An algorithmic account for how humans efficiently learn, transfer, and compose hierarchically structured decision policies

Jing-Jing Li (jl3676@berkeley.edu)
Helen Wills Neuroscience Institute, UC Berkeley

Anne G. E. Collins (annecollins@berkeley.edu)
Department of Psychology, Helen Wills Neuroscience Institute, UC Berkeley

Abstract

Learning structures that effectively abstract decision policies is key to the flexibility of human intelligence. Previous work has shown that humans use hierarchically structured policies to efficiently navigate complex and dynamic environments. However, the computational processes that support the learning and construction of such policies remain insufficiently understood. To address this question, we tested 1,052 human participants on a decision-making task where they could learn, transfer, and recompose multiple sets of hierarchical policies. We propose a novel algorithmic account for the learning processes underlying observed human behavior. We show that humans use meta-learning and Bayesian inference to expand compressed policies into hierarchical representations over learning. Furthermore, our modeling suggests that these hierarchical policies are structured in a temporally backward-looking or retrospective fashion.

Keywords: computational cognitive modeling; abstraction; hierarchy; meta-learning; decision-making; transfer; composition

Introduction

The ability of humans to learn, abstract, transfer, and compose complex decision policies between structurally related contexts is crucial to efficient and flexible generalization – a hallmark of human intelligence. Previous work has shown that humans can abstract states and actions hierarchically to effectively navigate complex and dynamic environments (Botvinick, Niv, & Barto, 2009; Xia & Collins, 2021), though existing frameworks fail to provide an account for how such hierarchical structures are learned, constructed, and organized at the algorithmic level. Here, we propose two algorithmic architectures, supported by data, that can capture human behavior on a decision-making task where participants can learn and transfer multiple sets of hierarchical policies.

Methods and results

1,052 undergraduate students completed the online behavioral experiment illustrated in Figure 1, which extends the paradigm used by Li, Xia, Dong, and Collins (2022). Participants who learned to perform better than chance in both stages during the training phase were included in the reported analyses (n=591). All model equations match or extend Xia and Collins (2021) unless otherwise noted.

Choice accuracy gradually increased and plateaued across training blocks (Figure 2). This slow learning was driven by a decrease in compression error, which implies the use of compressed policies that assume independence between both stages. We hypothesized that this behavior resulted from meta-learning of the hierarchical structures and tested it by fitting two models to human data: one with meta-learning and one fully hierarchical. The meta-learning of hierarchy is modeled as a mixture policy. On trial t, the meta-policy π_M is computed by

$$\pi_M(t) = P_t(\pi_C) \cdot \pi_C(t) + P_t(\pi_H) \cdot \pi_H(t)$$

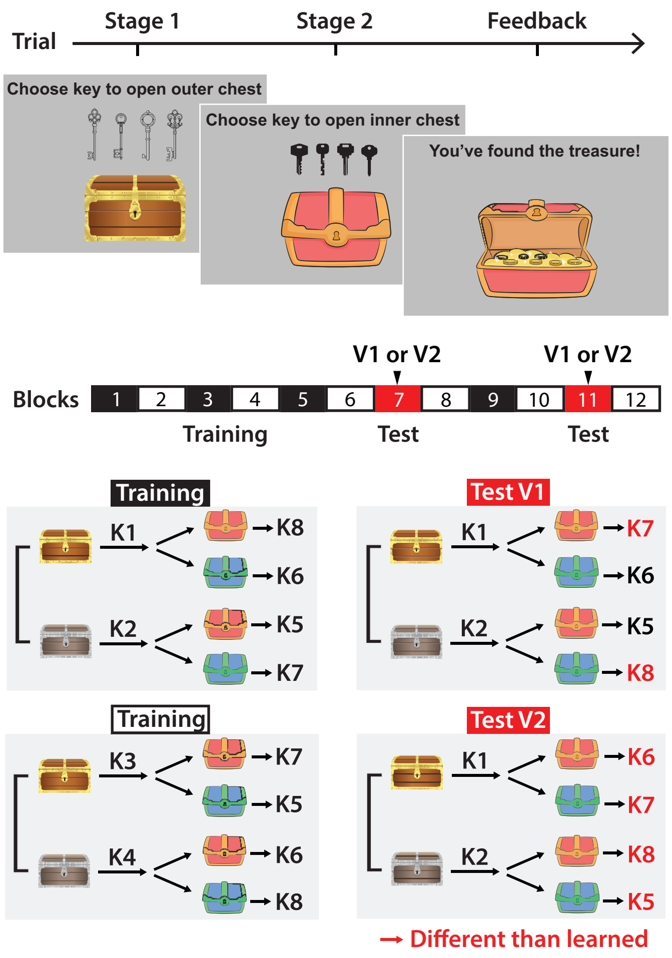

Figure 1: The task paradigm. Participants learned to unlock two nested chests (gold/silver in stage 1 followed by red/blue in stage 2) by finding the correct keys through trial-and-error via deterministic feedback. The hierarchically structured state-action mappings (context) changed every block: the correct key to the inner chest depended on the outer chest's color and the block context. Participants could only proceed to the next stage (pseudo-randomly determined) after they selected the correct key. During training, the block context alternated between two hierarchical structures (left). In test blocks, it switched to either V1 or V2 (right), which are partially similar to the first training structure (top left). The first two blocks included 60 trials and each following block included 32 trials.

where $P_t(\pi)$ denotes the probability of sampling some policy specified by the subscript (C for compressed and H for hierarchical). These probabilities are learned using Bayesian updates with a small, non-zero prior for the hierarchical policy, and the likelihood computed by marginalizing over the probabilities of sampling all compressed and hierarchical policies. Meta-learning substantially improved the model's ability to capture accuracy and compression error trends in human behavior.

As expected, participants who learned repeating test ver-

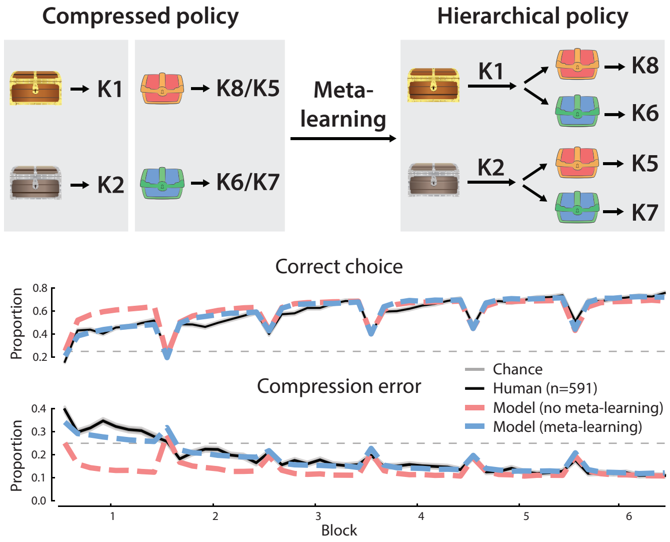

Figure 2: Compressed policies expand into hierarchical policies over learning. Using a separate policy for each stage would lead to compression error (e.g., choosing K5 instead of K8 for a red chest following a gold chest in the first training context). The rate of compression error decreases over training in humans, which is captured by the meta-learning model but not the fully hierarchical model.

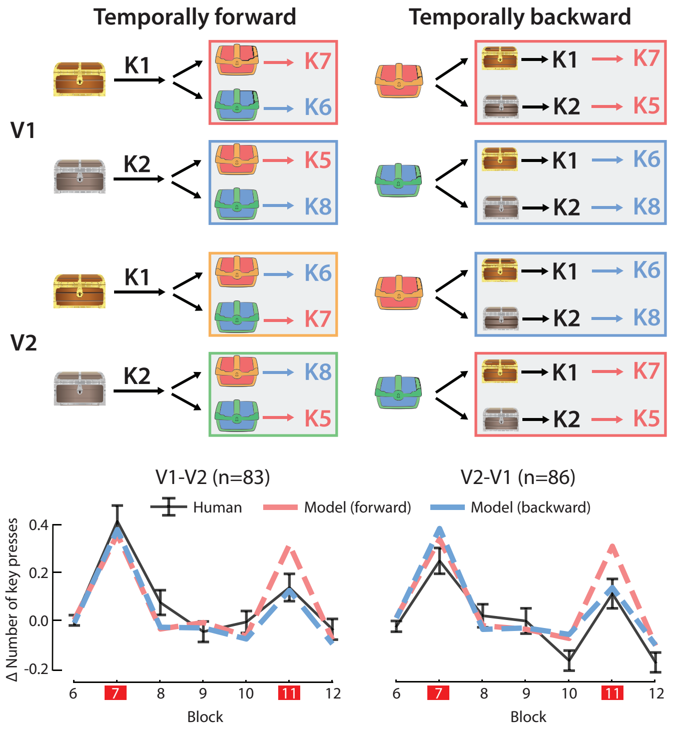

Figure 3: The learned hierarchical policies follow a temporally backward structure. A temporally forward structure implies prospective hierarchical construction (contextualized by stage 1 stimulus), while a temporally backward hierarchy is retrospective (contextualized by stage 2 stimulus). Participants who learned non-repeating test hierarchies (V1-V2 or V2-V1) showed improved performance between the two test blocks, as measured by the normalized average number of presses until finding the correct key in stage 2. This effect is captured by the temporally backward model but not the temporally forward model, since the former can compose learned structures between V1 and V2 while the latter cannot.

sions (V1-V1 and V2-V2) improved between test blocks (not reported here), indicating transfer of newly learned policies, which replicates Li et al. (2022). Surprisingly, participants who learned non-repeating test versions (V1-V2 and V2-V1) also improved between test blocks (Figure 3 bottom; one-tailed paired t-test p=3.84×10^{-5} for V1-V2 and p=1.03×10^{-2} for V2-V1). This improvement could not be explained by meta-learning only. We hypothesized that the structural similarity between V1 and V2 encouraged transfer and composition, since this effect was not observed in a control experiment where V1 was paired with a less structurally similar test block (not reported here). We compared two models with different hierarchical structures: our previous, *options*-inspired temporally forward model (Xia & Collins, 2021) that uses stage 1 (gold/silver) to contextualize the medium-level policies and a new temporally backward model that uses stage 2 (red/blue) as a context instead (Figure 3 top). The temporally backward model fitted human behavior better (one-tailed paired t-test on likelihood p=5.56×10^{-3} for V1-V2 and p=7.09×10^{-2} for V2-V1): it captured the transfer between V1 and V2, which the temporally forward model failed to (Figure 3 bottom).

Discussion

Our findings highlight processes at multiple timescales that support the acquisition of hierarchical policies in a complex, dynamic learning environment. Hierarchical policies are slowly constructed in a bottom-up fashion: simpler, compressed policies serve to bootstrap complex, hierarchical policies, which emerge through meta-learning. Furthermore, contrary to our expectations based on previous work (Botvinick et al., 2009; Xia & Collins, 2021), the structures learned by humans to represent hierarchical policies appear to be temporally "backward" rather than "forward": the *immediate* information before decision-making (stage 2 stimulus) contextualizes a policy over *earlier* information (stage 1 stimulus), which is held in memory. This structural organization is a departure from the standard options framework in hierarchical reinforcement learning, which holds the opposite (temporally forward) representation structure (Botvinick et al., 2009; Sutton, Precup, & Singh, 1999). Although both types of structures can be flexibly transferred and composed to facilitate new learning, a temporally backward one may be more resource rational, since it allows hierarchy to emerge without the effortful process of re-contextualizing compressed policies. Future research should explore the implications of our findings for human-inspired artificial intelligence and machine learning.

Acknowledgments

This work was supported by NIMH R01MH119383. We thank Liyu Xia, Milena Rmus, and Daniel Ehrlich for helpful discussions throughout the project.

References

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.

Li, J.-J., Xia, L., Dong, F., & Collins, A. G. (2022). Credit assignment in hierarchical option transfer. In *Cogsci... annual conference of the cognitive science society. cognitive science society (us). conference* (Vol. 44, p. 948).

Sutton, R. S., Precup, D., & Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, *112*(1-2), 181–211.

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