Humans flexibly transfer options at multiple levels of abstractions

Liyu Xia
Department of Mathematics
University of California, Berkeley
Berkeley, CA 94720
jimmyxia@berkeley.edu

Anne Collins
Department of Psychology
University of California, Berkeley
Berkeley, CA 94720
annecollins@berkeley.edu

Abstract

Humans are great at using prior knowledge to solve novel tasks, but how they do so is not well understood. Recent work showed that in contextual multi-armed bandits environments, humans create simple one-step policies that they can transfer to new contexts by inferring context clusters. However, the daily tasks humans face are often temporally extended, and demand more complex, hierarchically structured skills. The options framework provides a potential solution for representing such transferable skills. Options are abstract multi-step policies, assembled from simple actions or other options, that can represent meaningful reusable skills. We developed a novel two-stage decision making protocol to test if humans learn and transfer multi-step options. We found transfer effects at multiple levels of policy complexity that could not be explained by flat reinforcement learning models. We also devised an option model that can qualitatively replicate the transfer effects in human participants. Our results provide evidence that humans create options, and use them to explore in novel contexts, consequently transferring past knowledge and speeding up learning.

1 Introduction

Reinforcement Learning (RL) has greatly helped advance our understanding of human behavior. However, traditional RL algorithms suffer from the curse of dimensionality [1, 2], thus cannot account for the full complexity of human learning. To address this, [3] proposed the options framework. Options are temporally-extended multi-step policies assembled from simple actions or other options to achieve a meaningful sub-goal. Consider making coffee as an example option. We can break down the task into sub-options such as grinding coffee beans, boiling water, etc. These sub-options can be further divided until something as simple as reaching, grabbing, etc. The options framework provides many theoretical benefits, including more efficient exploration and longer-term planning [1, 2]. Moreover, recent literature [4, 5] provides potential evidence that humans may indeed use options in hierarchical tasks. In particular, [4, 5] showed that humans are able to infer meaningful sub-goals and form reward prediction error (RPE) signals for both the sub-goal and the overall goal.

However, how humans learn and transfer options remain unclear. There is little work on probing the learning dynamics in hierarchical tasks or directly testing the theoretical benefits of options in a behavioral setting. To address these questions, we took inspiration from the transfer learning paradigm [6, 7] proposed for a similar problem in the case of simpler one-step policies. [6, 7] showed that humans can create multiple policies over the same state space in a context-dependent manner; more importantly, humans can cluster different contexts together if the policy is successful. This clustering structure provides opportunities for transfer, since anything newly learned for one of the contexts can be immediately generalized to all the others in the same cluster. Moreover, human

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

participants are able to identify novel contexts as part of an existing cluster if there is enough overlap in behavioral strategies, resulting in more efficient exploration and faster learning.

Here we extended the transfer learning paradigm [6, 7] from one-step to multi-step options to behaviorally test option learning and transfer in human participants. We hypothesized that humans can learn and transfer options to guide exploration and speed up learning. These benefits can serve as behavioral signatures that test option transfer. Results support our hypotheses.

2 Experimental design

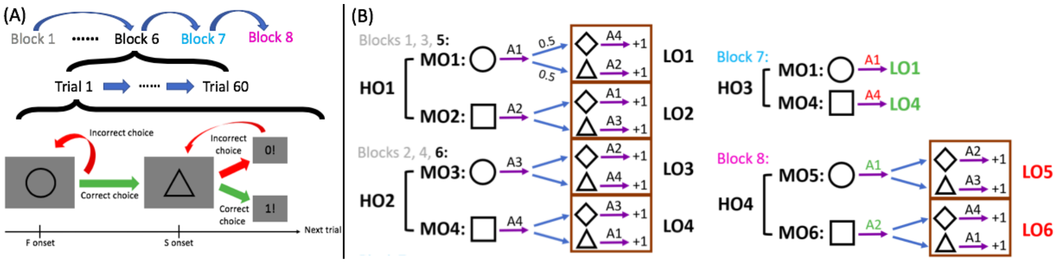

Figure 1: Experimental design. (A) Block and trial structure: Blocks 1-6 were learning blocks, followed by two testing blocks: Blocks 7 and 8. Each block had 60 trials. In each trial, participants needed to select the correct response for the first stage stimulus F in order to move on to the second stage stimulus S, where they could win points by selecting the correct response for S. (B) Stimulus-action assignments: In Blocks 1-6, participants had the opportunity to learn options at three levels of complexity: high, middle, and low-level options (HO, MO, and LO). In the testing phase, Block 7 tested participants' ability to reuse MO policies outside of their HO context; Block 8 tested negative transfer of MO policies in the second stage. Blocks were color coded for later analysis: Blocks 1-4 gray; Blocks 5-6 black; Block 7 blue; Block 8 magenta.

The experiment was approved by the UC Berkeley Institutional Review Board. 34 (22 female) UC Berkeley undergraduates participated in exchange for course credits. 7 were excluded for incomplete data or below chance performance, resulting in 27 subjects for data analysis.

The experiment consisted of eight blocks, with optional 20-second breaks in between (Fig. 1A). In each block, the participants used deterministic feedback to learn which of four keys to press for four different shapes. Two shapes were randomly chosen to be first stage stimuli (e.g. circle and square); and the other two were second stage stimuli (e.g. diamond and triangle). Each trial started with either of the first stage stimuli. Participants only moved to the second stage when they pressed the correct key for the first stage stimulus, after which either of the second stage stimuli would be randomly chosen to be presented. Both first stage stimuli led to both second stage stimuli equally often. Participants did not receive reward feedback after making choices in the first stage, other than seeing the same stimulus if they were incorrect, or switching to a different (second stage) stimulus if they were correct. In the second stage, participants also needed to select the correct key before they could move to another trial. Participants received 1 point for choosing the correct key, and 0 point for incorrect keys. Each block contained 60 trials, with each of first stage stimuli leading to each of second stage stimuli 15 times, all pre-randomized.

The correct stimulus-action assignments (Fig. 1B) changed across blocks. Blocks 1, 3, 5 shared the same assignments; Blocks 2, 4, 6 shared the same assignments. Blocks 7 and 8 tested transfer of options at various levels. Specifically, the protocol was set up so that participants could learn up to 3 levels of hierarchical task structure: low-level options (LO) for second stage policies; mid-level options (MO) for the pairing of first and second stage policies; high-level options (HO) for policies over MO's. Imagine in Blocks 1, 3, 5, the participants learned how to make breakfast (HO_1), consisting of toast (MO_1) and milk (MO_2). Making toast was broken down into cutting bread (the first stage) and spreading butter (the second stage, LO_1). In Blocks 2, 4, 6, participants learned how to make lunch (HO_2), consisting of coffee (MO_3) and spaghetti (MO_4). Coffee making was broken into grinding beans (the first stage) and boiling water (the second stage, LO_3). Block 7 combined the policies for toast from breakfast (MO_1) and spaghetti from lunch (MO_4) to form HO_3. Block 8 shared the same assignments as Blocks 1, 3, 5 in the first stage, but the second stage was new: for example, participants needed to use bread to make sandwiches (LO_5) instead of toast (LO_1).

3 Modeling

We simulated and compared 4 RL models with human data. All models were simulated 500 times with one set of chosen parameters.

3.1 The Option Model

We simulated the Option Model, an implementation of the options framework on the same task. In the first stage, the model tracks the probability P^1 of selecting each HO_i in different first stage contexts c_j^1, which encodes the current Block (temporal) context. In particular, the model runs a Chinese Restaurant Process (CRP) [8] to select HO: if contexts $\{c_{1:n}^1\}$ are clustered on $N^1 \leq n\ HO's$, when the model encounters a new context c_{n+1}^1, the prior probability of selecting a new HO_{n+1} is set to $P^1(HO_{n+1}|c_{n+1}^1) = \gamma^1/Z^1; P_1(HO_i|c_{n+1}^1) = N_i^1/Z^1$, where N_i^1 is the number of first stage contexts clustered on HO_i, and $Z^1 = \gamma^1 + \sum_i N_i^1$. The new HO_{n+1} is initialized with uninformative Q-values $1/\#\{possible\, actions\} = 1/4$. The model selects HO with the highest probability in the current context. Once an HO is selected, an action is sampled from the HO-specific policy based on the Q-values for first stage stimulus F: $P(A_i^1|F, HO) = exp(\beta_1 * Q_{HO}^1(F, A_i^1))/\sum_j exp(\beta^1 * Q_{HO}^1(F, A_j^1))$. After observing the outcome, the model uses Bayesian inference to update P^1: $P^1(HO_i|c_j^1) = P(r|F, A^1, HO_i)P(HO_i|c_j^1)/(\sum_k P(r|F, A^1, HO_k)P(HO_k|c_j^1))$, where r is 1 if A^1 is correct or 0 otherwise. Then the policy of the HO with the highest posterior probability is updated with Q-learning: $Q_{HO}^1(F, A^1) = Q_{HO}^1(F, A^1) + \alpha^1 * (r - Q_{HO}^1(F, A^1))$.

To simplify credit assignment, we assumed that selecting an action for a first stage stimulus activates an MO: for example, selecting A_1 for circle activates MO_1 (Fig. 1B). Each MO has an MO-specific probability table P_{MO}^2 similar to P^1 in the first stage, which guides LO selection in the second stage. The second stage is identical to the first stage, except that the second stage contexts for CRP are determined by which MO is activated. Thus the second stage parameters are α^2, β^2 and γ^2.

To better account for human behavior, we included two forgetting parameters, f^1 and f^2, and a meta-learning parameter m. At each choice, the model decays all Q-values for the first stage based on f^1: $Q_{HO}^1(F_i, A_j^1) = (1 - f^1) * Q_{HO}^1(F_i, A_j^1) + f^1 * 1/4$. Forgetting in the second stage was implemented similarly. The meta-learning parameter m discourages selecting the same action in the second stage as in the first stage. To closely capture human behavior, we simulated the Option Model with $\alpha^1 = 1, \beta^1 = 2, \gamma^1 = 14, f^1 = 0.0004, \alpha^2 = 0.8, \beta^2 = 3, \gamma^2 = 3, f^2 = 0.0002, m = 0.01$. However, the qualitative results of the model hold for a wide range of parameters.

3.2 The 1-Step Transfer Model

The 1-Step Transfer Model has the capability of transferring previously learned 1-step policies. It only differs from the Option Model in that the CRP in the second stage is independent of that in the first stage (in the Option Model, the second stage CRP depends on which MO is activated in the first stage). We simulated the 1-Step Transfer Model with the same parameters as the Option Model.

3.3 The Naive Flat Model

The Naive Flat Model learns the Q-values for the two first stage stimuli with learning rate α^1, inverse temperature β^1 and forgetting rate f^1. However, it disregards the non-Markovian nature of the task: it learns the Q-values for the two second stage stimuli (with α^2, β^2, f^2), without remembering the first stage stimulus. When entering in a new block, the Naive Flat Model has to re-learn all Q-values from scratch. The Naive Flat model also has a meta-learning parameter m. We simulated the Naive Flat Model with $\alpha^1 = 0.5, \beta^1 = 4, f^1 = 0.0025, \alpha^2 = 0.7, \beta^2 = 10, f^2 = 0.0001, m = 0.01$.

3.4 The Flat Model

The Flat Model is identical to the Naive Flat Model, except that in the second stage, the Flat Model treats the 4 combinations of the first and second stage stimuli as 4 distinct states. We simulated the Flat Model with the same set of parameters as the Naive Flat Model.

4 Results

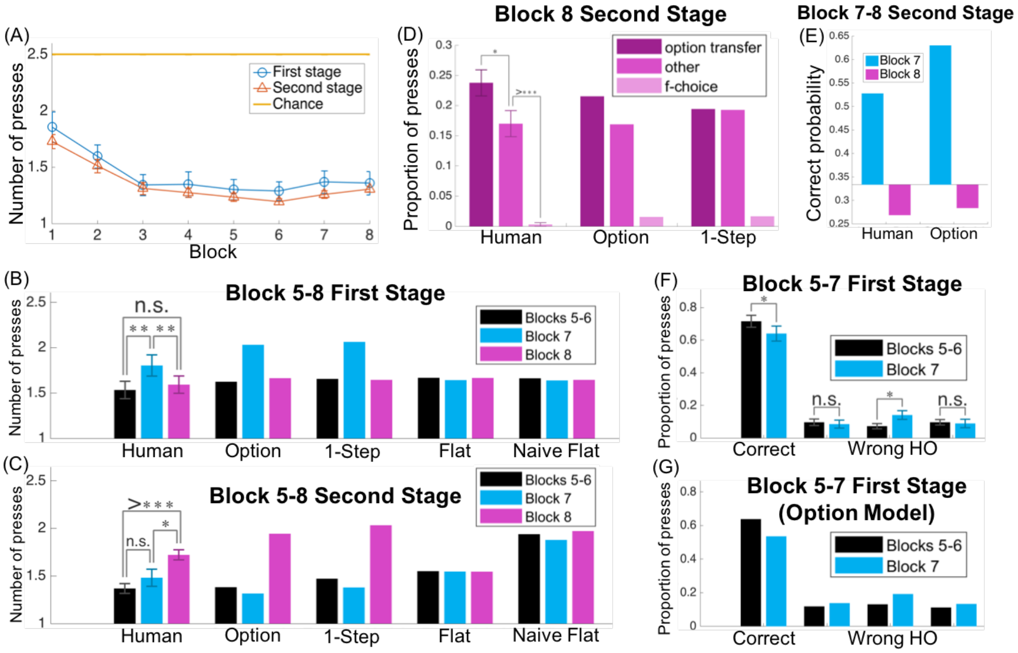

Figure 2: Behavioral and modeling results. (A) Block 1-8: average number of key presses. (B) Block 5-8 first stage: average number of key presses in the first 10 trials for human participants and 4 competing models. (C) Block 5-8 second stage: same as (B) for the second stage. (D) Block 8 second stage: proportion of choice types for human participants, the Option Model, and the 1-Step Transfer Model. (E) Block 7-8 second stage: correct probability of first press for human participants and the Option Model. (F) Block 5-7 first stage: proportion of choice types for human participants. (G) Block 5-7 first stage: same as (F) for the Option Model.

4.1 Participants do not use flat RL

We use the number of key presses until correct choice in each stage of a trial as an index of performance. Ceiling performance is 1 press per stage in a trial. Chance level is 2.5, assuming participants randomly choose each of the four keys without repetition. Participants learned the correct actions in both the first and second stage, as shown by better than chance and improving performance over Blocks 1-6 (Fig. 2A). We used the average of Blocks 5 and 6 as a benchmark for comparing against Blocks 7 and 8.

To probe potential transfer effects, we counted the average number of key presses in the first 10 trials of each block (Fig. 2B-C), since we only expected transfer effects at the beginning of each block, before behavior stabilized. In the first stage (Fig. 2B), participants pressed more times in Block 7 than Block 8 (paired t-test, P = 0.0014) and the average of Blocks 5 and 6 (paired t-test, P = 0.0038); there was no significant difference between Block 8 and the average of Blocks 5 and 6 (paired t-test, P = 0.3592). These results suggest participants negatively transferred HO in the first stage of Block 7, and positively transferred HO in the first stage of Block 8, as predicted.

In the second stage (Fig. 2C), participants pressed more times in Block 8 than Block 7 (paired t-test, P = 0.014) and the average of Blocks 5 and 6 (paired t-test, P < 0.001); there was no significant difference between Block 7 and the average of Blocks 5 and 6 (paired t-test, P = 0.1486). These results suggest participants positively transferred LO in the second stage of Block 7, and negatively transferred LO in the second stage of Block 8.

Among the simulations of 4 models (Fig. 2B-C), only the Option Model and the 1-Step Transfer Model could account for all transfer effects so far. The Naive Flat Model cannot achieve reasonable performance in the second stage because it ignores the non-Markovian nature of the task. The Flat Model achieves reasonable performance on the task, but does not demonstrate any transfer effects.

4.2 Second stage choices reveal option transfer

Specific errors participants make can reveal the structure they use to make decisions. To further disambiguate between the Option Model and the 1-Step Transfer Model, we categorized their errors into meaningful choice types [6]. For example, for the second stage of Block 8, consider the diamond following the circle in Block 8 (Fig. 1B): A_2 is the correct action; an A_1 error corresponds to the correct action in the first stage ("f-choice" type); an A_4 error would be the correct action if selecting MO_1 as a whole("option transfer" type); an A_3 error is labeled "other" type. Among the 3 error types (Fig. 2D), there were more "option transfer" type than the "other" type (paired t-test, P = 0.046), suggesting participants selected previously learned MO's as a whole. The Option Model can reproduce this effect. The 1-Step Transfer Model cannot: because the two CRPs in the first and second stages are independent, first stage choices do not inform second stage choices. Therefore, the choice type profile in Block 8 cannot be explained by transfer of 1-step policies alone.

We also computed the correct probability on the first press for the 4 branches in the second stage (2 branches following the circle and another 2 following the square, see Fig. 1B) (Fig. 2E), and compared to chance (1/3, accounting for meta-learning that the correct action in the second stage was always different from the first stage). Correct probability in Block 7 and the average of Blocks 5 and 6 were significantly higher than chance (sign-test, P(Block 7) = 0.0192, P(Blocks 5 and 6) < 0.001), without significant difference between the two (sign-test, P = 0.1892). Thus, there was positive transfer in the very first trials of Block 7. Block 8 was significantly below chance (sign-test, P = 0.0059), independently indicating negative transfer of learned mid-level options in the very first trials. The Option Model was able to qualitatively reproduce these transfer effects on the first press.

Positive transfer in the second stage of Block 7 was not affected by interference from the first stage of Block 7. Moreover, the worse performance in the second stage of Block 8 was mainly due to negative transfer of learned options (MO's) based on choice type analysis. Finally, the transfer effects on the first press of each block suggests that participants were using previously learned options to speed up learning in novel contexts without any experience in a new block.

4.3 First stage choices reveal transfer of policies over options

We hypothesized that the more key presses in Block 7 (Fig. 2B) was due to selecting the wrong HO. To test this, we also categorized error choices into 3 types. For example, consider the circle in Blocks 1, 3, 5 (Fig. 1B): A_1 is the correct action; an A_2 error corresponds to the correct action for the square in the same block ("wrong shape" type); an A_3 error corresponds to the correct action for the circle in Blocks 2, 4, 6 ("wrong HO" type), and an A_4 error is the "both wrong" type. We found that, compared to the average of Blocks 5 and 6 (Fig. 2F), the only error type that significantly increased in Block 7 was the "wrong HO" type (paired t-test, P = 0.0319). This suggests that the decreased performance in Block 7 was primarily due to selecting the wrong HO, highlighting negative transfer of options over options. The Option Model reproduces this transfer effect (Fig. 2G).

5 Conclusion

Our findings provide novel and strong support for the acquisition of option representation in human participants. The second stage of Block 7 provides a test of positive option transfer, whereas the second stage of Block 8 provides a test of negative option transfer. Moreover, participants were able to form policies over options (HO); this additional hierarchy built on top of MO did not interfere with the transfer of MO, demonstrating flexibility in transfer. Participants were also able to transfer previously learned options on the first press to speed up learning in novel contexts, further demonstrating the theoretical benefits of the options framework.

References

[1] Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. Cognition, 113(3), 262-280.

[2] Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. Current opinion in neurobiology, 22(6), 956-962.

[3] 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.

[4] Ribas-Fernandes, J. J., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. Neuron, 71(2), 370-379.

[5] Ribas-Fernandes, J. J., Shahnazian, D., Holroyd, C. B., & Botvinick, M. M. (2019). Subgoal-and goal-related reward prediction errors in medial prefrontal cortex. Journal of cognitive neuroscience, 31(1), 8-23.

[6] Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. Psychological review, 120(1), 190.

[7] Collins, A. G. E., & Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. Cognition, 152, 160-169.

[8] Pitman, J. (2006). Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002. Springer.