The options framework enables flexible transfer in humans

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

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

Abstract

Humans' ability to flexibly transfer previously learned skills to novel contexts is a fundamental ability that sets humans apart from state-of-the-art Artificial Intelligence (AI) algorithms. But human transfer is not well understood. Recent work proposed a theory for transferring simpler, one-step stimulus-action policies called task-sets. However, the daily tasks humans face are high dimensional and demand more complex skills due to curse of dimensionality. Hierarchical reinforcement learning's options framework provides a potential solution. Options are abstract multi-step policies, assembled from simple actions or other options, that can represent meaningful reusable skills. In this study, we extend the transfer learning paradigm that tests task-set transfer to the scenario of multi-step options, aiming to test if humans can indeed learn and transfer options at multiple levels. We developed a novel two-stage reinforcement learning protocol. Participants learned to choose the correct action in response to stimuli presented at two successive stages to receive reward in a trial. Crucially, we designed the contingencies leading to reward to provide participants opportunities to create options at multiple levels of complexity, and to transfer them in new contexts. Results from this experiment and another control experiment showed transfer effects at multiple levels of policy complexity that could not be explained by traditional flat reinforcement learning models. We also devised an option model that can qualitatively replicate the transfer effects in humans. Our computational and behavioral results provide evidence for option learning and flexible transfer at multiple levels of hierarchy. This has implications for understanding how humans learn flexibly, explore efficiently, and generalize knowledge.

Keywords: Hierarchical Reinforcement Learning, Options Framework, Transfer Learning

Acknowledgements

1 Introduction

Reinforcement Learning (RL) has helped advance our understanding of human behavior. However, traditional RL algorithms are unable to account for the full complexity of human learning; for example, they suffer from the curse of dimensionality [1, 2]. The *options* framework addresses some of these limitations [3]. 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 we reach something as simple as reaching, grabbing, etc.

The options framework provides many theoretical benefits, including more efficient exploration and longer-term planning [1]. For example, when we learn how to make a new kind of coffee, we already know how to engage in numerous related tasks and do not need to re-learn the entire process. In addition to the theoretical benefits, recent literature [2, 4] provides preliminary evidence that humans' hierarchical behavior may be well described by options. They showed that humans are able to infer meaningful sub-goals and form reward prediction error signals for both the sub-goal and the overall goal, as predicted by the options framework.

However, the fundamental questions of whether and how humans learn options and use options remain unanswered: there is little work on probing the learning dynamics in hierarchical tasks or directly testing the theoretical benefits of options in a behavioral setting. In particular, do humans create options in such a way that they can flexibly reuse them in new problems? If so, how flexible is this transfer? To answer these questions, we took inspiration from the paradigm in [5] that was designed to investigate these questions for simpler one-step policies called task-sets. [5] showed that humans can learn task-sets and cluster different contexts together if the same task-set applies. This clustering provides opportunities for transfer, since anything newly learned for one of the contexts can be immediately applied to the other contexts in the same cluster. Moreover, human participants were able to identify novel contexts as part of an existing cluster if there is enough overlap in behavioral strategies, resulting in more efficient exploration.

Here, we extend this to a multi-step paradigm in order to behaviorally test option learning and transfer in humans. Given that humans can transfer task-sets to novel contexts [5], we hypothesized that humans can learn and transfer options in a similar fashion to guide exploration and achieve faster learning. These benefits thus serve as behavioral signatures that test option acquisition.

This combination of the options framework and task-set transfer paradigm can provide great insights for studying complex human behavior. Thanks to the additional hierarchical structure in the action space, transfer of prior knowledge becomes possible at multiple hierarchies, providing rich opportunity for studying the flexibility of human transfer. For example, if we have learned water boiling while learning coffee making, we do not need to re-learn water boiling while learning tea making; this sub-skill can instead be incorporated into a tea-making option.

2 Methods

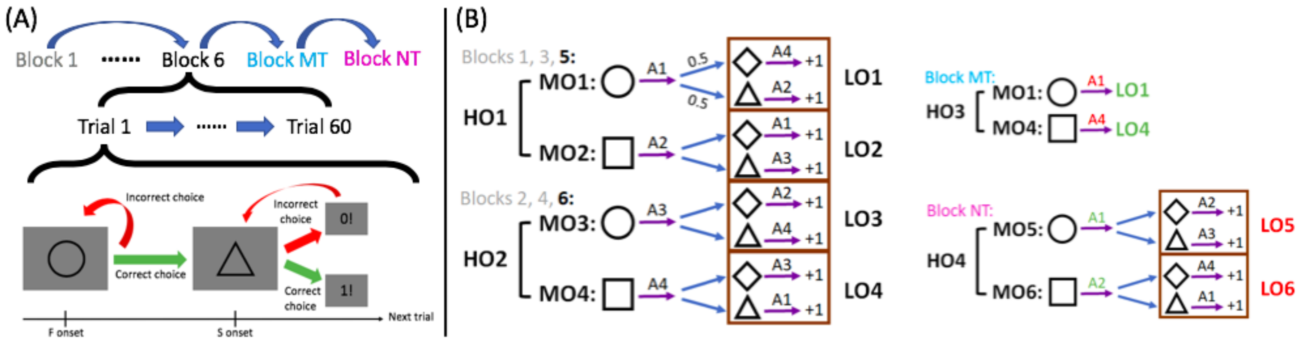

Figure 1: Experimental design. (A) Block and trial structure: Blocks 1-6 were learning blocks, followed by two testing blocks: Blocks MT and NT. Each block had 60 trials. In each trial, participants needed to select the correct response for the first stage stimulus F in order to move on to the second stage stimulus S, where they could win points by selecting the correct response for S. (B) Correct action assignment: 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 MT tested participants' ability to reuse MO policies outside of their HO context; Block NT 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 MT blue; Block NT magenta.

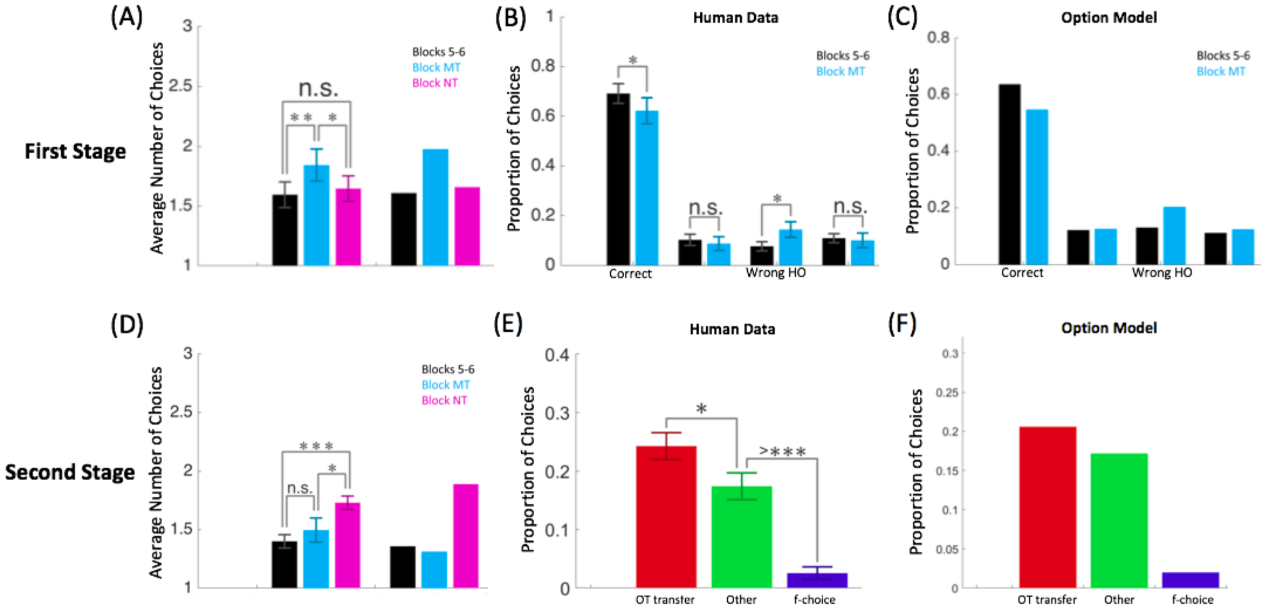

Figure 2: Behavioral and modeling results. (A)-(C) First stage. (A) Average number of key presses of the first 10 trials in the first stage for both human data and the Option Model. (B) Choice type proportions of participants' first key press of the first 10 trials in the first stage. (C) Same as (B) but for the Option Model. (D)-(F) Second stage. (D) Average number of key presses of the first 10 trials in the second stage for both human data and the Option Model. (E) Choice type proportions of the participants' first key press of each of the first 3 trials of each combination of first and second stage stimulus in the second stage (not summing to 1 because leaving out the correct proportion). (F) Same as (E) but for the Option Model.

2.1 Experimental Design

The experiment was approved by the UC Berkeley Institutional Review Board. It was administered to UC Berkeley undergraduates in exchange for course credits. 29 (19 females) undergraduates participated in the experiment. 6 were excluded for incomplete data or below chance performance, resulting in 23 subjects for data analysis.

The experiment consisted of eight blocks (Figure 1A). Within each block, the participants used deterministic feedback to learn which of four actions to select when presented with one of four different shapes. The participants chose an action by pressing one of four adjacent keys with the fingers of their dominant hand. Two of the shapes were randomly chosen to be first stage stimuli (e.g circle (F_1) and square (F_2)); the other two were second stage stimuli (e.g. diamond (S_1) and triangle (S_2)). Each block had 60 trials. Each trial started with either F_1 or F_2. The same first stage stimulus was repeated until the participant selected the correct action for that stimulus. This triggered either S_1 or S_2 with equal probability for the second stage of the trial. Participants did not receive explicit reward feedback for their choices in the first stage. Similarly, in the second stage, the same stimulus was repeated until participants selected the correct action. Participants received reward feedback of 1 point and heard an upward sound for choosing the correct choice, and no point with a downward sound for incorrect choices. F_1 and F_2 led to S_1 and S_2 15 times each, pre-randomized.

Importantly, the correct action for the same second stage stimulus was different depending on its preceding first stage stimulus (Figure 1B). For example, in Blocks 1, 3, and 5, the correct action for the diamond was A_4 if the first-stage stimulus was the circle, but A_1 if it was the square. This non-Markovian manipulation allowed us to test option learning: in the options framework, the correct action at the same state could be encoded by a different option-specific policy, with the appropriate option selected in the first stage.

Unbeknownst to the participants, the correct stimulus-action assignments systematically changed across blocks. Blocks 1, 3, and 5 shared the same assignments, as did Blocks 2, 4, and 6. We designed the experiment so that participants could potentially learn 3 levels of task structure: lower-level options (LO) for second stage policies, similar to task-sets; middle-level options (MO) for policies connecting first and second stages initiated by one of the first stage stimuli; and high-level options (HO) for task-sets over MOs. Expanding the coffee-making analogy, imagine in Blocks 1, 3, and 5, the participants learned how to make breakfast (HO_1) consisting of tea (MO_1) and toast (MO_2). Making toast was broken down into slicing bread (the first stage) and toasting (the second stage, LO_2). In Blocks 2, 4, and 6, participants learned how to make lunch (HO_2) consisting of coffee (MO_3) and spaghetti (MO_4). Coffee making is further decomposed into grinding beans (the first stage) and boiling water (the second stage, LO_3).

To test whether participants created and could transfer options at multiple levels, we designed Blocks MT (mixed transfer) and NT (negative transfer) (Figure 1B). In Block MT, we tested whether participants could reuse mid-level options outside of their high-level option contexts: the assignments were a combination of tea from breakfast (MO_1) and spaghetti from lunch (MO_4). In Block NT, we tested for negative transfer of mid-level options. Specifically, although the correct actions for both F_1 and F_2 were the same as in Blocks 1, 3 and 5, the correct actions for both S_1 and S_2 were new; e.g, instead of using the bread to make toast (LO_1), the participants needed to use bread to make a sandwich (LO_5).

To measure performance, we counted the number of key presses per trial. Since the experiment would not progress unless the participants chose the correct action, more key presses suggests worse performance. Ceiling performance is 1 press per stage within a trial, while the chance level is 2.5, assuming choosing 1 out of 4 keys randomly each time.

2.2 Modeling

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 to select HO [6]: 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 high-level option HO_{n+1} in this new context 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 high-level option's policy 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. The model also tracks HO-specific policies via RL. Once an HO is selected, a first stage policy is computed based on the HO's Q-values and the first stage stimulus F: $P(A_i^1|F, HO) = exp(\beta_1 * Q_{HO}^1(F, A_i^1))/\sum_j exp(\beta^1 * Q_{HO}^1(F, A_j^1))$. A first stage action A^1, ranging from A_1 to A_4, is then sampled from this computed policy. After observing the outcome (moving on to the second stage or not), 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 if 1 if A^1 is correct or 0 if it is not. Then the policy of the HO with the highest posterior probability is updated with Q-learning: $Q_{HO}^1(F, A^1) = Q_{HO}^1(F, A^1) + \alpha^1 * (r - Q_{HO}^1(F, A^1))$. Thus the parameters involved in the first stage are α^1, β^1 and γ^1.

To simplify credit assignment, we assumed that selecting a particular action for a particular first stage stimulus is equivalent to selecting an MO (e.g. selecting A_1 for circle activates MO_1 (Figure 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 similar to the first stage, except that the second stage contexts are determined by which MO is activated, instead of the current Block. All the equations of Chinese Restaurant Process, action selection and Q-learning remains the same, except for replacing c^1 by MO. The parameters involved in the first stage 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 is implemented similarly. The meta-learning parameter m discourages selecting the same action in the second stage as in the first stage. We simulated the Option Model 100 times with the following parameters: $\alpha^1 = 1, \beta^1 = 2, \gamma^1 = 13, f^1 = 0.0004, \alpha^2 = 0.8, \beta^2 = 3, \gamma^2 = 3, f^2 = 0.0002, m = 0.01$.

We also simulated 3 baseline models for comparison. Two were flat RL models: one learned action values for each of the 4 shapes; the other learned action values for the combination of the first and second stage stimuli. The two flat RL models were not able to reproduce any transfer effects. We also simulated a task-set model. The model can transfer HO and LO policies separately without implementing MO as in the Option Model. Since HO does not inform the selection of LO, this model can reproduce all the transfer effects except for Figure 2E.

3 Results

3.1 Overall performance

As the task progressed, performance saturated at Blocks 5 and 6. For the last 10 trials of Blocks 5 and 6, participants on average pressed 1.27 (SD = 0.58) times per trial for the first stage, and 1.15 (SD = 0.12) times per trial for the second stage. For transfer effects, we focused on the first 10 trials at each stage. For the first stage, participants pressed significantly more times in Block MT than Block NT (paired t-test, P = 0.015), and the average of Blocks 5 and 6 (paired t-test, P = 0.009); there was no significant difference between Block NT and the average of Blocks 5 and 6 (paired t-test, P = 0.478). The Option Model is able to reproduce these effects by erroneously activating HO_1 and HO_2 in Block MT. For the second stage, participants pressed significantly more times in Block NT than Block MT (paired t-test, P = 0.023) and the average of Blocks 5 and 6 (paired t-test, $P < 0.001$); there was no significant difference between Block MT and the average of Blocks 5 and 6 (paired t-test, P = 0.281). The Option Model again reproduces all these effects by leaving the second stage of Block MT intact, and erroneously reusing MO_1 and MO_2 as a whole in Block NT (Figure 2A, D).

3.2 Error types show evidence of transfer

We next investigated more precisely the actual actions that each participants selected, as specific errors can be informative about policy choice [5]. We only considered the first key press within each trial. We categorized choices in the first stage into 4 types. Consider the circle in Blocks 1, 3, 5 (Figure 1B): A_1 was the correct action; A_2 was the correct action for the square in the same block, thus the 'correct block wrong context' action; A_3 was the correct action for the circle in Blocks 2, 4, 6, thus the 'wrong HO' action, and A_4 was the 'wrong block wrong context' action. Our hypothesis suggested that the negative transfer in the first stage of Block MT should be a result of selecting the wrong HO. Indeed, we found that, compared to the average of Blocks 5 and 6 (Figure 2B), the only error type that significantly increased was the 'wrong HO' action (paired t-test, P = 0.0266). The Option Model reproduces this effect by erroneously selecting HO_1 or HO_2 in the first stage (Figure 2C).

We similarly probed the choice types in the second stage, For Block NT, consider the diamond following the circle in Block NT: A_2 was the correct action; A_1 was the same as the first stage action, thus 'f-choice' action; A_4 was the correct action if selecting MO_1 as a whole, thus the 'Option Transfer' action; A_3 is the 'other' action. Among the 3 error types (Figure 2E), 'Option transfer' choices were significantly more than the 'Other' type (paired t-test, P = 0.0414), suggesting the primary source of error was due to participants selecting previously learned MO as a whole. The Option Model reproduces this transfer effect, since the information from the first stage, carried by MO, misleads LO selection in the second stage. We also looked at the choice types of the second stage of Block MT. We found no significant difference between the second stage choice types of Block MT and the average of Blocks 5 and 6, suggesting that participants were still selecting MO as a whole in Block MT, even when interfered by the negative transfer in the first stage.

4 Conclusion

Our findings provide novel and strong support for the acquisition of option representation in human participants. The second stage of Block MT provides a test of positive option transfer, whereas the second stage of Block NT provides a test of negative option transfer. Moreover, participants were able to form task-sets of options (HO); this additional hierarchy built on top of MO did not interfere with the transfer of MO, demonstrating flexibility in transfer.

We also looked at reaction time to probe potential sequence learning effects. We did not find significant difference between the reaction time of choices that either follow the first stage action or not from learning, ruling out pure sequence learning. In addition, since there were two possible stimuli in the second stage, participants had to pay attention to the identity of the second stage stimuli, making it necessary to have an actual policy for the second stage.

One potential limitation in the protocol is the counterbalancing of Blocks MT and NT. In order to eliminate the potential impact of Block MT on Block NT, we ran a control experiment where we removed Block MT. We replicated all our previous conclusions of Block NT (details not shown).

In summary, we find compelling evidence of building options from scratch by probing the learning and transfer dynamics of a novel two-stage experimental protocol. Sequence learning and traditional flat RL models are not able to account for the positive and negative transfer effects shown here. We also took the first step in probing the flexibility of option transfer, and showed that human participants were able to flexibly transfer previously learned options without interference from higher levels. An interesting future question is whether participants learned a new MO_5 in Block NT or simply re-wrote the policy of MO_1.

References

[1] Matthew M Botvinick, Yael Niv, and Andrew C Barto. "Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective". In: *Cognition* 113.3 (2009), pp. 262–280.

[2] Carlos Diuk et al. "Divide and conquer: hierarchical reinforcement learning and task decomposition in humans". In: *Computational and robotic models of the hierarchical organization of behavior*. Springer, 2013, pp. 271–291.

[3] Richard S Sutton, Doina Precup, and Satinder Singh. "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning". In: *Artificial intelligence* 112.1-2 (1999), pp. 181–211.

[4] José JF Ribas-Fernandes et al. "Subgoal-and Goal-related Reward Prediction Errors in Medial Prefrontal Cortex". In: *Journal of cognitive neuroscience* 31.1 (2019), pp. 8–23.

[5] Anne GE Collins and Michael J Frank. "Cognitive control over learning: Creating, clustering, and generalizing task-set structure." In: *Psychological review* 120.1 (2013), p. 190.

[6] Jim Pitman. *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer, 2006.