Hierarchical Reinforcement Learning enables flexible transfer in humans

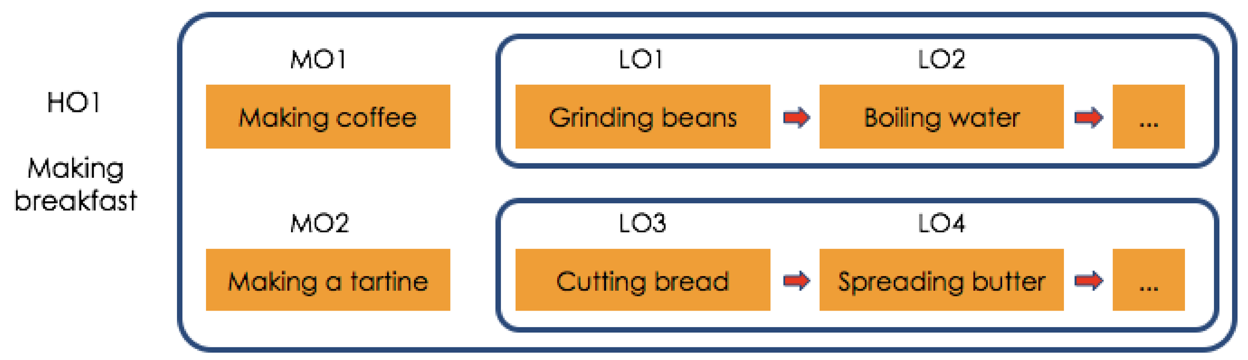
Liyu Xia, Anne G.E. Collins. University of California, Berkeley.

Poster C134

Introduction

Traditional reinforcement learning (RL) has 2 major limitations:
1. **Cannot** scale up to complex tasks that humans face.
2. **Cannot** explain how humans transfer previously learned skills to novel contexts.

With the observation that human behavior is hierarchical [1], recent studies proposed **the options framework** [2] from Hierarchical Reinforcement Learning (HRL) which provides many theoretical benefits [3]. Options are temporally-extended policies composed of primitive actions and/or smaller options.



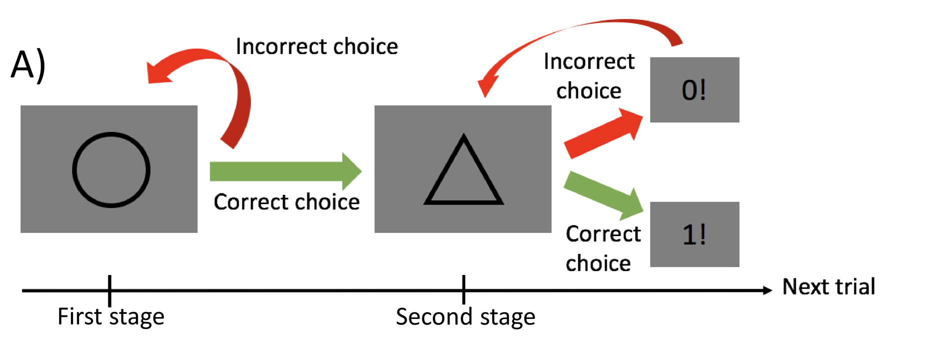
- HO: High-level option
- MO: Middle-level option
- LO: Low-level option

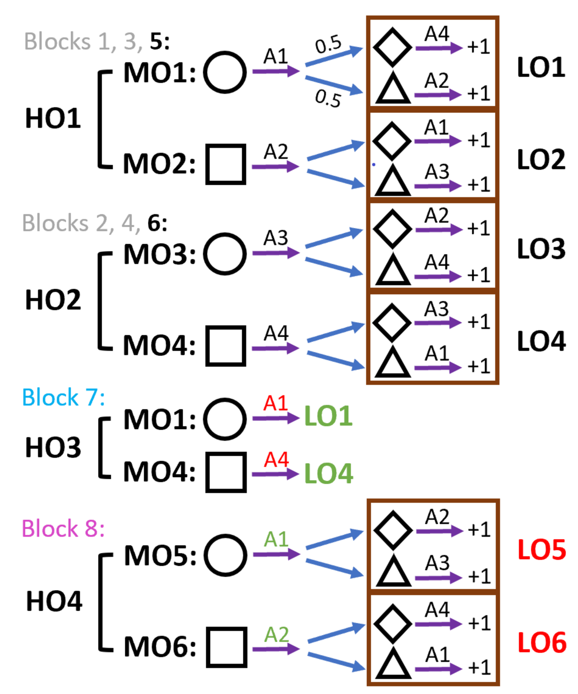
Prior work showed humans can learn 1-step policies (or task sets), and are able to transfer them to novel contexts [4].

Questions:

1. Do humans learn options? At multiple levels?
2. If so, can humans transfer learned options?

Experimental Protocol

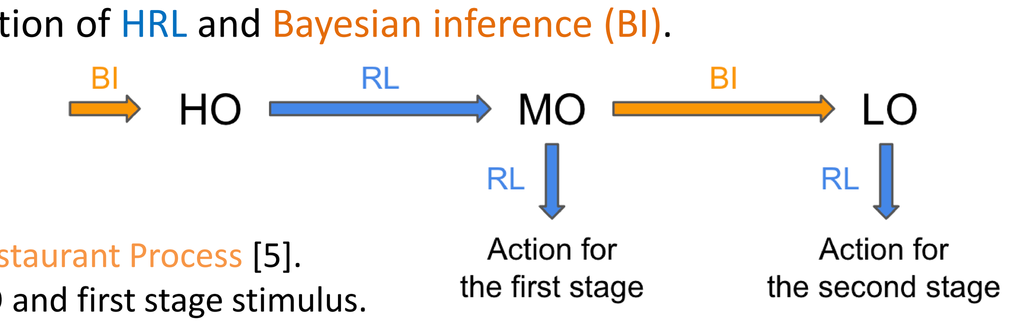




A) Participants select the correct action in **the first stage** to enter the **second stage**. They need to be correct again to win a point. Red means incorrect; green means correct.
B) Correct action changes across block. Red indicates potential negative transfer; green indicates potential positive transfer.

Option Model:

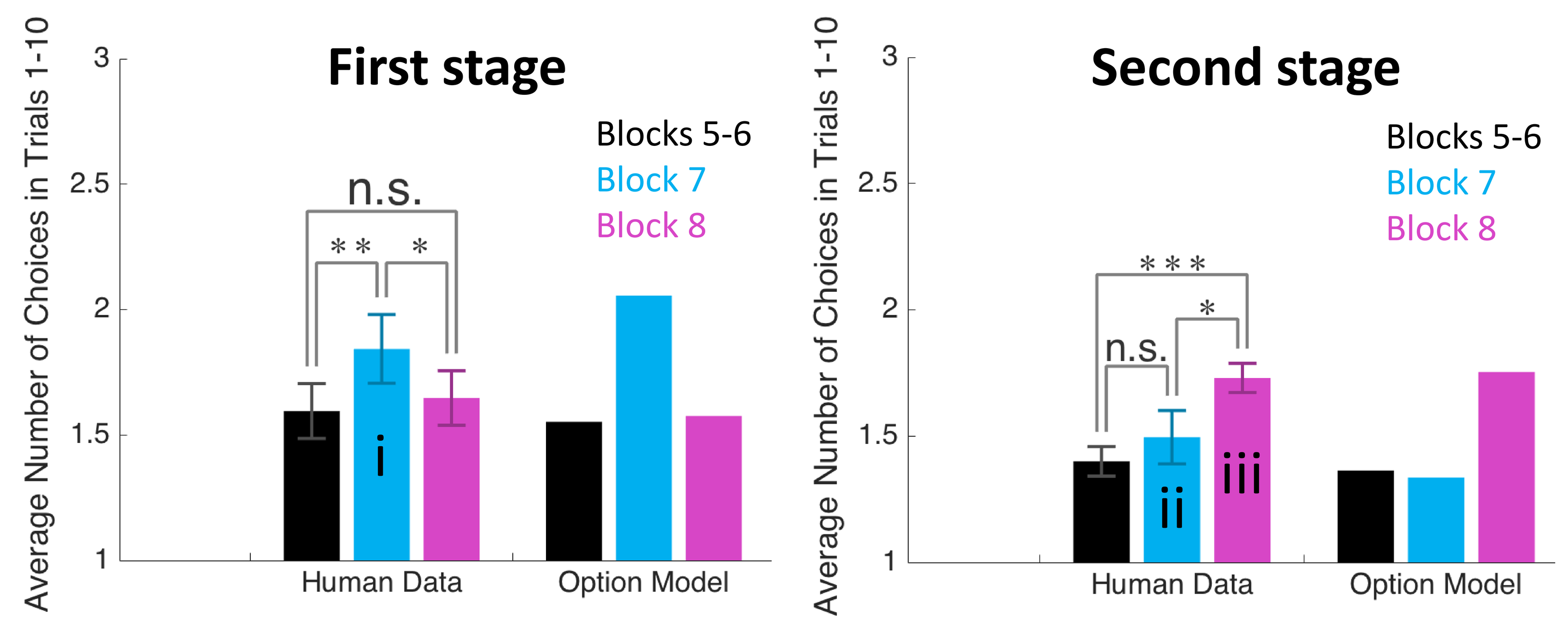
The option model is a combination of HRL and Bayesian inference (BI).



1. Choose an HO using Chinese Restaurant Process [5].
2. Choose an MO based on the HO and first stage stimulus.
3. Choose an **action for the first stage** based on the policy dictated by the MO.
4. Choose an LO based on the MO's policy. This policy is learned by BI.
5. Choose an **action for the second stage** based on the policy dictated by the LO.

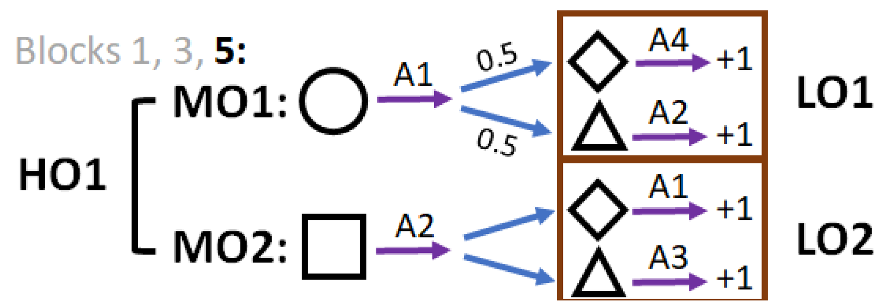
Behavioral results support model predictions

We counted **the number of key presses** in the **first 10** trials for transfer effects at the beginning of a block.



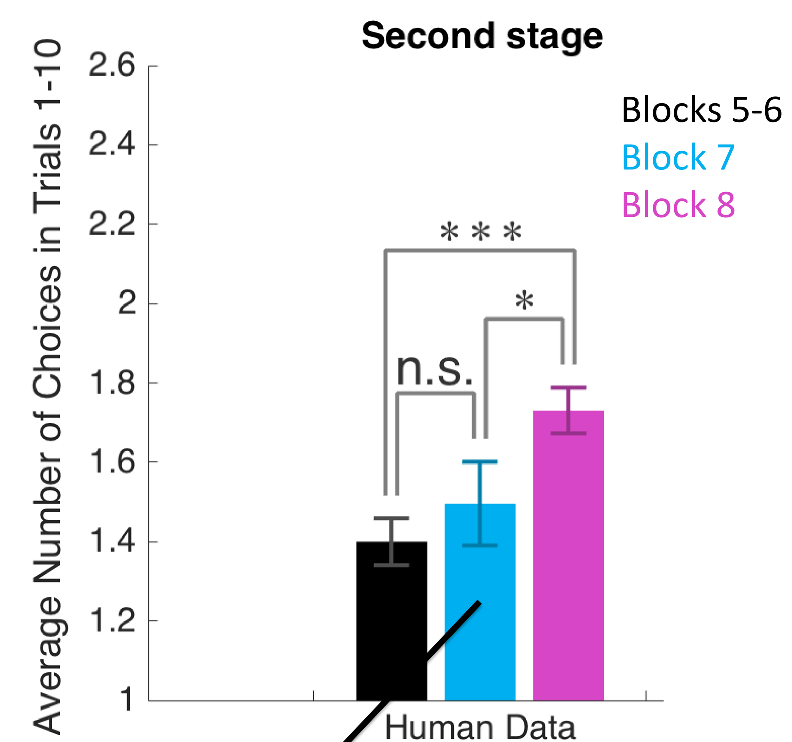
1. Behavioral data provides initial evidence for different transfer effect at both stages:
   (i) Negative transfer in Block 7 First stage
   (ii) No negative transfer in Block 7 Second stage
   (iii) Negative transfer in Block 8 Second stage
2. Option model simulations reproduce qualitative effects in behavioral data. No traditional flat RL can reproduce these transfer effects.

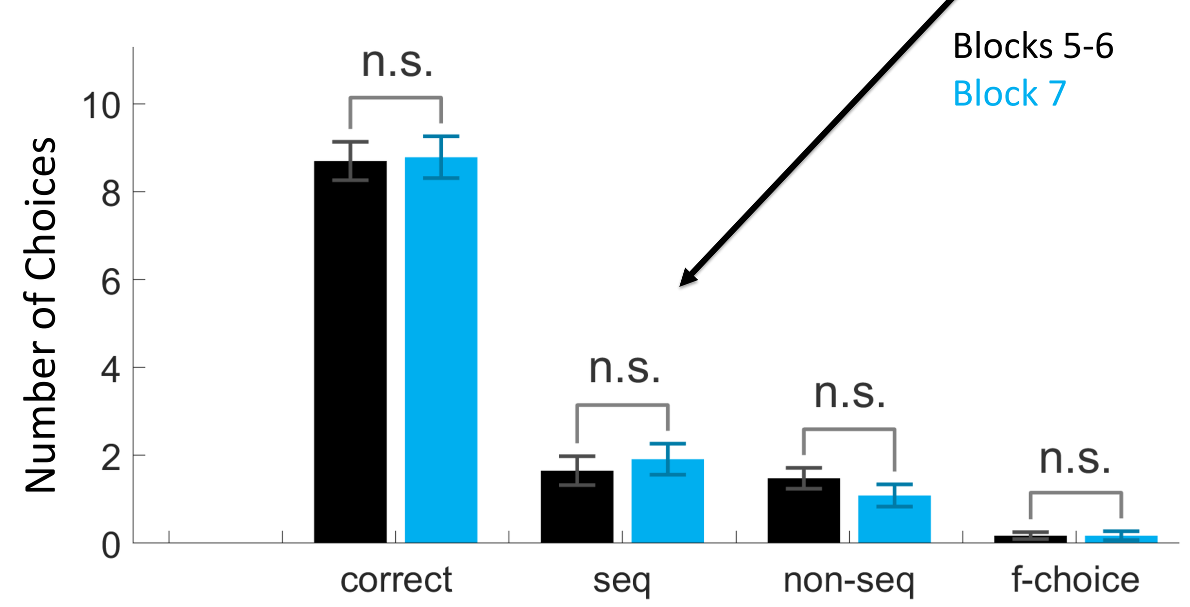
Positive transfer of middle-level options in Block 7



Consider Block 5 (◯, ◇) **:**

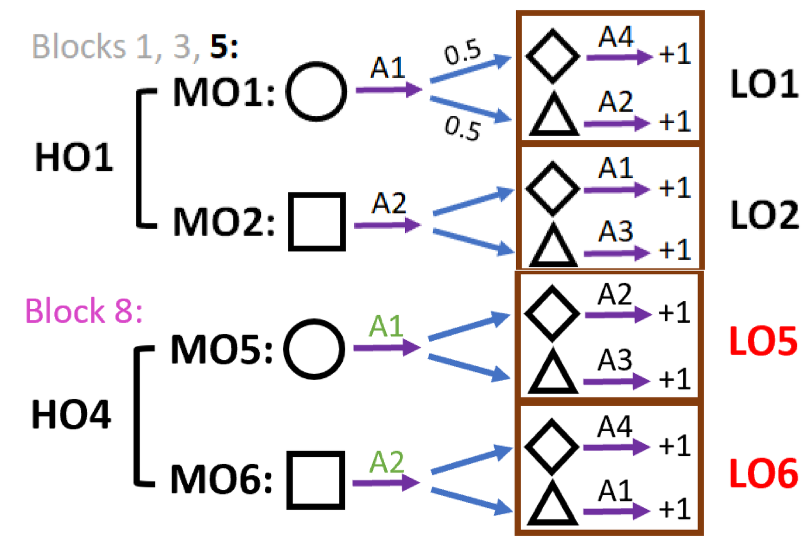
A4: Correct action (**correct**)
A1: Same as first stage choice (**f-choice**)
A2: Correct for the other second stage shape (**seq** for sequence learning)
A3: The other action (**non-seq**)

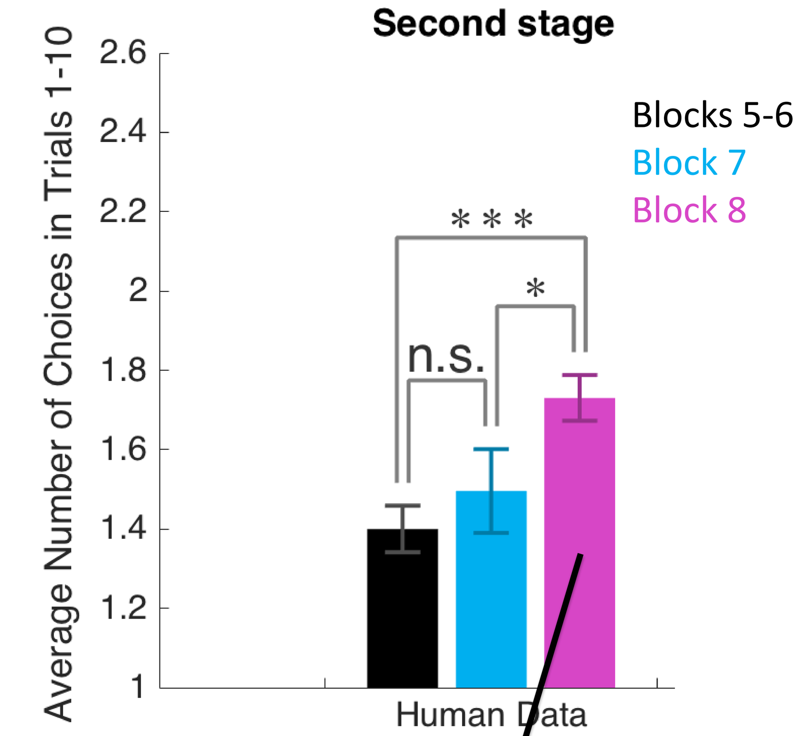




1. There is **no** significant difference between the second stage of Blocks 5-6 and Block 7 across all choice types, indicating that participants were able to flexibly transfer middle-level options as a whole even in the presence of interference from negative transfer in the first stage of Block 7.
2. We also compare the **RT** between **seq** and **non-seq** types and find no significant difference, indicating that the transfer effects cannot be explained by sequence learning alone.

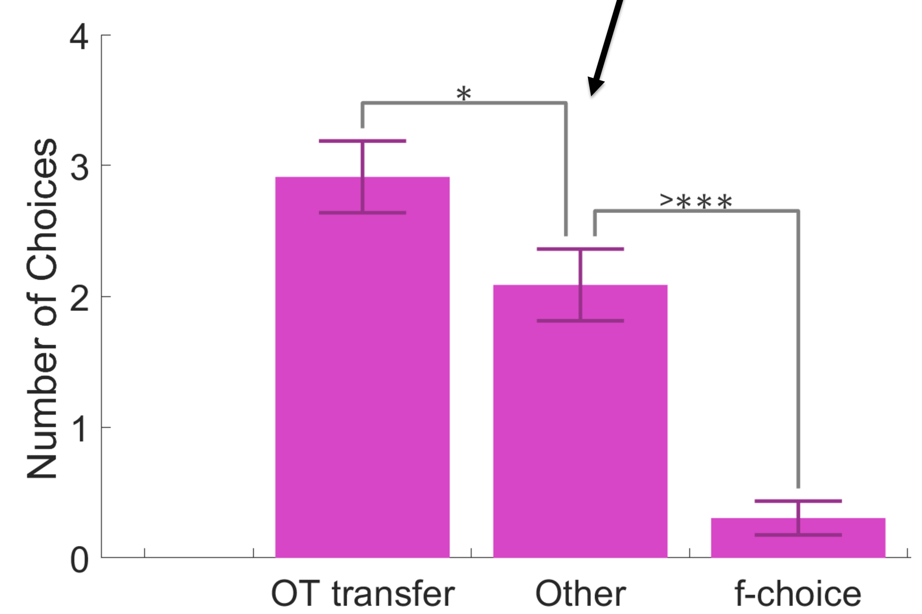
Negative transfer of middle-level options in Block 8





Consider Block 8 (◯, ◇) **:**

A2: Correct action
A1: Same as first stage choice (**f-choice**)
A4: Correct for the combination in Blocks 1, 3 and 5 (**OT transfer**)
A3: The other action (**other**)



The main source of negative transfer in the second stage of Block 8 comes from the OT transfer type.

Conclusions

Summary
- Humans learn temporally-extended policies called options, confirmed by both positive and negative transfer effects.
- Humans are able to flexibly transfer options at different levels.
- The Option Model captures transfers in human behavior qualitatively.
- Sequence learning alone cannot account for the transfer effects.

Future directions
- What is the neural underpinning of option learning? Is there any difference in the neural representation of 1-step policies and options?
- In novel contexts, do humans learn a new option, or rewrite an old one that is similar enough?

Bibliography

[1] Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences*, 12(5), 201-208.
[2] 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.
[3] 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.
[4] Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological review*, 120(1), 190.
[5] Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983* (pp. 1-198). Springer, Berlin, Heidelberg.

Comments, Questions: jimmyxia@berkeley.edu