conditional policy complexity for the Ns = 6 condition. (F) Average conditional policy complexity as a function of set size. All error bars indicate standard error.

in Random blocks (Fig. 6C) [mixed-effects ANOVA: $F(1300) = 20.84$, $p < 0.001$] and for the Ns = 4 versus Ns = 6 condition [$F(1300) = 21.75$, $p < 0.001$] (Fig. 6C). However, *conditional* policy complexity was *lower* on average in Structured versus Random blocks (Fig. 6F) [mixed-effects ANOVA: $F(1300) = 58.86$, $p < 0.001$], and higher for the Ns = 6 condition [$F(1300) = 174.98$, $p < 0.001$] (Fig. 6G).

Given these data, we argue that action chunking allows capacity-limited agents to leverage temporal structure in the environment to reduce their conditional policy complexity. Not only does this significantly reduce cognitive load, it also defines a new relationship between reward and complexity. As defined in the general policy compression framework, the reward-complexity trade-off curve defines the optimal frontier of performance for a range of policy complexities. As seen in Fig. 6A and B, policy complexity in both Random and Structured blocks fall on the same trade-off curve. This means that the only way to increase reward is to increase one's policy complexity. However, when considering conditional policy complexity, temporal structure changes the relationship between reward and complexity, making it possible to earn *more* reward with a *less complex* policy. By maximizing reward while minimizing conditional policy complexity, people leverage temporal structure to achieve the benefits of action chunking.

7. Chunking increases under a higher cognitive load

We next tested the prediction that action chunking increases under a higher cognitive load, which forces resources to be distributed over a greater number of items (Ma et al., 2014; Sims et al., 2012). In these situations, it is advantageous to chunk, as one can reduce demands on memory by eliminating the need to encode some state information: by selecting an action with high marginal probability *conditional on the previous state*, the agent no longer needs to pay attention to the upcoming state.

We expected several behavioral consequences resulting from higher cognitive load, which we validated in our initial analysis: (1) more overall error (which we quantified as lower accuracy) in Ns = 6, as a higher demand on cognitive resources means that each state is encoded with less precision, and (2) overall higher RTs in Ns = 6, because it takes longer to decode the correct action from the state when the state–action space is larger. We also expect several advantages of chunking to be more pronounced under higher load: (1) a greater reduction of intra-chunk errors in Ns = 6 Structured blocks and (2) a greater reduction in intra-chunk RTs in Ns = 6 Structured blocks, because action selection is faster and more accurate when the state sequence is predictable. We chose to specifically examine the errors and response times in the state that is fully predicted by the preceding state (hence, "intra-chunk") as opposed to looking at all of the errors/RTs, because behavior in the intra-chunk state directly captures the benefit of temporal predictability. For example, if the state sequence was $[s_2 \rightarrow s_1]$, the intra-chunk error would be the proportion of trials where the incorrect action was selected, and the intra-chunk RT would be the time it takes to select an action, in response to s_1.

In Fig. 7A and B, we plot the intra-chunk error as a function of intra-chunk trials for both set sizes. Intra-chunk error decreased overall over the course of learning in all task conditions. We found a weak effect of block type [mixed-effects ANOVA: $F(1300) = 12.69$, $p = 0.108$] and a significant effect of set size [$F(1300) = 12.69$, $p < 0.001$]. Contrary to our prediction, we did not find a significant difference in the decrease in intra-chunk error between the two set sizes [paired-sample t-test: $t(75) = -0.52$, $p = 0.60$]. As expected, the conditional policy compression model does predict a greater reduction of intra-chunk errors from the Random to Structured block in the higher load condition (Fig. 7E–H), while the unconditional policy compression model does not (Fig. 7I–L). We believe the lack of a significant decrease in intra-chunk errors between set sizes may reflect between-subject variability in how participants exploit temporal structure: some individuals may leverage chunking more effectively as cognitive load increases. Additionally, we point out that the group-averaged data in Fig. 7A–D resemble a hybrid

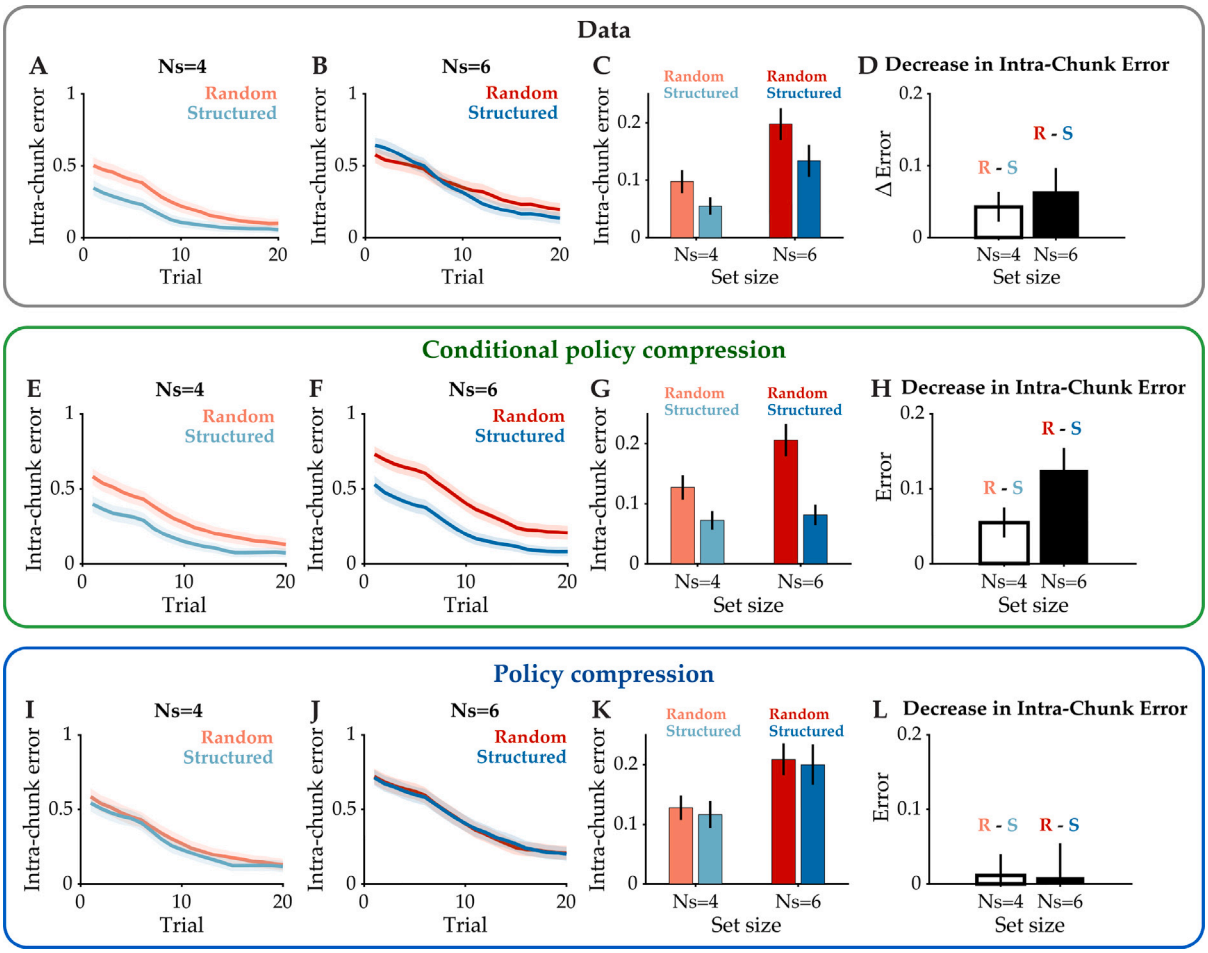

Fig. 7. Reduction in intra-chunk error during learning. (A) Participants' intra-chunk error as a function of intra-chunk trials for the Ns = 4 set size condition. Shaded region indicates standard error. (B) Participants' intra-chunk error as a function of intra-chunk trials for the Ns = 6 set size condition. (C) Asymptotic intra-chunk error (averaged over the last 5 trials) as a function of set size. (D) The decrease in intra-chunk error (Random-Structured) as a function of set size. (E–H) Same as (A–D) but for data stimulated from the conditional policy compression model. (I–L) Same as (A–D) but for data stimulated from the unconditional policy compression model.

of the conditional and unconditional policy compression models, which further supports the idea that participants differ in the extent to which they utilize temporal regularities to guide action selection.

Next, we analyzed intra-chunk RTs. In Fig. 8A and B, we plot the intra-chunk RT as a function of intra-chunk trials for both set sizes. When the stimulus sequence is predictable (Structured), intra-chunk RTs decrease with learning. In contrast, when the stimulus sequence is Random, reaction times do not decrease substantially during the course of learning, a result that is consistent with previous empirical findings and models (Dezfouli & Balleine, 2012). Overall, intra-chunk RT was significantly faster in Structured blocks [mixed-effects ANOVA: $F(1300) = 7.63$, $p = 0.006$] and in the Ns = 6 condition [$F(1300) = 24.42$, $p < 0.001$]. Additionally, the decrease in intra-chunk RTs from Random to Structured blocks was greater in Ns = 6 [$F(1300) = 4.39$, $p = 0.036$; paired-sample t-test: $t(75) = -2.06$, $p = 0.043$] (Fig. 8C and D), confirming our hypothesis. Only the conditional policy compression model was able to predict a greater reduction of intra-chunk RTs from the Random to Structured block in the higher load condition (Fig. 7E–H), while the unconditional policy compression model makes no such distinction (Fig. 7I–L).

In summary, we found partial evidence (from RTs but not errors) that participants chunked more under a higher memory load. Temporal structure in the state sequence helps people select actions faster, though not necessarily more accurately, under a higher cognitive load. This is uniquely predicted by the conditional policy compression model, in which actions that are commonly chosen after certain states are cognitively "cheaper," and therefore faster, to execute.

8. Chunking frees cognitive resources for not-chunked information

Previous work investigating the effects of chunking on cognitive load have suggested that by reducing overall working memory load, chunking benefits should be observed not only for the chunked, but also for other, non-chunked information (Kowialiewski et al., 2022; Mathy et al., 2024; Thalmann et al., 2019). Though neither of our policy compression models robustly predicts this,[3] we analyzed the proportion of errors and RTs in non-chunk trials to test the hypothesis that chunking confers benefits beyond just the intra-chunk states.

We first examined how non-chunk error changes as function of block type and set size (Fig. 9A and B). Non-chunk error decreased over

[3] In principle, the conditional model should be able to predict this result under certain parameters, because the conditional action probability is also informative for non-chunk action selection. Though the effect should be less drastic than in intra-chunk states, the conditional provides additional information in the policy about what action to not select (as only one action should be exclusively expressed in intra-chunk states, leaving a high likelihood for the other three possible actions), which should improve accuracy. This information should also reduce RT by bringing the policy slightly closer to the marginal in Structured, but not Random, blocks. Unfortunately, this effect is not clearly predicted with our current set of fitted parameters. We leave a more detailed investigation for future modeling work.

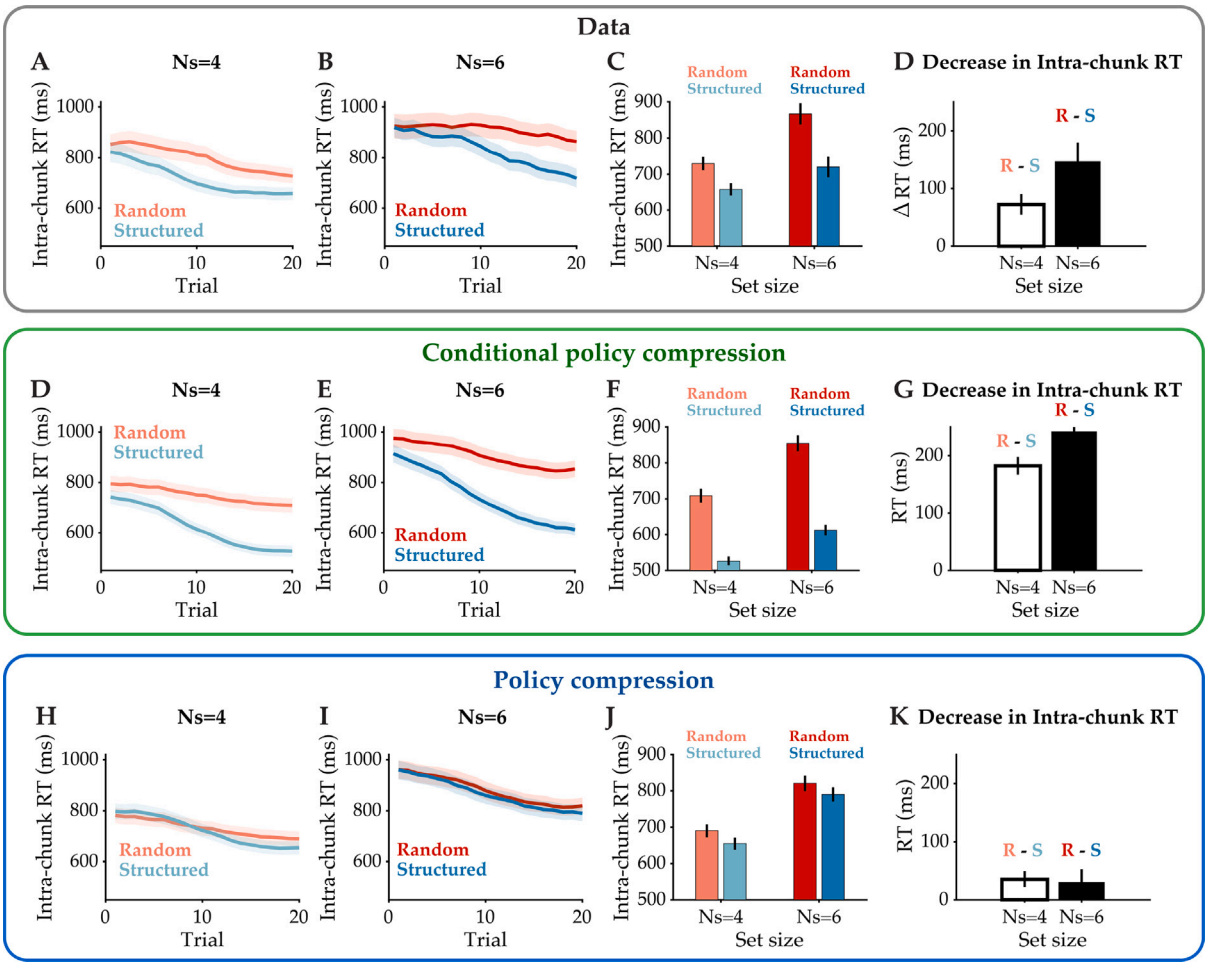

Fig. 8. Reduction in intra-chunk response times under higher cognitive load. (A) Participants' intra-chunk RT as a function of intra-chunk trials for the Ns = 4 set size condition. Shaded region indicates standard error. (B) Participants' intra-chunk RT as a function of intra-chunk trials for the Ns = 6 set size condition. (C) Asymptotic intra-chunk RTs (averaged over the last 5 trials) as a function of set size. (D) The decrease in intra-chunk RT (Random-Structured) as a function of set size. (E–H) Same as (A–D) but for data stimulated from the conditional policy compression model. (I–L) Same as (A–D) but for data stimulated from the unconditional policy compression model.

learning but, similar to intra-chunk error, was not significantly different between Structured and Random blocks [mixed-effects ANOVA: $F(1300) = 0.503$, $p = 0.478$]. As expected, intra-chunk error was overall higher in the Ns = 6 condition [$F(1300) = 17.94$, $p < 0.001$], and block type weakly interacted with set size: the decrease in non-chunk error from Random to Structured blocks was slightly greater in Ns = 6 [$F(1300) = 3.05$, $p = 0.081$; paired-sample t-test: $t(75) = 3.29$, $p = 0.071$].

We then analyzed non-chunk RTs: in Fig. 9E and F, we plot non-chunk RT as a function of non-chunk trials. Like intra-chunk RTs, non-chunk RTs also decrease with learning. Overall, non-chunk RT was significantly faster in Structured blocks [mixed-effects ANOVA: $F(1300) = 8.64$, $p = 0.003$] and in the Ns = 6 condition [$F(1300) = 45.96$, $p < 0.001$]. There was also a weak interaction between block type and set size: the decrease in non-chunk RTs from Random to Structured blocks was slightly greater in Ns = 6 [$F(1300) = 3.05$, $p = 0.081$; paired-sample t-test: $t(75) = -1.71$, $p = 0.092$].

Taken together, we found moderate evidence from both non-chunk errors and RTs that chunking frees cognitive resources for not-chunked information, and that people take more advantage of this benefit under a higher cognitive load (Ns = 6). While this hypothesis was based on the result of previous work and not directly on model predictions, our analysis of non-chunk states is consistent with the general idea that degree of policy compression should scale with the scarcity of cognitive resources.

9. Discussion

In this paper, we addressed a fundamental puzzle in action selection: why and under what circumstances does action chunking occur? The answer we provide here is that chunking is a natural consequence of limitations on policy complexity. We found that (1) people utilize structured temporal information to reduce their conditional policy complexity and response times, and (2) people chunk more under a higher working memory load. Our empirical findings were consistent with a model that optimizes reward subject to a constraint on conditional policy complexity.

Our results challenge influential models that attribute action chunking primarily to the cost of time. For instance, Dezfouli and Balleine (2012) proposed that action chunks form when the future reward gained from faster responding outweighs potential losses from chunk errors. Similarly, Wu et al. (2023) showed that human chunking behavior adapts to sequence structure and task demands in ways predicted by a rational model that optimizes for both speed and accuracy. However, these and similar models cannot explain two key features of our data: (1) longer average response times at larger set sizes, and (2) greater chunking under higher cognitive load. In these frameworks, time costs are assumed to remain constant across set sizes, and chunk formation is not sensitive to the number of states. Relatedly, Ramkumar et al. (2016) demonstrated that monkeys use motor chunking to reduce computational complexity—the time required to plan physical actions—by planning shorter motor trajectories early in learning. Across these

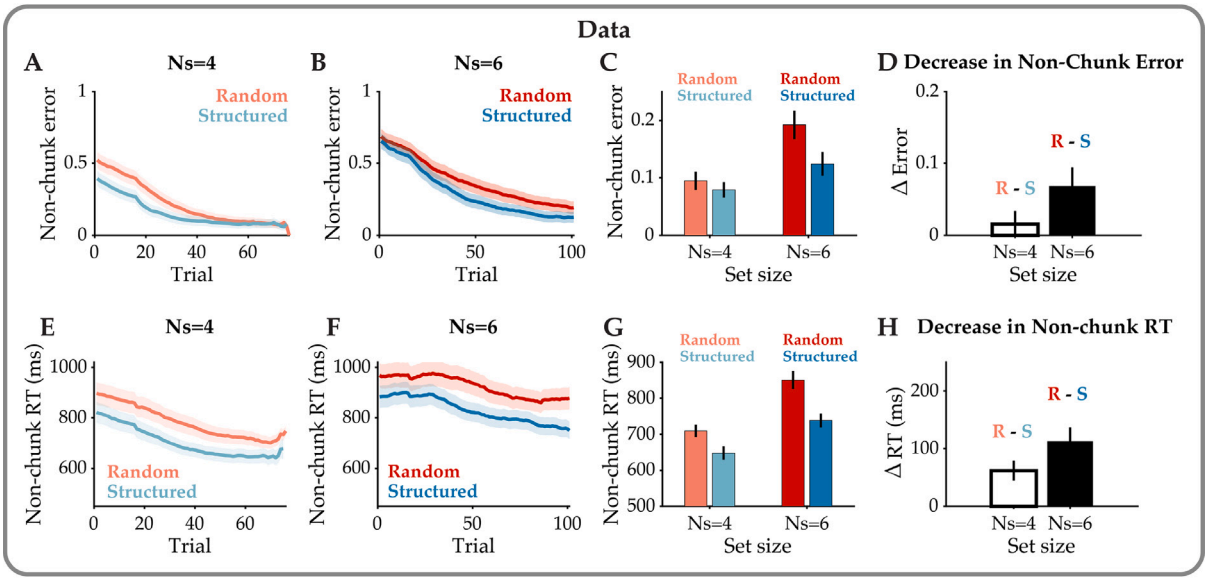

Fig. 9. Chunking frees resources to improve performance in non-chunk states. (A) Participants' non-chunk error as a function of non-chunk trials for the Ns = 4 set size condition. Shaded region indicates standard error. (B) Participants' non-chunk error as a function of non-chunk trials for the Ns = 6 set size condition. (C) Asymptotic non-chunk error (averaged over the last 10 trials) as a function of set size. (D) The decrease in non-chunk error (Random-Structured) as a function of set size. (E) Participants' non-chunk RT as a function of non-chunk trials for the Ns = 4 set size condition. Shaded region indicates standard error. (F) Participants' non-chunk RT as a function of non-chunk trials for the Ns = 6 set size condition. (G) Asymptotic non-chunk RTs (averaged over the last 10 trials) as a function of set size. (H) The decrease in non-chunk RT (Random-Structured) as a function of set size.

approaches, the central focus is on minimizing the *time* costs of action. In contrast, our model emphasizes minimizing the *memory* demands required to store and retrieve action policies, or policy complexity. Importantly, our framework not only addresses memory costs but also predicts response time benefits, which we confirmed empirically. In many real-world settings, both pressures likely interact: less memory-intensive policies are also faster to compute and execute, reflecting the fundamental relationship between response times and policy complexity (Bradmetz & Mathy, 2008; Lai & Gershman, 2021, 2024) (though see Lazartigues et al. (2021) for an alternative perspective).

As noted in the Introduction, the idea that chunking supports working memory efficiency through information compression is well established. Several models have leveraged this principle, either by compressing based on feature similarity (Chekaf et al., 2016; Kowialiewski et al., 2022; Mathy et al., 2024; Pothos, 2007) or by extracting structure from sequential input, as in the PARSER model (Perruchet & Vinter, 1998). These models effectively demonstrate how learners form chunks to reduce representational demands. Although they do not include mechanisms for action selection and therefore cannot be directly compared to our model, we suggest that these types of models could be extended by conceptualizing them as part of a two-stage process: an initial unsupervised chunking phase that forms state chunks based on structure in sequences, followed by a second phase that learns a policy conditioned on these learned state chunks. When a state chunk is encountered in the environment, a corresponding sequence of actions is produced. Such a model might be appropriate in tasks where people are exposed to sequences before engaging in action selection. Even so, it assumes that participants are ignoring actions entirely during chunk learning, which is inconsistent with our task design, where action selection and sequence learning occur simultaneously. The integrated nature of our paradigm requires models that address both processes concurrently.

Our work contributes to a growing body of research that applies information-theoretic principles specifically to action chunking. The options framework (Botvinick et al., 2009; Sutton et al., 1999) describes multi-step policies that group action chunks into higher-level units. Harb et al. (2018) demonstrated that meaningful options can emerge when deliberation costs for switching options are incorporated. Similarly, Jiang et al. (2022) showed that a compression objective can yield skills (similar to action chunks) that extract statistical regularities from offline data. Perhaps most closely related, the DADS framework (Sharma et al., 2019) optimizes conditional mutual information between states and skills given previous states, using a similar objective to ours but applied to a different problem of learning diverse skills. This builds on earlier work such as DIAYN (Eysenbach et al., 2018), which focused on optimizing unconditional mutual information to encourage skill discovery. However, a key distinction is that these studies *maximize* mutual information for skill discovery, while our approach *minimizes* it by grouping actions into structured chunks. Both methods improve task performance but with different goals: DADS focuses on enhancing skill diversity and predictability, while we aim to reduce policy complexity to improve efficiency and reduce cognitive load. Additionally, unlike previous studies that focused exclusively on theoretical frameworks or computational agents, we have provided empirical validation of these information-theoretic principles, linking capacity limits directly to observed human behavior. More generally, our results add to a larger body of work suggesting that chunking serves as a form of memory compression (Bates & Jacobs, 2020; Bates et al., 2019; Brady et al., 2009; Mathy & Feldman, 2012; Nassar et al., 2018; Norris & Kalm, 2021; Sims, 2016; Sims et al., 2012). By replacing highly correlated items with a compact chunk, agents reduce memory demands.

Our findings also align with and extend prior experimental work on hierarchical action composition. Eckstein and Collins (2020) and Xia and Collins (2021) showed that humans can learn and reuse hierarchical options to accelerate learning in new environments. Eckstein and Collins (2021) further demonstrated that humans can learn useful action chunks that predict task-relevant states, resulting in faster intra-chunk RTs and fewer errors over learning. Notably, they also found that disrupting the temporal structure of learned chunks impaired transfer learning. This result is consistent with our model's prediction that sudden changes in temporal structure should degrade performance, as an agent will lose out on reward if she continues to employ previously learned action chunks that do not match the structure of a new environment. However, our work offers an explicit connection between capacity limits and human behavior, as we empirically demonstrate

that action chunking emerges from an information-theoretic model that seeks to maximize reward while minimizing conditional policy complexity.

While few studies directly examine the relationship between working memory and action chunking within individuals, Bo and Seidler (2009) found that people with higher visuospatial working memory capacity formed longer chunks during sequence learning. Additionally, Bo et al. (2011) showed that greater capacity was also associated with more chunking during sequence blocks compared to random blocks. Though these results initially seem to contradict our findings—where greater memory load leads to more chunking—direct comparisons are difficult, as we manipulated memory load within participants rather than correlating it with individual capacity measures.

Our work is also closely related to sequential sampling models of decision making (Forstmann et al., 2016). In these models, prior probabilities can shift the starting point of an evidence accumulator, offering a potential explanation for the benefits of statistical learning (Dayan & Daw, 2008). The starting point reflects prior knowledge that can speed up decision-making (Kelly et al., 2021). Notably, Wu et al. (2023) used an accumulator model to explain response times in a similar chunking task, showing that chunk structure biases the starting point and leads to faster responses. A biased starting point effectively reduces policy complexity by decreasing the stimulus-dependency of choices, providing a complementary perspective on the relationship between policy complexity and response time. However, models like Wu et al. (2023)'s rely on ad hoc assumptions (e.g., biased starting points) to fit RT data, whereas our framework derives these behaviors from first principles and additionally offers a normative explanation for set size effects.

Several open questions remain. Our study focused on how memory limitations drive chunk formation and use but did not address how longer action chunks are formed during learning. Because our chunks were all of length two, future work could explore how capacity constraints affect the formation of more complex action sequences. Our model also assumes that agents condition actions on the previous state to reduce policy complexity. However, this strategy requires maintaining a history of states and storing a default policy—both of which place additional demands on memory resources. Although our experiment only required memory of a single state prior to action selection, we acknowledge that scaling this approach to longer state and action chunks would impose additional memory burdens not captured by our current framework. Despite these additional costs, our data remain consistent with a model in which people maintain a default policy and retain state history to exploit temporal structure. Future work should examine whether individuals also weigh the cognitive costs of maintaining historical dependencies and default policies, and how these costs factor into overall capacity constraints. Furthermore, our model primarily applies to environments with statistical regularities that exist independently of the agent's actions. In controllable environments, where the agent's actions determine the next states, a policy that depends only on the current state would be sufficient. However, even in such model-based scenarios where the policy itself may be relatively simple, maintaining a detailed world model can still be informationally costly. To generalize our theory to such settings, it must account for both policy complexity and the cost of storing the world model.

In sum, we have shown that conditional policy compression offers a compelling account of when and why action chunking occurs. Although our model is not a comprehensive theory of action sequence learning, it highlights how humans consider conditional policy complexity when learning cost-efficient behaviors.

CRediT authorship contribution statement

Lucy Lai: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ann Z.X. Huang:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Samuel J. Gershman:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are thankful to Chris Bates, Jay Hennig, and other members of the Computational Cognitive Neuroscience Laboratory for helpful comments. This research was supported by the Center for Brains, Minds and Machines (funded by NSF STC award CCF-1231216), the Multi-University Research Initiative Grant (ONR/DoD N00014-17-1-2961), and an NSF Graduate Research Fellowship.

Data availability

Data and code available at https://github.com/lucylai96/chunking/.

References

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.
Ballard, I. C., & McClure, S. M. (2019). Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, *317*, 37–44.
Banca, P., Ruiz, M. H., Gonzalez-Zalba, M. F., Biria, M., Marzuki, A. A., Piercy, T., Sule, A., Fineberg, N. A., & Robbins, T. W. (2023). Action-sequence learning, habits and automaticity in obsessive-compulsive disorder. *Elife*, *12*.
Bates, C. J., & 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., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2), 11–11.
Bo, J., Jennett, S., & Seidler, R. D. (2011). Working memory capacity correlates with implicit serial reaction time task performance. *Experimental Brain Research*, *214*(1), 73–81.
Bo, J., & Seidler, R. D. (2009). Visuospatial working memory capacity predicts the organization of acquired explicit motor sequences. *Journal of Neurophysiology*, *101*(6), 3116–3125.
Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262–280.
Bouchacourt, F., Palminteri, S., Koechlin, E., & Ostojic, S. (2020). Temporal chunking as a mechanism for unsupervised learning of task-sets. *Elife*, *9*.
Bradmetz, J., & Mathy, F. (2008). Response times seen as decompression times in Boolean concept use. *Psychological Research*, *72*(2), 211–234.
Brady, T., Konkle, T., & 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.
Chekaf, M., Cowan, N., & Mathy, F. (2016). Chunk formation in immediate memory and how it relates to data compression. *Cognition*, *155*, 96–107.
Chen, Z., & Cowan, N. (2005). Chunk limits and length limits in immediate recall: a reconciliation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*, 1235–1249.
Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*(e62), Article e62.
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., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working memory load strengthens reward prediction errors. *Journal of Neuroscience*, *37*, 4332–4342.
Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*, 1024–1035.
Collins, A. G., & Frank, M. J. (2018). Within-and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, *115*, 2502–2507.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114; discussion 114–85.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453.

Desmurget, M., & Turner, R. S. (2010). Motor sequences and the basal ganglia: kinematics, not habits. *Journal of Neuroscience*, *30*(22), 7685–7690.

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

Du, Y., Krakauer, J. W., & Haith, A. M. (2022). The relationship between habits and motor skills in humans. *Trends in Cognitive Sciences*.

Eckstein, M. K., & Collins, A. G. E. (2020). Computational evidence for hierarchically structured reinforcement learning in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(47), 29381–29389.

Eckstein, M. K., & Collins, A. G. E. (2021). How the mind creates structure: Hierarchical learning of action sequences. *CogSci*, *43*, 618–624.

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. arXiv [cs.AI].

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*, 641–666.

Franco, A., & Destrebecqz, A. (2012). Chunking or not chunking? How do we find words in artificial language learning? *Advances in Cognitive Psychology*, *8*(2), 144–154.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: a recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review (Washington, DC)*, *118*(4), 614–636.

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

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

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: exploring the effects of context. *Cognition*, *112*(1), 21–54.

Gray, R. M. (1972). *Conditional rate-distortion theory*. Information Systems Laboratory, Stanford Electronics Laboratories.

Graybiel, A. M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory*, *70*(1–2), 119–136.

Haith, A. M., & Krakauer, J. W. (2018). The multiple effects of practice: skill, habit and reduced cognitive load. *Current Opinion in Behavioral Sciences*, *20*, 196–201.

Harb, J., Bacon, P.-L., Klissarov, M., & Precup, D. (2018). When waiting is not an option: Learning options with a deliberation cost. *vol. 32*, In *Proceedings of the AAAI Conference on Artificial Intelligence*. (no. 1).

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

Jiang, Y., Liu, E., Eysenbach, B., Kolter, Z., & Finn, C. (2022). Learning options via compression. *Neural Information Processing Systems*, abs/2212.04590.

Jin, X., & Costa, R. M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, *466*(7305), 457–462.

Jin, X., Tecuapetla, F., & Costa, R. M. (2014). Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nature Neuroscience*, *17*(3), 423–430.

Kelly, S. P., Corbett, E. A., & O'Connell, R. G. (2021). Neurocomputational mechanisms of prior-informed perceptual decision-making in humans. *Nature Human Behavior*, *5*(4), 467–481.

Kowialiewski, B., Lemaire, B. t., & Portrat, S. (2022). Between-item similarity frees up working memory resources through compression: A domain-general property. *Journal of Experimental Psychology: General*, *151*(11), 2641–2665.

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In K. D. Federmeier (Ed.), *Psychology of learning and motivation*: *vol. 74*, *The psychology of learning and motivation* (pp. 195–232). Academic Press.

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

Lazartigues, L., Lavigne, F., Aguilar, C., Cowan, N., & Mathy, F. (2021). Benefits and pitfalls of data compression in visual working memory. *Attention, Perception, & Psychophysics*, *83*(7), 2843–2864.

Lengyel, G., Nagy, M., & Fiser, J. (2021). Statistically defined visual chunks engage object-based attention. *Nature Communications*, *12*(1), 272.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356.

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, *122*(3), 346–362.

Mathy, F., Friedman, O., & Gauvrit, N. (2024). Can compression take place in working memory without a central contribution of long-term memory? *Memory & Cognition*, *52*(8), 1726–1736.

Matsuzaka, Y., Picard, N., & Strick, P. L. (2007). Skill representation in the primary motor cortex after long-term practice. *Journal of Neurophysiology*, *97*(2), 1819–1832.

McDougle, S. D., & 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*, 1–20.

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, & Doya, K. (2006). Hierarchical chunking during learning of visuomotor sequences. In *The 2006 IEEE international joint conference on neural network proceedings* (pp. 249–253).

Nassar, M., Helmers, J., & Frank, M. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review*, *125*, 486–511.

Niu, X., Gündüz, D., Bai, B., & Han, W. (2023). Conditional rate-distortion-perception trade-off. In *2023 IEEE international symposium on information theory* (pp. 1068–1073). IEEE.

Norris, D., & Kalm, K. (2021). Chunking and data compression in verbal short-term memory. *Cognition*, *208*, Article 104534.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(7), 2745–2750.

Orhan, A., & Jacobs, R. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review*, *120*, 297–328.

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

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263.

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin*, *133*(2), 227–244.

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

Ramkumar, P., Acuna, D. E., Berniker, M., Grafton, S. T., Turner, R. S., & Kording, K. P. (2016). Chunking as the result of an efficiency computation trade-off. *Nature Communications*, *7*, 12176.

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: a MDL-based cognitive model of inductive learning. *Cognitive Science*, *35*(7), 1352–1389.

Sakai, K., Kitaguchi, K., & Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Experimental Brain Research*, *152*(2), 229–242.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *16*(4), 592–608.

Sharma, A., Gu, S., Levine, S., Kumar, V., & Hausman, K. (2019). Dynamics-aware unsupervised discovery of skills. In *International conference on learning representations*.

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

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807–830.

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

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*, 181–211.

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

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(1), 37–55.

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

Tosatto, L., Fagot, J., Nemeth, D., & Rey, A. (2022). The evolution of chunks in sequence learning. *Cognitive Science*, *46*(4), Article e13124.

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.

Wu, S., Éltető, N., Dasgupta, I., & Schulz, E. (2023). Chunking as a rational solution to the speed-accuracy trade-off in a serial reaction time task. *Scientific Reports*, *13*(1), 7680.

Xia, L., & Collins, A. G. E. (2021). Temporal and state abstractions for efficient learning, transfer, and composition in humans. *Psychological Review*, *128*(4), 643–666.