How the Mind Creates Structure: Hierarchical Learning of Action Sequences

Maria K. Eckstein (maria.eckstein@berkeley.edu) & Anne G.E. Collins (annecollins@berkeley.edu)
Department of Psychology, 2121 Berkeley Way West
Berkeley, California 94720, USA

Abstract

Humans have the astonishing capacity to quickly adapt to varying environmental demands and reach complex goals in the absence of extrinsic rewards. Part of what underlies this capacity is the ability to flexibly reuse and recombine previous experiences, and to plan future courses of action in a psychological space that is shaped by these experiences. Decades of research have suggested that humans use hierarchical representations for efficient planning and flexibility, but the origin of these representations has remained elusive. This study investigates how 73 participants learned hierarchical representations through experience, in a task in which they had to perform complex action sequences to obtain rewards. Complex action sequences were composed of simpler action sequences, which were not rewarded, but whose execution led to changes in the environment. After participants learned action sequences, they completed a transfer phase. Unbeknownst to them, we manipulated either complex or simple sequences by exchanging individual elements, requiring them to relearn. Relearning progressed slower when simple (rather than complex) sequences were changed, in accordance with a hierarchical representations in which lower levels are quickly consolidated, potentially stabilizing exploration, while higher levels remain malleable, with benefits for flexible recombination.

Keywords: Hierarchical cognition, reinforcement learning, action sequence learning, hierarchical reinforcement learning

Introduction

Solving complex problems requires abstraction. Even the most common every-day problems (e.g., crossing the road) are so high-dimensional that planning into the future rapidly results in a *combinatorial explosion* of possibilities (e.g. sequences of possible muscle movements; number of possible future states of the road). One way to alleviate this issue is to represent problems at a more abstract level in order to reduce the dimensionality of the problem (e.g., percepts clustered into objects such as green/red lights; muscle movements joined into meaningful actions such as crossing/waiting). Hierarchical representations of perception and/or action can provide such an abstraction. In real-world problems, agents also need to flexibly adjust their behavior to changing environmental circumstances, in a world that does not provide clear feedback as to which individual actions (e.g., typing a specific word) were adaptive, and often only rewards complex chains of actions upon completion (e.g., a completed essay). This problem of *sparse rewards* can also be alleviated by abstraction and hierarchical representations: Instead of relying solely on rewards at the end of extended action sequences, agents can set smart sub-goals along the way, and continuously adjust their behavior based on their performance on each one.

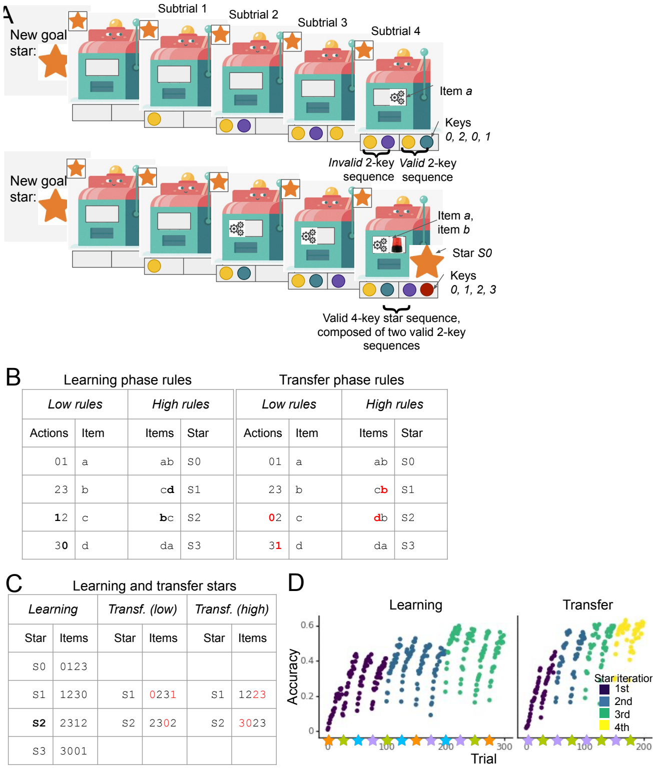
Previous research on abstraction. Indeed, previous research across disciplines has shown that hierarchy plays a central role for complex problem solving: Artificial Intelligence (AI) research has developed algorithms that abstract over time (Sutton et al., 1999), states (Finn et al., 2017; Vezhnevets et al., 2017), and learning itself (Wang et al., 2016) to solve increasingly difficult problems. In Psychology, decades of research have suggested that mental representations are hierarchical, most notably in the domains of cognitive control, expertise, and sequential action (Cohen, 2000; Newell, 1994). Recent research in psychology has increasingly tried to formalize this notion, using hierarchical Bayesian (e.g., Griffiths et al., 2019; Kemp and Tenenbaum, 2008; Solway et al., 2014) and hierarchical Reinforcement Learning (RL) models (e.g., Botvinick and Weinstein, 2014; Eckstein and Collins, 2020; Frank and Badre, 2012) to understand abstract cognitive processes. Lastly, neuroscience research has shown that the brain itself is organized hierarchically (Miller and Cohen, 2001), exhibiting “processing hierarchies” and “representational hierarchies” (Badre, 2008). In the former, superordinate levels (e.g. FPC, dlPFC) operate over longer timescales (e.g. general domain multi-step information) and asymmetrically modulate subordinate processing (e.g. striatal areas). In the latter, information gets increasingly abstract when moving from lower to higher level of hierarchy, such that higher levels favor generality over detail, and lower levels asymmetrically inherit information from higher ones.

The difficulty of learning abstraction. Even though it is broadly accepted that appropriate hierarchical representations are necessary to solve complex problems, it is still largely unknown how to *create* these representations; in AI, this issue is called the “option discovery problem” (“options” are targeted multi-step policies; Sutton et al., 1999). For example, humans have been shown to discover the Bayes-optimal task decomposition of a complex problem (Solway et al., 2014), but it is unclear how, given that they lack access to the full state space and have limited computational resources. AI research has investigated some promising avenues, for example equipping agents with intrinsic motivation so they can break down complex problems into simpler sub-problems, and receive teaching signals along the way. Intrinsic motivation is usually formulated as adding an artificial “intrinsic” reward signal to the “extrinsic” rewards provided by the environment (e.g., food; points in video games; Deci and Ryan, 1985). Intrinsic rewards then guide learning in the same way as extrinsic rewards. Intrinsic rewards often mimic novelty seeking and curiosity (Gershman and Niv, 2015; Lieshout et al., 2018; Pathak et al., 2017), in line with psychological theory that defines intrinsic motivation as doing actions for their own sake, rather than to achieve external goals (Deci and Ryan, 1985). One aspect that many approaches share when creating hierarchical representations is the identification of appropriate sub-goals, or other indices of existing structure in the environment (for review, see Konidaris, 2019).

Scrutinizing human abstraction. This study investigates how humans create hierarchical representations by characterizing how they slowly discover the hierarchical structure of a task. We hypothesized that participants would create hierarchical representations piece-by-piece by continuously learning how to perform new, ever more complex multi-step actions. Specifically, we hypothesized that participants would start by *exploring* their environment, randomly executing basic actions (e.g., individual key presses). Because some combinations of basic actions would provoke unexpected (non-rewarding) events in the environment (e.g., appearance of a novel item), this process would trigger participants' *curiosity*, and intrinsically motivate them to explore the event further. Participants' curiosity should only be satisfied once they know how to evoke the event reliably (i.e., executing the correct sequence of basic actions to create the item). At this point, the entire action sequence should have been consolidated as a new skill, laying the foundation for hierarchical structure. From then on, we hypothesized, participants would be able to employ these *learned skills* instead of basic actions to explore their environment in a more targeted way, speeding up the acquisition of even more abstract skills, more targeted exploration, and so on. This hypothesized creation of hierarchical structure allows participants to overcome reward sparsity because they can use environmental signals other than rewards to motivate learning. It also allows them to overcome combinatorial explosion because relying on a specific set of fixed multi-step actions is less planning-intensive than assessing all possible combinations of one-step actions.

Methods

Experimental task. To test these predictions, we created a task in which participants learned to execute complex action sequences, which were composed of simpler action sequences (which were composed of basic actions; Fig. 1). Participants were extrinsically motivated by points, which could be obtained by creating stars using a star-making machine. The machine accepted 4 key presses per trial, and created a star when a correct 4-key sequence was typed in. Participants were rewarded with 1 point only when the star matched the trial's goal star. Crucially, each star's 4-key sequence was composed of two 2-key sequences, each of which led the machine to create a specific item. Items were not rewarded with points, but signaled (unbeknownst to the participants) that a 2-key sequence was "valid" (part of a star-making 4-key sequence; see Fig. 1B). As such, they were a potential source of intrinsic motivation. Participants encountered 4 different stars, which required four different 4-key sequences; the goal star changed in each block. This task has a hierarchical structure: basic actions (individual keys) are at the lowest, valid



B

Learning phase rules				Transfer phase rules			
Low rules		*High rules*		*Low rules*		*High rules*	
Actions	Item	Items	Star	Actions	Item	Items	Star
01	a	ab	S0	01	a	ab	S0
23	b	c**d**	S1	23	b	c**b**	S1
12	c	**b**c	S2	**0**2	c	**d**b	S2
3**0**	d	da	S3	3**1**	d	da	S3

C

Learning and transfer stars

Learning		*Transf. (low)*		*Transf. (high)*	
Star	Items	Star	Items	Star	Items
S0	0123				
S1	1230	S1	0231	S1	1223
S2	2312	S2	2302	S2	3023
S3	3001				

Figure 1: Design of the star-making task. (A) At the beginning of each block, the new goal star is introduced, and remains visible on top of the screen throughout the block. On each trial, participants sequentially enter four key presses. Every time participants press a key, a colored circle appears at the corresponding position on the response board. If participant press a valid 2-key sequence (see rules in part C), a unique item appears. In the top row, keys 0 (subtrial 3; orange) and 1 (subtrial 4; teal) led to item *a* (gear). In the second row, the combination of 2-key sequences *a* (keys 0, 1) and *b* (keys 2, 3) led to the appearance of star *S*0. (B) Rules for valid sequences. Table "Learning phase rules" shows key sequences for the learning phase. The column "Low rules" shows valid 2-key sequences, with "Actions" referring to the identity and order of actions that need to be executed, and "Item" referring to the resulting item. Keyboard keys were randomly assigned to actions, and images were randomly assigned to items. The column "High rules" shows how valid 4-key sequences were composed of 2-key sequences ("Items") and to which star they lead. Table "Transfer phase rules": In the transfer phase, either low-level rules or high-level rules were manipulated. For low-level (high-level) manipulation, the low-level (high-level) rules in the Learning table were replaced by the low-level (high-level) rules in this table. Differences between learning and transfer rules are highlighted in red. (C) Overview of stars presented during learning and transfer phases. All four stars were learned, but only two were selected for the transfer phase. (D) Learning curves for learning (left) and transfer phase (right). Each dot represents the average accuracy over all participants on each trial. Trials were counted as accurate when the goal star (shown on the x-axis) was achieved.

2-key sequences at the intermediate, and star-making 4-key sequences at the most abstract level. We designed the task in this way to elicit learning of hierarchical structure that was motivated intrinsically, i.e., by observing non-rewarding events in the environment for the building blocks of complex sequences.

After learning, participants encountered an unsignaled transfer phase, in which some rules changed (Fig. 1B, 1C). The transfer phase investigated whether participants represented the task hierarchically. In the low-level transfer manipulation, we modified some 2-key sequences by replacing individual keys, such that previous valid 2-key sequences no longer created items, and were no longer part of 4-key star-making sequences. In the high-level manipulation, some 4-key sequences were modified by exchanging entire 2-key sequences with each other (Fig. 1B, right), such that all initial 2-key sequences stayed valid, but had to be combined differently to make stars. We controlled for the numbers of individual keys that were affected by each manipulation (Fig. 1C), and predicted that performance would differ between low- and high-level transfer if participants used a hierarchical representation: We argued that valid 2-key sequences were more consolidated than 4-key star-making sequences and should be more difficult to re-learn. 4-key sequences, on the other hand, should be more flexible and malleable and less affected by transfer.

Experimental details. All participants provided online informed consent in accordance with the Institutional Review Board of the University of California, Berkeley, and completed a demographics form. Participants then performed the task, which consisted of a tutorial, a learning and transfer phase with one machine using one hand, and another learning and transfer phase with a different machine using the other hand. After the task, participants completed a questionnaire about their task strategies.

On each trial, participants pressed four keys with the goal of finding the current trial's goal star, shown at the top of the screen, to receive a point (Fig. 1A). A point counter showed cumulative points. Each key press within a trial is called a "subtrial". Four keys were available on each trial, depending on the machine: Q, W, E, and R (left hand); or U, I, O, and P (right hand). Participants were allowed a maximum of 2.5 seconds for each trial; when the four key presses took longer, participants were told to respond faster next time and the trial was counted as missed. Each trial was followed by a 0.5-second inter-trial interval, after which the next trial started. Each key press was immediately visualized as a colored circle in a response box underneath the star machine, with a one-to-one match between key and color. When participants executed a valid 2-key sequence within the first (last) two slots, an item immediately appeared on the left (right) side of the machine's window. Each of the four valid 2-key sequences was represented by a unique item. When participants executed a valid 4-key sequence, a star immediately appeared. When the star coincided with the goal star, the point counter incremented by 1 point. When a trial did not form a valid 4-key sequence, no star appeared. Incorrect trials were not signaled otherwise.

Valid 2-key and 4-key sequences were constructed to maximize similarity between high-level and low-level transfer for experimental control. The same abstract rules were used for all participants (Fig. 1B-C). Systematic biases were avoided by randomizing the assignment of actions to keys, 2-key sequences to items, and 4-key sequences to stars.

For each machine, participants completed 12 blocks of 25 trials (with 4 key presses each) during the training phase, and 8 blocks of 25 trials during the transfer phase. Each block showed one goal star. Two of the four learning-phase stars were selected for the transfer phase (Fig. 1C, 1D). The transition between learning and transfer phase was not signaled.

After completing the first machine (learning and transfer), participants took a 1-minute break. After the break, they were presented with a new machine, and were instructed to use the opposite hand on a different set of keys to minimize carry-over between the machines. Hand order and machine order (low vs high transfer) were jointly randomized between participants. The new machine followed the same abstract rules as the old machine, but keys were randomly re-assigned to the new set of keys. A novel set of items indicated valid 2-key sequences, and a novel set of stars indicated star-making 4-key sequences. The task was written in jsPsych, a JavaScript library that facilitates online data collection.

Participants. Seventy-three undergraduate students completed the task online for course credit (58 females, 13 males, 2 declined to answer). Four were excluded because they did not meet demographic criteria (e.g., present or past psychological illness). Six were excluded because they missed more than 50 trials (mean missed trials after exclusion: 11.6, sd: 9.2, min: 1, max: 35). Two were excluded because they took more than 60 minutes for the task (mean duration after exclusion: 36 minutes, min: 26, max: 46, sd: 5.3). Eleven were excluded because they used pen and paper or other external devices to help with the task, which potentially obscured the cognitive processes we aimed to investigate. (Because the study was conducted online due to the Covid pandemic, we could not monitor the use of pen and paper directly, and asked participants in the post-experiment questionnaire.) In total, 17 participants were excluded, leading to a final sample of 56 participants (45 females, 10 males, 1 declined to answer; mean age: 20.6, min: 18.1, max: 31.8, sd: 1.96).

Data analysis. We used Python for data analysis and visualization. Regression models were conducted using the statsmodels package, which uses a Normal distribution to approximate p-values (and therefore does not implement the Satterthwaite correction; Seabold and Perktold, 2010). Unless otherwise specified, we used mixed-effects models and defined each participant as a group.

Results

Creating Hierarchy by Learning Action Sequences

We first investigated how participants learned new action sequences, focusing on just the learning phase.

Slowing after unexpected event. We had hypothesized that unexpected items would trigger participants' curiosity and motivate them to repeat the 2-key sequence which led to the item, thus facilitating learning. To assess this, we tested whether participants slowed down after discovering a new item for the first time. Slowing commonly arises after errors (Danielmeier and Ullsperger, 2011), rewards (Raio et al., 2020), or surprising events (Parmentier et al., 2019), and is usually interpreted as an orienting response, potentially related to learning and processing of prediction errors. We assessed response times for the third key press in a trial (subtrial 3) when an item was discovered for the first time in a block on subtrial 2, comparing trials in which an item was discovered to the trials preceding and subsequent to the discovery (Fig. 2A, red line). Repeated-measures t-tests, Bonferroni-corrected for multiple comparisons, revealed that participants were significantly slower on the trial of item discovery compared to both preceding ($t(54) = 4.1$, $p = 0.0003$) and subsequent trials ($t(54) = 6.9$, $p < 0.001$). This slowing was a specific post-item effect rather than general slowing, as it uniquely occurred on subtrial 3 when an item appeared on subtrial 2, but not on subtrial 4 (Fig. 2A, blue line).

Repetition of valid sequences. If the appearance of non-rewarding items indeed motivated participants to execute valid 2-key sequences, their frequency should increase over time. To tests this, we compared the frequency of all four valid 2-key sequences, aligned to their first discovery in a block, to four randomly selected invalid, but structurally-similar 2-key sequences (Fig. 2B). We then calculated the difference between the proportion of valid versus invalid sequences for each participant and each trial (Fig. 2B, inset), and used mixed-effects regression to predict this difference from the trial since sequence discovery. This analysis revealed a significant difference from zero (Intercept $\beta = 0.13$, $z = 14.7$, $p < 0.001$) with a negative effect of trial ($\beta = -0.01$, $z = -6.6$, $p < 0.001$), confirming that participants repeated valid 2-key sequences more often than invalid ones, with a negative effect of time since first sequence execution. This analysis was restricted to trials in which participants did not reach the goal star (incorrect trials) because correct trials naturally have a higher proportion of valid compared to invalid 2-key sequences (because all valid 4-key sequences are composed of valid 2-key sequences), and would therefore bias the result. In sum, participants repeated valid 2-key sequences more than invalid ones, suggesting that the appearance of items triggered intrinsic motivation.

Increased use of valid 2-key sequences. We next assessed whether the overall proportion of valid compared to invalid

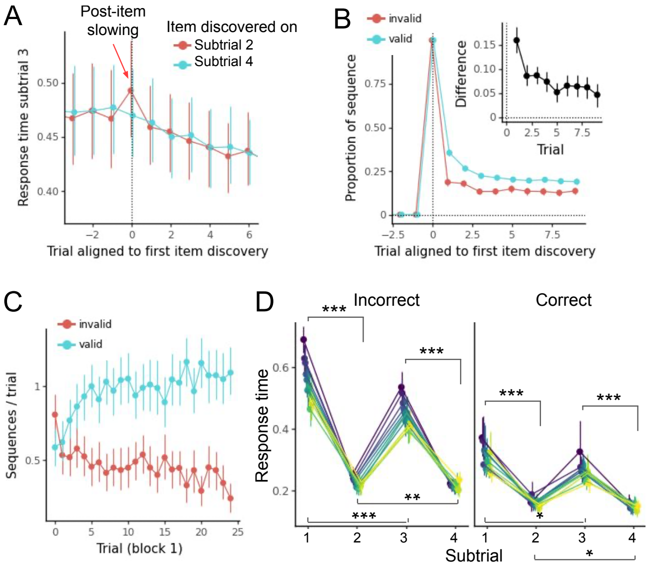


Figure 2: (A) Post-item slowing. Response times were elevated on subtrial 3 when a 2-key sequence was discovered before (on subtrials 1 and 2; red), but not after (subtrials 3 and 4; blue), revealing specific post-item slowing. Dots represent means; error bars between-participant 95% confidence intervals. (B) Repetition of valid and invalid sequences after first discovery. Trial 0 shows the first execution of a 2-key sequence in a block. Subsequent trials show the proportion of trials on which the same sequence was executed, separately for valid (blue) and invalid (red) 2-key sequences. The inset shows within-participant difference between valid and invalid sequences. (C) Number of 2-key sequences executed per trial (block 1 only). The blue line shows the average of the four valid 2-key sequences (signaled by item appearance), and the red line shows the average of four matched invalid 2-key sequences (not signaled by items). The maximum number of 2-key sequences per trial is two because each trial allows for four key presses. The red and blue lines do not add up to two because only matched invalid sequences were included in the analysis. (D) Response time for each key press within a trial. Colors indicate block number (dark to light). Stars show results of repeated-measures t-tests described in the main text (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$).

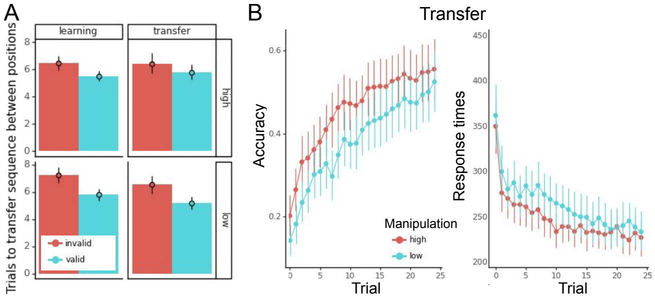


Figure 3: (A) Number of trials to use a 2-key sequence that was first discovered in one position in the opposite position of a trial. (B) Performance in the transfer phase. Accuracy (left) and response times (right) over trials, averaged over blocks, for both high (red) and low (blue) transfer phases.

sequences increased over the course of the first learning block, in accordance with participants' hypothesized expansion of the action repertoire (Fig. 2C). We used mixed-effects regression to predict the number (0, 1, or 2) of valid and invalid 2-key sequences on each trial from sequence validity (valid vs invalid) and trial number (1-25), as well as their interactions. The significant interaction between sequence validity and trial ($\beta = 0.036$, $z = 7.1$, $p < 0.001$) confirmed that the trajectories of valid and invalid sequences differed. Follow-up models revealed a positive slope for valid sequences ($\beta = 0.011$, $p = 0.04$), indicating increase in use, and a negative slope for invalid sequences ($\beta = -0.025$, $p < 0.001$), indicating decreased in use. For the same reason as above, the analysis was limited to incorrect trials only. These results suggest that participants expanded their action repertoire by adding temporally-extended actions to the initial set of individual keys (basic actions). Rather than continuing to explore their environment using just basic actions, participants seemed to shift their exploration strategy toward using 2-key sequences.

Patterned response times. We next assessed the temporal structure of participants' key presses, hypothesizing that if participants treated valid 2-key sequences like stand-alone actions, the two keys of the sequence would be executed in quick succession, compared to slower execution at sequence boundaries. Indeed, participants typed faster at sequence completion than initiation, with faster response times on subtrial 2 compared to 1, and subtrial 4 compared to 3, for both correct and incorrect trials, as revealed by repeated-measures t-tests using 8-way Bonferroni correction (all $t(55)s > 2.9$, all $ps < 0.04$; Fig. 2D). Interestingly, participants also responded faster on subtrial 3 compared to 1, and 4 compared to 2, suggesting that participants might have frontloaded processing, such that the second 2-key sequence was already prepared before or during the first sequence. This pronounced slow-fast-slow-fast response pattern suggests that participants executed two distinct 2-key actions rather than four individual actions, supporting our hypothesis that participants chunked pairs of key presses into a single unit, consolidating 2-key sequences into distinctive, temporally-extended actions.

Using Hierarchy for Exploration and Planning

We next investigated whether and how participants used their learned hierarchical action space for exploration and planning, analyzing the transfer phase.

Using 2-key sequences for exploration. We first tested whether participants actively moved 2-key sequences between "slots" (first slot: subtrials 1 and 2; second slot: subtrials 3 and 4). This would indicate that they did not just learn 2-key sequences as distinct, stand-alone actions, but also actively explored how to reach stars using them. Specifically, we assessed the number of trials that passed between the first discovery of a new 2-key sequence and its first use in the opposite slot, and compared it between valid and invalid sequences in each block (Fig. 3A). On average, participants took 5.6 trials after discovery to transfer valid sequences but 6.2 for invalid ones. Mixed-effects regression on the difference revealed a significant intercept ($\beta = -0.87$, $se = 0.33$, $z = -2.63$, $p = 0.008$), with no effect of block ($\beta = -0.004$, $p = 0.94$), confirming that participants transferred valid 2-key sequences faster than expected based on baseline (invalid sequences). In sum, participants quickly transferred valid action sequences from the position in which they were originally discovered to the opposite position, suggesting flexible reuse and exploration.

Differences between high and low transfer. Finally, we assessed the transfer phase of the experiment, comparing the impact of modifying 2-key sequences (low-level manipulation) versus 4-key sequences (high-level manipulation; Fig 1C). We predicted that the low-level manipulation would impair performance more because 2-key sequences were more consolidated and less accessible to change once they became part of participants' abstract action space. We hypothesized that the high-level manipulation would affect performance less because the combination of 2-key sequences was less consolidated than the 2-key sequences themselves, such that new associations can be re-learned more easily. We tested this prediction by probing differences between accuracy during high-level and low-level manipulation (Fig. 3B), using mixed-effects regression to predict accuracy from transfer type (high versus low), trial number (1-25), and their interaction. The model showed main effects of transfer type ($\beta = 0.10$, $z = 6.4$, $p < 0.001$) and trial ($\beta = 0.1$, $z = 15.7$, $p = 0.10$) and no interaction ($\beta = 0.002$, $z = 1.6$, $p = 0.11$), revealing that performance was indeed more impaired during lower-level manipulation, and that this effect did not diminish over time. In sum, performance suffered more when the lower level was manipulated compared to the higher one, supporting the role of 2-key sequences as building blocks for planning and complex action.

Discussion

This study provides an analysis of how humans create hierarchical action spaces, shedding light both on how hierarchy is created, and how it is subsequently used: Our results suggest that participants create hierarchical action spaces by continuously learning new action sequences, as evidenced by the slowing after events that were unexpectedly triggered by their own actions (Gershman and Niv, 2015; Parmentier et al., 2019), the subsequent repetition of these actions to reproduce the event, and the increased use of the learned sequences for future exploration. The use of the learned sequences was characterized by its active, hypothesis-driven nature: Participants moved sequences between trial slots, and showed more difficulty re-learning lower-level than higher-level sequences, suggesting increased consolidation of lower levels (Goodman et al., 2011).

Limitations of the current task design. One limitation of the study is the inherent difficulty to directly compare low-level and high-level manipulations. By definition, action sequences are affected differently, and we experienced that controlling one aspect of experimental design (e.g., positions of manipulated keys) often led to disturbances in others (e.g., number of manipulated keys). We aimed to address this issue by choosing imbalances that worked against our hypothesis (e.g., minimizing change in low-level manipulation), thereby ascertaining that an existing effect (e.g., better performance in high-level manipulation) was due to hierarchical structure, rather than design imbalances.

Ambiguities around intrinsic motivation. The definition of "intrinsic motivation" differs between fields (education, psychology, AI), as well as between specific studies within each field. The common denominator is that intrinsic motivation cannot rely on extrinsic rewards, but what information it can employ is largely undefined. We reasoned that, despite being intrinsic to agents, intrinsic motivation is likely still tied to external events (e.g., inherent enjoyment of watching a movie, reading a book, listening to music, going on a hike), even though exceptions might exist (e.g., pure joy about a novel thought that occurred unrelated to current surroundings). Our experiment aimed to operationalize this intuition through action-triggered, but non-rewarding events. However, some might argue that these events are too closely related to the reward structure of the task to count as purely "intrinsic" motivators. Future research might help settle this question. For example, one could conduct a task variant in which the appearance of items is not tied to the reward structure (e.g., 2-key sequences that create items are never part of star-producing 4-key sequences, and vice versa). If participants still show signs of intrinsic motivation to perform item-producing 2-key sequences in this variant (e.g., slowing down, increase in frequency, shuffling between slots), this would prove that the appearance of items provides motivation that is independent of rewards, and therefore intrinsic.

Differences between intrinsic and extrinsic motivation. An interesting question for future research is whether intrinsic and extrinsic motivators affect learning differently. In AI, intrinsic and extrinsic rewards are often combined additively, i.e., treated as the same entity. However, it is also possible that each affects learning in a different way. To test this, we are planning to use a task version that replaces all items with explicit rewards (e.g., points). One possible outcome of this study is that participants are less motivated to learn how to create stars if they can also obtain rewards by performing much simpler 2-key sequences, depending on the relative worth of 2-key sequences compared to stars. If this is the case, it might be a fundamental role of intrinsic motivation to continuously adjust the "rewardingness" of intermediate actions in order to continuously enable learning of increasingly complex actions. Indeed, intrinsic motivation intuitively decreases with increased familiarity and skill, whereas extrinsic rewards stay at the same levels infinitely (at least in theory; Singh et al., 2009), making the former a more promising candidate for "life-long" learning.

Another important control will be to remove the intermediary items altogether, to shed light on the importance of intermediary feedback for learning complex actions. A likely outcome of this study is that participants will have increased difficulties to learn 4-key sequences without the scaffolding provided by intermediary feedback for 2-key sequences. This would support our hypothesis that non-rewarding environmental information can be crucial for the acquisition of complex action sequences, even when it is not itself rewarding.

Computational modeling. We have previously presented an algorithm that formalizes our hypotheses about the creation of hierarchical action spaces and makes predictions about task variants with more (or fewer) levels and more (or fewer) basic actions (Eckstein and Collins, 2017). Our next step will be to apply this algorithm to humans. The curiosity-driven hierarchical reinforcement learning ("CHaRLy") algorithm is based on the options framework (Sutton et al., 1999): Whenever CHaRLy experiences an unexpected event (e.g., new item appears), the creation of a new "option" is triggered (in this case, re-creating the unexpected event). Initiated uniformly, options are learned through RL value updating based on experienced outcomes (Sutton and Barto, 2017). In this way, each option eventually is a reliable policy to trigger one particular event (item). Options make up the lower level of hierarchy in CHaRLy. The high level is needed to choose between all options and basic actions. This choice is guided by another set of RL values, which operate at the high level. The values of basic actions are updated based on whether an extrinsic reward was obtained (e.g., goal star appears). The values of options are updated based on both the achieved extrinsic and an additional intrinsic reward that reflects surprise. CHaRLy uses the successor representation (Dayan, 1993) to accurately represent the temporal structure of the task and learn values that are appropriate for sequential actions.

The CHaRLy model contains the main features of hierarchy present in previous models of human hierarchical thought (e.g., Collins and Koechlin, 2012; Eckstein and Collins, 2020; Xia and Collins, 2021), including the two-level structure and existence of distinct low-level policies. However, it is unique in that it also combines more complex features, including complex, temporally extended actions (Xia and Collins, 2021) and curiosity (Singh et al., 2005), making it a more comprehensive hierarchical RL model of human behavior. It is a good candidate to capture human behavior in this task because it makes the right qualitative predictions, and quantitatively describes our hypotheses.

Conclusion. Humans' astonishing ability to learn from sparse rewards in highly-flexible and ever-changing environments might therefore rely on an ability to create increasingly complex building blocks of behavior, based on curiosity about non-rewarding environmental events. These building blocks—shaped by the environment—can eventually facilitate the efficient discovery of reward.

Acknowledgements

Zisu Dong was involved in an earlier version, Aram Moghadassi helped with task design, and Amy Zou implemented the task and collected data. This work was supported by NIMH grant RO1MH119383 to AGEC.

References

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200. https://doi.org/10.1016/j.tics.2008.02.004

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Phil. Trans. R. Soc. B*, *369*(1655), 20130480. https://doi.org/10.1098/rstb.2013.0480

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality? *European Journal of Cognitive Psychology*, *12*(1), 1–36. https://doi.org/10.1080/095414400382181

Collins, A. G. E., & Koechlin, E. (2012). Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLOS Biology*, *10*(3), e1001293. https://doi.org/10.1371/journal.pbio.1001293

Danielmeier, C., & Ullsperger, M. (2011). Post-Error Adjustments. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00233

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624.

Deci, E., & Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer US. https://doi.org/10.1007/978-1-4899-2271-7

Eckstein, M. K., & Collins, A. G. E. (2017). CHRL: Combining intrinsic motivation and hierarchical reinforcement learning. *Advances in Neural Information Processing Systems, workshop*.

Eckstein, M. K., & Collins, A. G. E. (2020). Computational evidence for hierarchically structured reinforcement learning in humans. *Proceedings of the National Academy of Sciences*, *117*(47), 29381–29389. https://doi.org/10.1073/pnas.1912330117

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135.

Frank, M. J., & Badre, D. (2012). Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 1: Computational Analysis. *Cerebral Cortex*, *22*(3), 509–526. https://doi.org/10.1093/cercor/bhr114

Gershman, S. J., & Niv, Y. (2015). Novelty and Inductive Generalization in Human Reinforcement Learning. *Topics in Cognitive Science*, *7*(3), 391–415. https://doi.org/10.1111/tops.12138

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–119. https://doi.org/10.1037/a0021336

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30. https://doi.org/10.1016/j.cobeha.2019.01.005

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Konidaris, G. (2019). On the necessity of abstraction. *Current Opinion in Behavioral Sciences*, *29*, 1–7. https://doi.org/10.1016/j.cobeha.2018.11.005

Lieshout, L. L. F. v., Vandenbroucke, A. R. E., Müller, N. C. J., Cools, R., & Lange, F. P. d. (2018). Induction and relief of curiosity elicit parietal and frontal activity. *Journal of Neuroscience*, 2816–17. https://doi.org/10.1523/JNEUROSCI.2816-17.2018

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Newell, A. (1994). *Unified Theories of Cognition*. Harvard University Press.

Parmentier, F. B. R., Vasilev, M. R., & Andrés, P. (2019). Surprise as an explanation to auditory novelty distraction and post-error slowing. *Journal of Experimental Psychology: General*, *148*(1), 192–200. https://doi.org/10.1037/xge0000497

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *arXiv preprint arXiv:1705.05363*.

Raio, C. M., Konova, A. B., & Otto, A. R. (2020). Trait impulsivity and acute stress interact to influence choice and decision speed during multi-stage decision-making. *Scientific Reports*, *10*(1), 7754. https://doi.org/10.1038/s41598-020-64540-0

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python, 5.

Singh, S., Barto, A. G., & Chentanez, N. (2005). *Intrinsically Motivated Reinforcement Learning:* (tech. rep.). Defense Technical Information Center. Fort Belvoir, VA. https://doi.org/10.21236/ADA440280

Singh, S., Lewis, R. L., & Barto, A. G. (2009). Where Do Rewards Come From?, 6.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. (2014). Optimal behavioral hierarchy. *PLoS computational biology*, *10*(8), e1003779.

Sutton, R. S., & Barto, A. G. (2017). *Reinforcement Learning: An Introduction* (2nd ed.). 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*(1), 181–211. https://doi.org/10.1016/S0004-3702(99)00052-1

Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., & Kavukcuoglu, K. (2017). FeUdal Networks for Hierarchical Reinforcement Learning. *arXiv:1703.01161*.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

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