Title
Mathematical Modeling of Human Learning and Decision Making

Author
Xia, Liyu

Publication Date
2021

Peer reviewed|Thesis/dissertation

Mathematical Modeling of Human Learning and Decision Making

by

Liyu Xia

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Applied Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jim Pitman, Co-chair
Assistant Professor Anne Collins, Co-chair
Professor Per-Olof Persson

Spring 2021

Mathematical Modeling of Human Learning and Decision Making

Abstract

Mathematical Modeling of Human Learning and Decision Making

by

Liyu Xia

Doctor of Philosophy in Applied Mathematics

University of California, Berkeley

Professor Jim Pitman, Co-chair

Assistant Professor Anne Collins, Co-chair

Reinforcement learning (RL) has been widely used to study and model human, animal and artificial intelligence. In this thesis, we focus on modeling human learning and decision making, and exemplify two ways mathematical RL modeling can add to our existing knowledge of human cognition: (1) provide a quantitative tool for parametrizing individual differences in human behavior, and (2) augment existing models to account for complex multi-step human learning.

Contents

Acknowledgments

I have been extremely fortunate to be able to pursue my PhD study in Applied Mathematics at UC Berkeley. I would like to thank Professor Jim Pitman, the co-chair of my dissertation committee from the Mathematics and Statistics Department, who was extremely supportive when I decided to take the initiative to go down an atypical route for most Mathematics PhD students — studying human learning and decision making. I would also like thank Professor Anne Collins, the co-chair of my dissertation committee from the Psychology Department and Helen Wills Neuroscience Institute, who offered constant guidance and mentorship to help me transition into the field of mathematical modeling of human cognition and make this thesis possible.

I am very thankful for Professor Per-Olof Persson, a member of my dissertation committee and also the chair of my qualifying committee. I would also like to thank Professor Friedrich Sommer and Professor Bruno Olshausen for being on my qualifying committee and motivating my early interests in computational neuroscience.

I would like to thank Professor Linda Wilbrecht for being a close senior coauthor. Also, this thesis would not be possible without the support of my amazing labmates, Beth Baribault, Maria Eckstein, Saran Master, Milena Rmus, and Amy Zou, and my diligent research assistants, Katya Brooun, Flora Dong, Kshitiz Gupta, Yi Liu, Sabrina Ni, and Wendy Shi.

Finally, this thesis is dedicated to my husband, Josue Melendez Rodriguez, who has been extremely supportive throughout my PhD; and my parents, Lingling Xiong and Yongqiang Xia, for all the years of unconditional love and support.

Chapter 1

Introduction

1.1 Reinforcement Learning

What constitutes a complete understanding of how the brain works? According to the well-known framework proposed by David Marr [108], the answer should encompass the understanding of the brain at three levels: computational, algorithmic, and implementational. The computational level tries to find a set of objective functions that the brain optimizes for; the algorithmic level studies the mathematical algorithms used by the brain to carry out the optimization of the objective functions; finally, the implementational level addresses how these mathematical algorithms can be physically carried out by the brain.

Reinforcement Learning (RL) has attracted much attention recently because it encompasses all three levels [125]. For the computational level, RL postulates that the main objective of the brain is to maximize cumulative future rewards [159]. At the algorithmic level, RL, with a wide range of variations, tries to learn about the values of different states and actions using local prediction error signals [159]. Finally, relating to the implementational level, there has been neuroscientific evidence of how the prediction error signals are carried by the neurotransmitter, dopamine, in the brain [144].

Therefore, RL appears to be a promising approach to explain brain function and understand human behavior, and has been widely applied to study various aspects of human cognition [123, 75, 102, 57, 39, 38]. Note that the computational and algorithmic principles of RL can be easily applied to any problem domain related to decision making and optimization. For example, RL has been successfully applied to healthcare [76, 137, 177, 103], education [107, 136], control [168, 72, 13], and so on. It has proven especially successful in artificial intelligence (AI) recently [147, 146], demonstrating the generality of the theoretical framework and principles of RL in both biological and artificial agents.

In this thesis, we focus on using RL and mathematical modeling for explaining human behavior in two different ways: one to quantify human cognition by extracting trial-by-trial human learning and decision making and thus make inferences and predictions about human choices in a simple one-step probabilistic task; and the other one to develop new mathemati-

cal models that augment existing RL models to explain transfer and generalization behavior in more complex multi-step human cognition. We also point out potential connections with and inspirations from research ideas in AI.

1.2 Markov Decision Process

To facilitate the discussion on human learning and decision making later on, we first introduce the mathematical formalism of Markov Decision Process (MDP), the general set of problems that RL is designed to solve. An MDP consists of a state space, denoted $\mathcal{S}$, which describes the set of states that the agent can occupy in an environment; an action space, denoted $\mathcal{A}$, which describes the set of actions that the agent can take; a transition function, denoted P, that maps state transitions triggered by the actions that the agent takes; a reward function, denoted R, that assigns reward for the actions that the agent takes.

Specifically, in this thesis, we only consider discrete-time, finite MDPs [158, 159]. Finite refers to the fact that the the state and action spaces are finite. Discrete-time means that we assume the agent is interacting with the environment at some discrete time scale, $t = 0, 1, 2, ...$ At each step t, the agent observes the environmental state, $s_t \in \mathcal{S}$, based on which it chooses an action $a_t \in \mathcal{A}$. The environment then shifts the agent to state s_{t+1} based on the transition function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, where $\sum_{s_{t+1}} P(s_t, a_t, s_{t+1}) = 1$ for any (s_t, a_t) pair. After the transition, the environment also delivers the reward $r_{t+1} = R(s_t, a_t)$. The goal of RL is to learn a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ that specifies a probabilistic distribution over action a_t at state s_t, in order to maximize the expected cumulative discounted future reward from each state $s \in \mathcal{S}$:

$$V^{\pi}(s) := \mathbb{E}[r_{t+1} + \tau r_{t+2} + \tau^2 r_{t+3} + \cdots | s_t = s, \pi], \tag{1.1}$$

where $\tau \in [0, 1)$ is a discount factor (e.g. $\tau = 0$ indicates a perfectly shortsighted agent than only cares about the next-step reward, r_{t+1}). $V^{\pi}(s)$ is often referred to as the state value function. The discount factor is usually below 1 because (1) reward delivered in the future is usually not as valuable as the present reward ($100 today is more valuable than $100 tomorrow), and (2) the mathematical derivations and computations of RL are more well-behaved since the infinite geometric sum in Eq 1.1 is convergent. In essence, the state value function models our expectation of how rewarding a certain state is in the long term, and thus the agent should choose actions that would take them to the most rewarding states.

In other words, RL tries to address Marr's computational level question by specifying the objective of finding an optimal policy $\pi^* = \max_{\pi} V^{\pi}(s)$ that optimizes for the state value function for all states $s \in \mathcal{S}$. However, there have been debates on how general this reward objective is in human and animal behavior, generally referred to as "the reward hypothesis" [159, 86]. While we certainly find money and food rewarding, arguments can be made that our behavior is sometimes motivated by things that are not immediately rewarding on the surface. One great example is curiosity, although one might also argue that curiosity-driven behavior helps us gather more information and reduce uncertainty in the environment, and

thus is rewarding. Another example is goal-directed behavior, where we are motivated to finish an abstract goal and it's unclear how rewarding each step along the way is. The latter example actually motivated the introduction of pseudo-reward functions for accomplishing sub-goals in hierarchical reinforcement learning (HRL), which will be discussed more in-depth later. While this thesis continues with this reward objective of RL, see [159, 125, 86, 150] for a more comprehensive discussion of the reward hypothesis and how general the reward objective of RL is.

1.3 Model-free Reinforcement Learning

With the computational level objective of RL explained, we move on to discussing the algorithmic aspect of RL. There are various RL algorithms to solve for the optimal policy and thus optimize the state value function and maximize cumulative discounted future reward. One common way of categorizing different RL algorithms is whether the algorithm learns a model of the environment, which includes the transition function P and the reward function R. The RL algorithms that explicitly try to approximate P and R are termed model-based, whereas the ones that directly optimize the value function or approximate the optimal policy are termed model-free. Since model-based RL learns more things, it tends to be more computationally intense, but is more flexible to environmental changes. On the other hand, model-free RL requires less computational resource, but is less adaptive to environmental changes, since it doesn't actively model the environment.

The interplay between model-free and model-based algorithms has proven to be very helpful not only in solving MDPs [157, 77, 147, 29, 134], but also in understanding human cognition [75, 92, 37]. It turns out that human participants employ both algorithms in learning and decision making [75], including algorithms that sit in the middle (successor representation, [115]). Model-based reasoning in humans tends to be described as the deliberative system that tries to make predictions about state transitions and reward, from which we plan our course of actions into the future. The model-free system is often associated with being more habitual and only learns about the values of the states, without explicit understanding of the environmental dynamics. For a more comprehensive overview of the dichotomy between model-free and model-based RL, see [37].

For the mathematical modeling used in this thesis, we primarily used model-free RL algorithms for simplicity. However, the mathematical methods we adopted can be and have been applied to studying the model-based/planning aspects of human learning ([75, 44, 117]).

Note that instead of learning state value functions, for problems of smaller size, we often prefer to learn action value functions, defined as below:

$$Q^{\pi}(s,a) := \mathbb{E}[r_{t+1} + \tau r_{t+2} + \tau^2 r_{t+3} + \cdots | s_t = s, a_t = a, \pi], \tag{1.2}$$

where $Q^{\pi}(s,a)$ is often referred as Q-values. The main reason is that even if we learned the state value function, and thus which states are most rewarding in the long term, without knowledge of the state transition function, P (restricting ourselves to be model-free), we still

do not know which actions to take in order to reach those states. On the other hand, once we have learned the action value function, we can survey the Q-values of all actions at the current state, and then pick the action that achieves the maximum.

One fundamental example of model-free RL algorithm is Q-learning [159], that tries to learn Q-values from local prediction errors. It bootstraps from the recursive nature of the Q-value function:

$$Q^{\pi}(s,a) = \mathbb{E}[r_{t+1} + \tau r_{t+2} + \tau^2 r_{t+3} + \cdots | s_t = s, a_t = a, \pi] \tag{1.3}$$

$$= r_{t+1} + \tau \sum_{s'} P(s,a,s') \max_{a'} Q^{\pi}(s',a') \tag{1.4}$$

Once the agent experiences a transition (s,a,s') by following policy π, it calculates the reward prediction error (RPE):$= r_{t+1} + \tau \max_{a'} Q^{\pi}(s',a') - Q(s,a)$, and updates our Q-value estimate accordingly:

$$Q^{\pi}(s,a) = Q^{\pi}(s,a) + \alpha * RPE, \tag{1.5}$$

where α is the learning rate parameter. We then use the new Q-value estimate to better our policy π, and iterate. Note that RPE reduces to $r_{t+1} - Q(s,a)$ if the environment/task is one-step.

Let's illustrate Q-learning with a simple one-step traffic example. Suppose we arrive in a new town, which we do not know the traffic rules for. We are at an intersection and need to decide whether we should cross the street. For this simple MDP, the state space is $\mathcal{S} = \{red, yellow, green\}$, representing the three colors of the traffic light. And the action space is $\mathcal{A} = \{cross, not\ cross\}$. Our goal is to learn the Q-values $Q(s,a)$ from trial-and-error interaction with the environment, i.e. crossing the streets several times. We assume all interactions terminate after crossing one intersection (i.e. we only consider one-step decision making and ignore state transitions).

We start out by initializing all Q-values to 0, assuming we do not have any prior knowledge of the traffic rule at town X, i.e. $Q(s,a) = 0,\ \forall s \in \mathcal{S}, a \in \mathcal{A}$. Suppose we encounter a red light, we query our current Q-values for the red light. It turns out that $Q(red, cross) = Q(red, not\ cross) = 0$, so we throw a fair coin. Suppose that the coin randomly decides that we should cross, but we almost get hit by a car while crossing, resulting in a negative reward of $R(red, cross) = -1$. Note that conceptually the Q-values, $Q(s,a)$, reflect our expectation of how much reward we will receive by taking action a in state s. At the beginning, we expect nothing in particular (i.e. $Q(red, cross) = 0$); however, the actual reward that the environment delivers is $r = -1$, resulting in a discrepancy, i.e. $RPE = r - Q(s,a)$. RPE can then by used as learning signal to update our current expectation, so that it better matches the actual reward we received:

$$Q(s,a) = Q(s,a) + \alpha * RPE, \tag{1.6}$$

where α is the learning rate parameter between 0 and 1, reflecting how much new information we integrate into the updated Q-values. For example, with $\alpha = 0$, we essentially disregard new reward information, whereas if $\alpha = 1$, we update the new $Q(red, cross)$ to $r = -1$ in one-shot. Note that, higher learning rate does not always translate into better learning performance in the long term, especially when the environment is stochastic and/or volatile [169, 12, 43].

Now that we have updated (most likely, decreased) $Q(red, cross)$ with RPE, next time when we run into another intersection with red light and query Q-values, we would have $Q(red, cross) < Q(red, not\ cross)$, and thus we would more likely decide not to cross. A common way of translating Q-values to actual policies is through softmax:

$$\pi(a) = \frac{exp(\beta Q(s, a))}{\sum_{i=1}^{2} exp(\beta Q(s, a_i))}, \tag{1.7}$$

where β is the inverse temperature parameter, and a_1, a_2 refer to cross and not cross respectively. β ranges from 0 to infinity and reflects how stochastic the choice is. For example, if $\beta = 0$, all actions have equal probability of being chosen, disregarding the learned Q-values. On the other hand, if $\beta = \infty$, then policy π deterministically chooses the action that maximizes $Q(s, \cdot)$ (e.g. in this case, we would have picked not cross for the red light with probability 1).

Eq 1.6 and 1.7 combined is often referred to Q-learning [159], where Eq 1.6 is the learning step (updating Q-value estimate), whereas Eq 1.7 is the decision step (updating policy). Q-learning provides a partial answer to Marr's algorithmic level, i.e. what algorithms do we actually use to optimize for the long term reward objective. Moving a step further, there has actually been neuroscientific evidence that RPE signal is carried by neurotransmitter dopamine in midbrain [144], thus addressing the implementational level of Marr's original proposal.

A unique advantage of using Q-learning to model human behavior, is that it provides a fully generative and mechanistic understanding of trial-by-trial human learning and decision making. It also acts as a potential bridge between human behavior that we can experimentally observe and the underlying neural processes that give rise to such behavior [124]. Later in Chapter 2 we will show an example use of Q-learning to understand the development of the RL system during adolescence.

While this thesis primarily focuses on using Q-learning to quantify human cognition, which indeed touches upon Marr's all 3 levels, we do not attempt to provide a comprehensive answer to all of human cognition. Aside from the aforementioned model-free vs. model-based dichotomy, RL is unlikely to be the only learning system in the brain. There have been recent evidence that other memory processes (such as working memory and episodic memory) also substantially contribute to learning, and even interact with model-free RL, see [39, 17] for examples.

1.4 Hierarchical Reinforcement Learning

The previous traffic example only illustrates simplistic one-step decision making scenarios. In real world problems that we face daily, one-step strategies clearly fall short. Moreover, basic RL algorithms, such as Q-learning, suffer from the so-called curse of dimensionality [159]. The main issue is that, as the state space and action space become increasingly more high dimensional and complex, sometimes it is unrealistic to fully explore every possible state and action, plus rewards (and thus RPE and learning signals) can be sparse and hard to come by (e.g. need to work hard for many years to get a PhD).

To address this, [158] proposed the hierarchical reinforcement learning (HRL) options framework. Options are temporally-extended multi-step policies assembled from simple actions or other options to achieve meaningful sub-goals. 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.

Formally [158], each option o contains three components: an initiation set $\mathcal{I} \subseteq \mathcal{S}$, a termination function $\xi : \mathcal{S} \to [0, 1]$, and an option-specific policy $\pi_o : \mathcal{S} \times \mathcal{A} \to [0, 1]$. The initiation set, a subset of the original state space, describes the set of states where the option can be initiated. The termination function maps each state to the probability of terminating the current option. The option-specific policy guides decision making once the option is initiated. For example, the initiation set for the option of making potatoes might be kitchen, and the option might terminate when the potatoes are cooked. Agents learn when to select options in the same way they learn to select actions (e.g. make coffee for breakfast, not for dinner) by using normal reinforcement signals. Agents learn the option-specific policies using pseudo-rewards obtained when reaching the sub-goal.

The options framework provides many theoretical benefits for learning [22, 20], assuming that useful options are available. Unlike traditional flat RL algorithms that only learn step-by-step policies, options help explore more efficiently and plan longer term. For example, when we learn how to cook a new kind of potato, we already know how to cut potatoes. Moreover, we can plan with high-level behavioral modules such as cutting potatoes, instead of planning in terms of reaching, grabbing, and peeling. If non-useful options are available, the options framework predicts that learning can be instead slowed down [22]. The question of how to identify and create useful options has been a topic of active and intense research in AI [111, 112, 149, 105, 104, 84, 64, 83, 119, 174].

Recent studies [52, 53, 138, 139, 143] show early evidence that the options framework could be a useful model of human learning and decision making. [52, 143] showed that participants were able to spontaneously identify bottleneck states from transition statistics, which aligned with graph-theoretic objectives for option discovery developed in AI [112]. In addition, in hierarchical decision-making tasks, [53, 138, 139] showed that human participants signaled RPE for both sub-goals and overall goals. These results indicate that humans are able to identify meaningful sub-goals, and to track sub-task progression, both key features of the options framework. [23, 80] have also suggested potential neural correlates for

implementing the computations required to use options.

However, the fundamental question of whether and how humans learn and use options during learning remains unanswered [52]: there is little work probing the learning dynamics in tasks with a temporal hierarchy, or directly testing the theoretical benefits of options in a behavioral setting. In Chapter 3, we aim to (1) characterize how humans learn representations that support hierarchical and compositional behavior, and (2) investigate whether an expanded options framework can account for it. In particular, do humans create options in such a way that they can flexibly reuse learned options in new problems? If so, how flexible is this transfer?

1.5 Modeling Human Behavior with RL

The key strength of RL comes not just from its broad theoretical framework and ability to achieve state-of-the-art performance in AI [147]. For the purpose of this thesis, our main focus is on its ability to account for a wide range of human behavior [123, 102, 39, 44, 52, 115, 55]. Therefore, in order to practically apply RL models to explain and predict human behavioral data, and better understand the underlying cognitive processes in learning and decision making, mathematical and statistical techniques for model fitting become crucial. For example, the Q-learning model explained in Sec 1.3 has two free parameters, namely α (Eq 1.6) and β (Eq 1.7). Assuming Q-learning is the right model for a behavioral dataset, model fitting aims to find the right configuration of α and β that maximizes the likelihood of observing the data collected, which is usually nonlinear and non-convex, and thus difficult to optimize.

The problem becomes even more statistically challenging when we want to account for behavior from a population of human participants jointly instead of for just one participant. To meet this challenge, hierarchical Bayesian modeling [73] and state-of-the-art sampling methods [27] have been developed. These methods provide better and more stable parameter estimates, and go beyond point estimates to provide uncertainty estimates, which allow more robust statistical inference [87]. In Chapter 2, for example, we demonstrate the effectiveness of these applied mathematical techniques at quantifying individual differences in learning during adolescence.

1.6 Contributions of This Thesis

In Chapter 2, we present an example study that used RL as a mathematical model to understand human learning and decision making under uncertainty within a one-step probabilistic task. We also used hierarchical Bayesian modeling coupled with sampling-based methods to fit RL model parameters, such as the learning rate parameter α, to a population of participants across a wide age range (8-30). These mathematical and statistical modeling techniques helped provide a quantitative and mechanistic understanding of human learning

under uncertainty by compressing human behavior into just a few RL model parameters. These model parameters then enabled us to parametrize individual differences in learning and understand the development of model-free RL system through adolescence.

In Chapter 3, we move beyond simple one-step decision making tasks and aim to capture multi-step learning using options (Sec 1.4). To account for this more complex human cognition and the rich transfer and generalization effects shown by human, we came up with a novel trial-by-trial mathematical model that combined Chinese Restaurant Process (a non-parametric Bayesian clustering algorithm, [133]) and the HRL options framework [158]. The options framework let the model learn temporally extended multi-step policies, while Chinese Restaurant Process allowed the model to reuse previously learned policies or creating new ones if necessary. We also designed a series of novel multi-step decision making tasks to test various empirical predictions of our new model on human participants. We demonstrated that this novel augmentation to the existing HRL option model helped explain hierarchical human behavior and more importantly, transfer and generalization of previously learned options at multiple levels of policy complexity.

Chapter 2

Mathematical Modeling of Learning Changes During Adolescence

In this chapter, we provide an example study that used RL as the mathematical model to obtain a quantitative and mechanistic understanding of human learning and decision making under uncertainty. In particular, we coupled hierarchical Bayesian modeling and sampling based methods [73] to extract trial-by-trial learning dynamics of human participants, make inferences and predictions of participants choices, and more importantly, trace the developmental trajectory of learning through adolescence. This chapter is adapted from a paper submitted to PLOS Computational Biology [172]. I would like to acknowledge my co-authors, Sarah and Maria, who coded the experimental task and spent much time collecting a large population of human participant data; Beth, who provided immense suggestions on modeling; and senior co-authors, Ron, Linda and Anne, who helped secure funding and provided guidance throughout. Permission to use this study as part of this thesis has been obtained from all co-authors.

2.1 Introduction

In the everyday world, perfectly predictable outcomes are rare. Yet, we need to track important events and their relationships to other events and actions. For example, we might want to learn where the best place to obtain food is, or where a potential mate likes to hang out – this might help us decide where to go, expecting a positive outcome to occur frequently, but not always. Our ability to learn about these probabilistic relationships is crucial for our daily life and decision making.

This challenge needs to be met by the developing brain, especially during adolescence [126, 47, 155, 24, 152, 164]. Naively, one might assume that the brain simply gets better at this (and possibly all) forms of learning with brain maturation. However, what does *better* mean in this context? Most learning mechanisms are subject to tradeoffs between speed and stability. Fast learning may be suitable for a highly certain environment with

deterministic relationships/statistics, but can lead to impulsive behavior in a more uncertain environment with probabilistic relationships/statistics [12, 43]. By contrast, slower learning that integrates over a longer time scale may lead to more robust and stable performance in probabilistic environments. During development, there may be periods where one form of learning is emphasized over the other. Changes could be gradual and monotonic, or show sharp steps when driven by factors such as hormonal changes at puberty onset [94]. There may also be inverted U shapes [109, 56, 24], that peak to support a sensitive period when specific information is available in the environment and/or when an organism needs to accomplish its transition to independence [41, 132, 68].

To study how learning changes across adolescence, we used the theoretical framework of RL. Mathematical RL models assume that we estimate the long-term values of an action in a given state by integrating over time the feedback we receive for choosing this action in this state, through a trial-and-error process (Sec 1.3, [159]). RL has greatly enhanced our understanding of human behavior and the neural processes that underlie learning and decision-making in both certain and uncertain environments [123, 75, 39, 102, 57]. Moreover, RL processes offer a quantitative parameterization of individual differences: for example, RL decision noise may capture exploratory choice [126]; RL learning rates control the time scale of integration of rewards, with potential asymmetries between positive and negative outcomes [28, 100]; and RL forgetting parameters may capture memory dependent processes [109].

For these reasons, RL has been previously used to probe developmental changes in learning and decision making, including during adolescence [126, 47, 43, 78, 109, 56, 128]. While there has been some consensus on certain developmental trends, such as lower decision noise with age [126], in general, developmental results in both how learning behavior changes and in how RL processes (and parameters) change are highly variable and dependent on the specific tasks used [126, 162].

To study how learning under uncertainty changes during adolescence, we used the *Butterfly task* [43], where participants needed to learn probabilistic associations that were stable throughout the task. We collected data from a population of 297 participants across a wide age range (8-30), over-sampling participants age 8-18 to focus on the adolescent period (see Fig 2.2 for detailed breakdown of age group by sex). In fact, in this same population, we conducted a total of four tasks that varied across multiple dimensions (such as deterministic/probabilistic feedback, stable/volatile contingencies, memory load, etc.), with the initial motivation to address the issue introduced by task heterogeneity [126]. However, the focus of this study is on the Butterfly task alone. While we mainly present results from the Butterfly task, we also discuss comparisons and relationships with two other tasks in this sequence of four tasks [109, 56].

The Butterfly task tests participants' ability to learn probabilistic associations between four butterflies and two possible preferred flowers from reward feedback. This task has been used in developmental studies before [43], and produced an intriguing result showing that adolescent performance was greater than adults in a two group design ($N = 41$ adolescents age 13-17 and $N = 31$ adults age 20-30). We sought to further investigate performance in this task during development with a larger sample that would enable evaluation of the trajectory

of development from age 8-30 and examine the role of puberty in changes in performance.

To evaluate the potential role of gonadal hormones and pubertal development in driving changes in learning, we also measured pubertal development and saliva testosterone (see Sec 2.2.1, Sec 2.2.4). We expected to observe an inverted U shape in performance that peaked in mid adolescence [142], coinciding with previously observed peaks in nucleus accumbens activation in response to rewarding outcomes [71, 24]. Previous studies have also found positive relationships between adolescent testosterone levels and nucleus accumbens activation [24, 106, 153]; therefore, we expected that pubertal development might explain the timing of any observed peak. A further motivation to conduct this study was to investigate the possibility that participants of different ages were differently sensitive to positive and negative outcomes, something that has been observed in other studies [18], but was not investigated previously using the Butterfly task [43].

Contrary to our predictions, we found no evidence for adolescent performance advantage in our version of the Butterfly task. Instead, we found performance increased through early adulthood, then stabilized. We used hierarchical Bayesian methods to fit mathematical RL models to the trial-by-trial data (see Sec 2.2.7) and examined how participants integrated information across trials and made decisions. Increases in performance with age were explained by an increase in learning from rewarded outcomes and a decrease in exploration. These findings are largely consistent with studies of learning and decision making in other tasks that show steady improvement in performance across adolescent development [43, 109, 126]. We compare and contrast with findings that show adolescents outperforming adults [43, 56] to shed light on the conditions when adolescents vs. adults may show performance advantages in learning.

2.2 Methods

2.2.1 Participants

All procedures were approved by the Committee for the Protection of Human Subjects (CPHS number, community participants: 2016-06-8925; student participants: 2016-01-8280) at the University of California, Berkeley (UCB). A total of 297 (151 female) participants completed the task: 187 children and adolescents (age 8-18) from the community, 55 UCB undergraduate students (age 18-25), and 55 adults (age 25-30) from the community. Participants under 18 years old and their guardians provided their informed assent or written permission; participants over 18 provided informed written consent themselves.

We assessed pubertal development for children and adolescents through saliva samples and the pubertal development questionnaire, from which we calculated testosterone levels (T1, see Sec 2.2.4) and Puberty Development Score (PDS, [131]) respectively.

Community participants were compensated with a $25 Amazon gift card for completing the experimental session, and an additional $25 for completing optional take-home saliva

samples; undergraduate participants received course credit for participation. All participants were pre-screened for the absence of present or past psychological and neurological disorders.

2.2.2 Experimental design

To probe how learning under uncertainty changes during adolescence, we used the Butterfly task [43, 62, 61, 74], which was the third in a sequence of four tasks that participants completed in the experimental session [109]. The Butterfly task has previously shown great effectiveness in studying learning and decision making under uncertainty and individual differences [43, 62, 61, 74]. The task was a contextual two-armed bandit task with binary feedback: there were four stimuli (blue, purple, red, and yellow butterflies) and two bandits (pink and white flowers). Participants needed to figure out the preferred flower for each of the four butterflies through trial and error. Each butterfly had a preferred flower, which remained fixed throughout the experiment.

On each trial (Fig 2.1A), participants were presented one butterfly and two flowers. They needed to choose a flower within 7s using a video game controller. They were instructed to respond as quickly as possible. The chosen flower would stay on the screen for 1s. If participants correctly chose the preferred flower, they would receive positive feedback (*Win!*) with 80% chance; however, 20% of the time the other flower would be the rewarding one, resulting in negative feedback (*Lose!*). The feedback stayed on the screen for 2s. If participants received the 'Win!' feedback, they received 1 point, whereas 'Lose!' meant 0 points. Participants were instructed to figure out the preferred of each butterfly and to obtain as many points as possible (they could always see the total number of points earned so far on the upper right corner of the screen). The total amount of points won was not translated to a real-life reward such as money. There were 30 trials for each butterfly, resulting in a total of 120 trials. The butterfly-flower mapping, position of flowers, sequence of butterflies and the probabilistic feedback were pre-randomized and counterbalanced across participants.

2.2.3 Exclusion criteria

We collected data from 297 participants on the Butterfly task, with one participant excluded for only having 18 trials of data, resulting in 296 participants. We also excluded participants who were overall more likely to change their choice of flower for a given butterfly than repeat it after receiving positive feedback, which suggested that they either did not understand the task or were not engaged in it. 20 participants under 18 and one undergraduate participant were excluded due to this criterion. Note that all behavioral results presented later in Sec 2.3.1, Sec 2.3.2, and Sec 2.3.3 hold with the original population of 296 participants.

So far 275 participants remained. To further identify participants who were not engaged in the task without excluding participants solely on a pure performance criterion, we instead implemented the following, less stringent, *conjunctive* exclusion criteria:

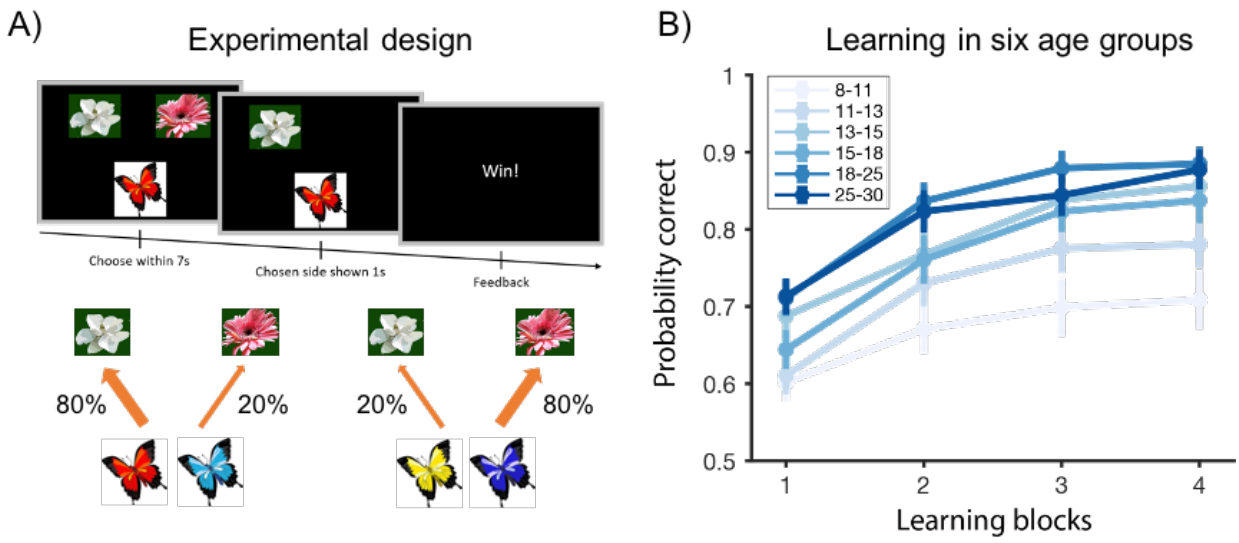

Figure 2.1: **Experimental design and overall performance.** (A) On each trial, participants needed to choose the butterfly's preferred flower. Each butterfly's preferred flower stayed the same throughout the experiment. If participants correctly picked the butterfly's preferred flower, they observed a *Win!* feedback with 80% chance, and *Lose!* otherwise. The other choice delivered positive feedback only with 20% chance. (B) Average probability of a correct choice over four 30-trial learning blocks. Learning curves showed all age groups learned the task, and that performance generally improved with age group.

- Criterion 1. Proportion of stay trials (the participant picked the same flower as the previous trial regardless of the butterfly) was higher than $median + 2 * sd$, across the entire population.
- Criterion 2. Proportion of stay trials was lower than $median - 2 * sd$, across the entire population.
- Criterion 3. The number of contiguous stay trials was higher than 12 in a row.
- Criterion 4. Any number of missing trials: available data less than the full 120 trials, indicating that participants stopped before the end of the experiment.
- Criterion 5. Performance was not better than chance (50%): performance was based on proportion of trials where participants correctly chose the preferred flower of the butterfly.

We excluded participants who fit **both** Criterion 5 and one of Criteria 1-4. Criteria 1-3 revealed participants who were either choosing the same action (e.g. always choosing left) or constantly switching between the two actions without a relationship to the butterfly on the

screen or the task at all. Together, Criteria 1-4 were likely to include participants who either did not understand the instructions or were not engaged for a significant proportion of the task. Finally, we took the intersection with Criterion 5 (at or below chance performance), so that we only excluded participants whose lack of understanding or disengagement (one of Criteria 1-4) significantly impacted their performance to be at or below chance (Criterion 5). Note that there were many participants who had worse than chance performance (Criterion 5), but were not excluded because they did not meet any of Criteria 1-4, which were indicators of not paying attention/misunderstanding the task. Therefore, this final criterion was less stringent than a pure performance criterion. Table 2.1 shows the detail breakdown of participants' age group for each criterion.

Age group	8 - 13	13 - 18	18 - 25	25 - 30
Criterion 1	3	2	1	0
Criterion 2	4	0	1	0
Criterion 3	6	4	2	4
Criterion 4	6	1	0	1
Criterion 5	14	5	2	2
Criterion 1-4	14	5	3	5
Criterion 5 and 1-4	8	2	0	1

Table 2.1: **Exclusion criteria breakdown by age.** Number of participants excluded due to each exclusion criterion for each of the four age groups.

After applying this conjunctive criterion, we further eliminated 11 participants for later analysis. In the end, we analyzed $N = 264$ participants, with 157 participants younger than 18 (Fig 2.2). All results in Sec 2.3 remain qualitatively similar with weaker exclusion criteria ($N = 275$).

2.2.4 Saliva collection and testosterone testing

In addition to self-report measures of pubertal development, we also collected saliva from each of our participants under 18 to quantify salivary testosterone, following methods reported in [109]. Testosterone is a reliable measure of pubertal status in boys and girls and is associated with changes in brain and cognition in adolescence [79, 130]. Participants refrained from eating, drinking, or chewing anything at least an hour before saliva collection. Participants were asked to rinse their mouth out with water approximately 15 minutes into the session. At least an hour into the testing session, they were asked to provide 1.8 mL of saliva through a plastic straw into a 2 mL tube. Participants were instructed to limit air bubbles in the sample by passively drooling into the tube, not spitting. Participants were allotted 15 minutes to provide the sample. After the participants provided 1.8 mL of saliva, or 15 minutes had passed, the sample was immediately stored in a -20°F freezer. The date and time were

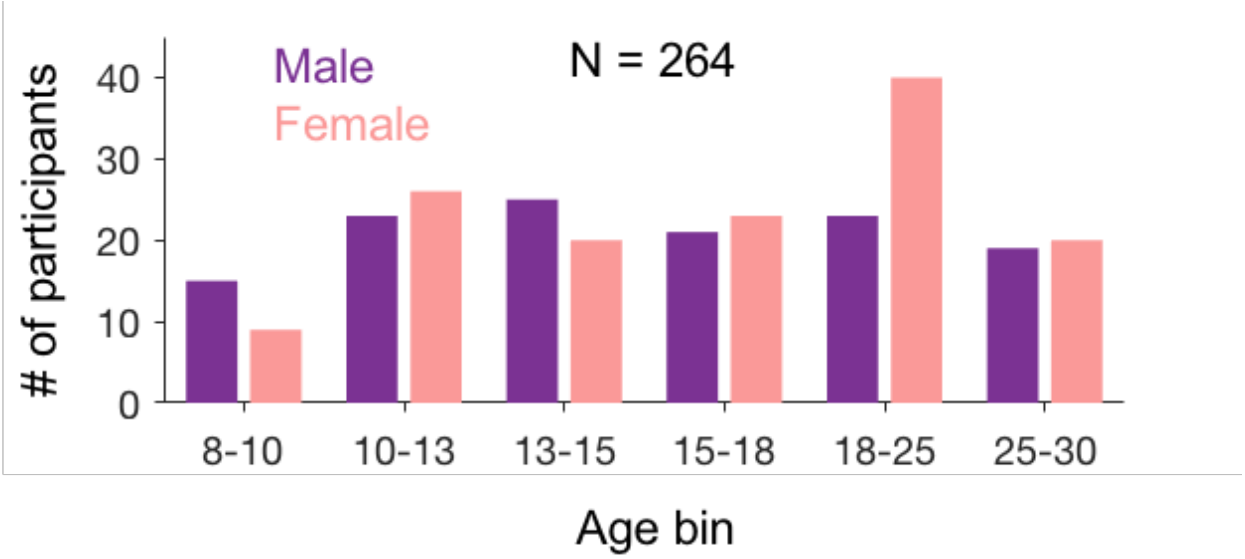

Figure 2.2: **Demographics of participants by age and sex.** We analyzed $N = 264$ participants across a wide age range (8-30), over-sampling participants under age 18 to focus on the adolescence period.

noted by the experimenter. The participants then filled out a questionnaire of information which might affect the hormone concentrations measured in the sample (i.e. whether the participants had recently exercised).

Salivary testosterone was quantified using the Salimetrics Salivatory Testosterone ELISA (cat. no. 1-2402, Bethesda, MA). Intra- and inter- assay variability for testosterone were 3.9% and 3.4%, respectively. Samples below the detectable range of the assay were assigned a value of 5 pg/mL, 1 pg below the lowest detectable value. Final testosterone sample concentration data were cleaned with a method developed in [9]. There were no participants with any samples above the detectable range, i.e. the measured values of the testosterone concentration for all participants were within the range that we validated as detectable by our technique. Within participants aged 8 to 18 only, outliers greater than three standard deviations above the group mean were fixed to that value, then incremented in values of +0.01 to retain the ordinality of the outliers.

We also visualized the relationships between pubertal measures and age for our population of participants. The correlation between pubertal measures and age was very strong as expected (Fig 2.3).

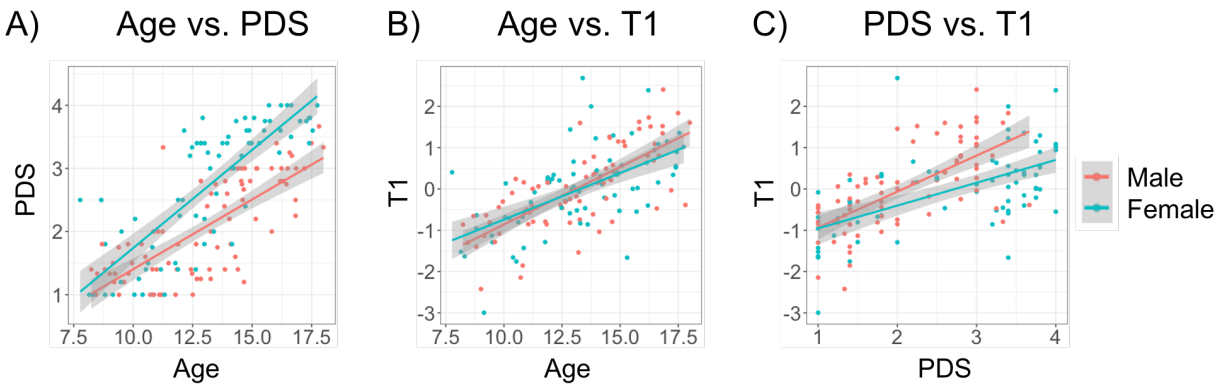

Figure 2.3: **Relationship between pubertal measures (PDS and T1) and age.** Scatter plot of (A) age vs. PDS, (B) age vs. T1, and (C) PDS vs. T1. Color indicates sex.

2.2.5 Model independent analysis

For each participant in each trial, we recorded whether they chose the butterfly's preferred flower (*correct* choice) or not, and whether they received reward or not (win vs. lose), which were different due to the probabilistic nature of the task. As an aggregate measure of performance, we computed average accuracy within each of the four 30-trial learning blocks for each participant. We also computed median and standard deviation of reaction time within each learning block. We ran (linear and quadratic) regression to assess whether those behavioral metrics changed with age and pubertal measures.

We also calculated the proportion of trials (p) among all 120 trials where participants correctly chose each butterfly's preferred flower as an overall performance measure. Because this proportion was not normally distributed across participants (Kolmogorov–Smirnov test, $p = 0.003$), we instead used log odds ($\log \frac{p}{1-p}$) for all later statistical tests. The log odds were normally distributed (Kolmogorov–Smirnov test, $p = 0.26$).

We used the median reaction time for each participant as a speed measure. Because reaction time was not normally distributed across participants (Kolmogorov–Smirnov test, $p = 0.02$), for all later statistical tests, we used log-transformed reaction time, which was normally distributed (Kolmogorov–Smirnov test, $p = 0.8$).

To visualize age effects (Fig 2.1, 2.4), we broke participants under age 18 into four equal-sized groups within each sex respectively, and then combined both sexes. The boundaries for the four age groups were approximately 8-11, 11-13, 13-15, and 15-18 (the exact boundaries for each age group and sex can be found in Table 2.2). Together with two age groups above 18 (18-25 for students and 25-30 for community participants), we had a total of six age groups.

We also binned pubertal measures to control for pubertal development in certain analysis. For PDS, we considered all participants with PDS score of 1 as one group, and broke down

Age group	First quartile	Second quartile	Third quartile	Fourth quartile
Male	8.2 - 10.9	11.0 - 13.3	13.3 - 14.7	14.9 - 18.0
Female	7.8 - 11.2	11.3 - 13.1	13.2 - 15.0	15.1 - 17.7

Table 2.2: **Age boundaries for each of the four age groups under 18.**

the rest into three equal-sized bins within each sex and then combined. For T1, we first log-transformed raw testosterone levels, and then broke down participants into four equal-sized bins within each sex and then combined the bins across sex.

Going beyond aggregate measures across trials, we also ran a mixed effect logistic regression to predict participants' choices on a trial-by-trial basis and tested how previous reward history and delay affected learning and decision making. Specifically, for each trial, we defined the *reward history*, h, as the number of trials that participants had previously received a *Win!* feedback for the current butterfly; we also defined *delay*, d, as the number of intervening trials since the last time the participant encountered the same butterfly and got rewarded. We then used the lme4 package in R to test:

$$p(correct) = logit(1 + h + d + (1 + h + d|sub)), \tag{2.1}$$

where *sub* represented random effects of individual participants respectively. All regressors were z-scored. We analyzed whether the random effects varied with age using linear and quadratic regressions.

2.2.6 Mathematical models

We used mathematical RL modeling to obtain a more quantitative and mechanistic understanding of participants' trial-by-trial learning. We applied six variants of RL models, then used the parameter estimates of the best fitting model as the basis for inference.

2.2.6.1 Classic RL ($\alpha\beta$)

Our simplest RL model was the $\alpha\beta$ model, with just two free parameters, α (learning rate) and β (inverse temperature/decision noise). The $\alpha\beta$ model uses Q-learning to compute $Q(s, a)$, as the expected value of choosing flower a for butterfly s. On trial t, the probability of choosing a is computed by transforming the Q-value with a softmax function:

$$P(a|b) = \frac{exp(\beta Q_t(s, a))}{\sum_{i=1}^{2} exp(\beta Q_t(s, a_i))}, \tag{2.2}$$

where $Q_t(s, a)$ is the Q-value until trial t. The inverse temperature parameter β thus controls how exploratory/stochastic the decision making process is, with higher β resulting

in more deterministic choices. After observing reward r_t (0 for "Lose!" or 1 for "Win!"), the Q-value $Q(s, a)$ is updated through the classic delta rule:

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha RPE, \tag{2.3}$$

where $RPE = r_t - Q_t(s, a)$ is the reward prediction error. Note that this delta rule can also be rewritten as:

$$Q_{t+1}(s, a) = (1 - \alpha)Q_t(s, a) + \alpha r_t, \tag{2.4}$$

which shows the updated Q-value (Q_{t+1}) as a linear combination of past estimates (Q_t) and the most recent reward (r_t). Thus, the learning rate parameter α is often interpreted as a time integration constant, controlling how much of the past estimate contributes to the current estimate. For example, $\alpha = 1$ would result in one-shot learning, i.e. set Q-value to be identical to the reward feedback each trial, resulting in an integration time scale of one trial (and no information about any other past trials). Smaller α results in integrating reward information across more trials from present into the past.

We initialized all Q-values to the uninformative value of 0.5 (the average of positive and negative feedback) for this model and all other models under consideration.

2.2.6.2 RL with asymmetric learning rates ($\alpha^+\alpha^-\beta$)

The $\alpha^+\alpha^-\beta$ model differed from the $\alpha\beta$ model by using two distinct learning rate parameters, α^+ and α^-. Recent literature suggests that humans learn from positive and negative feedback to different degrees, and even with potentially different neural mechanisms [100]. RL models with asymmetric learning rates have also been widely used and examined in theoretical [28, 88] and developmental [18, 30, 85, 78, 118] contexts, especially in studies with probabilistic tasks.

Having both α^+ and α^- allowed the model to have different sensitivity to positive and negative RPE [67]. Specifically, in Eq 2.3, α^+ was used when $RPE > 0$, and α^- otherwise.

2.2.6.3 Asymmetric RL with $\alpha^- = 0$ ($\alpha^+0\beta$)

The $\alpha^+0\beta$ model was the same as the $\alpha^+\alpha^-\beta$ model, except that the α^- parameter was set to 0. This change made the model insensitive to negative feedback. We included this model because of the observation that the fitted values of the α^- parameter from the $\alpha^+\alpha^-\beta$ model were very small and not recoverable (see Sec 2.3.4.1).

2.2.6.4 RL with forgetting ($\alpha\beta f$)

The $\alpha\beta f$ model builds upon the $\alpha\beta$ model by including an additional forgetting parameter, f. On each trial, after applying the delta learning rule in Eq 2.3, Q-values decay toward the uninformative starting value of 0.5, implementing a forgetting process:

$$Q_{t+1}(s,a) = (1-f) * Q_{t+1}(s,a) + f * 0.5. \tag{2.5}$$

Eq 2.5 is implemented for all butterfly-flower pairs except the butterfly and the selected flower on the current trial.

2.2.6.5 Asymmetric RL with $\alpha^- = 0$ and forgetting ($\alpha^+ 0 \beta f$)

The $\alpha^+ 0 \beta f$ model builds upon the $\alpha^+ 0 \beta$ and $\alpha \beta f$ models by including the forgetting parameter f and setting $\alpha^- = 0$.

2.2.6.6 RL with asymmetric learning rates and forgetting ($\alpha^+ \alpha^- \beta f$)

For full factorial design, we also included the $\alpha^+ \alpha^- \beta f$ model, which has both asymmetric learning rates for positive and negative feedback and the forgetting parameter.

2.2.7 Hierarchical Bayesian model fitting

We fitted all RL models using hierarchical Bayesian methods [73] jointly to all participants, instead of to each participant independently. The hierarchical model made the likelihood intractable, but it can be well approximated by sampling. We used no-U-Turn sampling, a state-of-the-art Markov Chain Monte Carlo (MCMC) sampler, implemented in the probabilistic programming language STAN [27], to sample from the joint posterior distribution of model parameters for all participants.

Compared to the classic participant-wise maximum likelihood estimation approach, hierarchical model fitting with MCMC provides better point estimates for individual participants and allows natural inference of effects on parameters at the group level [87]. Moreover, the empirical distribution of the samples approximates the true posterior, which provides a measure of uncertainty, beyond point estimates of individual model parameters. This allows more robust statistical inference.

To illustrate the model fitting procedure (Fig 2.4A), we use the simplest model, $\alpha\beta$, as an example. We assumed that parameters of individual participants came from the same group level distribution. We further assumed that the group level distribution is truncated normal, parametrized by a group-level mean and standard deviation (μ_α and σ_α for learning rate α; μ_β and σ_β for inverse temperature β). The group-level mean and standard deviation followed weakly informative priors (uniform and bounded). The parameters for each participant were sampled from the group level distributions: for example, $\alpha[j]$ for participant j was sampled from the truncated normal distribution $Normal(\mu_\alpha, \sigma_\alpha), T[0,1]$, since we know the learning rate parameter is restricted between 0 and 1. The individual participants' parameters were then used to calculate the likelihood of each participant's actions on each trial ($a[j][t]$) according to the $\alpha\beta$ model, where j and t indicate participant number and trial number, respectively.

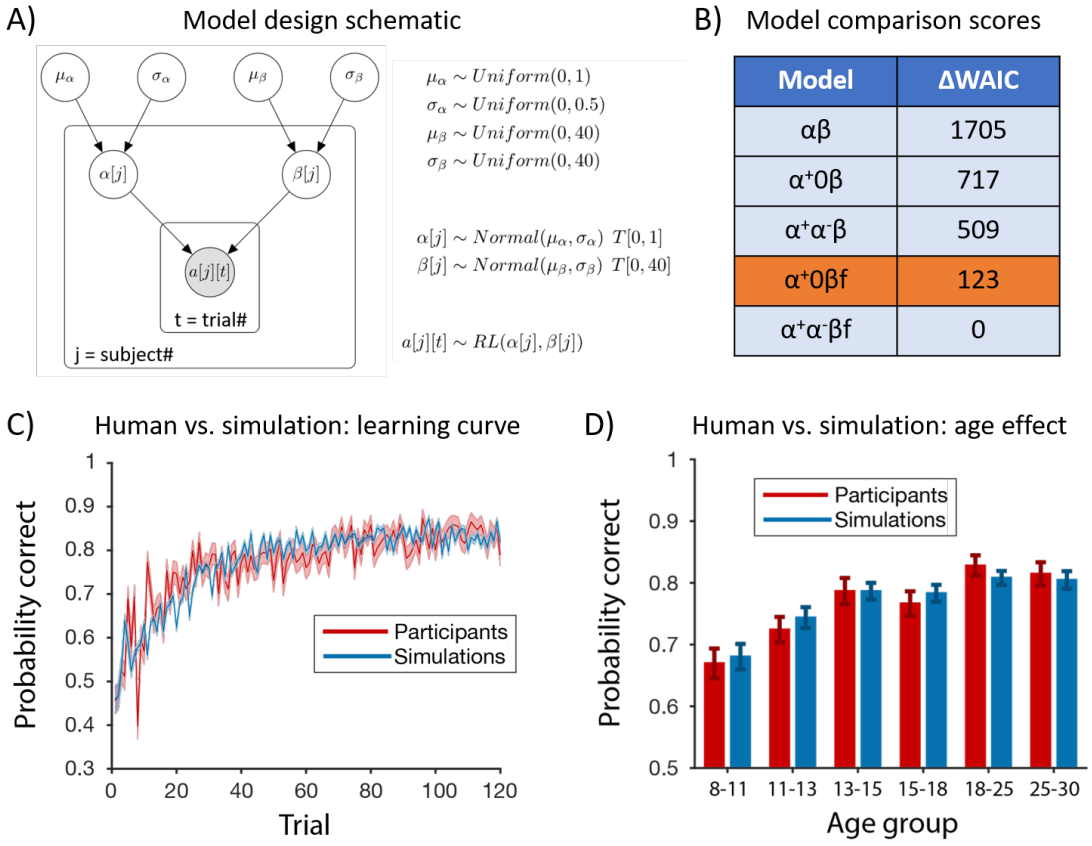

Figure 2.4: **Hierarchical Bayesian modeling and model comparison.** (A) We illustrate with the $\alpha\beta$ model as an example, We assumed that parameters of individual participants came from the same group level distribution, which is truncated normal parametrized by the group-level mean and standard deviation (μ_α and σ_α for α; μ_β and σ_β for β). The group-level mean and standard deviation followed weakly informative priors (uniform and bounded). With parameters for each individual participant, we calculated the likelihood of each action on every trial based on the $\alpha\beta$ model. $T[m, n]$ indicates truncation of distribution (e.g. the learning rate, α, is bounded by [0, 1]). The filled circle represented observed variable (in this case, participants' choices on each trial); unfilled circles represented latent variables (in this case, group and individual model parameters). (B) We calculated WAIC for model comparison. While the $\alpha^+\alpha^-\beta f$ model had the lowest (i.e. best) WAIC score, the α^- parameter was not recoverable. We thus focused on the $\alpha^+0\beta f$ model. (C-D) We used fitted parameters from the $\alpha^+0\beta f$ model to generate simulated trajectories. The simulated performance captured the average learning curve throughout the experiment (C), and replicated the age effect (D).

For each model, we ran 4 MCMC chains, with each chain generating 4000 samples (after 1000 warmup samples), resulting in 16000 samples per model for later inference. We assessed convergence for all models using the *matstanlib* library [11]. In particular, we ensured that $\hat{R}$ statistics for all free parameters were below 1.05; that the effective sample sizes (ESS) for all free parameters were more than 25 $\times$ the number of chains; and that samples generally were not the result of divergent transitions. Note that these criteria are more stringent than the standard criteria for convergence of $\hat{R} \leq 1.1$ and ESS $> 5\times$ number of chains as per [73]. Among all 6 models, only the $\alpha\beta f$ model was unable to converge when fitted hierarchically, thus we fitted the $\alpha\beta f$ model independently for each participant. The results presented later for the $\alpha\beta f$ model all came from non-hierarchical fitting. We also fitted the other 5 models non-hierarchically in order to compare with the $\alpha\beta f$ model.

Hierarchical Bayesian modeling also provides a natural way to test for potential age effects on model parameters. Specifically, we incorporated the regression model using age to predict model parameters into the original graphical model (Fig 2.4A), and directly sampled regression coefficients for age jointly with other model parameters.

Same as before, we assumed that the model parameters for individual participants followed a truncated normal distribution with a group-level standard deviation, but now we replace the prior on the group-level mean with a regression statement with respect to age. For example, to probe linear effects of age on α, we assumed that the parameter $\alpha[j]$, used to compute the likelihood of participant j's choices, followed $Normal(\alpha_{intercept} + \alpha_{linear} * age[j], \sigma_\alpha), T[0, 1]$, where $age[j]$ was the z-scored age of participant j, and $\alpha_{intercept}, \alpha_{linear}$ were regression coefficients for which we set weakly informative priors. To probe quadratic effects, we just further included $\alpha_{quadratic} * age[j]^2$.

To test for effects of age on the model parameters, we examined whether the posterior distribution of all 16000 samples for the linear (α_{linear}) and quadratic ($\alpha_{quadratic}$) regression coefficients were significantly different from 0.

2.2.8 Parameter and model identifiability and validation

We verified that model parameters and models themselves were identifiable using generate and recover procedures (see Sec 2.3.4.1, [169]). We validated models by simulating models with fitted parameters 100 times per participants, and comparing model simulations with behavior.

2.3 Results

2.3.1 Overall performance

To assess learning progress and potential age effects, we first calculated the proportion of correct trials within each of the four 30-trial learning blocks for six age groups (Fig 2.1B). As indicated in Sec 2.2.5, we grouped all participants under 18 into four equal-sized bins

($N = 39, 39, 39, 40$). The other two groups were undergraduate participants (age 18-25, $N = 53$) and adult community participants (age 25-30, $N = 54$).

All age groups exhibited learning over the course of the experiment. Specifically, we found a significant main effect of age group and block on participants' performance (two-way mixed-effects ANOVA, age group: $F(5, 255) = 8.5, p < 0.0001$; block: $F(3, 765) = 136, p < 0.0001$). There was no interaction between age group and block (two-way mixed-effects ANOVA: $F(15, 765) = 1, p = 0.49$). This showed that participant's performance improved as the experiment progressed, and older participants generally outperformed younger participants.

To further characterize the effect of age on overall performance, we computed the proportion of correct trials over all 120 trials. We found that the overall performance of 13-18 year-olds (top two quartiles) was significantly higher than that of 8-13 year-olds (bottom two quartiles; unpaired t-test, $t(1, 155) = 3.5, p = 0.0001$), and significantly lower than that of 18-25 year-olds (unpaired t-test, $t(1, 130) = 2.5, p = 0.01$). However, there was no significant difference (unpaired t-test, $t(1, 105) = 0.2, p = 0.8$) between the performance of 25-30 year-olds and 18-25 year-olds (Fig 2.1B, Fig 2.4D).

To examine the continuous relationship between participants' performance and age, we ran a regression analysis using age to predict performance (Fig 2.5A). We found that including a quadratic term was justified in addition to the linear term (sequential ANOVA: $F(1, 261) = 7.6, p = 0.006$). The regression analysis revealed linear and quadratic effects of age on performance (linear: $\beta_{age} = 0.05$, 95% CI $= [0.03, 0.07]$; quadratic: $\beta_{age^2} = -0.004$, 95% CI $= [-0.007, -0.001]$). There was no effect of sex or its interaction with age (multiple linear regression, both p's > 0.45).

To identify whether the quadratic effect was indicative of an inverse U shape, we conducted the two-line regression [148]. We found the break point at around 19 years old, before which $\beta_{age} = 0.11, z = 5.24, p < 0.0001$, and after which $\beta_{age} = 0, z = 0.14, p = 0.9$. This indicates that performance linearly increased with age through adolescence and saturated in early adulthood, showing that the quadratic effect reflected a saturating process rather than an inverse U shape.

2.3.2 Reaction time

We also computed the median (Fig 2.5B) and standard deviation of reaction time for each participant. We found a linear effect of age on median reaction time ($\beta_{age} = -0.01$, 95% CI $= [-0.02, -0.006]$). This suggests that participants reacted faster with age, confirming previous results [109]. Adding a quadratic term was not justified (sequential ANOVA: $F(1, 261) = 1, p = 0.3$). We also found a linear effect of age on the standard deviation of reaction time (linear regression: $\beta_{age} = -0.02$, 95% CI $= [-0.03, -0.01]$); adding a quadratic term was justified (sequential ANOVA: $F(1, 261) = 17, p < 0.0001$; quadratic regression: $\beta_{age^2} = 0.003$, 95% CI $= [0.002, 0.005]$). Two-line regression revealed a break point at 19 years old, before which we found $\beta_{age} = -0.05, z = -4.15, p < 0.0001$, and after which we found $\beta_{age} = 0.03, z = 1.88, p = 0.06$. This indicated that the variability in reaction time decreased with age, and this decrease saturated in early adulthood and might even invert, consistent with

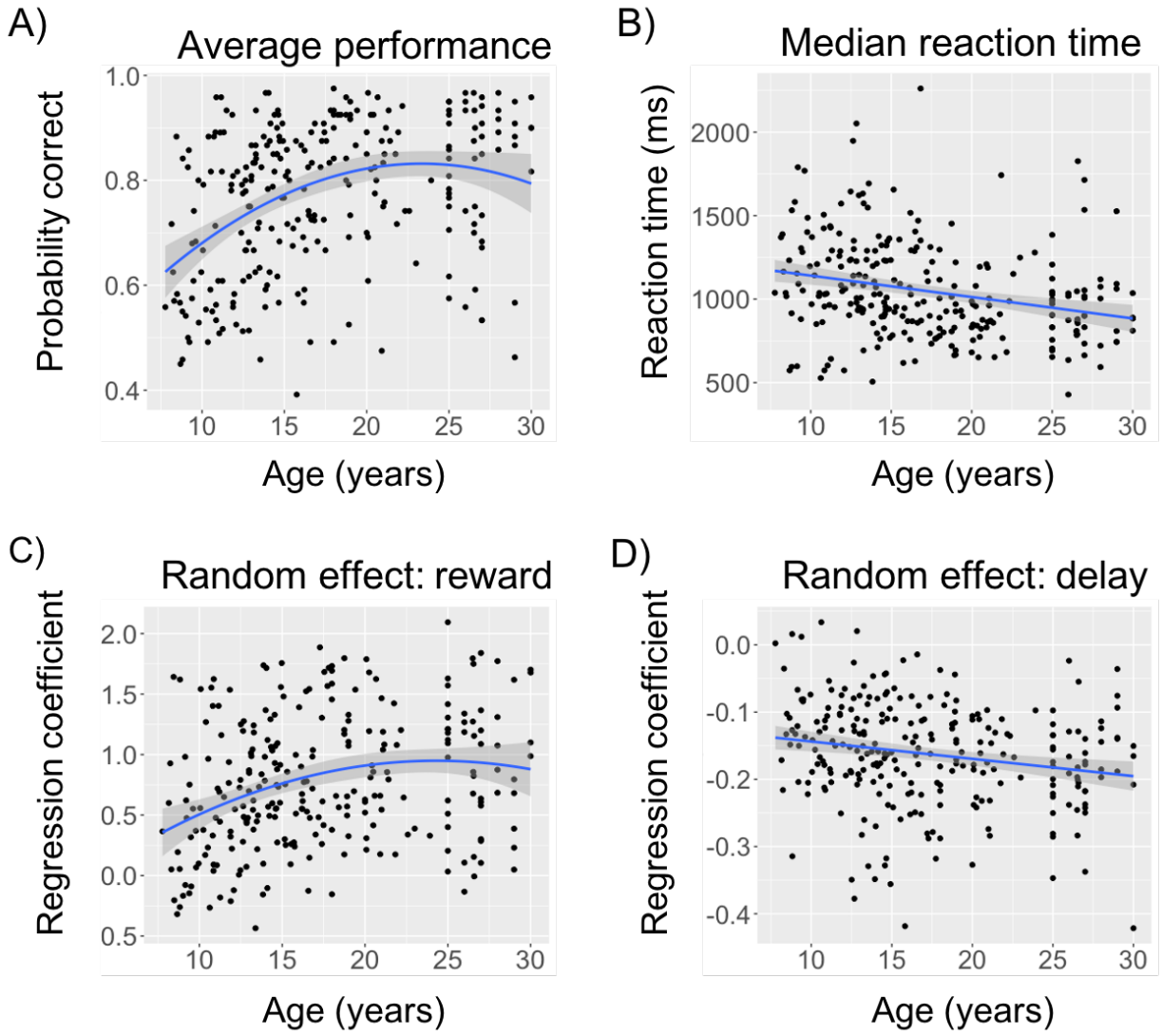

Figure 2.5: **Age effects on participants' behavior.** Scatter plot of age (x-axis) and (A) probability of choosing the correct response, (B) median reaction time, (C) random effect for reward history, and (D) random effect of delay. Each black dot represents one participant. The blue curve represents linear/quadratic regression line. There was no effect of sex in any analysis. Shaded region represents 95% confidence interval.

previous findings [109, 98]. There was no significant effect of sex on the median reaction time (unpaired t-test: $t(1, 262) = 0.4, p = 0.7$), but female participants had a significantly smaller standard deviation than male participants (unpaired t-test: $t(1, 262) = 2.72, p = 0.0085$).

These results indicated better performance and faster responses in older participants, ruling out speed-accuracy tradeoffs. Both age group (Fig 2.1B) and continuous age (Fig 2.5A) analyses revealed a nonlinear saturating relationship between age and performance.

2.3.3 Mixed-effect logistic regression

To better probe trial-by-trial learning dynamics, we used reward history and delay to predict the probability of a correct choice on each trial in a mixed-effect logistic regression. We found significant fixed effects of reward history and delay ($\beta_h = 0.8, \beta_d = -0.17$, both p's < 0.0001). This suggests that participants were more likely to pick the preferred flower as they received more reward feedback for the butterfly (reinforcement learning effect), and encountered the butterfly more recently (forgetting effect).

We found linear and quadratic effects (linear: $\beta_{age} = 0.02$, 95% CI $= [0.01, 0.03]$; quadratic: $\beta_{age^2} = -0.002$, 95% CI $= [-0.004, -0.0003]$; sequential ANOVA: $F(1, 261) = 5, p < 0.02$) of age on the random effect of reward history (Fig 2.5C). Two-line regression revealed a break point at around 21 years old, before which we found $\beta_{age} = 0.04, z = 3.84, p = 0.0001$, and after which we found $\beta_{age} = 0.02, z = 0.75, p = 0.45$. Therefore, similar to the trend we observed for overall performance (Fig 2.5A), participants' sensitivity to reward increased with age and saturated in early adulthood. We also found that participants became more sensitive to delay with age, shown by the linear effect ($\beta_{age} = -0.003$, 95% CI $= [-0.004, -0.001]$) of age on the random effect of delay (Fig 2.5D). Adding a quadratic term was not justified (sequential ANOVA, $F(1, 261) = 3, p = 0.08$).

2.3.4 Mathematical modeling

We used mathematical modeling and model comparison to obtain a quantitative and mechanistic understanding of participants' trial-by-trial learning and decision making. For all six models except the $\alpha\beta f$ model, we fitted all participants jointly using hierarchical Bayesian modeling [73] combined with sampling [27] for approximating the likelihood function (see Sec 2.2.7). We also fitted all six models non-hierarchically to compare with the $\alpha\beta f$ model.

2.3.4.1 Model comparison

For hierarchical Bayesian modeling, we used WAIC to compare the relative fit of models at the population level [166], an information criterion that penalizes model complexity appropriately for hierarchical Bayesian models. WAIC is fully Bayesian and invariant to reparametrization. Smaller WAIC indicates a better fit to the data, controlling for complexity. Since the $\alpha\beta f$ model was unable to converge when fitted hierarchically, we compared the other five hierarchical models using WAIC.

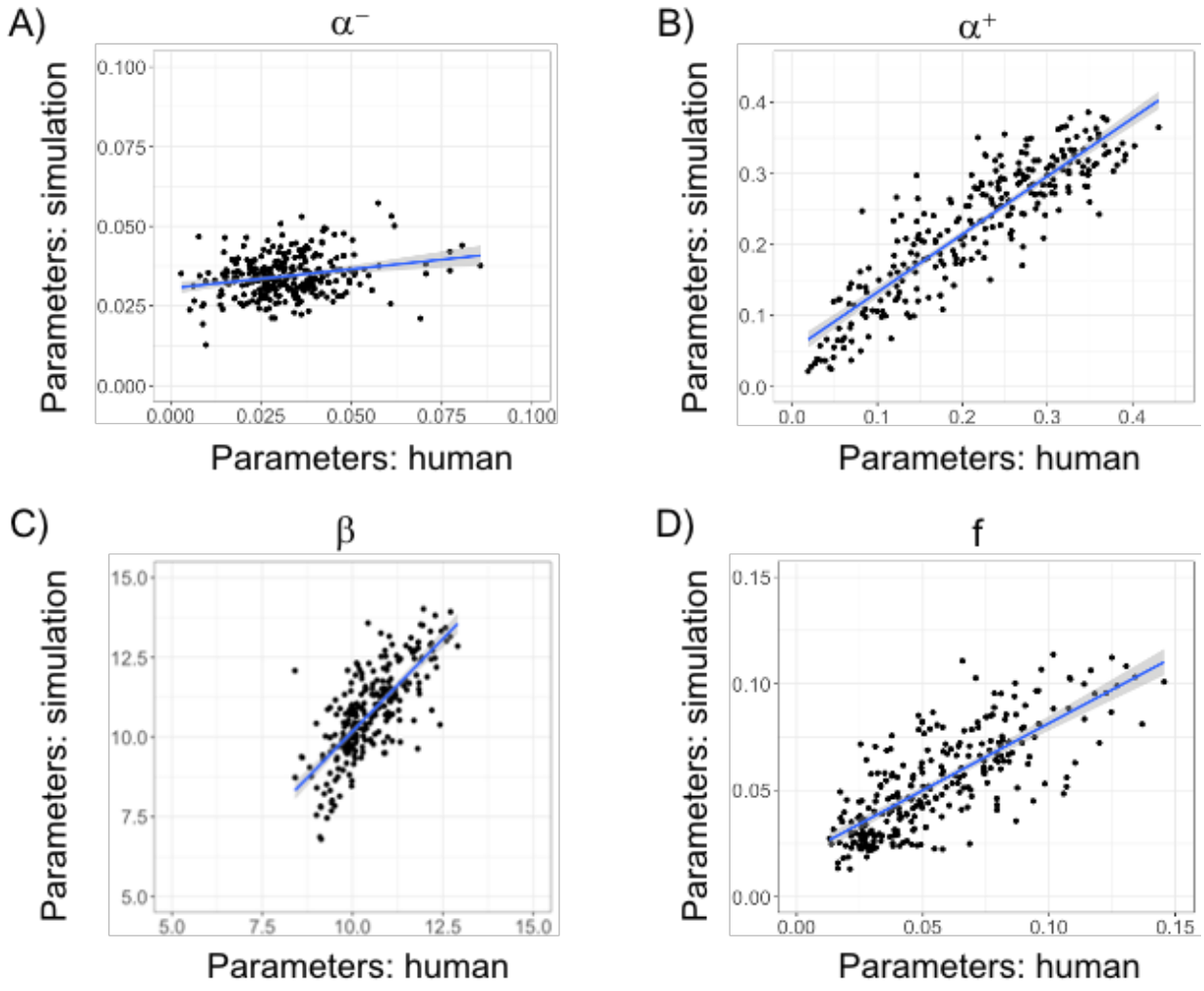

Figure 2.6: **Generate and recover.** Scatter plot of fitted parameters used to simulate data vs. recovered parameters for (A) α^-, (B) α^+, (C) β, and (D) f. All parameters except α^- can be reasonably recovered.

The $\alpha^+\alpha^-\beta f$ model with asymmetric learning rates and the forgetting parameter had the lowest (best) WAIC score (Fig 2.4B). However, a generate and recover procedure [169] showed that the α^- parameter values in the $\alpha^+\alpha^-\beta f$ model were very close to 0, and that they were not adequately recoverable (and therefore unsuitable to use as the basis for inference, Fig 2.6A).

Note that this poor recoverability of α^- might be primarily due to its small values. To test this hypothesis, we fitted the $\alpha^+\alpha^-\beta f$ model to participants data, and then fixed the α^+, β and f parameters, but substituted new values for the α^- parameter, sampled from a

broader range (Gaussian $N(0.5, 0.2)$ truncated $T[0.1, 0.9]$). We then performed the generate and recover analysis with this set of parameters. As shown in Fig 2.7, the α^- values could be well recovered in this higher range of values.

We further validated the fitted models by visualizing the learning curves of simulated data and comparing with participants' data (Fig 2.8). Note that since the $\alpha\beta f$ model did not converge hierarchically (see Sec 2.2.7), the simulation for the $\alpha\beta f$ model came from fitting each participant independently (i.e. flat instead of hierarchical fitting), while all the other five model simulations came from hierarchically fitted parameters.

Overall, the $\alpha^+0\beta$ and $\alpha^+0\beta f$ models both tracked participants' learning curve well throughout the experiment, whereas all the other models overshot to various extent. This combined with the fact that α^- was not recoverable suggested that the $\alpha^+\alpha^-\beta f$ model, despite having a better model comparison score (Fig 2.4B), had a risk of overfitting.

Since the values of the forgetting parameter were fairly small, it was not surprising that the simulations of the $\alpha^+0\beta$ and $\alpha^+0\beta f$ models did not differ a lot (Fig 2.8CF). The forgetting parameter might also capture patterns of data not readily visible in the learning curve, since delays between same butterflies were randomized across participants.

Consequently, although the $\alpha^+\alpha^-\beta f$ model has better model comparison score, our analysis focuses on the simpler $\alpha^+0\beta f$ model as the winning model because (1) α^- was not recoverable (Fig 2.6A) and (2) $\alpha^+0\beta f$ validated participants' behavior better (Fig 2.8). This is in following with a large literature indicating that quantitative criteria for model comparison are not the single factor that should guide model selection [127, 16, 93, 121]. In general, the $\alpha^+0\beta f$ model simulation captured the average learning curve throughout the entire experiment (Fig 2.4C) and age effects on overall performance (Fig 2.4D).

Note that conclusions for the α^+, β, and f parameters remained the same if we used the $\alpha^+\alpha^-\beta f$ model instead. Furthermore, the α^- parameters were generally very small for the fitted $\alpha^+\alpha^-\beta f$ model (Fig 2.6A). This suggested that participants were learning either very little from negative feedback or not at all.

We also compared all six models when fitted non-hierarchically using BIC [145], since we fitted all participants independently instead of hierarchically. We used the $\alpha^+0\beta f$ model as the baseline to calculate the difference between the BIC of all other models with the $\alpha^+0\beta f$ model. Fig 2.9 showed that the $\alpha^+0\beta f$ model was a decent model in all age groups consistently.

2.3.4.2 Model identifiability

While the generate and recover analysis so far showed good recoverability of fitted parameters in our winning model $\alpha^=0\beta f$, we further validated model identifiability/recoverability, i.e. we asked the question of when we generate data from, say, Model A, whether Model A would still be the winning model among others for explaining this data.

We focused only on the $\alpha\beta$, $\alpha^+0\beta$, and $\alpha^+0\beta f$ models, since our model comparison results mostly relied the identifiability of (1) asymmetric learning rates with $\alpha^- = 0$ and (2) the forgetting parameter. We simulated a dataset from each of the three fitted flat

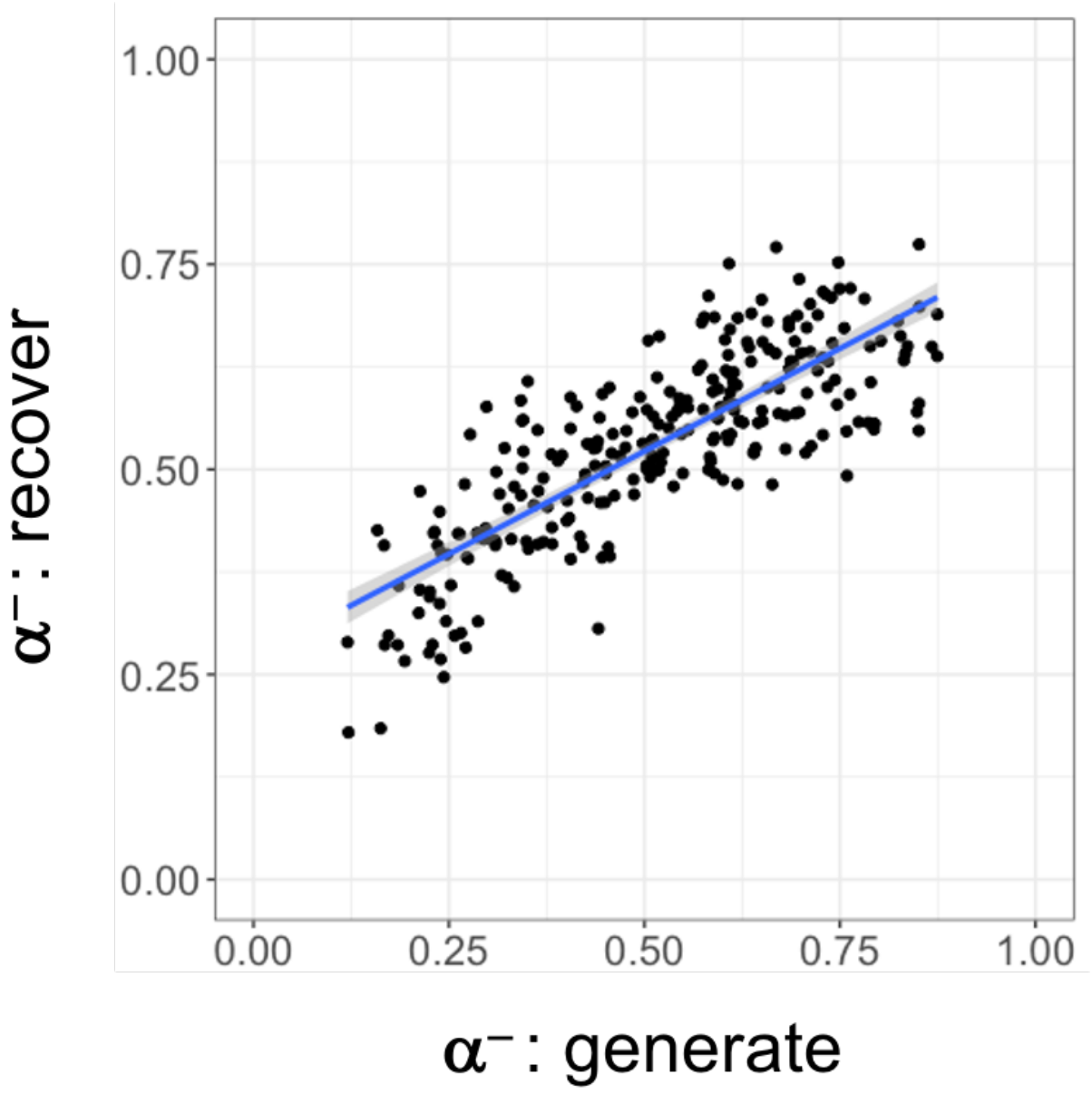

Figure 2.7: **Generate and recover of α^- in a higher and healthier range.** The α^- values could be well recovered in this higher range of values.

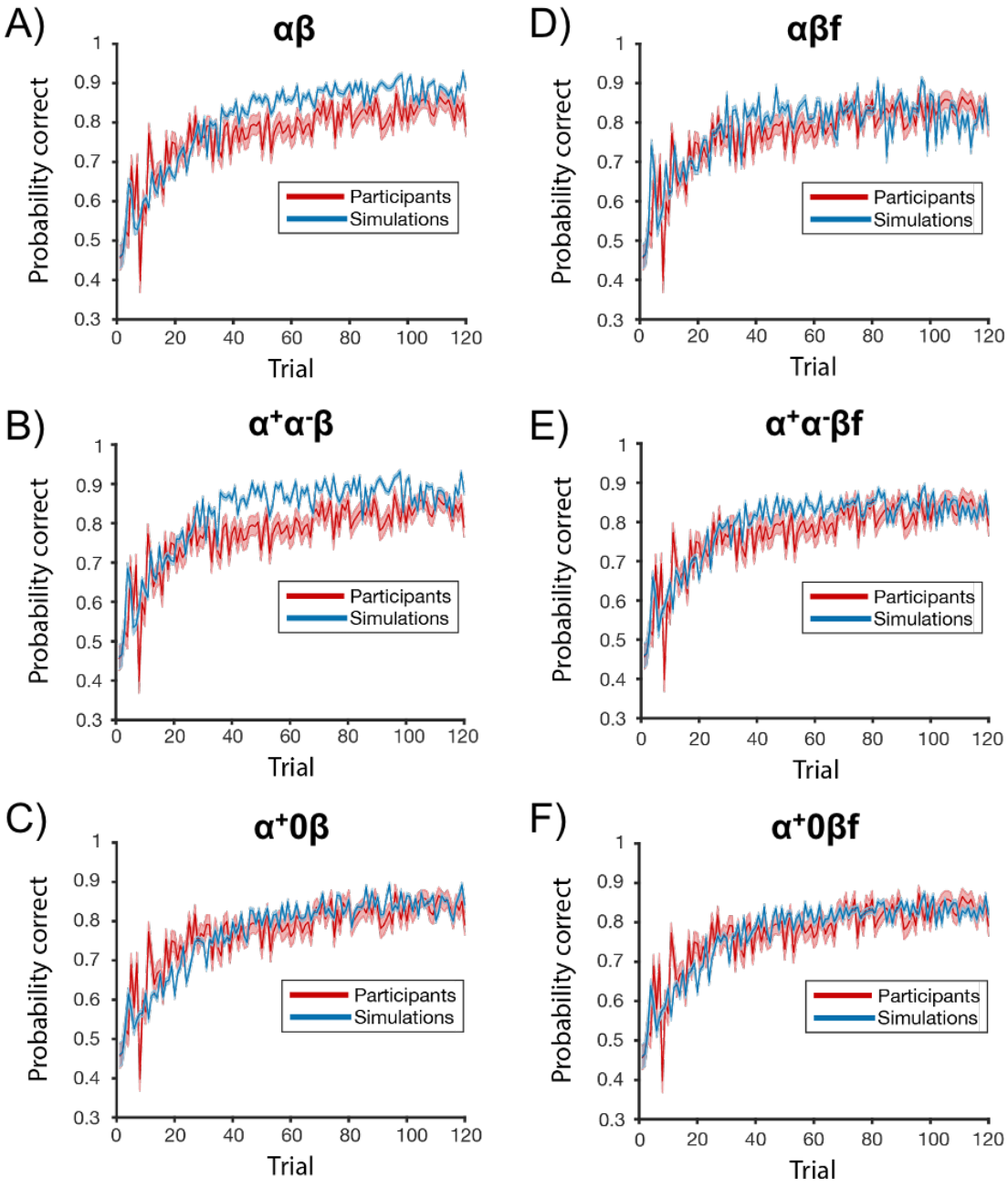

Figure 2.8: **Model validation.** Simulated performance from all six fitted models. The $\alpha\beta f$ model was fitted non-hierarchically; all other five model simulations came from hierarchical fitting.

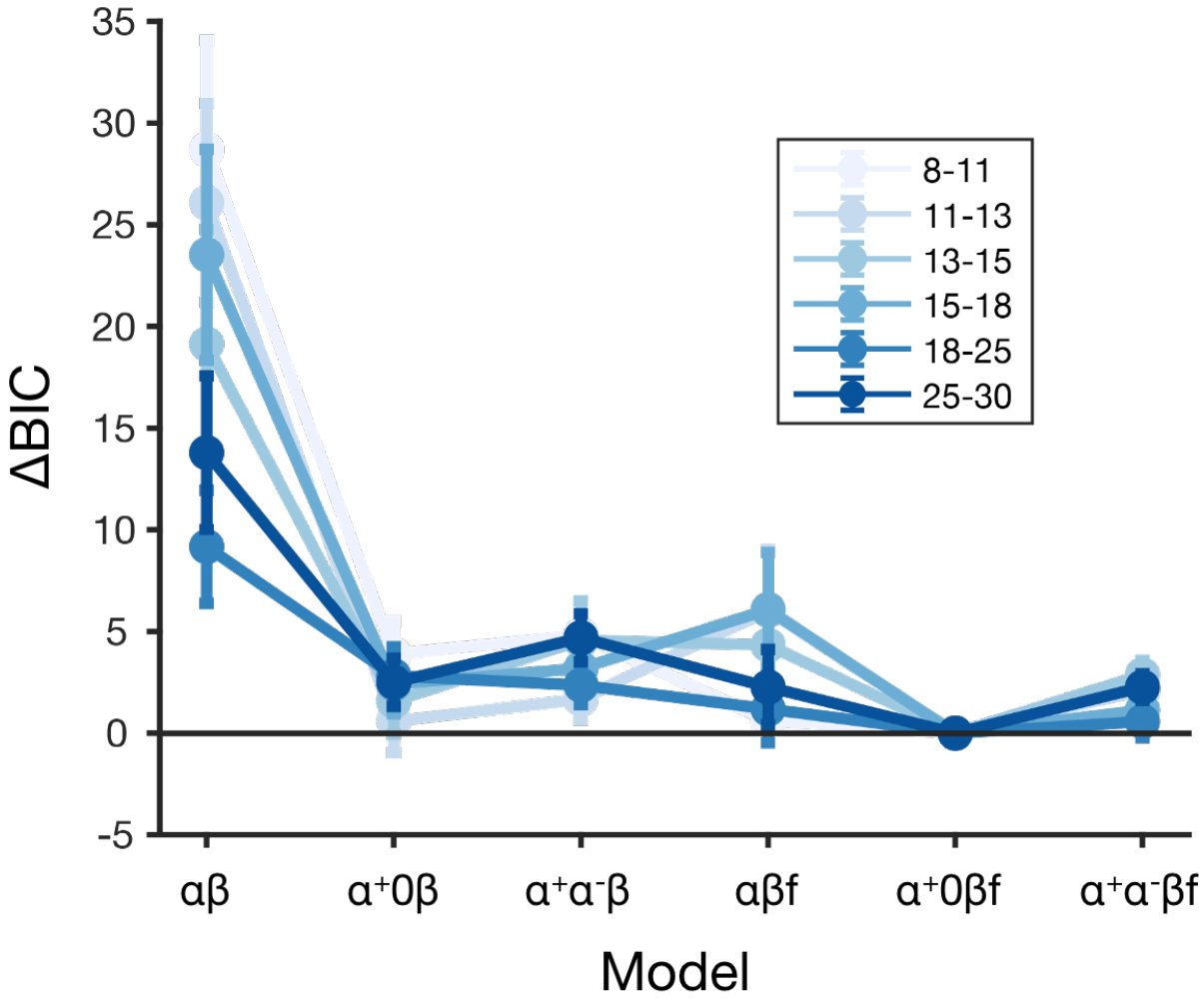

Figure 2.9: **Flat model comparison of all six models per age group.** We calculated the difference between the BIC for each of the six models with $\alpha^{+}0\beta f$ model per participant. Color represents six age groups. Results showed that the $\alpha^{+}0\beta f$ model was a decent model in all groups consistently.

models and used all three models to fit these three datasets, again non-hierarchically. For the data simulated from each model, we calculated the BIC for each of the three models used for recovery and each participant separately, and used BIC as an approximation of model evidence [100] to calculate protected exceedance probabilities [156, 140]). The exceedance probabilities were summarized in the following confusion matrix (Table 2.3), showing great model identifiability.

Model name	$\alpha\beta$	$\alpha^+0\beta$	$\alpha^+0\beta f$
$\alpha\beta$	1	0	0
$\alpha^+0\beta$	0	1	0
$\alpha^+0\beta f$	0	0	1

Table 2.3: **Model identifiability analysis.** The rows indicated the model where the dataset was generated from, whereas the columns indicated the model used for recovery. Each entry indicated protected exceedance probability, showing great model identifiability.

2.3.4.3 Nonlinear relationship between performance and model parameters

The model's overall performance in this task depends nonlinearly on the interaction of all parameters. To better illustrate this nonlinear relationship, we simulated the $\alpha + \alpha - \beta f$ model with the following parameter values: α^+ (0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7), α^- (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7), β (5, 7.5, 10, 12.5, 15), and f (0, 0.05, 0.1, 0.15, 0.2). Each parameter combination was simulated 100 times.

For simplicity of visualization, we first considered the winning model $\alpha^+0\beta f$, i.e. by setting $\alpha^- = 0$, and focused on the discrete grouping of participants. Fig 2.10 showed overall simulated performance changes with respect to α^+ (y-axis) and β (x-axis) for different forgetting values. Increasing learning rate α^+ did not always improve performance; in fact, too high of a learning rate could result in sub-optimal behavior ([169]). All six age groups (four under 18, two above 18) had average β values around 10 (from young to old: 10.12, 9.95, 10.54, 10.46, 10.85, 10.99) and average f values around 0.05 (0.0669, 0.0555, 0.054, 0.0553, 0.0554, 0.0545), roughly corresponding to the second subplot's third column (Fig 2.10). The average α^+ values for each age group are 0.14, 0.17, 0.22, 0.22, 0.26, 0.25 (plotted on the heat map as colored circles, darker means older age group), where higher α^+ generally improved overall performance. The simulations thus showed that all parameter changes with age in our population go towards more "optimal" performance in this task, but remained fairly far from it.

Another interesting observation was that no forgetting was not always optimal, especially in scenarios where the learning rate was high. This was more visible in Fig 2.11, where we interchanged the role of β and f from the previous heat map (now each subplot was indexed by β values from 5 - 15 and the x-axis of each subplot represents forgetting). This might be

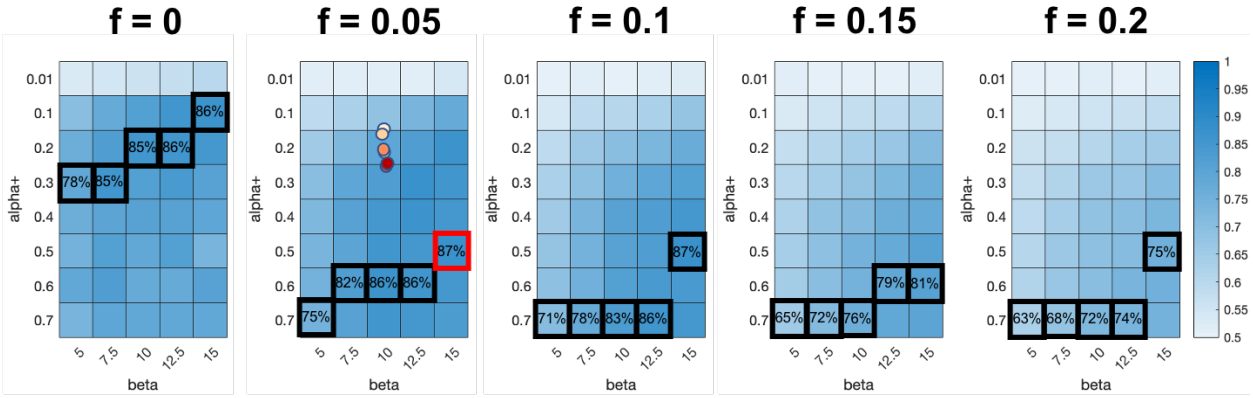

Figure 2.10: **Heat map for simulated performance of the $\alpha^+0\beta f$ model.** Overall simulated performance changes with respect to α^+ (y-axis) and β (x-axis), where each subplot corresponded to $f = 0-0.2$ from left to right. Black rectangle highlighted the local maximum within each column of each subplot (i.e. fixed β value), whereas the red rectangle highlighted the global maximum. Each colored circle indicated one of the six age groups; darker meant older age group.

due to the fact that forgetting could actually help unlearn the sub-optimal win-stay/lose-shift behavior created by high learning rates and/or highly exploitative policies.

In general, within the range of 5 - 15 for β, overall performance increased with higher β. And the optimal α^+ seemed to be inversely related to β but positively related to f.

Similar observations to α^+ could be made for α^-. In Fig 2.12, we set $f = 0$ for simplicity. Each subplot corresponded to one specific combination of (α^+, β) values, where the x-axis indicated α^- values and the y-axis overall performance. We also add a vertical bar for the corresponding α^+ value. For all combinations, while increasing α^- helped performance in the low range, high α^- always hurt performance. Note that optimal performance was rarely obtained with symmetric values of learning rates (i.e. the peak of the curve did not match the vertical bar). Since [43] employed symmetric learning rates, adult participants with high α^+ would have the same high α^- in their sample, which could also hurt performance.

2.3.4.4 Age differences in model parameters

With the winning model $\alpha^+0\beta f$, we next asked which computational processes drove the changes in performance over age by testing how model parameters changed with age. We adapted hierarchical Bayesian modeling to probe effects of age on model parameters. Specifically, we incorporated the regression of age as a predictor of model parameters into the hierarchical Bayesian model (Fig 2.4A), and directly sampled regression coefficients for age jointly with other model parameters (see Sec 2.2.7).

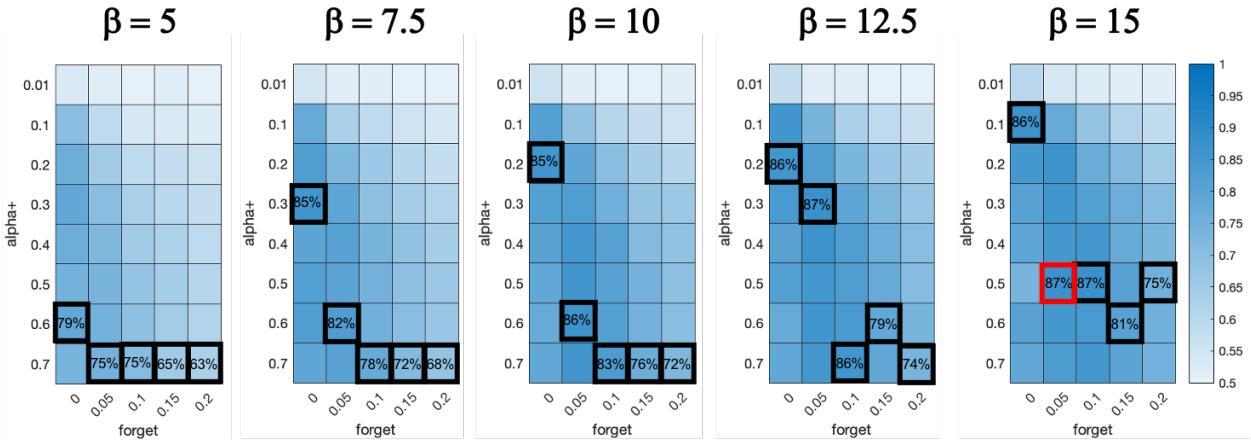

Figure 2.11: **Heat map for simulated performance of the $\alpha^+0\beta f$ model.** Overall simulated performance changed with respect to α^+ (y-axis) and f (x-axis), where each subplot corresponded to $\beta = 5 - 15$ from left to right. Black rectangle highlighted the local maximum within each column of each subplot (i.e. fixed f value), whereas the red rectangle highlighted the global maximum.

To test for effects of age on the model parameters, we examined whether the 95% credible interval (CI) of the posterior samples for each of the linear and quadratic regression coefficients did or did not include 0, where 0 indicates no effect (Fig 2.13). We found linear and quadratic effects of age on α^+ (linear coefficient 95% CI = $[0.05, 0.11]$; quadratic coefficient 95% CI = $[-0.1, -0.03]$) and β (linear CI = $[1.6, 3.4]$; quadratic CI = $[-1.9, -0.3]$). The trajectory of quadratic change over age for α^+ and β closely mimicked that for overall performance (Fig 2.5A). We also found marginally linear (Fig 2.5C), but not quadratic effects of age on f (linear CI = $[-0.04, 0.001], p = 0.066$; quadratic CI = $[-0.01, 0.02]$), with the forgetting parameter potentially decreasing over age.

We also performed generate and recover procedure to validate the significant age regression coefficients from hierarchical modeling (i.e. linear and quadratic age effects on α^+ and β; Fig 2.13). Fig 2.14 showed that the posterior of the samples from the recovery matched the posterior of the samples fitted on actual participants data fairly well for all significant age regressors.

2.3.5 Pubertal effects

To study whether pubertal development also affected participants' learning and decision making, we used pubertal measures (pubertal development score PDS and testosterone level T1) to predict fitted model parameters. We found a marginally significant effect of PDS on α^+, but not on β or the forgetting parameter f (two-way ANOVA, α^+: $F(3, 148) = 0.05$; β:

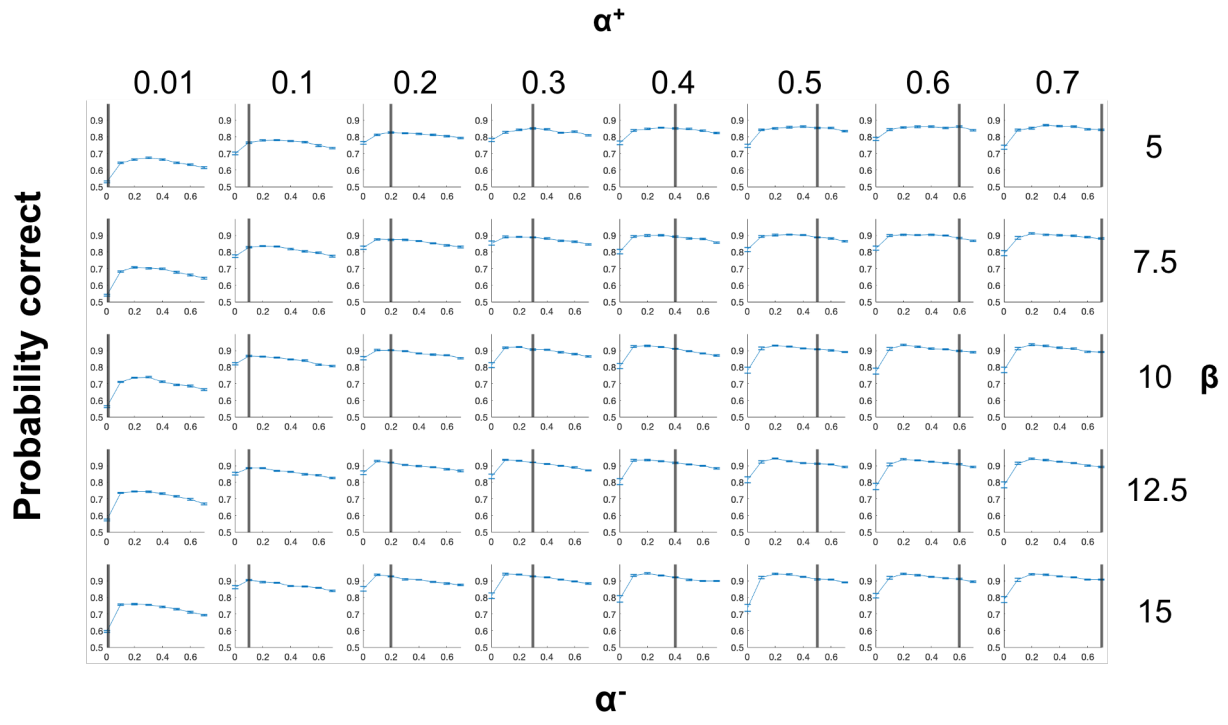

Figure 2.12: **Simulated performance of the $\alpha^+\alpha^-\beta$ model.** Overall simulated performance (y-axis) changed with respect to α^- (x-axis), where each subplot corresponded to a combination of (α^+, β) values. The vertical bar corresponded to the α^+ value. The error bars showed standard error across 100 simulations.

$F(3, 148) = 0.19$; f: $F(3, 148) = 0.7$). There was no significant effect of sex or interaction with sex (two-way ANOVA, all p's > 0.17).

We found a main effect of T1 on α^+, but not on β or f (two-way ANOVA, α^+: $F(3, 141) = 0.001$; β: $F(3, 141) = 0.1$; f: $F(3, 141) = 0.27$). There was again no effect of sex or interaction with sex (two-way ANOVA, all p's > 0.4).

We found that both PDS and T1 had a linear effect on α^+ (linear regression. PDS: $\beta_{PDS} = 0.02$, 95% CI $= [0.01, 0.04]$; T1: $\beta_{T1} = 0.02$, 95% CI $= [0.01, 0.04]$), and no effect on the forgetting parameter f (linear regression. Both p's > 0.5). There was a linear effect of PDS, but not T1, on β (linear regression. PDS: $\beta_{PDS} = 0.15$, 95% CI $= [0.003, 0.29]$; T1: $\beta_{T1} = 0.13$, 95% CI $= [-0.01, 0.27]$). However, the effects of PDS and T1 on α^+ and β disappeared when adding age into the regression (multiple linear regression, all p's > 0.55), while age remained the only significant predictor (multiple linear regression, all p's < 0.027).

To further explore the effect of PDS and T1 on model parameters while controlling for age, we performed the same regression using PDS or T1 to predict model parameters within each of the four age groups under 18 (Fig 2.15, Fig 2.16). We found that within the third

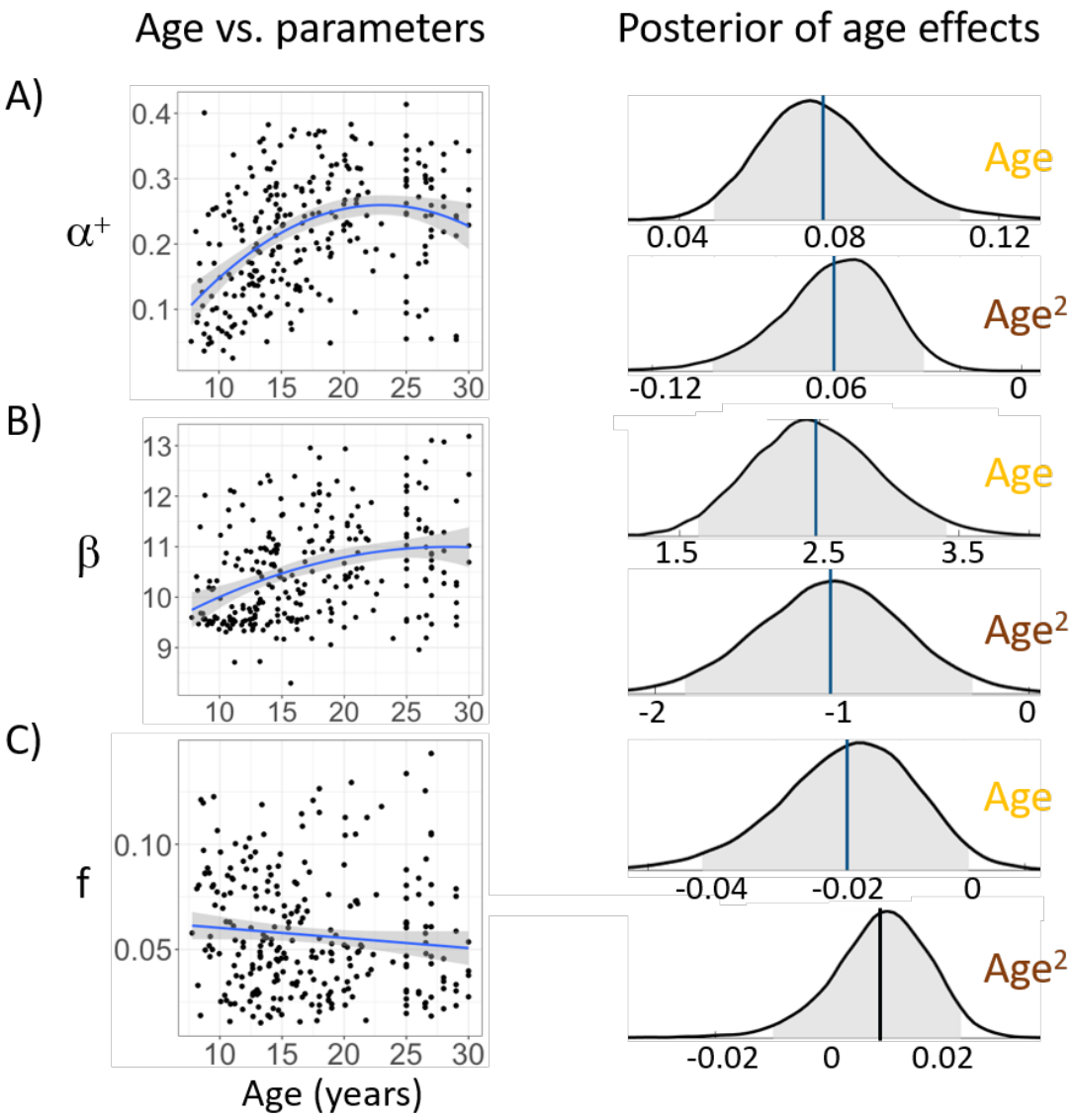

Figure 2.13: **Age effects on model parameters.** We directly incorporated age-related parameters into MCMC sampling to test within the hierarchical Bayesian modeling framework whether age had a linear or quadratic effect on all three model parameters: α^+ (A), β (B), f (C). Left panel: individual parameters from the original $\alpha^+ 0 \beta f$ model plotted against age. For visualization, we included a regression line; the shaded region indicates 95% CI. Right: distribution of posterior samples for linear (top, yellow) and quadratic (bottom, brown) regression coefficients. The vertical line represents the mean of all samples, with blue indicating an effect being present (i.e., 95% CI not including 0), and black indicating no effect. Shaded region shows 95% confidence interval.

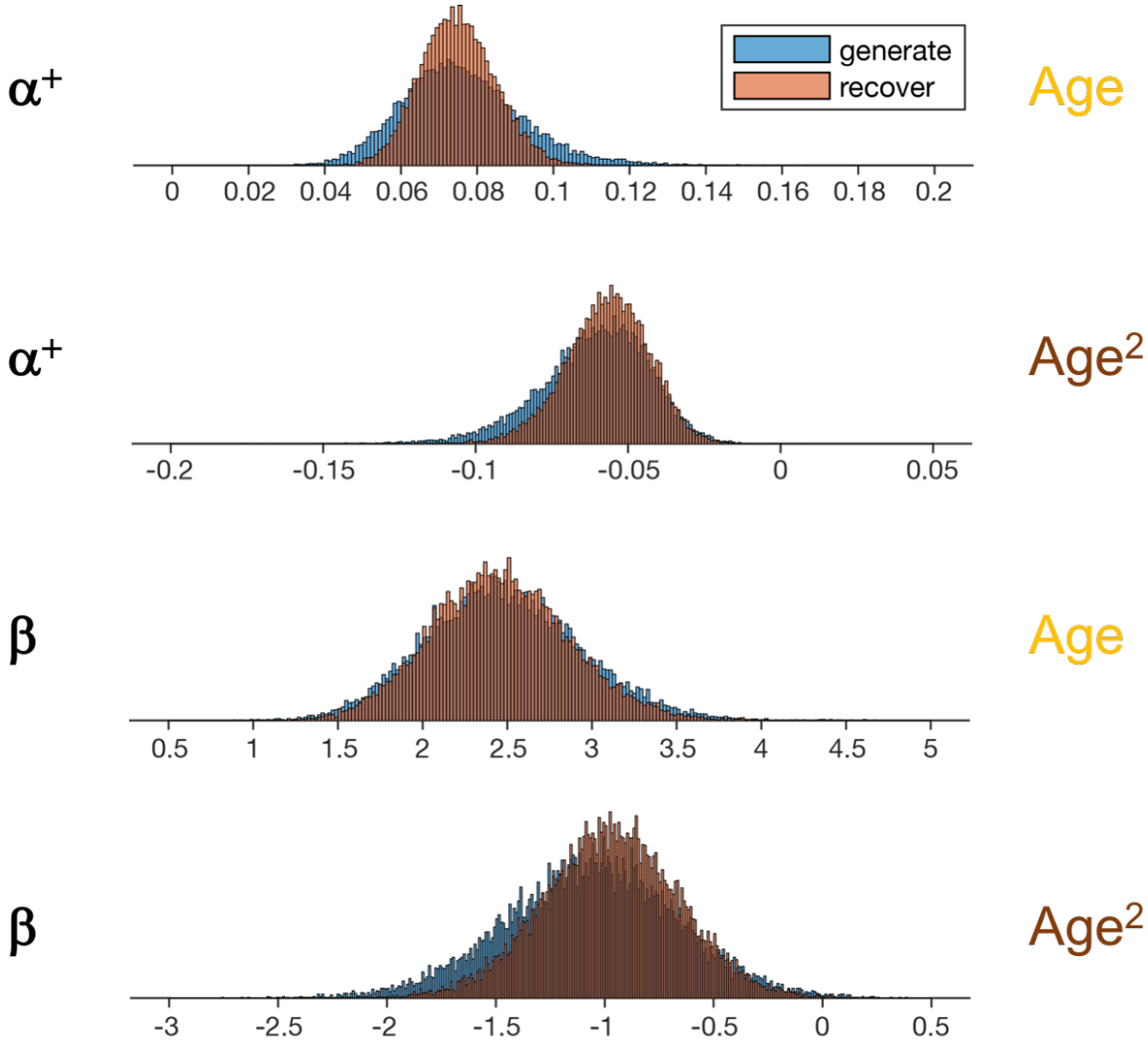

Figure 2.14: **Generate and recover for the age regression coefficients in hierarchical modeling.** All linear and quadratic age regressors for α^+ and β could be recovered well.

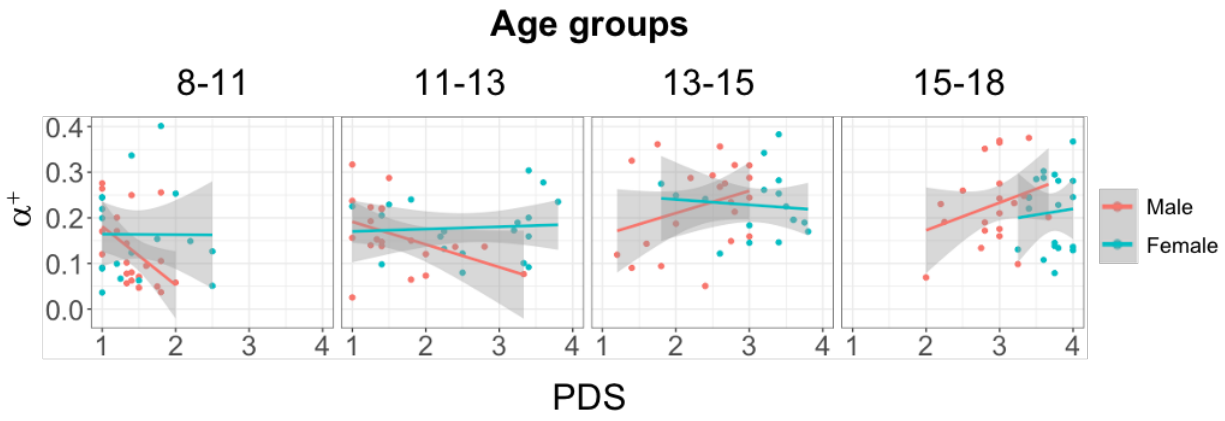

Figure 2.15: **PDS effect on α^+ in each of the four age groups younger than age 18.** There was not significant effects of PDS on α^+ in any of the four age bins.

age group (age 13-15), there was a positive effect of T1 on α^+ (linear regression: $\beta_{T1} = 0.05$, 95% CI $= [0.01, 0.08]$, $p = 0.005$). This remained significant when correcting for multiple comparisons (two parameters by four groups). This T1 effect remained when controlling for age in the regression (multiple linear regression: $\beta_{T1} = 0.05, p(T1) = 0.006$, $p(age) = 0.95$). This effect was not found in the other three age bins (linear regression, all p's > 0.17). We did not find significant effects of PDS on α^+ in any of the four age bins (linear regression, all p's > 0.37).

2.4 Discussion

How do humans learn to make choices when the outcome is uncertain? To learn probabilistic contingencies, humans need to integrate information over multiple trials to avoid overreacting to noise in the environment. But to learn efficiently, they also need to pay attention to recent information. Here, we investigated how humans trade off these constraints across development, what the underlying computational mechanisms that support such learning are, and how they change during adolescence.

At the population level, mathematical model comparison (Fig 2.4B) suggested that two mechanisms modulated learning of probabilistic contingencies. First, participants did not treat positive and negative feedback identically; rather, they had a strong bias to learn more from positive, and little to none from negative feedback. This asymmetry has been widely observed in previous studies [109, 18, 78], potentially due to differential mechanisms integrating positive and negative feedback [65]. Second, we found that learning was better explained by including a forgetting mechanism: more intervening trials between two iterations of a choice decreased the strength of past information [109].

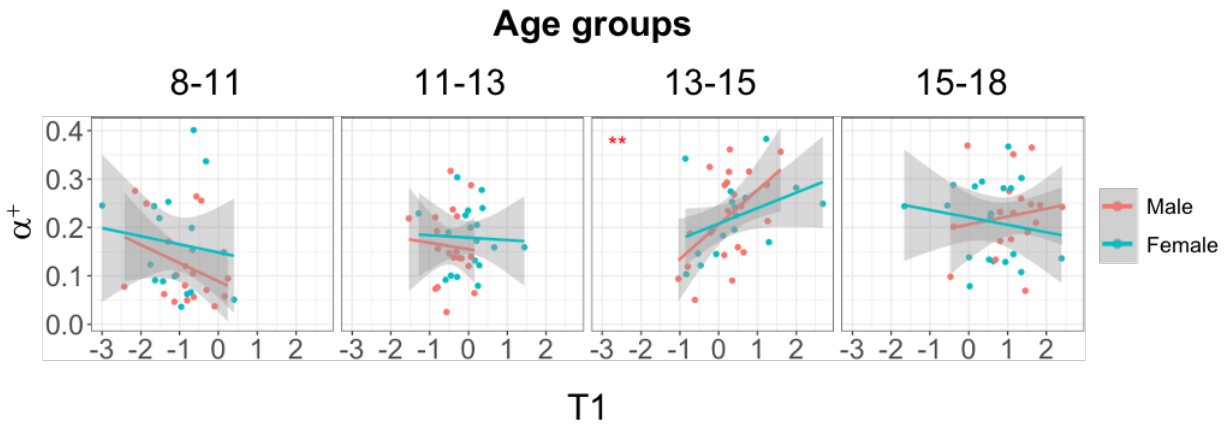

Figure 2.16: **T1 effect on α^+ in each of the four age groups younger than age 18.** There was a positive effect of T1 on α^+ in the third age group (red star indicated significance), but not in the other three age groups.

Our behavioral and modeling results suggest that learning in a stable probabilistic task environment changed markedly from childhood to adulthood. In particular, we found that overall performance increased with age, stabilising in early adulthood. This behavioral pattern was mirrored by the learning rate parameter (α^+) as well as inverse temperature (β), a parameter indicating a decrease in noise or exploration in choice. These results are consistent with the age effects observed in previous work using tasks with probabilistic [56] and deterministic [109] feedback,

Our observations that learning rate α^+ and inverse temperature β increase with development are generally consistent with previous work using the deterministic learning task *RLWM*, tested in the same participants as shown here [109], and a probabilistic task with same the same overall task structure as the Butterfly task, but different feedback methods [43]. However, we did not find higher performance in adolescents than adults, as had been observed in this previous Butterfly task study [43] (Fig 2.5A). Even when using the same age bins as [43], which limited our number of participants to $N = 84$ 13-18-year-old adolescents and $N = 86$ 20-30-year-old adults, we instead found that the performance in 20-30-year-olds was significantly higher than 13-18-year-olds (unpaired t-test, $t(168) = 2.3, p = 0.02$). Because our 18-25 year-olds were differently recruited than the rest of our sample, we additionally compared the 13-18-year-old adolescents to $N = 54$ 25-30-year-old adults (i.e. not including the undergraduate participants between 20-25 years old). We still found significantly better performance in adults (unpaired t-test, $t(136) = 2.2, p = 0.03$).

The finding in [43] was interpreted as "an upside" to slower learning that led to more robust integration over time of information, and thus higher overall performance under uncertainty at younger ages. Indeed, lower learning rates can be more optimal in probabilistic

tasks than higher learning rates. However, the relationship between learning rates and performance when learning probabilistic contingencies is complex and non-monotonic: it follows an inverse U shape, as very low learning rates lead to integrating information too slowly, but high learning rates lead to being too susceptible to noisy feedback (see Sec 2.3.4.3, [169]). Furthermore, the inverse U shape itself is dependent on the degree of exploration and forgetting ([126, 43, 169], see simulations in Fig 2.10, Fig 2.11). Learning rates were smaller in our study compared to [43]: the group level mean for α^+ in our sample was 0.18, whereas in [43], the mean was around 0.3 and 0.6 for adolescents and adults respectively (Fig 2B in [43]). In higher ranges of learning rates [43], an increase in learning rate could result in a decrease in performance (right side of the inverse U shape), while in our lower range, it could lead to an increase in performance (left side of the inverse U shape). Thus, the two studies are consistent in identifying an increase in learning rate with age, but over a different range of learning rate values (0.3 vs. 0.6), leading to opposite effects on performance. Indeed, in our study, the parameter trajectory with age corresponded to a slow improvement towards more "optimal" behavior, as defined by correct performance in the task (Fig 2.10).

Moreover, we modeled learning from positive and negative feedback asymmetrically [126], as opposed to the symmetric learning rate in [43]. In particular, our winning model $\alpha^+ 0 \beta f$ did not learn from negative feedback at all. A high α^- can also result in worse asymptotic performance in the Butterfly task (see Fig 2.12), resulting in more switching from the preferred flower.

Therefore, while we found a similar trend as in [43] that learning rates increased with age (Fig 2.13A), our learning rate values were much smaller, and the resulting trend in overall performance was different. Note that this difference in the range of learning rates could be a result of differences in the task specifics (our experiment did not have a memory retrieval aspect with novel images or brain imaging; our task was also the third in a sequence of four tasks). Differences in performance could also stem from differences in socioeconomic status (SES) and education level between the groups recruited to each study. For example, our 18-25 year-olds were undergraduate students, who may have a different education level than the 25-30 year-old community participants in our study or the adults sampled in [43]. The incentivization for undergraduate participants (course credit) and community participants (monetary) was also different. Finally, since the study is cross-sectional rather than longitudinal, age could also be confounded by birth year.

Nevertheless, our results support other previous developmental findings. In particular, we also found a decrease in exploration with age [109, 30], and an increase in learning rate previously observed in both deterministic [109] and probabilistic learning tasks [56]. Note that other studies have observed a decrease in learning rates (e.g., single-learning-rate models: [45, 82, 128]; models with asymmetric learning rates: [78, 25, 18]) or no change [85, 118]. These differences are potentially due to different task structures, samples, and modeling choices. For a more comprehensive review, see [126]).

While we found that performance increased during adolescence and saturated in early adulthood in this stable probabilistic learning task, a probabilistic switching task in the same population of participants [56] found a pronounced inverse U shape in overall performance,

which peaked at age 13-15. We conclude that this difference in age of peak performance in these two tasks stems from the reliance or lack of reliance on negative outcomes. The stable associations in the Butterfly task might encourage the participants to focus mostly on positive feedbacks (although this is not optimal based on the simulations in Fig 2.12, as α^- in the low range can improve performance), whereas in a volatile task setting [56], negative feedback was crucial for identifying when the correct action switched. This suggests that even with the same population of participants in two probabilistic tasks, task stability / volatility greatly changed participants' use of neural systems and behavioral strategies. For this population of participants, the volatile condition in [56] gave the 13-15 year old adolescents an edge over adults, while the stable condition in the Butterfly Task gave young adults an edge over adolescents.

While we found that random effects of delay on performance became more pronounced with age (Fig 2.5D), mathematical model fitting in contrast showed that forgetting parameters became weaker (Fig 2.13C). One possible interpretation for this apparent contradiction might relate to two simultaneous changes. First, adults might rely more on working memory processes [39] for probabilistic tasks [110], which manifested in a strengthened effect of delay. However, the decay of these memory processes might also decrease with age [109], which could be captured here by the decrease in the forgetting parameter. Thus, younger participants might show a weaker effect of delay not because their memory system was forgetting less (it was forgetting more), but because they used their working memory system less in this task [109, 54], and instead relied more on slower but more robust learning systems.

While we found that pubertal measures did not explain much additional variance compared to age in model parameters (see Sec 2.3.5), we found that testosterone level T1 had a significant positive effect on α^+ within the third age group of 13-15 years. Several explanations for this time limited observation are possible: a) gonadal hormone effects are stronger at this time of mid puberty, b) the other individual drivers of variation are more consistent at this time allowing detection of puberty related effects, or c) this result is a type I error. We favor hypothesis a) and b) because this observation about learning rate is broadly consistent with several studies [24, 106, 153] which report a positive relationship between testosterone levels and nucleus accumbens bold activity in response to rewards in mid adolescence. These data combined suggest a putative link between testosterone, nucleus accumbens activity, and learning rate in mid adolescence.

Another way of interpreting the relationship between testosterone levels and learning rate in age 13-15 can potentially be from the perspective of social learning [122, 63]. One of the most important sources of uncertainty during adolescence comes from social experiences (the rapid changes of social roles and contexts). A previous study [26] indeed showed that testosterone level affects social learning in adolescents, while another longitudinal study found a relationship between testosterone gains and nucleus accumbens activity gains in response to threatening face presentation [153]. Even though our study did not directly address social learning or social stimuli, the relationship between testosterone and learning might translate to the current paradigm.

Overall, work on the role of puberty and learning is currently in an early phase of under-

standing. It is likely that there are gonadal hormone dependent and independent aspects of development in the brain that will need to be disentangled [46]. A longitudinal design will have stronger statistical power to isolate puberty dependent effects [94]. Other sources of hormones and neuropeptides may also contribute to coordinate developmental change across the body and cumulative experience may also contribute. Age is less noisy to measure than pubertal development or hormones, but age is not a satisfactory explanation at the level of biological mechanism.

2.5 Conclusion

In conclusion, we sought to examine the development of learning in a stable probabilistic environment using a large adolescent and young adult sample with continuous age in the 8-30 range. Combining behavioral analysis and mathematical modeling, we showed developmental gains in performance through early adulthood that were explained by an increase in learning from rewarded outcomes (corresponding to a narrower time scale of information integration) and a decrease in exploration. These data and models help explain why learning and decision making differ during development and why a 'one-size-fits-all' approach may not equally serve youth at different stages.

Chapter 3

Augment Existing Mathematical Models to Explain Complex Human Cognition

While in Chapter 2 we demonstrated the usefulness of applying mathematical RL models to quantify human learning and decision making under uncertainty, the Butterfly task only involved one-step decision making. In this chapter, we provide an example study where HRL option is used for theory development, augmenting existing theories of human learning to account for more complex, multi-step human cognition. We also introduced Chinese Restaurant Process, a nonparametric clustering algorithm, to further account for transfer and generalization shown by human participants that cannot be explained by any existing RL models. This chapter is adapted from a paper submitted to Psychological Review [171]. I would like to acknowledge my senior co-author, Anne, who helped secure funding and provided constant guidance throughout the study. Permission to use this study as part of this thesis has been obtained from all co-authors.

3.1 Introduction

Recent advances have shown that RL algorithms [159] can give rise to extremely powerful AI systems [114, 146]. However, despite tremendous recent progress, artificial RL agents are unable to mimic and capture humans' ability to learn fast, efficiently, as well as transfer and generalize knowledge [97, 22, 52].

Human behavior and cognition possesses two key features that are essential to humans' efficient and flexible learning: cognitive representations are hierarchical [151, 90, 89, 4] and compositional [97]. Hierarchy has been identified as a crucial element of cognition in multiple domains such as perception [163, 99, 167, 15], decision making [176, 95, 5, 4, 6, 8, 48, 49, 160, 55], and learning [66, 7, 38, 36, 55]. Hierarchy in choices is often temporal [19, 21]: choices may be described at multiple degrees of granularity by breaking them down into more and

more basic chunks. For example, the task of making dinner can be broken down to making potatoes and making black beans; making potatoes can be broken down into sub-tasks such as cutting potatoes, roasting, etc. However, hierarchical levels may also represent different degrees of state abstractions at a similar time scale[90, 4, 38, 34]: for example, you may decide to make dinner (highest, most abstract level), which will consist of a salad, which will specifically be a Cesar salad (lowest, most concrete level).

Human behavior is also compositional: humans are able to compose simpler skills together in novel ways to solve new tasks in real life. For example, we can combine cutting potatoes with different routines to accomplish various tasks including fried potatoes, meshed potatoes, etc. Compositionality goes hand in hand with hierarchy, as it assumes the existence of different levels of skills. It has also been central to the study of human cognition [14, 96, 69] and artificial agents [170, 3, 173, 129].

The HRL options framework [158] (see Sec 1.4), originally proposed in AI, incorporates both hierarchy and compositionality features in an effort to make learning more flexible and efficient. Note that the options framework is not the first attempt to incorporate hierarchy and compositionality to model complex human cognition. Within psychology in particular, the concept of "options" echoes the idea of "chunking" in the cognitive architecture literature [101, 2]. Cognitive architectures models such as ACT-R [2] rely strongly on the hierarchical representation of behaviors, whereby procedures frequently executed in successions can become chunks that can be selected at a higher level of abstraction. However, we were not able to find examples of such cognitive models that focused on how humans might rapidly learn and transfer hierarchical representations. Furthermore, a distinct aspect of the HRL options framework (compared to cognitive architectures) is its objective of reward maximization [22], which is inherited as an augmentation of traditional flat RL (e.g. the RL models used in Chapter 2). In that sense, the options framework proposes a mathematical framework at Marr's computational level of analysis [125], not only at the algorithmic one. In our model, this reward objective also allows us to naturally include Bayesian inference as a way of optimal option selection and transfer (see Sec 3.2.1.5). However, there have also been initial attempts to combine ideas from reward maximization of RL with cognitive architectures [120, 70]. It would be especially interesting to consider potential connections between the options framework and various cognitive architectures, which were designed to explain a wide range of human cognition and not limited to structural learning from trial-by-trial interactions with the environment and reward feedbacks.

While there have been recent neuroscientific evidence of option learning in human participants, the fundamental question of whether and how humans learn and use options during learning remains unanswered [52]: there is little work probing the learning dynamics in tasks with a temporal hierarchy, or directly testing the theoretical benefits of options in a behavioral setting. In this study, we aim to 1) characterize how humans learn representations that support hierarchical and compositional behavior, and 2) investigate whether an expanded options framework can account for it. 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? In order to address these questions, we need to first identify aspects of human learning and

transfer that reflect the use of options, but cannot be explained by traditional RL, from a modeling perspective.

Previous research [38, 36, 33] showed evidence for flexible creation and transfer of a simple type of options that operate in non-sequential environments: one-step policies, also called task-sets [116]. While a vanilla flat RL model learns about state-action mappings (policies) as they are, such as cutting, boiling and stir frying potatoes (Fig 3.1), RL models that learn task-sets achieve transfer by learning state abstractions. For example, the model, after learning the policy of cutting potatoes, can generalize to cutting other vegetables by clustering the vegetables that it has never encountered before to the context of potatoes. [38, 36, 33] showed that humans can create multiple task-sets over the same state space in a context-dependent manner in a contextual multi-armed bandit task; furthermore, humans can cluster different contexts together if the task-set 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 (Fig 3.1). Moreover, human participants can identify novel contexts as part of an existing cluster if the cluster-defined strategy proves successful, resulting in more efficient exploration and faster learning.

However, the task-sets framework only supports hierarchy in state abstraction, not hierarchical structure in time (also called temporal abstraction, Fig 3.1), an essential component of the options framework. Since most real world tasks require multiple steps, RL models that only learn one-step task-sets are not sufficient. In particular, note that RL models that only learn task-sets might still get confused about whether it should boil or stir fry after the vegetable is cut. This is due to the non-Markovian (or semi-Markovian [158]) aspect of the environment: for the same observed state (cut vegetable), the optimal action might be different depending on the overarching goal, that cannot be currently observed. An RL model that further learns temporal abstractions such as options would instead combine one-step task-sets together as one abstract behavioral module. Once a specific option is activated, it resolves the ambiguity regarding the optimal action following cutting vegetable.

Here, we propose that combining state abstraction from task-set transfer [38, 36, 33] and temporal abstraction from the options framework [158] can provide important insights into complex human cognition. The additional temporal hierarchical structure offered by options should enable transfer of prior knowledge at multiple levels of hierarchy, providing rich opportunity for capturing the flexibility of human transfer. For example, in addition of being able to resolve the optimal action in a non-Markovian task (Fig 3.1), if humans have learned the simple sub-option of boiling water while learning how to make coffee, they do not need to re-learn it for learning to make tea or steamed potatoes; this sub-option can instead be naturally incorporated into a tea-making option, speeding up learning.

In this study, we present a new experimental protocol that allows us to characterize how humans develop hierarchical, compositional representations to guide behavior during trial-by-trial learning from reward feedback. In particular, it allows us to test whether humans create options during learning, and whether they use them in new contexts to explore more efficiently and transfer learned skills, at multiple levels of hierarchy. Our new two-stage learning game provides participants opportunities to create and transfer options at multiple

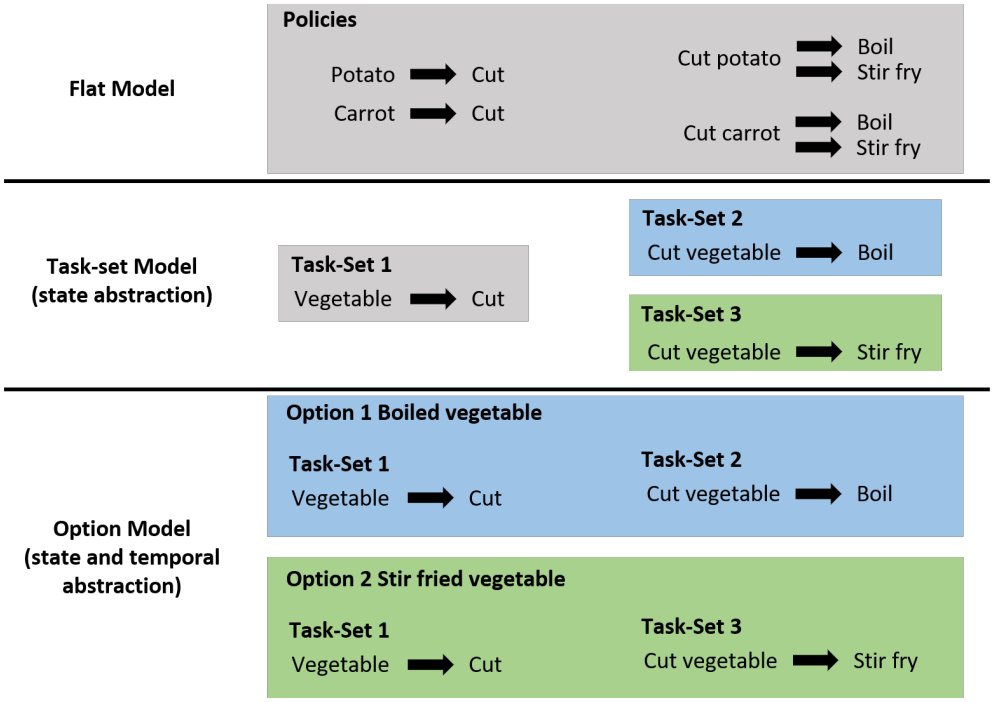

Figure 3.1: **Model schematics**. Schematics of how state and temporal abstractions can be used to describe increasingly more complex human cognition. **Flat Model.** The usual flat RL model learns one-step policies for different vegetables (potatoes and carrots) separately as different states (gray), with potentially multiple actions leading to reward in a given state (e.g. boil or stir fry potatoes). **Task-set Model.** The task-set model clusters both potatoes and carrots into the same state abstraction, namely, vegetable, thus everything learned about one vegetable will be immediately transferable to all the other vegetables. However, the task-set model only learns one-step policies, and in this non-Markovian task is unable to resolve the optimal action after the vegetable is cut, since it can be either boiled (blue) or stir fried (green). **Option Model.** The option model learns state abstractions, but also temporal abstractions by combining one-step rules into temporally-extended policies, resolving the action selection after the vegetable is cut by activating a temporal abstraction from the beginning. Now one activates either the option of boiling vegetable (blue) or stir frying vegetable (green) from the start of cutting vegetable.

levels of complexity.

To characterize how humans learn hierarchical and compositional representations to interact with the world and to test various predictions of learning and transferring temporal abstractions, we conducted a series of four experiments. The structure of the environment in Experiment 1 was non-Markovian, encouraging participants to learn option-like temporally-extended policies, and included test phases in which options could be transferred or re-composed; indeed, we found evidence of participants learning and transferring options at multiple levels. Experiment 2 provided a replication of Experiment 1 and further revealed interesting interaction between option transfer and meta-learning, as well as the complexity of credit assignment in hierarchical tasks. Experiment 3 mimicked Experiment 1, but removed the non-Markovian feature of Experiment 1: because all relevant information was observable, there was no additional benefit to creating options. Thus, Experiment 3 allowed us to test whether participants would spontaneously learn and transfer options even when there was no behavioral benefit to do so. Last, Experiment 4 aimed to test whether participants could compose options learned at different time and different levels. Given that humans can transfer task-sets to novel contexts [38, 36, 33], we hypothesized that humans would learn and transfer options to guide exploration and achieve better learning performance. The results of these four experiments (3 replicated in an independent sample) showed that human participants are able to learn, flexibly transfer and compose option-like temporally-extended policies at multiple levels.

We also present a formal mathematical RL model that brings together aspects of the classic HRL options framework with the task-set model's Bayesian inference mechanisms for clustering and transfer. The model combines the benefits of both frameworks. Specifically, the model relies on HRL-like options at three levels of hierarchy, and uses HRL-like learning mechanisms (using both rewards and pseudo-rewards) to learn policies and option-specific policies, respectively. Furthermore, our model uses Bayesian inference with a non-parametric prior to guide exploration and selection of options, inspired by the task-set model, and in that sense departing from traditional HRL framework. Our model makes specific predictions about learning, transfer, exploration, and error types in the four experiments. Our mathematical model captured the observed patterns of behavior, supporting the importance of hierarchical representations of choices for flexible, efficient, generalizable learning and exploration. Additionally, we showed that other models, including flat RL models or RL models with state abstraction but no temporal abstraction are insufficient in explaining the learning and transfer patterns we observe in human participants. Thus, our new experimental and theoretical framework characterizes how humans learn hierarchical and compositional representations to interact with our environment, and shows how this supports flexible transfer and efficient exploration.

Exp	18-25	26-30	31-35	36-40	41+	Unknown	Total
Exp 1	14	18	26	23	33	2	116
Exp 3	4	9	18	9	25	0	65
Exp 4	14	17	24	15	40	0	110

Table 3.1: Age range distribution for Mturk participants in Experiments 1, 3, and 4.

3.2 Experiment 1

Experiment 1 was designed to test if human participants are able to learn and flexibly transfer options. We designed a sequential 2-step decision-making paradigm (where each step was a contextual 4-armed bandit) to allow participants to learn options at multiple levels of complexities. Options changed between blocks, but the design provided participants with opportunities to practice reusing previously learned options. In two final test blocks, we directly tested creation and transfer of options by changing and/or combining previously learned options in novel ways.

3.2.1 Methods

3.2.1.1 Participants

All experiments were approved by the Institutional Review Board of the University of California, Berkeley. Experiment 1 was administered in-lab to UC Berkeley undergraduates who received course credit for their participation. 34 (22 female; age: mean = 20.6, sd = 1.6, min = 18, max = 24) UC Berkeley undergraduates participated in Experiment 1, and 9 participants were excluded due to incomplete data or poor learning performance (see results), resulting in 25 participants for data analysis.

For replication purposes, we also recruited participants through Amazon Mechanical Turk (MTurk) who performed the same experiment online. Participants were compensated a minimum of $3 per hour for their participation, with a bonus depending on their performance to incentivize them. 116 participants (65 female; see age range distribution in Table 3.1) finished the experiment. 61 participants were further excluded due to poor performance (Sec 3.2.1.4), resulting in 55 participants for data analysis.

3.2.1.2 Experiment 1 in-lab Protocol

Experiment 1 consisted of eight 60-trial blocks (Fig 3.2), with optional 20-second breaks in between blocks. In each block, the participants used deterministic truthful feedback to learn which of four keys to press for four different shapes. Each trial included two stages; each stage involved participants making choices in response to a single stimulus (Fig 3.2A) by pressing one of four keys. Each trial started with one of two possible stimuli, henceforth the first stage stimuli (e.g. circle or square). Participants had 2 seconds to make a choice.

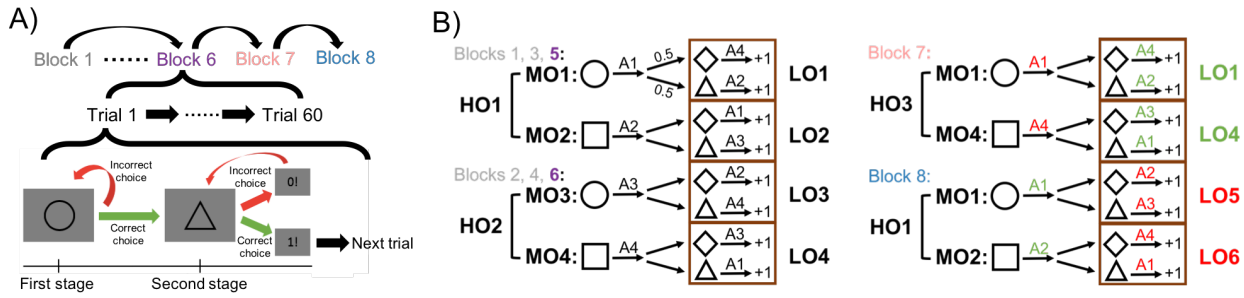

Figure 3.2: **Experiment 1 protocol.** (A) Block and trial structure: Blocks 1-6 were learning blocks, followed by two testing blocks: Blocks 7 and 8. Each block had 60 trials. In each trial, participants needed to select the correct response for the first stage stimulus (e.g. circle) in order to move on to the second stage stimulus (e.g. triangle), where they could win points by selecting the correct response. (B) Stimulus-action assignments: In Blocks 1-6, participants had the opportunity to learn options (extended policies) 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, potentially eliciting positive transfer (green) of LOs in the second stage, and negative transfer (red) of choices in the first stage. Block 8 tested predicted positive transfer in the first stage, but negative transfer of MO policies in the second stage, by replacing old LOs by new ones. Blocks were color coded for later result figures: Blocks 1-4 gray; Blocks 5-6 purple; Block 7 rose; Block 8 blue.

Participants only moved on to the second stage of the trial when they pressed the correct key for the first stage stimulus, or after 10 unsuccessful key presses, which enabled them to potentially try all four keys for a given stimulus in a single trial. Successful key press for the first stage of a trial did not result in reward feedback, but triggered a transition to the second stage, where participants saw one of the two other stimuli, henceforth labeled second stage stimuli (e.g. diamond and triangle). Both first stage stimuli led to both second stage stimuli equally often, and shapes were randomly assigned to either first or second stage across participants. In the second stage, participants also could not move on until they selected the correct choice (or selected wrong 10 times in a row for the same image). Participants received explicit feedback after each second stage choice: the screen indicated 1/0 point for pressing the correct/incorrect key, displayed for 0.5 second (Fig 3.2A). After a correct second stage choice, participants saw a fixation cross for 0.5 second, followed by the next trial's first stage stimulus. Each block contained 60 trials, with each first stage stimulus leading to each second stage stimulus 15 times in a pseudo-randomized sequence of trials.

Crucially, the correct stimulus-action assignments were designed to create a non-Markovian environment, and thus to encourage the creation of multi-step policies. In particular, second stage correct choices were dependent on what the first stage stimulus was. This encouraged

participants to make temporally extended choices (potentially options): their second stage strategies needed to depend on the first stage. Assignments, illustrated in Fig 3.2B, changed across blocks. Blocks 1, 3, 5 shared the same assignments; Blocks 2, 4, 6 shared the same assignments; this encouraged participants to not unlearn policies, but rather discover that they could reuse previously learned multi-level policies as a whole in new blocks.

Assignments in Blocks 7 and 8 intermixed some of the learning blocks assignments with new ones to test (positive and negative) transfer of options at various hierarchy levels. Specifically, the protocol was set up so that participants could learn up to 3 levels of hierarchical task structure (low, mid, and high level policies). More precisely, low-level options (LO) corresponded to second stage policies (a pair of stimulus-action associations, commonly labelled a *task-set*) [116]. Mid-level options (MO) were policies over both first and second stage stimuli. High-level options (HO) were policies over MOs (a pair of stimulus-MO associations in the first stage, which could be thought of as a *task-set over options*). As a concrete analogy, in Blocks 1, 3, 5, the participants learned how to make breakfast (HO_1), consisting of potatoes (MO_1) and eggs (MO_2). Making potatoes (MO_1) was broken down into cutting potatoes (the first stage) and then roasting (the second stage, LO_1). In Blocks 2, 4, 6, participants learned how to make lunch (HO_2), consisting of carrots (MO_3) and sandwich (MO_4). Making carrots (MO_3) was broken down into washing carrots (the first stage) and then steaming (the second stage, LO_3).

Block 7 tested positive transfer of second stage policies and negative transfer of first stage policies. In particular, we combined the policies for potatoes from breakfast (MO_1) and sandwich from lunch (MO_4) to form a new policy HO_3 (dinner). If participants build three levels of options, we expect positive transfer of mid-level options MO_1 and MO_4: participants should be unimpaired in making potatoes or a sandwich. However, we expect negative transfer of high-level options HO_1 and HO_2: participants seeing that making potatoes was rewarded might start making eggs as usual in breakfast (HO_1), instead of sandwich as rewarded here.

Block 8 tested positive transfer of first stage policies and negative transfer of second stage policies. In particular, the first stage of Block 8 shared the same assignments as Blocks 1, 3, 5 in the first stage, allowing participants to immediately transfer HO_1. However, the second stage policies (LO_5 and LO_6) were novel, which might potentially result in negative transfer: for example, participants might try to transfer LO_1 (roasting) following MO_1 (make potatoes), but the second stage policy was changed to LO_5 (e.g. frying).

3.2.1.3 Experiment 1 MTurk Protocol

To replicate our findings, we ran a minimally modified version of Experiment 1 online via MTurk. The task was slightly shortened, due to evidence that in-lab participants reached asymptotic behavior (Fig 3.4) early in a block, and to make the experiment more acceptable to online workers. Blocks 1 and 2 had a minimum of 32 and a maximum of 60 trials, but participants moved on to the next block as soon as they reached a criterion of less than 1.5 key presses per second stage trial in the last 10 trials (the 55 Mturk participants included

for data analysis on average used 42 (SD = 10, median = 37, min = 32, max = 60) trials in Block 1 and 39 (SD = 10, median = 33, min = 32, max = 60) trials in Block 2). Blocks 3-8 were all shortened to 32 trials, with each first stage stimulus leading to each second stage stimulus 8 times.

3.2.1.4 Data analysis

We used the number of key presses until correct choice in each stage of a trial as an index of performance. Since the experiment would not progress unless the participants chose the correct action, more key presses indicates worse performance. Ceiling performance was 1 press per stage within a trial. Chance level was 2.5, assuming choosing 1 out of 4 keys randomly, unless indicated otherwise. To probe for any potential transfer effects, we calculated the average number of key presses at the beginning of each block (trials 1-10), before learning has saturated. As a stronger test of option transfer, we also calculated the probability that the first press for a given stimulus at each stage of a trial was correct in different blocks.

To rule out participants who were not engaged in the task, we excluded any participant who did not complete Blocks 5-8 within an allotted amount of time (6 minutes each) - indeed this could only happen if participants often reached the 10 key presses needed to move on to the next stage without the correct answer, a clear sign of no engagement.

We additionally excluded any participant whose average performance in the last 10 trials of either first or second stage in either Block 5 or 6 was at or below chance, since it indicated a lack of learning and engagement in both stages of the task. These exclusion criteria were applied to all experiments, including Mturk participants.

Note that the analysis of the *first* 10 trials and the *last* 10 trials served different purposes, since they reflected different stages of learning. The beginning of each block when participants had not yet integrated all the new block information was where we expected to see transfer effects. On the other hand, the last 10 trials of a block showed asymptotic performance and were used to ensure learning had occurred, in particular for exclusion criteria. In short, the performance in the last 10 trials answered the question of how participants made choices after repeated exposure to the same environments for many iterations, while the first 10 focused on learning (and potentially transfer) in a new environment.

Among 116 Mturk participants in Experiment 1, 104 were above chance in the second stage (the more difficult one), but only 55 were above chance in the first stage (the easier one). Thus most participants were excluded due to the first stage performance criterion. The same trend was true for the other two Mturk experiments: most Mturk participants were excluded due to performance in the first stage in Experiment 3 and Experiment 4. We hypothesize that the poor first stage performance in many is due to the task's incentive structure - participants knew they only earned points (which were converted to monetary bonus for MTurk participants) in the second stage. All second stage results were qualitatively similar to the ones reported in this chapter for all experiments when we relaxed the exclusion criterion to include participants at chance in the first stage.

The options framework makes predictions about the specific choices made in response to a stimulus, beyond whether a choice is correct: the nature of the errors made can be informative [38]. We categorized the specific choices participants made into meaningful choice types, to further test our predictions about potential option transfer effects. As the choice types were stage and experiment dependent, we describe the choice type definitions in the result sections where necessary. When performing choice type analysis, we only considered the first key press of the first or second stage in each trial. We also compared reaction time of different choice types to test potential sequence learning effects.

For statistical testing, we used parametric tests (ANOVAs and paired t-test) when normality assumptions held, and non-parametric tests (Kruskall-Wallis and sign test) otherwise.

3.2.1.5 Mathematical modeling

To quantitatively formalize our predictions, we designed a mathematical model for learning and transferring options, inspired by the classic HRL framework as well as other HRL literature [38, 158]. We simulated this model, as well as three other learning models that embody different hypotheses about learning in this task, to compare which model best captures patterns of human learning and transfer. All models were simulated 500 times. We did not fit the model to the trial-by-trial choices of participants because computing the likelihood of the hierarchical models is intractable. In flat reinforcement learning models, state, action and rewards on each trial are fully observed. However, for the main HRL model used in this study, we assume that participants first select an option, conditioned on which they select a primitive action. Note that we only observed the primitive action from participants' key presses, not the selection of options. Therefore, in order to calculate the full likelihood, one would have to marginalize the option choices for each trial, resulting in the integration of exponentially many trajectories throughout the experiment. Even if participants only needed to choose between 2 hidden options, participants often made more than 1000 key presses in our experiment, which would require summing over $2^{1000} (> 10^{300})$ trajectories, rendering the calculation of the likelihood function intractable.

All results presented in the main text figures were simulated with parameters chosen to match participants' behavioral patterns qualitatively and quantitatively well (Table 3.2). However, our qualitative predictions are largely independent of specific model parameters: we show in Sec 3.6 that a single set of parameters (Table 3.3), consistent across all experiments, makes the same qualitative predictions regarding transfer effects.

3.2.1.5.1 The Naive Flat Model

The Naive Flat Model is a classic reinforcement learning model that learns Q-values to guide action selection in response to stimuli. In the first stage, it learns a Q-value table $Q^1(F_i, A_j^1)$, where F_1 and F_2 are two first stage stimuli, $A_1, \ldots, A_4$ are four possible actions. We use superscript to index stage (1 means first stage, 2 means second stage). The Q-values are initialized to uninformative Q-values $1/\#\{\textit{possible actions}\} = \frac{1}{4}$, since each of the four

actions has an equal probability of resulting in a pseudo-reward of 1 for transitioning into the second stage. On each choice, a first stage policy is computed based on the first stage stimulus, F_i, with the softmax function:

$$P(A_j^1|F_i) = \frac{exp(\beta^1 * Q^1(F_i, A_j^1))}{\sum_k exp(\beta^1 * Q^1(F_i, A_k^1))}, \tag{3.1}$$

where β^1 is the inverse temperature parameter. A first stage action A^1, ranging from A_1 to A_4, is then sampled from this softmax policy. After observing the outcome (moving on to the second stage or not), the Q-values is updated with Q-learning [159]:

$$Q^1(F_i, A^1) = Q^1(F_i, A^1) + \alpha^1 * (r - Q^1(F_i, A^1)), \tag{3.2}$$

where α^1 is the learning rate parameter, and r is 1 if A^1 is correct and 0 otherwise.

In the second stage, the model similarly learns another Q-value table $Q^2(S_i, A_j^2)$, where S_1 and S_2 are two second stage stimuli, with learning rate α^2 and inverse temperature β^2. Note that it disregards the non-Markovian nature of the task: it learns the Q-values for the two second stage stimuli without remembering the first stage stimulus. As such, this model is a straw man model that cannot perform the task accurately, but exemplifies the limitations of classic RL in more realistic tasks, and serves as a benchmark.

At the start of a new block, the Naive Flat Model resets all Q-values to 1/4, and thus has to re-learn all Q-values from scratch. To better account for human behavior, we also included two forgetting parameters, f^1 and f^2. After each choice, the model decays all Q-values for the first stage based on f^1:

$$Q^1(F_i, A_j^1) = (1 - f^1) * Q^1(F_i, A_j^1) + f^1 * 1/4. \tag{3.3}$$

Forgetting in the second stage is implemented similarly.

Participants very quickly learned that the correct second stage action was different from the first stage one (see results). To account for this meta-learning heuristic, we add a free meta-learning parameter, m, that discourages selecting the same action in the second stage as in the first stage. Specifically, if π is the second stage policy as computed from softmax, we set $P(A^1|S_i) = m$, where A^1 is the action chosen in the first stage, and re-normalize:

$$P(A^{other}|S_i) = (1 - m) \times \pi(A^{other})/(1 - \pi(A^1)), \tag{3.4}$$

where A^{other} is any action other than A^1.

Parameters f^1, f^2 and m, which capture memory mechanisms and heuristics orthogonal to option learning, are included in all models and implemented in the same way. In total, the Naive Flat Model has 7 parameters: $\alpha^1, \beta^1, f^1, \alpha^2, \beta^2, f^2, m$.

3.2.1.5.2 The Flat Model

The Flat Model extends the Naive Flat Model with a single addition of first-stage memory, which makes this model able to perform the task well in both stages. Specifically, in the second stage, the Flat Model remembers the first stage stimulus by treating each of the 4 combinations of the first and second stage stimuli as a distinct state and learns Q-values for all 4 combinations. The Flat Model has the same 7 parameters as the Naive Flat Model.

3.2.1.5.3 The Task-Set Model

The Task-Set Model is given the capability of transferring previously learned task-sets (one-step policies) with Bayesian inference. In the first stage, the model tracks the probability P^1 of selecting each first stage task-set HO_i in different first stage contexts c_j^1, which encodes the current temporal (block) context (e.g. 8 contexts in the first stage of Experiment 1). In particular, the model uses a Chinese Restaurant Process (CRP) prior to select HO [133]: 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:

$$P^1(HO_{n+1}|c_{n+1}^1) = \frac{\gamma^1}{Z^1}; \tag{3.5}$$

and the probability of reusing a previously created high-level option HO_i is set to:

$$P^1(HO_i|c_{n+1}^1) = \frac{N_i^1}{Z^1}, \tag{3.6}$$

where γ^1 is the clustering coefficient for the CRP, N_i^1 is the number of first stage contexts clustered on HO_i, and $Z^1 = \gamma^1 + \sum_i N_i^1$ is the normalization constant. The new HO_{n+1} policy is initialized with uninformative Q-values $1/\#\{possible\,actions\} = \frac{1}{4}$. The model samples HO based on the conditional distribution over all HO'S given the current temporal context. The model also tracks HO-specific policies via Q-learning. Once an HO is selected, a first stage policy is computed based on the HO's Q-values and the first stage stimulus F_i with softmax:

$$P(A_j^1|F_i, HO) = \frac{exp(\beta^1 * Q_{HO}^1(F_i, A_j^1))}{\sum_k exp(\beta^1 * Q_{HO}^1(F_i, A_k^1))}, \tag{3.7}$$

where β^1 is the inverse temperature. A first stage action A^1, ranging from A_1 to A_4, is then sampled from this softmax policy. After observing the outcome (moving on to the second stage or not), the model uses Bayes' Theorem to update P^1:

$$P^1(HO_k|c_j^1) = \frac{P(r|F_i, A^1, HO_k)P(HO_k|c_j^1)}{(\sum_l P(r|F_i, A^1, HO_l)P(HO_l|c_j^1))}, \tag{3.8}$$

where r is 1 if A^1 is correct and 0 otherwise, and $P(r|F_i, A^1, HO_l) = 1 - Q_{HO_l}^1(F_i, A^1)$ if $r = 0$, or $Q_{HO_l}^1(F_i, A^1)$ if $r = 1$. Then the Q-values of the HO with the highest posterior probability is updated:

$$Q_{HO}^1(F_i, A^1) = Q_{HO}^1(F_i, A^1) + \alpha^1 * (r - Q_{HO}^1(F_i, A^1)), \tag{3.9}$$

where α^1 is the learning rate.

The second stage runs a separate CRP with P^2, similar to P^1 in the first stage, which guides selection of task-sets LO over second stage stimuli. All other are identical to the first stage except that the second stage contexts are determined by both temporal (block) context and the first stage stimulus (e.g. 16 contexts in the second stage of Experiment 1). All the equations of CRP, action selection and Q-learning remain the same. The Task-Set Model has 9 parameters: $\alpha^1, \beta^1, \gamma^1, f^1, \alpha^2, \beta^2, \gamma^2, f^2, m$.

3.2.1.5.4 The Option Model

The Option Model extends the task-set model to include multi-step decisions (options MO). The first stage is identical to the Task-Set Model. However, instead of just choosing an action for the first stage, a whole MO is activated. For example, if the circle is observed in Block 1, HO_1 may trigger the model to select MO_1, which triggers the selection of A_1. The selection of MO_1 would then make the model likely to select LO_1 for the second stage (Fig. 3.2B). To simplify credit assignment, we make the simplifying assumption - warranted in our task - that there is a one-on-one mapping between first-stage actions and options, allowing us to index MOs by their first-stage action. This is meant as a technical simplification, rather than a theoretical assumption.

The second stage is similar to the second stage of the Task-Set Model. The only difference is that each MO has an MO-specific probability table P^2_{MO}. In the Task-Set Model, the CRP in the second stage using P^2 is independent of the first stage choices. In contrast, in the Option Model, the first stage choice determines which MO is activated (e.g. choosing A_1 for the circle in Experiment 1 is equivalent to choosing MO_1 as a whole, Fig 3.2B), which then determines which probability table, P^2_{MO}, to use for running the CRP in the second stage and to select LOs. This implementation captures the essence of options in the HRL framework, in that selection of MO in the first stage constrains the policy chosen until the end of the second stage (where the option terminates). Specifically, for the P^2_{MO} activated by the MO chosen in the first stage, there are 16 contexts in the second stage of Experiment 1 (8 blocks and 2 first stage stimuli). If contexts $\{c^2_{1:n}\}$ are clustered on $N^2 \leq n$ LOs, when the model encounters a new context c^2_{n+1}, the prior probability of selecting a new low-level option LO_{n+1} in this new context is set to:

$$P^2_{MO}(LO_{n+1}|c^2_{n+1}) = \frac{\gamma^2}{Z^2}; \tag{3.10}$$

and the probability of reusing a previously created low-level option LO_i is set to:

$$P^2_{MO}(LO_i|c^2_{n+1}) = \frac{N^2_i}{Z^2}, \tag{3.11}$$

where γ^2 is the clustering coefficient for the CRP, N^2_i is the number of second stage contexts clustered on LO_i for the current MO, and $Z^2 = \gamma^2 + \sum_i N^2_i$ is the normalization constant.

The new LO_{n+1} policy is initialized with uninformative Q-values $1/\#\{possible\,actions\} = \frac{1}{4}$. The model samples LO based on the conditional distribution over all LOs given the current context and MO. The model also tracks LO-specific policies via Q-learning. Once an LO is selected, a second stage policy is computed based on the LO's Q-values and the second stage stimulus S_i with softmax:

$$P(A_j^2|S_i, LO) = \frac{exp(\beta^2 * Q_{LO}^2(S_i, A_j^2))}{\sum_k exp(\beta^2 * Q_{LO}^2(S_i, A_k^2))}, \tag{3.12}$$

where β^2 is the inverse temperature. To account for the meta-learning heuristic, we add a free meta-learning parameter, m, that discourages selecting the same action in the second stage as in the first stage. Specifically, if π is the second stage policy as computed from softmax, we set $P(A^1) = m$, where A^1 is the action chosen in the first stage, and re-normalize:

$$P(A^{other}) = (1 - m) \times \pi(A^{other})/(1 - \pi(A^1)), \tag{3.13}$$

where A^{other} is any action other than A^1. A second stage action A^2, ranging from A_1 to A_4, is then sampled from this policy. After observing the outcome (moving on to the second stage or not), the model uses Bayes' Theorem to update P_{MO}^2:

$$P_{MO}^2(LO_k|c_j^2) = \frac{P(r|S_i, A^2, LO_k)P_{MO}^2(LO_k|c_j^2)}{(\sum_l P(r|S_i, A^2, LO_l)P_{MO}^2(LO_l|c_j^2))}, \tag{3.14}$$

where r is 1 if A^2 is correct and 0 otherwise, and $P(r|S_i, A^2, LO_l) = 1 - Q_{LO_l}^2(S_i, A^2)$ if $r = 0$, or $Q_{LO_l}^2(S_i, A^2)$ if $r = 1$. Then the Q-values of the LO with the highest posterior probability is updated:

$$Q_{LO}^2(S_i, A^2) = Q_{LO}^2(S_i, A^2) + \alpha^2 * (r - Q_{LO}^2(S_i, A^2)), \tag{3.15}$$

where α^2 is the learning rate. This implementation captures the essence of options in the HRL framework, in that selection of MO in the first stage constrains the policy chosen until the end of the second stage (where the option terminates). The Option Model has the same 9 parameters as the Task-Set Model.

Note that in our Option Model, there are two ways in which the option selection is instantiated. (1) One way is to use inference with a CRP prior [133]: instead of estimating the values of different HO's through incremental Q-learning, we estimated the likelihood of reward after selecting each HO's using Bayes' formula. This is inspired from our previous task-set model [38], and equips our Option Model with a level of flexibility in transfer (by inferring which option is likely to be useful in a new environment), something that traditional HRL options framework cannot achieve. We discuss this departure from classic HRL options framework further in the discussion (Sec 3.7.2). (2) We also implemented the option value functions by learning the values of different MO's within each HO's. Since MO are indexed by their first-stage action, the Q-values that participants learned for actions in the first stage correspond to MO option values. This is in line with the classic option values in the HRL options framework [158].

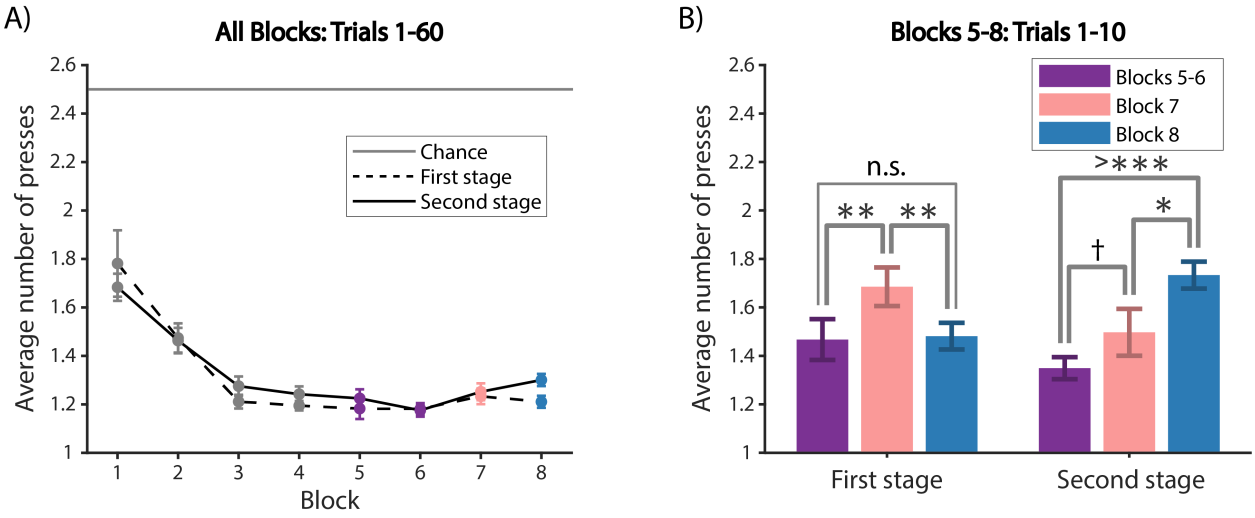

Figure 3.3: **Experiment 1 general behavior.** (A) Average number of key presses in the first and the second stages per block. Chance is 2.5, ceiling is 1 press. (B) Average number of key presses for the first 10 trials of Blocks 5-8 for the first (left) and second stages (right). We use n.s. to indicate $p \geq 0.1$; † for $p < 0.1$; $*$ for $p < 0.05$; $**$ for $p < 0.01$; $***$ for $p < 0.001$; and $>***$ for $p < 0.0001$. We indicated all statistical significance with these notations from now on.

3.2.2 Experiment 1 Results

3.2.2.1 Participants do not use flat RL

Participants' performance improved over Blocks 1-6 (Fig 3.3A) and within blocks (Fig 3.4). This improvement may reflect the usual process of learning the task observed in most cognitive experiments, as indicated by the improvement between Block 1 and 2 (paired t-test, first stage: $t(26) = 2.2, p = 0.03$; second stage: $t(26) = 3.9, p = 0.0006$). However, it could also reflect participants' ability to create options at three different levels in Blocks 1 and 2, and to successfully reuse them in Blocks 3-6 to adapt to changes in contingencies more efficiently. Below, we present specific analyses to probe option creation in test blocks. We used participants' performance averaged over Blocks 5 and 6 as a benchmark for comparing against performance in test Blocks 7 and 8.

We probed potential option transfer effects over the first 10 trials for each block (Fig 3.3B), before behavior reached asymptote (Fig 3.4). In the first stage, there was a main effect of block on number of key presses (1-way repeated measure ANOVA, $F(2, 48) = 6.9, p = 0.002$). Specifically, participants pressed significantly more times in Block 7 than Blocks 5-6 and Block 8 (paired t-test, Blocks 5-6: $t(24) = 3.0, p = 0.006$; Block 8: $t(24) = 3.0, p = 0.006$).

We checked whether the performance of circle and square in Block 7 was asymmetrically affected due to the interleaving of odd and even blocks (Fig 3.2B). Specifically, participants

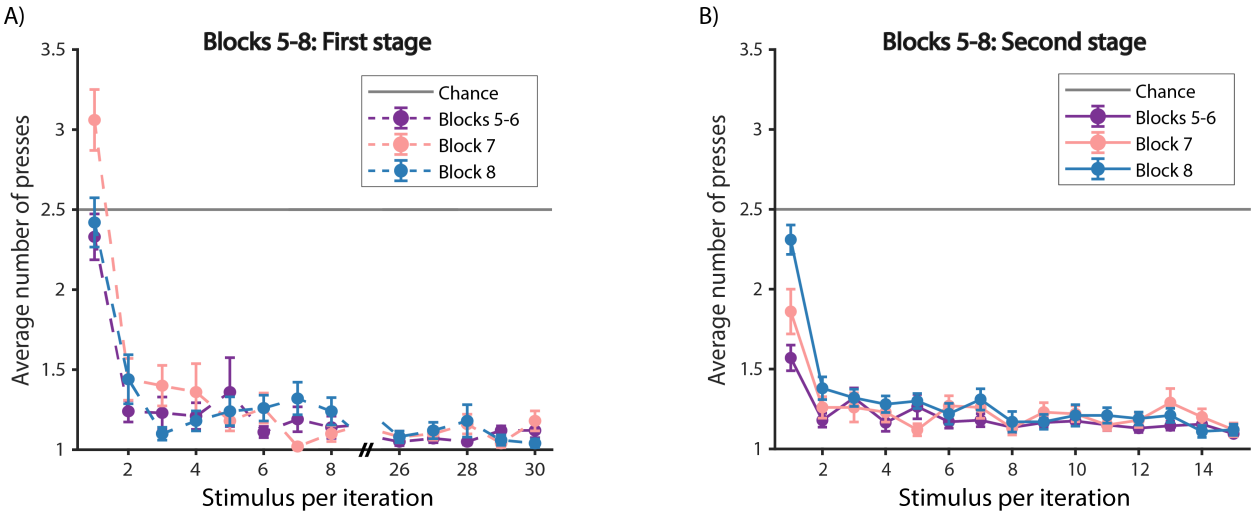

Figure 3.4: **Experiment 1 performance within Blocks 5-8 for in-lab participants.** (A) First stage. (B) Second stage.

might start Block 7 by using HO_1 in odd blocks; thus the negative transfer in the first stage of Block 7 would be primarily due to more key presses from the square, not the circle. To test this possibility, we calculated average number of key presses in the first 5 trials for circle and square respectively in Block 7. However, we found no significant difference between the performance of circle and square in the first stage (paired t-test, $t(24) = 1.38, p = 0.18$); we also found no significant difference between the performance in the second stage following circle and square (paired t-test, $t(24) = 0.44, p = 0.66$).

These results provide preliminary evidence for negative transfer of previously learned HO in Block 7: participants might attempt to reuse HO_1 or HO_2, since either policy is successful for half the trials, but is incorrect and thus results in more key presses in the first stage for the other half of the trials. There was no significant difference between Block 8 and Blocks 5-6 (paired t-test, $t(24) = 0.25, p = 0.81$). This provides initial evidence for positive transfer of HO_1 in Block 8, since performance in the first stage of Block 8 was on par with Blocks 5-6.

In the second stage (Fig 3.3B), there was also a main effect of block in number of key presses (1-way repeated measure ANOVA, $F(2, 48) = 11, p < 0.0001$). Specifically, participants pressed significantly more times in Block 8 than Block 7 and Blocks 5-6 (paired t-test, Block 7: $t(24) = 2.4, p = 0.025$; Blocks 5-6: $t(24) = 5.8, p < 0.0001$). The difference between Block 7 and Blocks 5-6 was marginally significant (paired t-test, $t(24) = 2.0, p = 0.06$). These results suggest participants positively transferred MO in the second stage of Block 7, where such generalization was helpful, since their performance was nearly not impaired compared to Blocks 5-6 where participants were able to reuse full HO. Furthermore, it suggests that they negatively transferred MO in the second stage of Block 8, where the first stage

choice that respected the current MO was followed by a new LO for correct performance, and thus necessitated to create a new MO.

Behavioral results in both the first and second stages provide initial evidence for option learning and transfer at distinct levels, both positive – when previous policies can be helpfully reused – and negative – when they impair learning. To further validate our hypothesis that participants learned options, we compared the simulations of four models with human behavior (Table 3.2).

Among the four models (Fig 3.5), only the Option Model and the Task-Set Model could account for the results. The Naive Flat Model could not achieve reasonable performance in the second stage because it ignored the non-Markovian aspect of the task - it was unable to learn two different sets of correct choices for a given second stage stimulus, because this required conditioning on the first stage stimulus (Fig 3.2B). Thus, it serves to illustrate the limitations of classic RL, but is a straw man model in this task. The Flat Model achieved reasonable performance in both the first and second stages, being able to take into account the first stage in second stage decisions, but did not demonstrate any transfer effects. Thus, results so far replicate previous findings that participants create one-step policies or task-sets, that they can reuse in new contexts, leading to positive and negative transfer [38, 36, 33]. We now present new analyses to show that the findings extend to creating multi-step policies or options.

3.2.2.2 Second stage choices reveal option transfer

To strengthen our results, we further examined the specific errors that participants made as they can reveal the latent structure used to make decisions. To further disambiguate between the Option Model and the Task-Set Model, we categorized errors into meaningful choice types [38]. We focused on the second stage choices for model comparison (Fig 3.6), the part of the experiment designed so that temporally extended policies could have an impact on decision making.

We hypothesized that participants learned MO's that paired the policies in the first and second stages. Therefore, positive transfer in the second stage of Block 7 and negative transfer in the second stage of Block 8 should be due to participants selecting the entire MO that was previously learned in response to a first stage stimulus, including the correct key press for the first level stimulus as well as the corresponding LO for the second level. We defined choice types based on this hypothesis. For example, for the second stage of Block 8, consider the diamond following the circle in Block 8 (Fig 3.2B): 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.

We computed the proportion of the 3 error types for the first 3 trials of each of the 4 branches in the second stage of Block 8 (Fig 3.6A). There was a main effect of error type (1-way repeated measure ANOVA, $F(2, 48) = 44, p < 0.0001$). In particular, we found more *option transfer* errors than the *other* errors (paired t-test, $t(24) = 2.5, p = 0.02$), suggesting

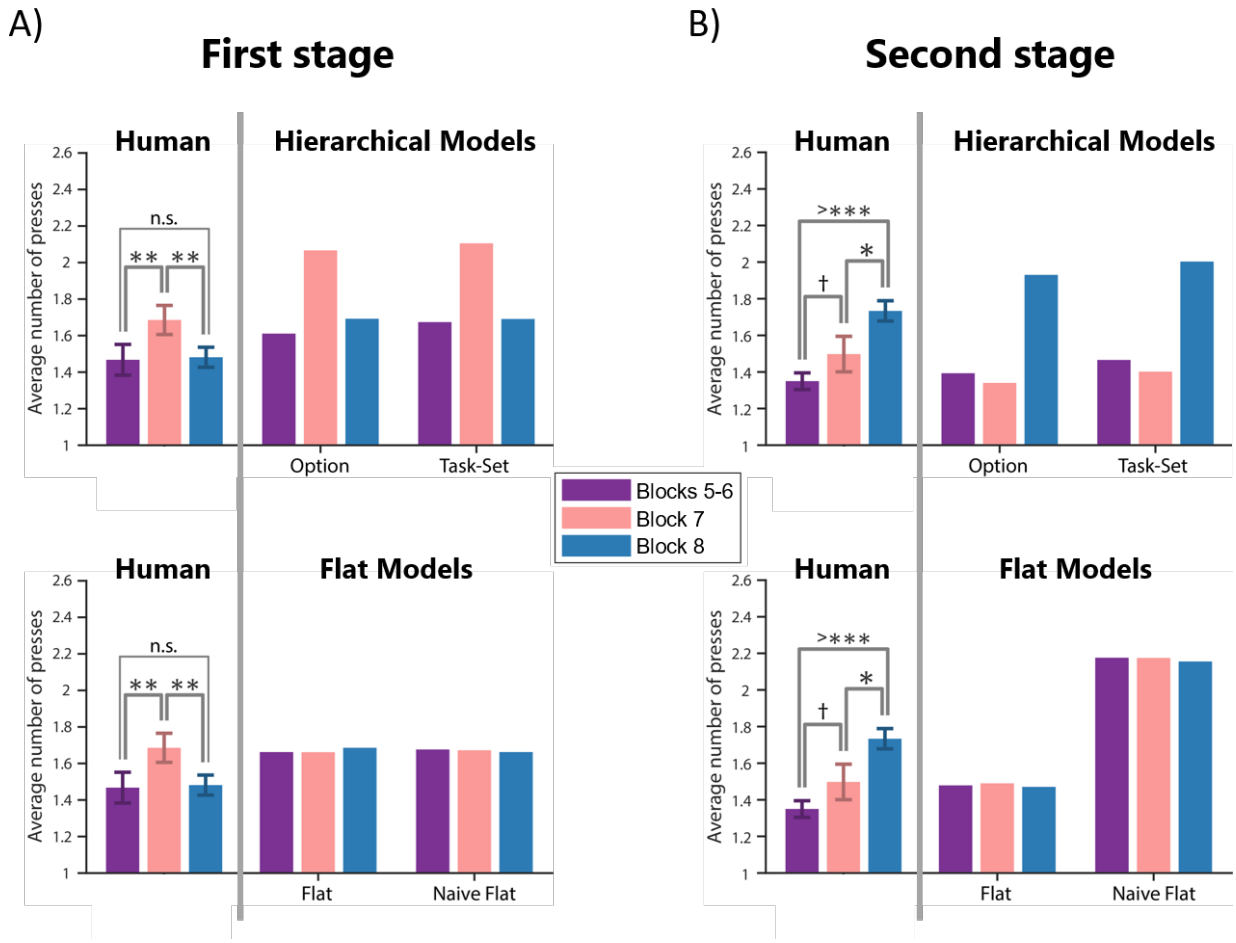

Figure 3.5: **Experiment 1 transfer effects.** Average number of first (A) and second (B) stage key presses in the first 10 trials of Block 5-8 for participants as well as model simulations. We ran 500 simulations of each hierarchical model (top) and flat model (bottom). See Table 3.2 for model parameters. Behavioral results show patterns of positive and negative transfer predicted by hierarchical, but not flat RL models, in both stages.

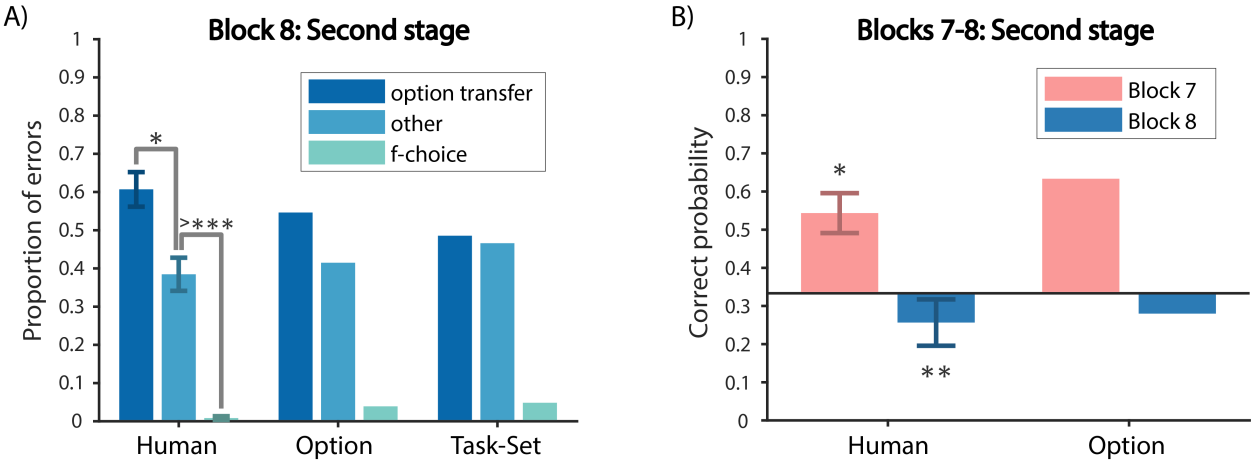

Figure 3.6: **Experiment 1 second stage choices.** (A) Error type analysis of the second stage in Block 8 for participants, the Option Model and the Task-Set Model. Participants made significantly more option transfer errors than other errors. This was predicted by the Option Model, but not by the Task-Set Model. (B) Probability of a correct first key press for the second stage of the first trial of each of the 4 branches in Blocks 7-8 reveals positive and negative transfer prior in first attempt (left), as predicted by the Option Model (right).

that participants selected previously learned MO's as a whole at the beginning of the second stage of Block 8. The Option Model could reproduce this effect because the agent selects an entire option (MO) in the first stage: not only its immediate response to the first stage stimulus, but also its policy over LO choice in the second stage. The Task-Set Model could not reproduce this effect, because the first stage choice was limited to the first stage, and the second stage did not use any choice information from the first stage. Therefore, the error type profile in Block 8 could not be accounted for by transfer of one-step task-sets alone, ruling out the Task-Set Model.

There was also more *other* type than *f-choice* errors (paired t-test, $t(24) = 8.8, p < 0.0001$). There were few *f-choice* errors, likely due to meta-learning [165]: participants observed that the correct action in the second stage was always different from the first stage (Fig 3.2B). We included a mechanism in all models to capture this heuristic and quantitatively capture behavior better.

The same choice type definitions were not well-defined for the second stage of blocks other than Block 8. Therefore, we categorized errors differently in Blocks 1-7. For example, consider the diamond following the circle in Blocks 1, 3, and 5 (Fig 3.2B): A_4 is the *correct* choice; an A_1 error corresponds to the correct choice in the first stage (*f-choice* type); an A_2 error corresponds to the correct action for the other second stage stimulus, triangle, in the same LO, thus we defined it to be the *sequence* type, because A_2 followed the first stage correct action A_1 half of the time, as opposed to the *non-sequence* action A_3, which never

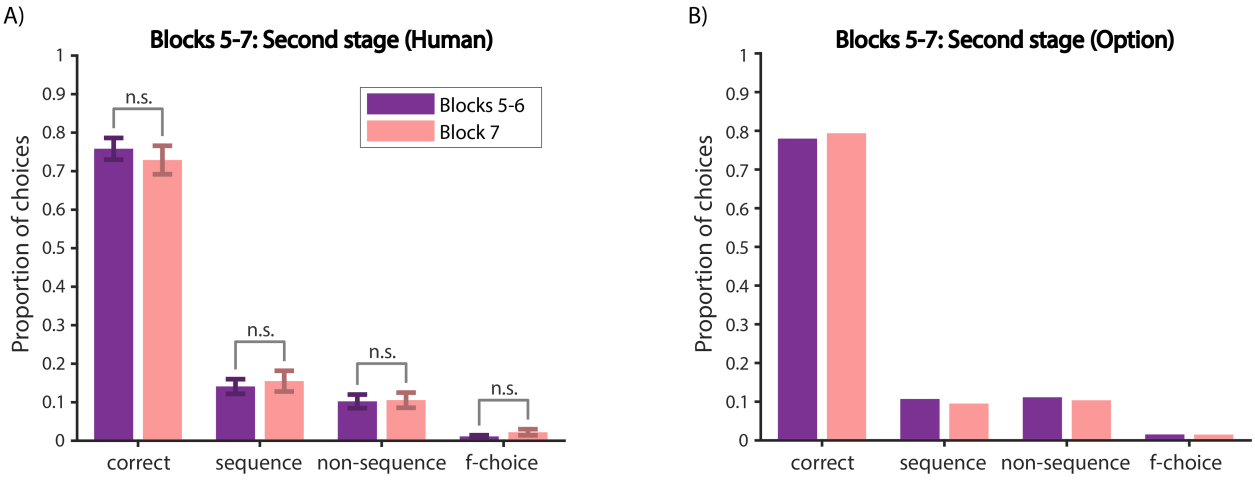

Figure 3.7: **Experiment 1 second stage choices.** Choice type analysis of the second stage comparing Blocks 5-6 and Block 7 for (A) participants and (B) the Option Model. There was no significant difference across all choice types, indicating positive transfer in the second stage of Block 7.

happened after A_1. Aggregating the first 3 trials for each of the 4 branches in the second stage of Blocks 5-7 (Fig 3.7A), we did not find any significant difference in any of the 4 choice types between the second stage of Block 7 and that of Blocks 5-6 (paired t-test, all ($t(24) \leq 1$, p's> 0.30). This indicates that the positive transfer in the second stage of Block 7 was not interfered by the negative transfer in the first stage of Block 7, further confirming that participants were selecting learned MO's as a whole, but re-composing them together into a new HO. The Option Model is also able to quantitatively capture the similarity of the choice type profiles between Block 7 and Blocks 5-6 (Fig 3.7B).

3.2.2.3 Second stage reaction time and sequence learning effects

Sequence learning [31] predicts that the reaction time of the "sequence" type to be faster than the "non-sequence" type. Therefore, we calculated the average reaction time (Fig 3.8, for both "sequence" and "non-sequence" error types

We broke down each block to 2 different time periods: early (trials 1-7 for each of the 4 branches in the second stage) and late (trials 8-15 for each of the 4 branches). Aggregating Blocks 3-6, we found a marginal effect of time period (2-way repeated measure ANOVA, $F(1, 21) = 3.0, p = 0.099$), which might be due to participants generally becoming faster as they progressed within a block. We also found a main effect of error type (2-way repeated measure ANOVA, $F(1, 21) = 4.5, p = 0.046$) on reaction time. Specifically, we found no significant difference ($t(23) = 1.3, p = 0.2$) between the reaction time of the "sequence" and "non-sequence" error types in the early time periods (Fig 3.8A). The "sequence" type

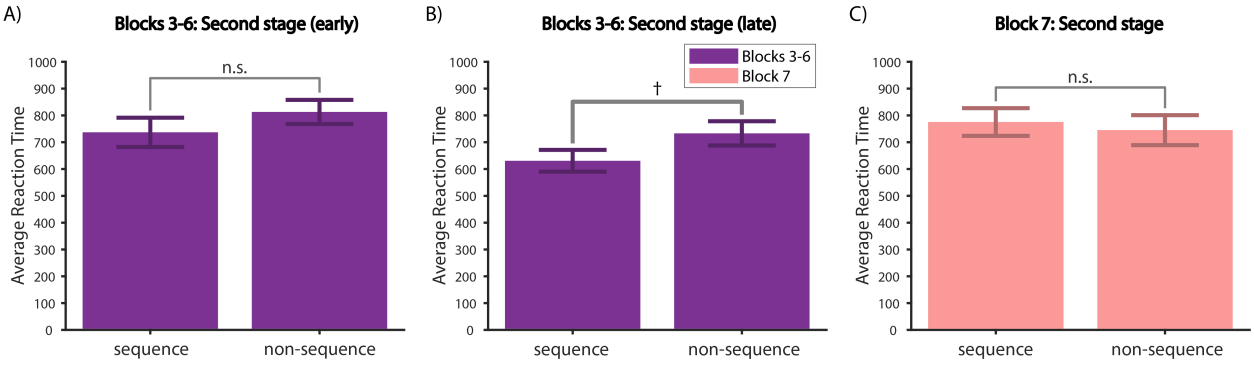

Figure 3.8: **Experiment 1 reaction time.** (A) Average reaction time for trials 1-7 for each of the 4 branches in the second stage for Blocks 3-6 for sequence (left) and non-sequence (right) error types. (B) Same as (A) for trials 8-15. (C) Average reaction time for sequence (left) and non-sequence (right) error types in the second stage of Block 7.

was marginally faster (paired t-test, $t(22) = 1.9, p = 0.072$) than the "non-sequence" type in the late time period (Fig 3.8B). We also found no significant difference (paired t-test, $t(20) = 1.1, p = 0.3$) between the "sequence" and "non-sequence" types in the entire Block 7 (Fig 3.8C). These results suggest that the transfer effects we observed at the beginning of each block could not be due to pure sequence learning, which only start to take effect during learning saturation.

3.2.2.4 The first press in the second stage reveals theoretical benefit of options

While the first several trials demonstrated transfer effects, the Option Model predicts immediate transfer effect on the first press in the second stage of a new block without any experience. Therefore, we computed the probability of a correct choice on the first press for the 4 branches in the second stage (Fig 3.6B), and compared to chance ($\frac{1}{3}$, accounting for the meta-learning effect that the correct action in the second stage was always different from the first stage). The probability of a correct first key press in Block 7 and Blocks 5-6 was significantly above chance (sign test, Block 7: $p = 0.015$; Blocks 5-6: $p < 0.0001$), without significant difference between the two (sign test, $p = 0.26$). These positive transfer effects on the first press supports our prediction that participants were using previously learned MO to guide exploration and thus speed up learning even without any experience in Blocks 5-7. Block 8 was significantly below chance (sign test, $p = 0.004$), independently indicating, via negative transfer, exploration with previously learned MO in the very first trials. The Option Model was able to quantitatively reproduce these positive and negative transfer effects evident in the first press in the second stage, since the first stage choice can immediately help inform which LO to use in the second stage.

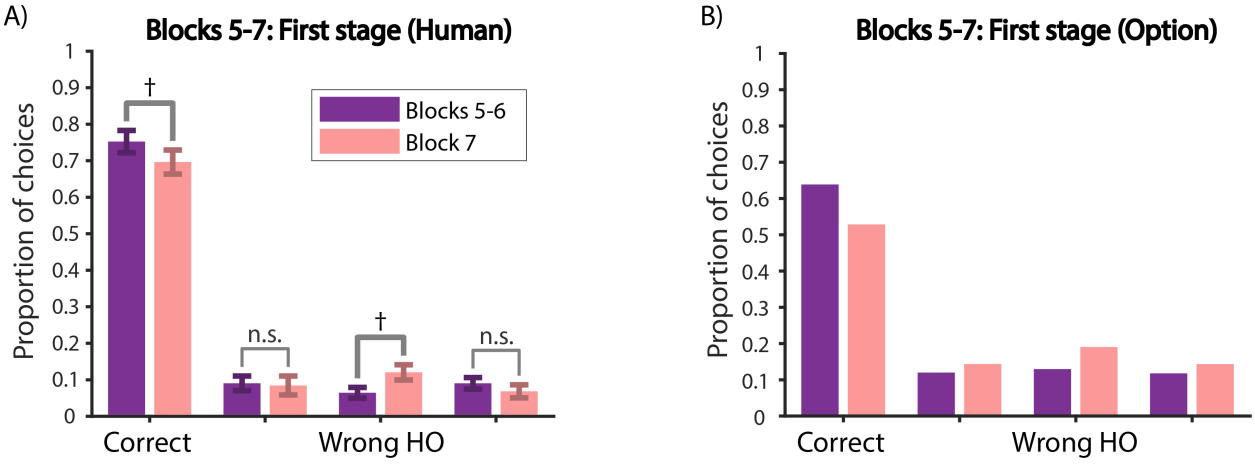

Figure 3.9: **Experiment 1 first stage choices.** Choice type analysis of the first stage in Blocks 5-7 for participants (A) and the Option Model (B). Participants made significantly more wrong *HO* errors in Block 7 than in Blocks 5-6, but no change for the other two error types. This suggests that participants were negatively transferring *HO* in the first stage of Block 7, as predicted by the Option Model.

3.2.2.5 First stage choices reveal transfer of policies over options

To test whether participants learned *HO*'s in the first stage, we investigated errors in the first stage. We hypothesized that the increase in key presses in the first stage of Block 7 (Fig 3.3B) was due to selecting a previously learned but now wrong *HO* in the first stage, which would be characterized by a specific error. We categorized first stage errors into 3 types (*wrong shape*, *wrong HO*, and *both wrong*), which we exemplify for the circle in Blocks 1, 3, and 5 (Fig 3.2B): 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, and 6 (*wrong HO* type); and A_4 is the *both wrong* type. According to our hypothesis, we expected that the worse performance in the first stage of Block 7 (Fig 3.5B) should be primarily due to the *wrong HO* errors. We found a main effect of choice type (2-way repeated measure ANOVA, $F(3, 72) = 195, p < 0.0001$) and a significant interaction between block and choice type ($F(3, 72) = 2.9, p = 0.04$). In particular, we found that in Block 7 (Fig 3.9A), compared to Blocks 5-6, only the *wrong HO* error type marginally increased (paired t-test, $t(24) = 1.9, p = 0.07$) in Block 7. The Option Model reproduced this choice type profile in the first stage(Fig 3.9B), by attempting to transfer previously learned *HO*, which would hurt performance in the first stage.

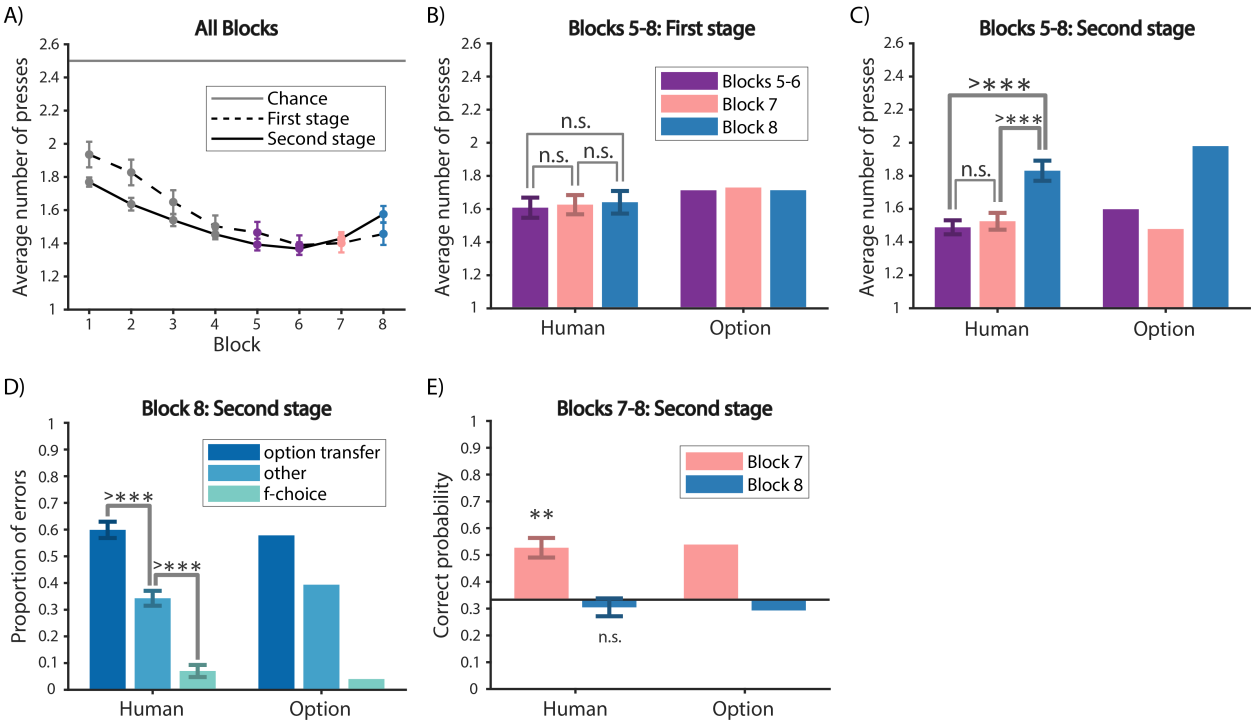

Figure 3.10: **Experiment 1 Mturk results.** (A) Average number of key presses in the first and the second stages per block. (B) Average number of key presses for the first 10 trials of Blocks 5-8 for the first stage for participants (left) and the Option Model (right). (C) Same as (B) for the second stage. (D) Error type analysis of the second stage in Block 8 for participants (left) and the Option Model (right). We replicated the same pattern as the in-lab population (Fig 3.6A). (E) Probability of a correct first key press for the second stage of the first trial of each of the 4 branches in Blocks 7-8 for participants (left) and the Option Model (right).

3.2.2.6 Experiment 1 Mturk replicates option transfer in the second stage

While in-lab participants' behavior showed promising evidence in favor of transferring multi-step options, we sought to replicate our results in a larger and more diverse population. Therefore, we ran a shorter version of Experiment 1 on Mturk (Fig 3.10A, Fig 3.11). In the second stage, we replicated the main effect of block on the number of presses (1-way repeated measure ANOVA, $F(2, 108) = 19, p < 0.0001$). Specifically, the average number of key presses (Fig 3.10C) in the first 10 trials of Block 7 was not significantly different from that of Blocks 5-6 (paired t-test, $t(54) = 0.72, p = 0.47$). Participants pressed significantly more times in Block 8 compared to Block 7 and Blocks 5-6 (paired t-test, Block 7: $t(54) = 4.5, p < 0.0001$; Blocks 5-6: $t(54) = 5.3, p < 0.0001$), replicating results from in-lab participants (Fig 3.3B).

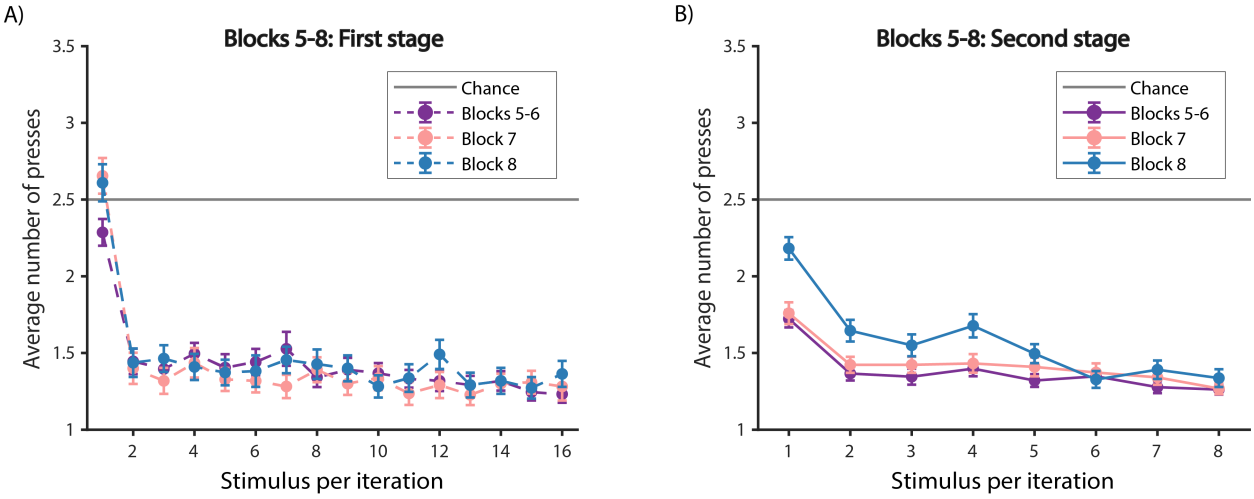

Figure 3.11: **Experiment 1 performance within Blocks 5-8 for Mturk participants.** (A) First stage. (B) Second stage.

In the second stage of Block 8 (Fig 3.10D), there was a main effect of error type (1-way repeated measure ANOVA, $F(2, 108) = 62, p < 0.0001$). The *option transfer* errors were significantly more frequent than the *other* type errors (paired t-test, $t(54) = 4.7, p < 0.0001$), and the *other* type was significantly more frequent than the *f-choice* type (paired t-test, $t(54) = 6.7, p < 0.0001$). This also replicates the error type profile of in-lab participants.

For the probability of correct choice in the first press (Fig 3.10E), we also found participants were performing significantly above chance in the second stage of Blocks 3-4, Blocks 5-6 and Block 7 (sign test, Blocks 3-4: $p = 0.001$; Blocks 5-6: $p = 0.003$; Block 7: $p = 0.001$), but not significantly different from chance in Block 8 (sign test, $p = 0.18$). There was also no significant difference between Block 7 and Blocks 5-6 (sign test, $p = 1$). This supported the previous finding that participants used temporally extended MOs to explore in a new context.

We did not replicate the negative transfer in the first stage of Block 7 (Fig 3.10B) shown in in-lab participants (Fig 3.3B). There was no main effect of block on the number of presses (1-way repeated measure ANOVA, $F(2, 108) = 0.19, p = 0.83$). Mturk participants did not press significantly more times in the first stage of Block 7 than Block 8 or Blocks 5-6 (paired t-test, Block 7: $t(54) = 0.30, p = 0.77$; Blocks 5-6: $t(54) = 0.32, p = 0.75$). This is potentially due to the lack of motivation among Mturk participants to exploit structure in the first stage, since participants did not receive points for being correct in the first stage. On the other hand, participants received points for choices in the second stage, which, as indicated by the Mturk experiment instruction, would impact their bonus. This might explain why the transfer effects in the first stage did not replicate, but the second stage transfer did. Note that in this case, the absence of transfer allowed the Mturk participants to make fewer errors

in Block 7 than they might otherwise, highlighting the fact that engaging in a cognitive task and building and using structure is not always beneficial.

The option model was able to account for Experiment 1 Mturk data, despite the lack of transfer in the first stage, by assuming either a faster forgetting of HOs (higher f^1) or a lower prior for reusing them (higher γ^1) (Table 3.2). Indeed, simulations reproduced the lack of transfer in the first stage (Fig 3.10B), and also captured all option transfer effects demonstrated by Mturk participants in the second stage(Fig 3.10C-E).

We conclude that, in the Mturk sample, similar to the in-lab sample, we successfully replicated the main option transfer effects in the second stage due to selecting a temporally extended policy MO as a whole. This is reflected by number of presses, proportion of error types in Block 8, and the probability of correct choice in the first press (Fig 3.10C-E). While we did not replicate transfer of high level-options (task-sets of options), this could be accommodated by the model, and understood as a lack of motivation at learning the highest level of hierarchy HO.

3.3 Experiment 2

Experiment 2 was administered to UC Berkeley undergraduates in exchange for course credit. 31 (21 females; age: mean = 20.2, sd = 1.8, min = 18.3, max = 26.3) UC Berkeley undergraduates participated in Experiment 2. 4 participants in Experiment 2 were excluded due to incomplete data or below chance performance, resulting in 26 participants for data analysis.

3.3.1 Experiment 2 Protocol

Experiment 1's Block 8 comes after a first testing block that includes re-composing of previous options, which could interfere with our interpretation of positive and negative transfer results in Block 8, for example by making participants aware of the potential for structure transfer. In Experiment 2, we removed Block 7 of Experiment 1 to eliminate this potential interference (Fig 3.12). Therefore, Block 7 in Experiment 2 was identical to Block 8 in Experiment 1. In addition, to limit experiment length and loss of motivation at asymptote in each block, we decreased the length of Blocks 3-7 to 32 trials each, with each first stage stimulus leading to each second stage stimulus 8 times. All other aspects were identical to Experiment 1.

3.3.2 Experiment 2 Results

3.3.2.1 Second stage choices replicate option transfer

Participants were able to learn the correct actions in both the first and second stages and their performance improved over Blocks 1-6, (Fig 3.13A). The within-block learning curves also showed that participants performance improved and then reached asymptote as they progressed within a block (Fig 3.14).

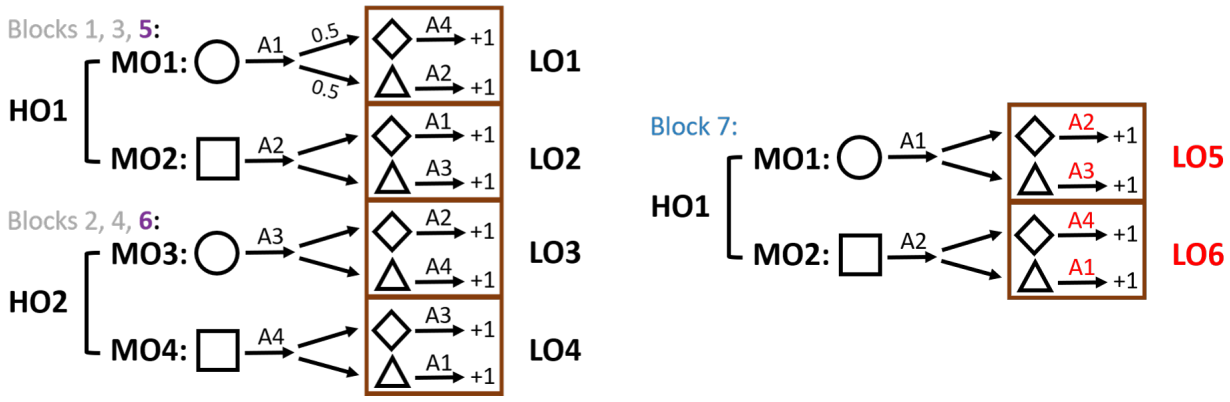

Figure 3.12: **Experiment 2 protocol.** To eliminate potential interference of Block 7 on Block 8 in Experiment 1, Block 7 of Experiment 1 was removed in Experiment 2. Therefore, Block 7 in Experiment 2 was identical to Block 8 in Experiment 1.

We replicated the negative transfer effects in the second stage of Experiment 1 (Fig 3.3B) both in terms of number of presses (Fig 3.13C) and error types in Block 7 (Fig 3.13D). Participants pressed significantly more times in the second stage of Block 7 compared to Blocks 5-6 (paired t-test, $t(25) = 6.4, p < 0.0001$). In Block 7 specifically, there was a main effect of error type (1-way repeated measure ANOVA, $F(2, 50) = 30, p < 0.0001$). The proportion of the error type *option transfer* was significantly higher than the error type *other* (paired t-test, $t(25) = 3.2, p = 0.004$).

We also observed transfer effects on the first press in the second stage (Fig 3.13E). We found that the probability of a correct choice was significantly above chance in Blocks 3-4 and Blocks 5-6 (sign test, Blocks 3-4: $p = 0.0094$; Blocs 5-6: $p < 0.0001$), and significantly below chance in Block 7 (sign test, $p < 0.0001$). This replicates results in Blocks 3-6 and 8 in Experiment 1 (Fig 3.6B). The Option Model could quantitatively reproduce all these transfer effects (Fig 3.13B-D).

We also analyzed the reaction time (Fig 3.15) of the "sequence" and "non-sequence" error types in Blocks 5-6 in Experiment 2. As in Experiment 1 (Sec 3.2.2.3), we broke down each block into 2 halved time periods: early (trials 1-4 for each of the 4 branches in the second stage) and late (trials 5-8 for each of the 4 branches). We found a main effect of time period and error type, and a significant interaction (2-way repeated measure ANOVA, time period: $F(1, 16) = 8, p = 0.012$; error type: $F(1, 16) = 16, p = 0.0009$; interaction: $F(1, 16) = 15, p = 0.0013$). Specifically, there was no significant difference (Fig 3.15A) between the reaction time of the "sequence" and "non-sequence" types in the early time period (paired t-test, $t(21) = 0.61, p = 0.55$). However, the "sequence" type was significantly faster (Fig 3.15B) than the "non-sequence" type in the late period (paired t-test, $t(17) = 4.8, p = 0.0002$). These results replicated the trend observed in the second stage of Experiment 1 (Fig 3.8): sequence learning might take effect during learning saturation, but

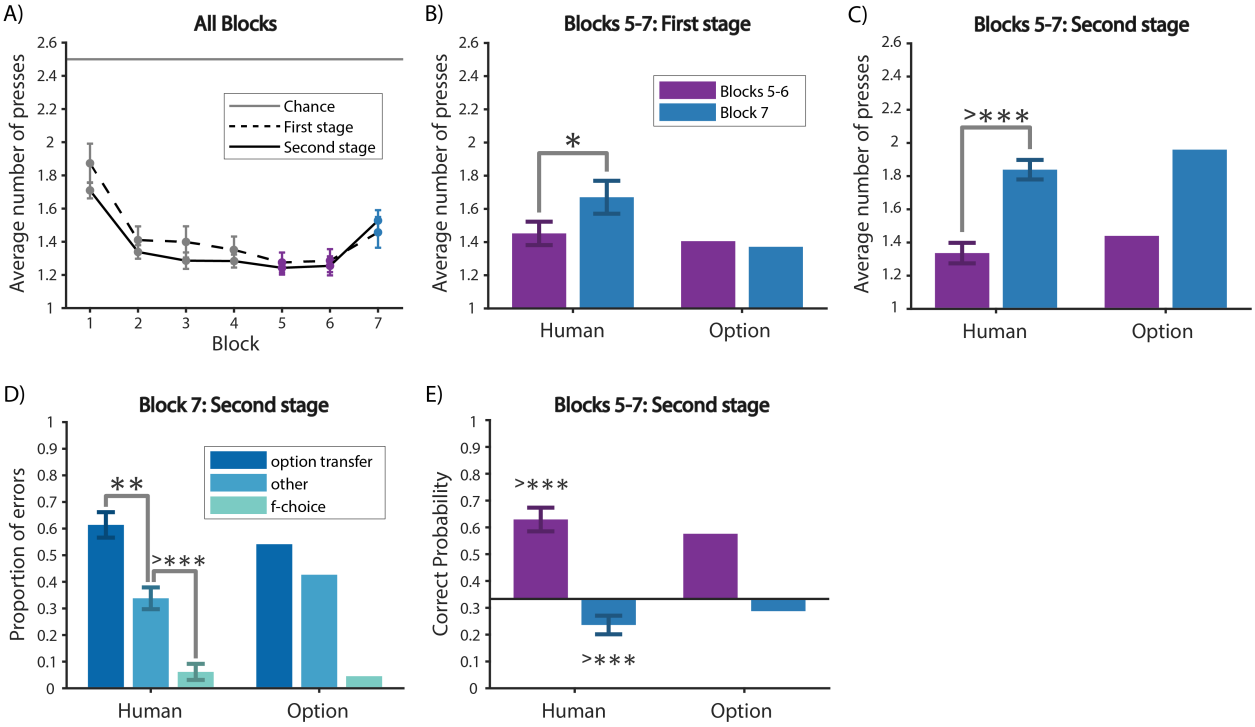

Figure 3.13: **Experiment 2 results.** (A) Average number of key presses in the first and the second stages per block. (B, C) Average number of key presses for the first 10 trials of Blocks 5-7 for the first (B) and second (C) stage for participants (left) and the Option Model (right). (D) Error type analysis of the second stage in Block 7 for participants (left) and the Option Model (right). We replicated the same pattern as in Block 8 of Experiment 1 (Fig 3.6A, Fig 3.10D). (E) Probability of a correct first key press for the second stage of the first trial of each of the 4 branches in Blocks 5-7 for participants (left) and the Option Model (right).

not the beginning of blocks, where we typically expect to observe transfer effects.

3.3.2.2 Second stage choices in Block 7 reveal interaction between meta-learning and option transfer

Because there was no Block 7 from Experiment 1, we had a less interfered test of negative transfer in the second stage of Block 7 of Experiment 2. Therefore, we further broke down the second stage choice types for each of the 4 branches in the second stage of Block 7 in Experiment 2 (Fig 3.16A). Consider (Fig 3.2B) the two first stage stimuli as F_1 (circle) and F_2 (square), and the two second stage stimuli as S_1 (diamond) and S_2 (triangle). We found a main effect of error type on proportion of errors and a marginally significant interaction between branch and error type (2-way repeated measure ANOVA, error type: $F(2, 36) =$

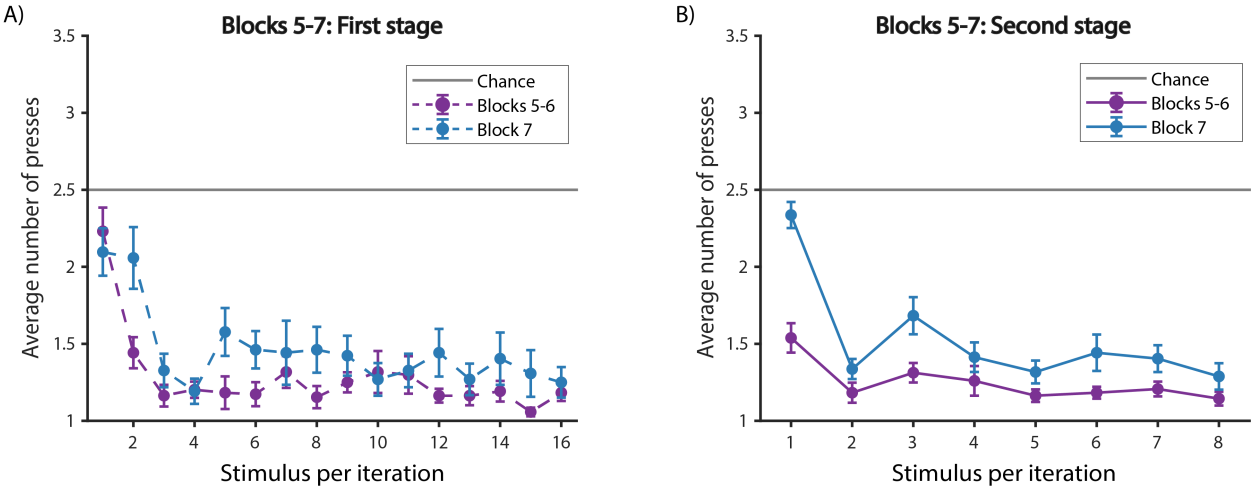

Figure 3.14: **Experiment 2 performance within Blocks 5-7.** (A) First stage. (B) Second stage.

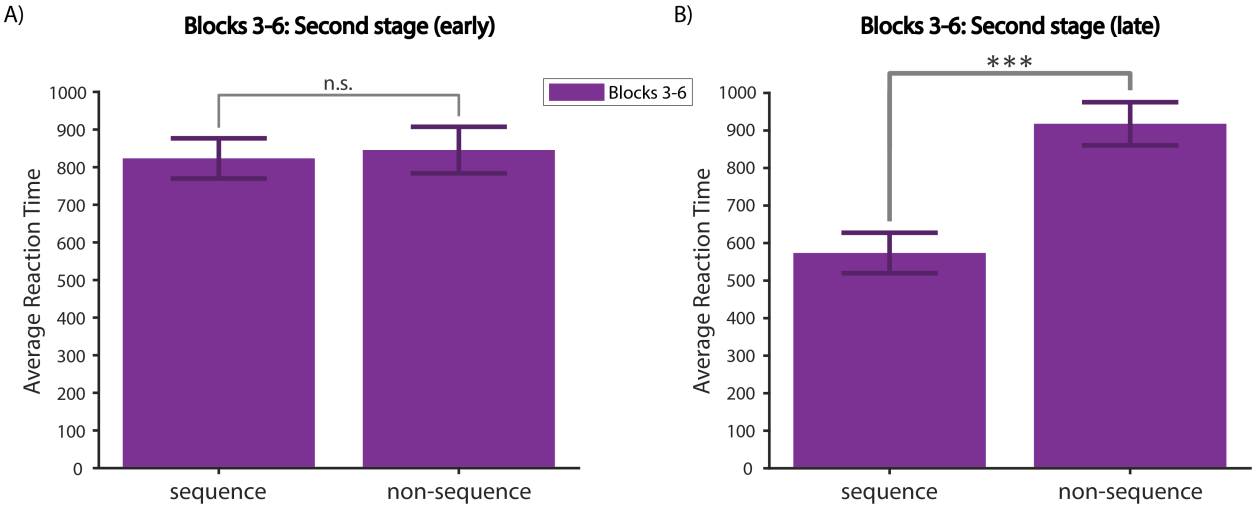

Figure 3.15: **Experiment 2 reaction time.** (A) Average reaction time for trials 1-4 for each of the 4 branches in the second stage for Blocks 3-6 for sequence (left) and non-sequence (right) error types. (B) Same as (A) for trials 5-8.

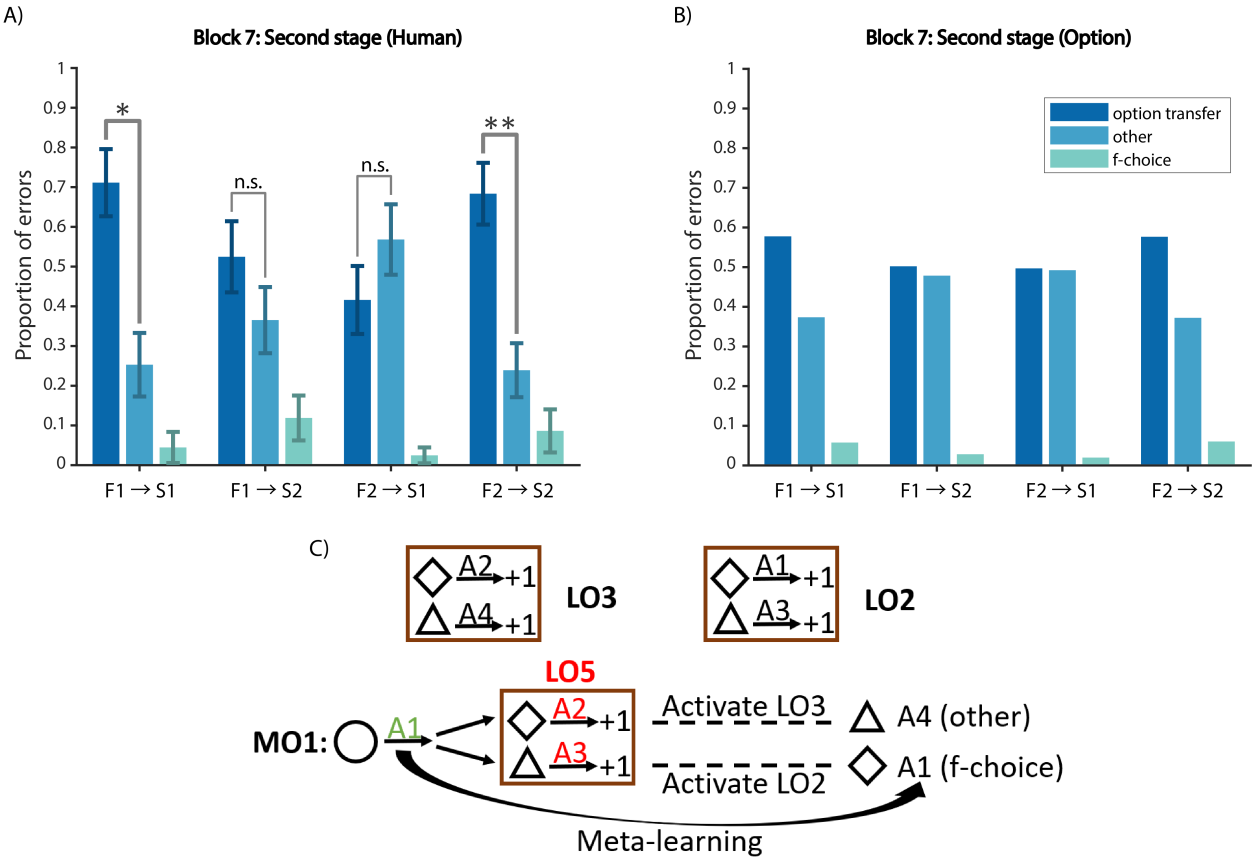

Figure 3.16: **Experiment 2 second stage choice shows interaction between option transfer and meta learning.** Error type analysis for each of the 4 branches in the second stage of Block 8 for participants (A) and the Option Model (B). The option transfer error was more than other error only for $F_1 \to S_1$ and $F_2 \to S_2$, which was predicted by the Option Model. (C) Example schematic for the interaction: learning A_2 for the diamond activates LO_3; learning A_3 for the triangle activates LO_2; meta-learning only suppresses LO_2 but not LO_3.

$20, p < 0.0001$; interaction: $F(6, 108) = 2.1, p = 0.055$). Specifically, we found the error type profile in Fig 3.13C was mainly contributed by $F_1 \to S_1$, i.e. circle in the first stage followed by diamond in the second stage, and $F_2 \to S_2$ (paired t-test, $F_1 \to S_1$: $t(23) = 2.7, p = 0.013$; $F_2 \to S_2$: $t(23) = 3.1, p = 0.005$). On the other hand, there was no significant difference between the *option transfer* and *other* error types for $F_1 \to S_2$ and $F_2 \to S_1$ (paired t-test, $F_1 \to S_2$: $t(22) = 0.9, p = 0.38$; $F_2 \to S_1$: $t(22) = 0.81, p = 0.43$). It is striking that this highly non-intuitive result is perfectly predicted by the Option Model (Fig 3.16B).

The Option Model offers an explanation as the interaction between option transfer and

meta-learning (Fig 3.16C). Meta-learning discourages participants from selecting second-stage actions that repeat the correct first-stage action, and as such, discourage them from sampling some, but not other LOs (e.g. LO_2 in the example of Fig 3.16C). This interference in the exploration of potential LO's leads to some transfer errors to be more likely, in an asymmetrical way.

3.3.2.3 Influence of the second stage on the first stage

For the first stage choices (Fig 3.13B), we found that participants pressed significantly more times in the first 10 trials of Block 7 compared to Blocks 5-6 (paired t-test, $t(25) = 2.4, p = 0.024$). This effect was not found in Experiment 1 between Block 8 and Blocks 5-6 (Fig 3.3B), and was not predicted by the model.

One potential explanation for this surprising result is that the error signals in the second stage propagated back to the first stage. Specifically, the errors participants made by selecting the wrong LO in the second stage are credited to the chosen LO's policy, but participants might also credit these errors to using the wrong HO in the first stage. Going back to our example, if your meal is not tasty, it might not be because you roasted the potatoes instead of boiling them, but it might be because you needed vegetables instead of potatoes in the first place. To test this explanation, we further probed choice types in the first stage of Experiment 2 (Fig 3.17). Indeed, we found significantly more *wrong HO* errors in Block 7, compared to Blocks 5-6 (paired t-test, $p = 0.045$). Therefore, the increase in number of key presses in the first stage of Block 7 was mainly contributed by more *wrong HO* errors, indicating that participants explored another high level option (cooking vegetables). The same effect was not seen in the first stage of Experiment 1 between Block 8 and Blocks 5-6 (Fig 3.3B), potentially due to the interference of Block 7 in Experiment 1.

The Option Model could not capture this effect, since the selection of HO was only affected by learning in the first stage (Sec 3.2.1.5), as a way of simplifying credit assignment (see Sec 3.7.3 for a more detailed discussion on credit assignment).

3.4 Experiment 3

Experiment 3 was administered to UC Berkeley undergraduates in exchange for course credit. 35 (22 females; age: mean = 20.5, sd = 2.5, min = 18, max = 30) UC Berkeley undergraduates participated in Experiment 3. 10 participants in Experiment 3 were excluded due to incomplete data or below chance performance, resulting in 25 participants for data analysis.

An additional 65 (37 female; see age range distribution in Table 3.1) Mturk participants finished the experiment. 34 participants were further excluded due to poor performance, resulting in 31 participants for data analysis (Sec 3.2.1.4).

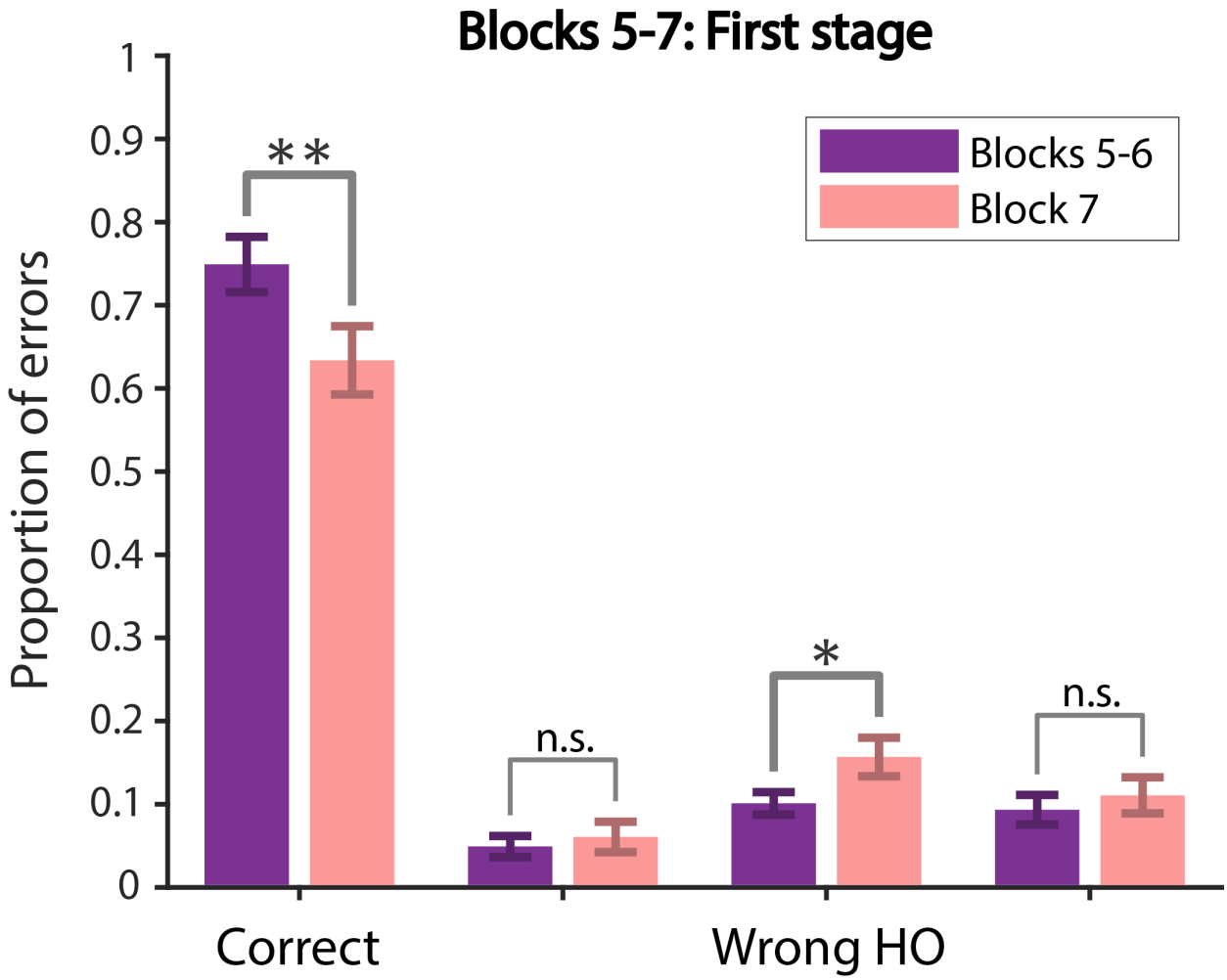

Figure 3.17: **Experiment 2 first stage choices.** Choice type analysis of the first stage comparing Blocks 5-6 and Block 7. The only error type that significantly increased was the wrong *HO* error, suggesting that participants were perseverating in the first stage while learning the new mappings in the second stage of Block 7.

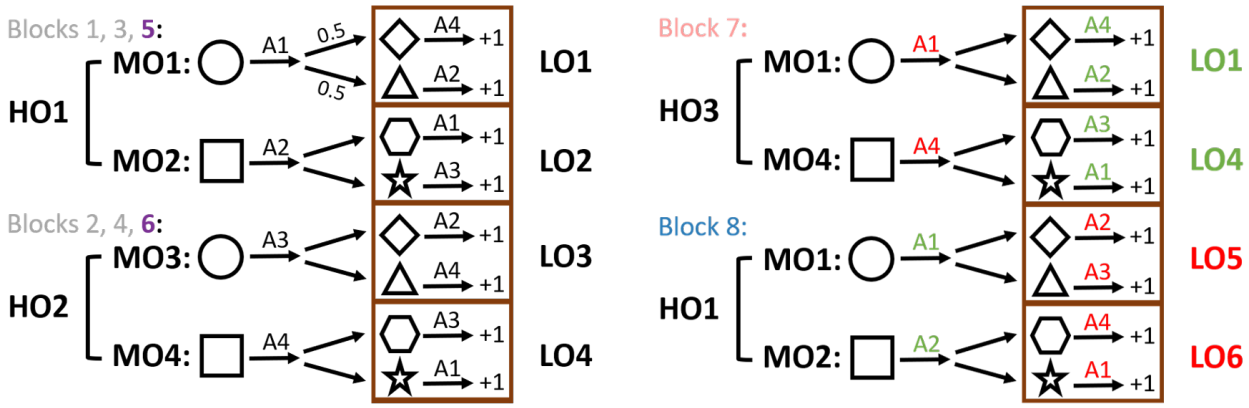

Figure 3.18: **Experiment 3 protocol.** The second stage stimuli following each first stage stimuli were different: diamond and triangle followed circle; hexagon and star followed square. All state-action assignments remained the same as Experiment 1. This manipulation allowed us to test whether participants would naturally learn and transfer options in the second stage even when they could simply learn the correct key for each of the 4 second stage stimuli individually, rather than needing to take into account first stage information.

3.4.1 Experiment 3 in-lab Protocol

In Experiment 1, to perform well in the second stage, participants had to learn option-specific policies, due to the non-Markovian nature of the task (the correct action for the same second stage stimulus was dependent on the first stage stimulus). In Experiment 3, we removed this non-Markovian feature of the protocol and tested whether the removal would reduce or eliminate option transfer. Based on previous research on task-sets showing that participants build structure when it is not needed [36, 32], we predicted that participants might still show some evidence of transfer. However, we predicted that any evidence of transfer would be weaker than in previous experiments.

In Experiment 3, the second stage stimuli following the two first stage stimuli were different (Fig 3.18). For example, diamond and triangle followed circle, whereas star and hexagon followed square. This eliminated the key non-Markovian feature from Experiment 1, since participants could simply learn the correct key for each of the 4 second stage stimuli individually without learning option-specific policies. Blocks 1 and 2 had 60 trials; we shortened Blocks 3 to 8 to 32 trials for the same reason as in Experiment 2. All other aspects of the protocol were identical to Experiment 1.

3.4.2 Experiment 3 Mturk Protocol

In the Mturk version, Blocks 1 and 2 had a minimum of 32 and a maximum of 60 trials, but participants moved on to the next block as soon as they reached a criterion of less than 1.5

key presses per second stage trial in the last 10 trials (the 31 Mturk participants included for data analysis on average used 36 (SD = 7, median = 32, min = 32, max = 60) trials in Block 1 and 35 (SD = 4, median = 32, min = 32, max = 59) trials in Block 2). Blocks 3 to 8 all had 32 trials each. Experiment 3 MTurk was thus perfectly comparable to Experiment 1 MTurk, as such, we focus first on MTurk results, since the same comparison could not be drawn between Experiments 1 and 3 for in-lab participants.

3.4.3 Experiment 3 Results

3.4.3.1 Mturk participants show reduced option transfer

Mturk participants were able to learn the correct actions in both the first and second stages, and their performance improved over Blocks 1-6, (Fig 3.19A). The within-block learning curves also showed that participants performance improved and then reached asymptote as they progressed within a block (Fig 3.20).

We first analyzed the average number of key presses in the first 10 trials of each block and stage. For the first stage (Fig 3.21A), we found no effect of block on number of presses across Blocks 5-8 ($F(2, 60) = 0.13, p = 0.88$), as in Experiment 1 MTurk. For the critical second stage (Fig 3.19B), there was a main effect of Block ($F(2, 60) = 3.3, p = 0.043$). Specifically, there was no significant difference between Block 7 and Blocks 5-6 (paired t-test, $t(30) = 0.25, p = 0.81$). Participants pressed significantly more times in Block 8 than in Block 7 and Blocks 5-6 (paired t-test, Block 7: $t(30) = 2.1, p = 0.048$; Blocks 5-6: $t(30) = 2.2, p = 0.036$).

The negative transfer effect observed in the first stage of Block 7 in Experiment 1 (Fig 3.5A) was not present here in Experiment 3 (Fig 3.19). In addition to the fact that the first stage was never explicitly rewarded, as in Experiment 1, participants in Experiment 3 were even less motivated to exploit structure in the first stage. This is because the first stage in Experiment 3 was not necessary for resolving the second stage actions (Fig 3.18), while the non-Markovian aspect of Experiment 1 (Fig 3.2B) forced participants to incorporate first stage information to resolve the correct choice for the second stage.

We calculated the proportion of error types in the second stage of Block 8 (Fig 3.19C). Unlike in Experiment 1, we did not observe significantly more *option transfer* error than *other* error (paired t-test, $t(30) = 1.6, p = 0.11$). This choice type profile, compared to that in Experiment 1 and Experiment 2 (Fig 3.6A, Fig 3.10D, Fig 3.13D) suggests reduced option transfer in the second stage.

We also calculated the probability of a correct second stage first press for each of the 4 branches in the second stage (Fig 3.19D). The probability was significantly above chance in Blocks 3-4 and Blocks 5-6 (sign test, Blocks 3-4: $p = 0.0002$; Blocks 5-6: $p < 0.0001$). It was marginally above chance in Block 7 (sign test, $p = 0.07$) and not significantly different from chance in Block 8 (sign test, $p = 1$). Compared to the results in Experiment 1 (Fig 3.6B, Fig 3.10E). These results suggest participants were still taking advantage of pre-

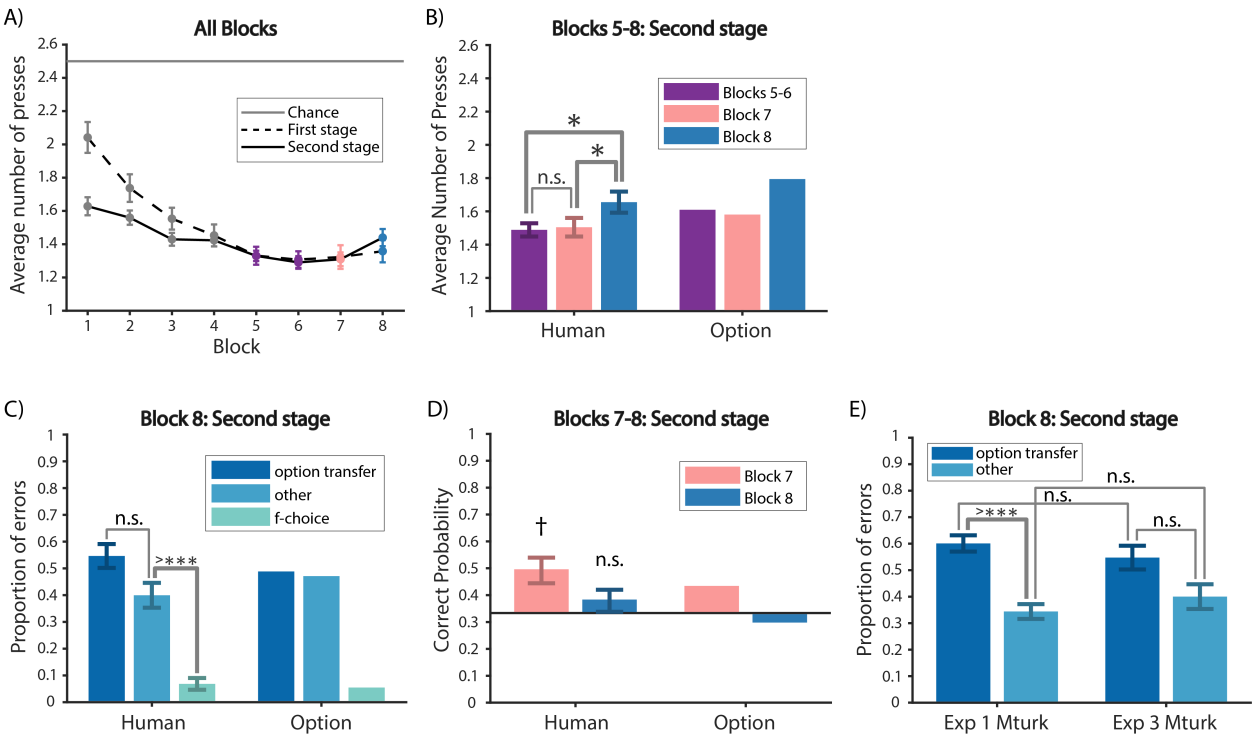

Figure 3.19: **Experiment 3 Mturk results.** (A) Average number of key presses in the first and the second stages per block. (B) Average number of key presses for the first 10 trials of Blocks 5-8 for the second stage for participants (left) and the Option Model (right). (C) Error type analysis of the second stage in Block 8 for participants (left) and the Option Model (right). The proportion of option transfer error was not significantly different from other error, different from Experiment 1 and Experiment 2, suggesting reduced option transfer. (D) Probability of a correct first key press for the second stage of the first trial of each of the 4 branches in Blocks 7-8 for participants (left) and the Option Model (right). (E) Comparison of Experiment 1 Mturk and Experiment 3 Mturk participants in terms of error types in the second stage of Block 8: There was no significant effect of experimental condition.

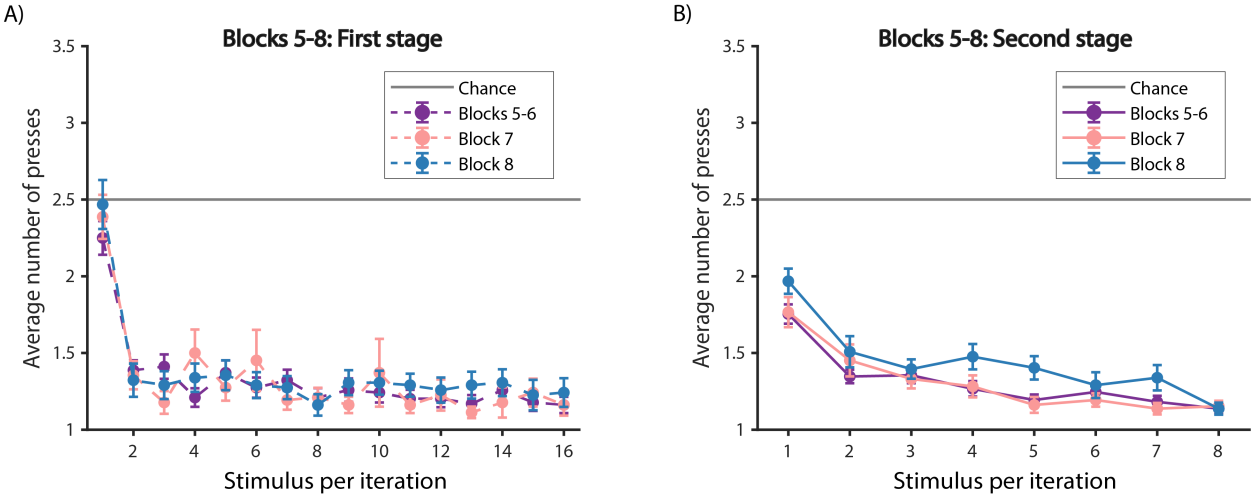

Figure 3.20: **Experiment 3 performance within Blocks 5-8 for Mturk participants.** (A) First stage. (B) Second stage.

viously learned options to speed up learning at the beginning of each block, but potentially to a lesser extent compared to Experiment 1 and Experiment 2.

To formally quantify the effect of the experimental manipulation, we compared Experiment 1 and Experiment 3 for Mturk participants. In particular, we compared the proportion of *option transfer* and *other* error types in the second stage of Block 8 between the two experiments (Fig 3.19E). We found a main effect of error type (2-way mixed ANOVA, $F(2, 168) = 76, p < 0.0001$), but there was no interaction between experiment and error type (2-way mixed ANOVA, $F(2, 168) = 0.89, p = 0.41$). In particular, the proportion of *option transfer* error type was not significantly higher in Experiment 1, compared to that in Experiment 3 (unpaired t-test, $t(84) = 1, p = 0.32$). This further shows that while there might be reduced option transfer in the second stage of Block 8 based on the error type profile (Fig 3.19C), we could not rule out option transfer in Experiment 3.

The Option Model could capture a reduction in option transfer (Fig 3.19B-D), with an increase in the second stage clustering coefficient γ^2, which controls how likely the model is to select a new blank policy compared to previously learned ones in the second stage, as well as the forgetting parameter in the second stage, f^2, which increases the speed at which the model forgets previously learned LO (Table 3.2.

3.4.3.2 In-lab participants replicate results from Mturk participants

In-lab participants replicated all aforementioned trends shown in Mturk participants (Fig 3.22, Fig 3.23). In particular, there was a main effect of block on number of choices in the second stage ($F(2, 46) = 7.2, p = 0.002$). In-lab participants also pressed significantly more times

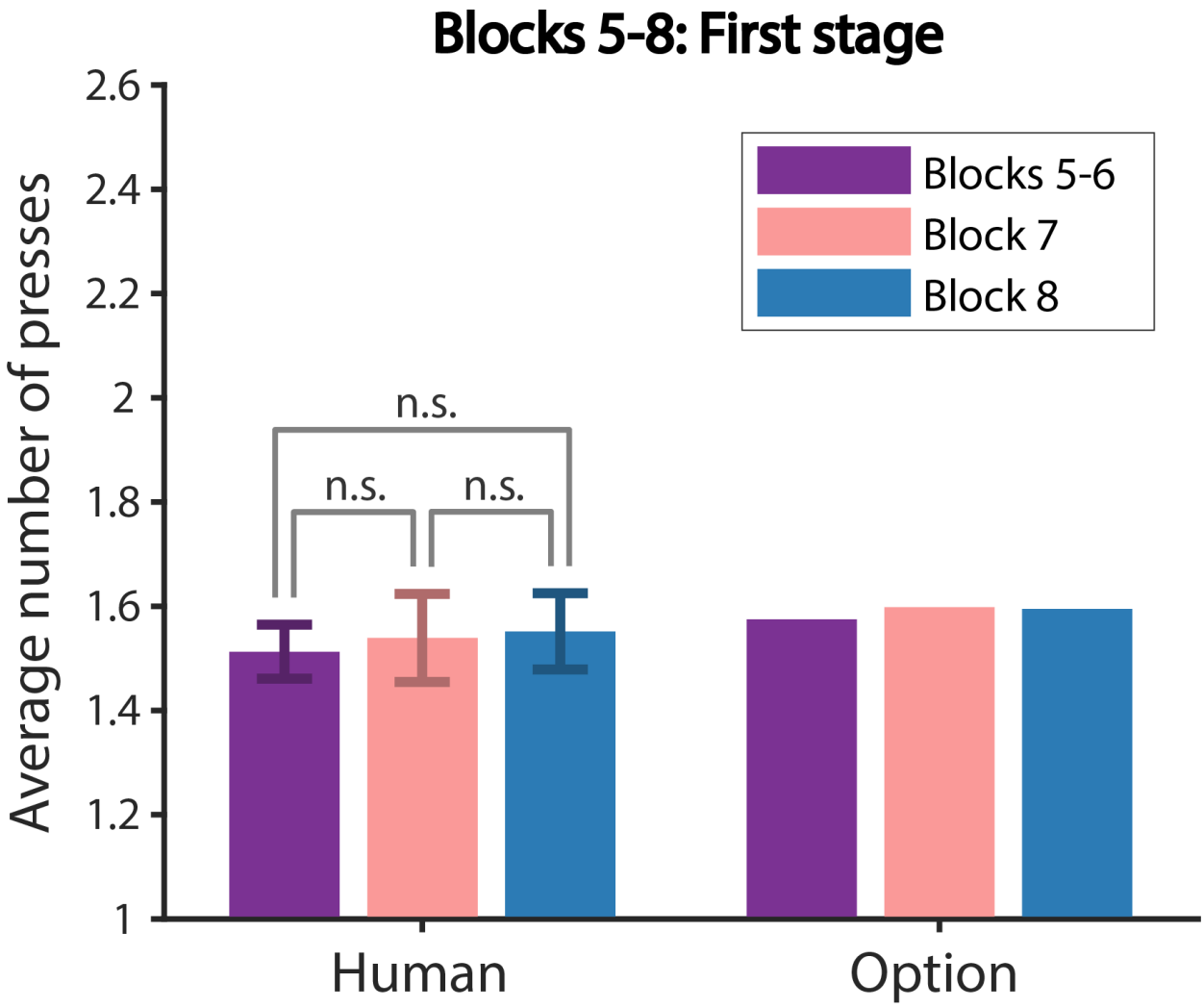

Figure 3.21: **Experiment 3 Mturk first stage choices.** Average number of presses in the first 10 trials of Blocks 5-8 in the first stage for participants (left) and the Option Model (right). This shows a lack of transfer in the first stage, representative of Experiments 3-4 first stage for both in-lab and Mturk populations.

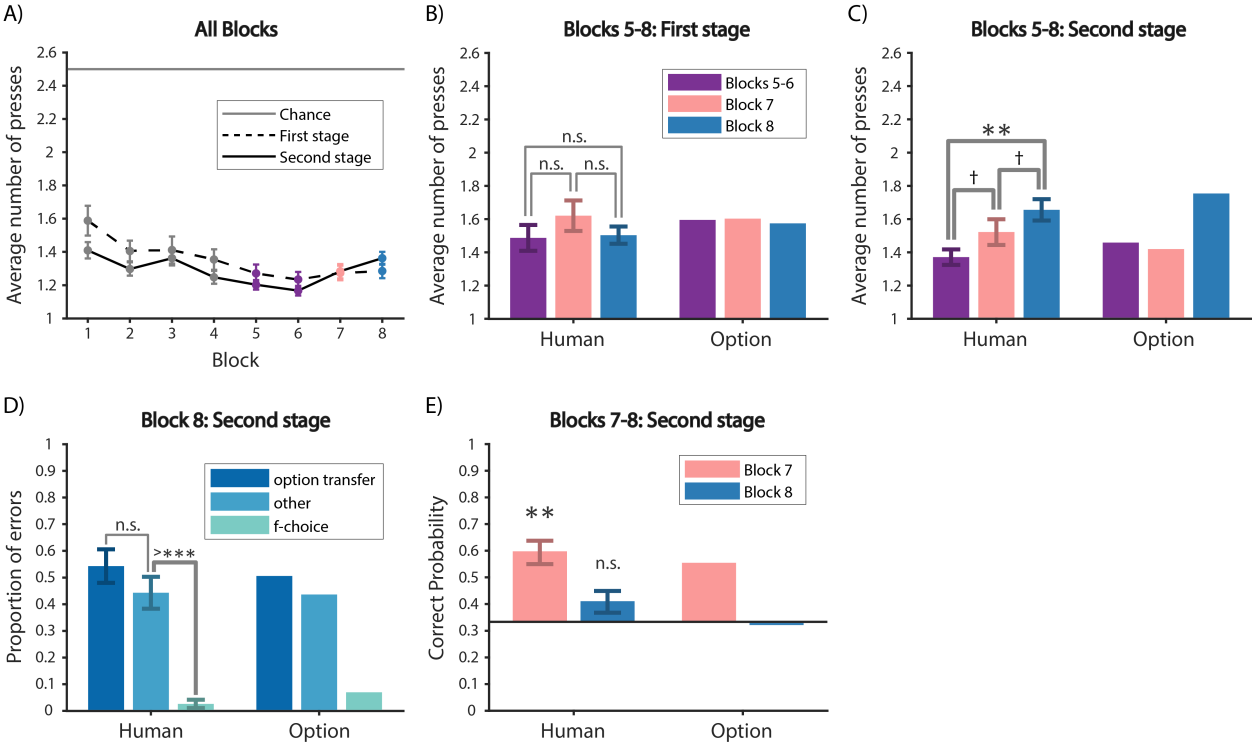

Figure 3.22: **Experiment 3 summary.** (A) Average number of key presses in the first and the second stages per block. (B) Average number of key presses for the first 10 trials of Blocks 5-8 for the first stage for participants (left) and the Option Model (right). (C) Same as (B) for the second stage. (D) Error type analysis of the second stage in Block 8 for participants (left) and the Option Model (right). The proportion of option transfer error was not significantly different from other error, different from Experiment 1 and Experiment 2, suggesting reduced option transfer. (E) Probability of a correct first key press for the second stage of the first trial of each of the 4 branches in Blocks 7-8 for participants (left) and the Option Model (right).

in the second stage of Block 8 than Blocks 5-6 (paired t-test, $t(23) = 3.6, p = 0.0017$), and marginally more than Block 7 (paired t-test, $t(23) = 1.9, p = 0.067$). Moreover, similar to Mturk participants, the proportion of *option transfer* error type was not significantly different from *other* error type (paired t-test, $t(23) = 0.8, p = 0.43$). These results replicated reduced option transfer in the second stage in a separate in-lab population. Note that we could not do the same comparison between Experiment 1 and Experiment 3 for in-lab participants, because the number of trials per block for Experiment 1 and Experiment 3 was different in-lab.

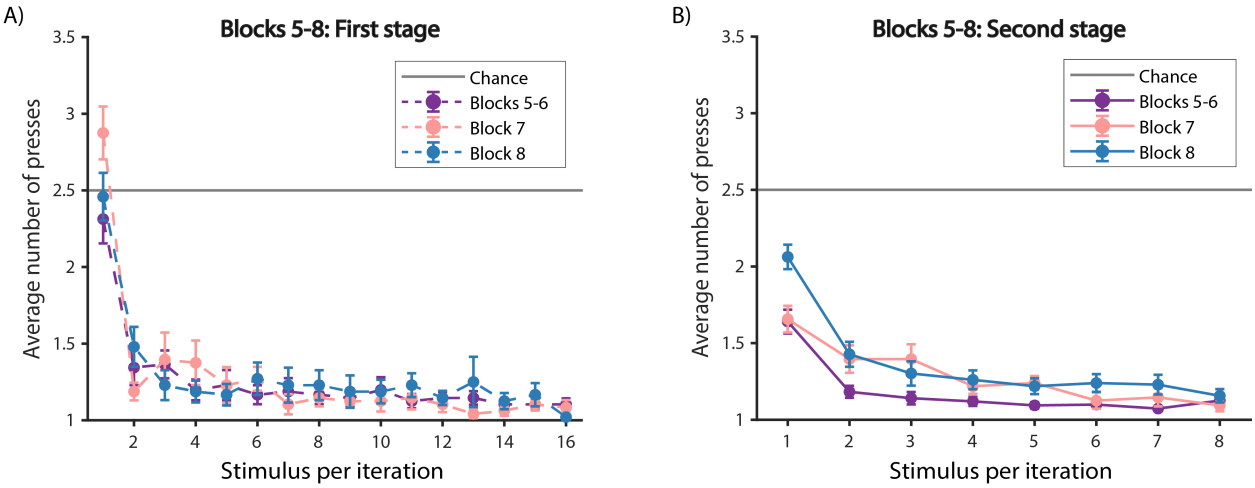

Figure 3.23: **Experiment 3 performance within Blocks 5-8 for in-lab participants.** (A) First stage. (B) Second stage.

3.5 Experiment 4

Experiment 4 was administered to UC Berkeley undergraduates in exchange for course credit. 31 (23 females; age: mean = 20.2, sd = 1.4, min = 18, max = 23) UC Berkeley undergraduates participated in Experiment 4. 12 participants were excluded due to incomplete data or below chance performance, resulting in 19 participants for data analysis.

An additional 110 (50 females; see age range distribution in Table 3.1) Mturk participants finished the experiment. 49 participants were excluded due to poor performance, resulting in 61 participants for data analysis (Sec 3.2.1.4).

3.5.1 Experiment 4 in-lab Protocol

Experiment 4 (Fig 3.24) was designed to test whether participants were able to compose options learned at different levels. Specifically, the protocol was identical to Experiment 1, except for Blocks 7 and 8. Block 8 in Experiment 4 was similar to Block 8 in Experiment 1, introducing two new LO's (LO_{new}) at the second stage as a benchmark for pure negative transfer.

The main difference between Experiment 4 and Experiment 1 was Block 7. In Block 7, one of the first stage stimuli (e.g. square) elicited the same extended policy MO_2 (A_2 followed by LO_2 in the second stage), allowing positive MO transfer (*match* condition LO_{match}). In contrast, the other first stage stimulus (e.g. circle) elicited a new policy recomposed of old subpolicies: participants needed to combine what they learned in the first stage of MO_1 in Blocks 1, 3, and 5 (A_1) (allowing for first stage transfer of HO_1), and the second stage of Blocks 2, 4, and 6 (LO_3; *mismatch* condition $LO_{mismatch}$). Extending the food analogy, in

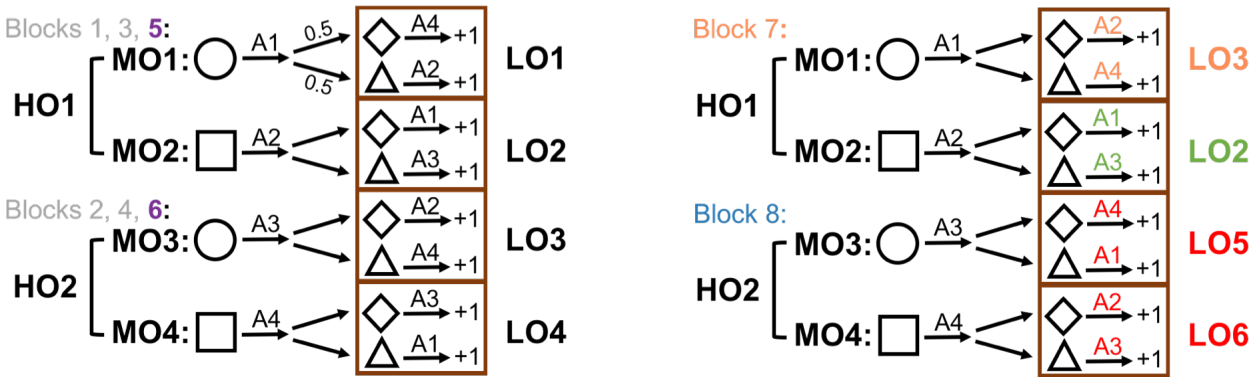

Figure 3.24: **Experiment 4 protocol.** In Experiment 4, we tested participants' ability to recompose LO policies within MO policies. Blocks 1-6 were identical to Experiment 1. In Block 7, green indicates positions of potential positive transfer: MO_2 followed by LO_2 was learned in Blocks 1, 3, 5. Orange indicates positions of option composition: although MO_1 previously included LO_1 for second stage stimuli, it was modified to LO_3 in Block 7. In Block 8, red indicates positions of negative transfer: LO_5 and LO_6 were completely novel to participants. Blocks were color coded for later analysis: Blocks 1-4 gray; Blocks 5-6 purple; Block 7 orange; Block 8 blue.

Blocks 1, 3, 5, participants learned to make potatoes (MO_1) by cutting potatoes (the first stage) and then roasting (LO_1). In Block 7, participants also needed to cut potatoes, but then steam them (LO_3), which was already learned as part of MO_3 (make vegetables) in Blocks 2, 4, 6. All blocks had 60 trials each.

3.5.1.1 Experiment 4 Mturk Protocol

The Mturk version was shortened for online workers. Blocks 1 and 2 had a minimum of 32 and a maximum of 60 trials, but participants moved on to the next block as soon as they reached a criterion of less than 1.5 key presses per second stage trial in the last 10 trials (the 61 Mturk participants included for data analysis on average used 46 (SD = 11, median = 42, min = 32, max = 60) trials in Block 1 and 43 (SD = 11, median = 38, min = 32, max = 60) trials in Block 2). All other blocks had 32 trials each.

3.5.2 Experiment 4 Results

3.5.2.1 Mismatch impacted performance of in-lab participants

Participants' performance improved over Blocks 1-6 (Fig 3.25A) and within each block (Fig 3.26). First stage performance was similar in Blocks 5-8, as expected by the model (Fig 3.21). To test more specifically whether participants were able to compose options,

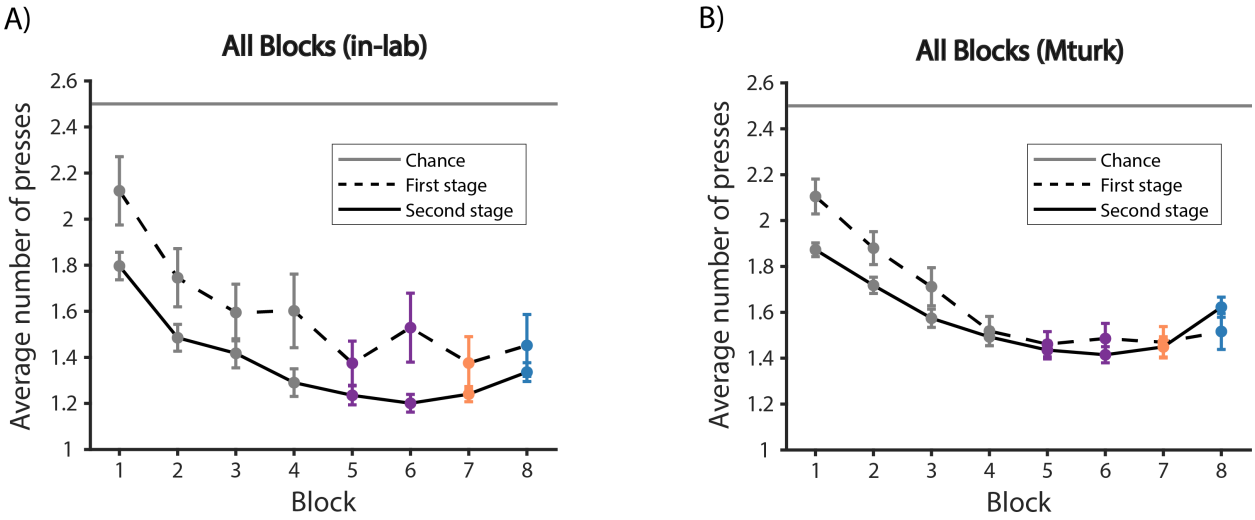

Figure 3.25: **Experiment 4 number of presses.** Average number of key presses in the first and the second stages per block for (A) in-lab participants and (B) Mturk participants.

we focused on comparing the second stage behavior for old LOs (LO_{match} and $LO_{mismatch}$) and the average of LO_5 and LO_6 (LO_{new}) in Blocks 7-8. The Option Model predicted that performance for LO_{match} in Block 7 should be the best due to positive transfer, since participants should have learned the extended MO_2 policy whereby LO_2 followed A_2 in Blocks 1, 3, and 5 (Fig 3.24). LO_{new} should be the worst due to negative transfer, with all 4 stimulus-action assignments in the second stage novel. Performance for $LO_{mismatch}$ in Block 7 should fall in between (as observed in the number of key pressed, Fig 3.27A). While there should be negative transfer, as MO_1 was usually followed by LO_1, LO_3 had been previously learned, so its performance should still surpass the performance in the second stage of Block 8, where LO_5 and LO_6 were completely novel to the participants. Therefore, we predicted $LO_{match} > LO_{mismatch} > LO_{new}$ in terms of performance.

In the second stage (Fig 3.27A), there was a main effect of block on number of presses (1-way repeated measure ANOVA, $F(2, 36) = 9.9, p = 0.0004$). Specifically, the average number of key presses in LO_{new} (Block 8) was significantly more than Blocks 5-6 and LO_{match} (paired t-test, Blocks 5-6: $t(18) = 4.1, p = 0.0007$; LO_{match}: $t(18) = 3.6, p = 0.002$). There was no significant difference between Blocks 5-6 and LO_{match} (paired t-test, $t(18) = 0.7, p = 0.49$), supporting the model's prediction of positive MO transfer in this condition. The model predicted that $LO_{mismatch}$ performance should be between LO_{new} and LO_{match}: $LO_{mismatch}$ performance should reflect positive LO transfer but negative MO transfer. This was observed qualitatively, though the results did not reach significance (paired t-test, LO_{match}: $t(18) = 1.6, p = 0.13$; LO_new: $t(18) = 1.4, p = 0.18$). These results replicate the negative transfer effects in the second stage of Block 8 shown in Experiment 1 (Fig 3.6A) and Experiment 2 (Fig 3.13D). In addition, they provide initial support for the compositionality hypothesis of

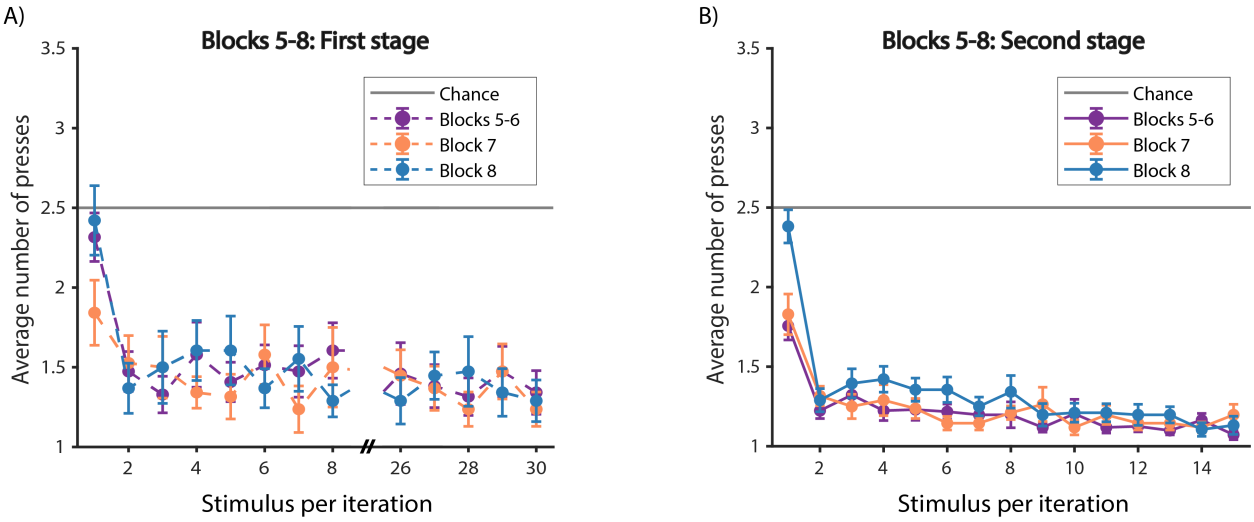

Figure 3.26: **Experiment 4 performance within Blocks 5-8 for in-lab participants.** (A) First stage. (B) Second stage.

the model, with intermediary transfer in the mismatch condition.

We confirmed the previous results by analyzing the proportion of trials in which the first key press was correct. We found that, in the first 3 trials for each of the 4 branches in the second stage (Fig 3.27B), there was a main effect of LO condition (1-way repeated measure ANOVA, $F(2, 36) = 7.2, p = 0.002$) on the proportion of correct choices for the first press of each trial. In particular, we found no significant difference between $LO_{mismatch}$ and LO_{new} (paired t-test, $t(18) = 0.56, p = 0.58$), while the performance of LO_{match} was significantly higher than $LO_{mismatch}$ and LO_{new} (paired t-test, $LO_{mismatch}$: $t(18) = 2.6, p = 0.017$; LO_{new}: $t(18) = 4.4, p = 0.0003$). These results suggested that the mismatch between MO_1 and LO_3 impacted participants' performance, a marker of negative option (MO) transfer. In the first three iterations, participants' first presses indicated that they were not able to efficiently re-compose the $LO_{mismatch}$ into a new mid-level option.

To better investigate participants' choices before they experienced any new information in a new block, we also computed the probability of a correct first key press for the second stage of the first trial of each of the 4 branches in the Blocks 5-8 (Fig 3.27D). We found a main effect of block (Friedman Test, $\chi^2(2, 36) = 20, p < 0.0001$). Specifically, Blocks 5-6 and LO_{match} were significantly above chance (sign test, both $p < 0.0001$); $LO_{mismatch}$ was not significantly different from chance (sign test, $p = 0.34$); LO_{new} was significantly below chance (sign test, $p = 0.0007$). There was a marginal difference between LO_{match} and $LO_{mismatch}$ (sign test, $p = 0.09$), but no significant difference between $LO_{mismatch}$ and LO_{new} (sign test, $p = 0.24$). These results further showed that the mismatch condition impacted participants' performance on the first press due to negative option (MO) transfer, and replicated the strong negative transfer in Block 8 in Experiment 1 and Experiment 2. The Option Model

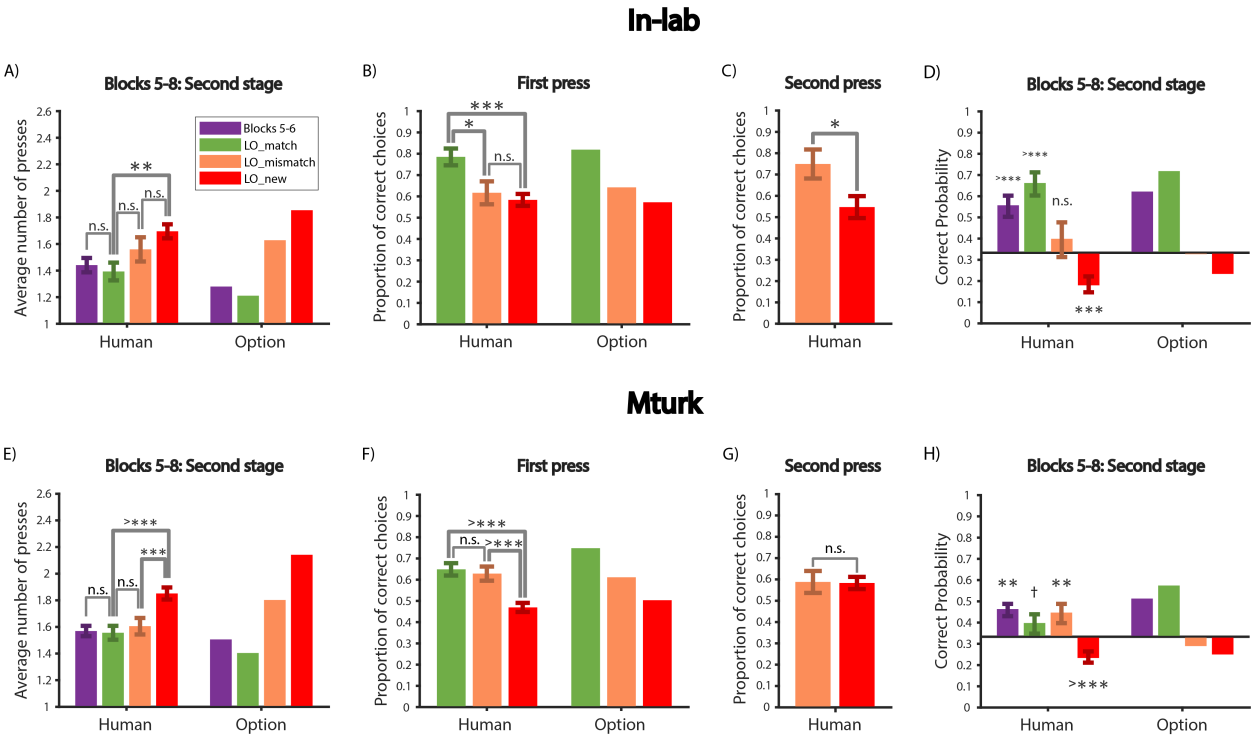

Figure 3.27: **Experiment 4 results show re-composition of options.** (A)-(D) In-lab participants. (A) Average number of key presses for the first 3 trials for each of the 4 branches in the second stage of Blocks 5-8 for participants (left) and the Option Model (right). Block 7 was split into LO_{match} and $LO_{mismatch}$; Block 8 corresponded to LO_{new}. (B) Proportion of correct choices on the first press of trials 1-3 for each of the 4 branches in the second stage for LO_{match}, $LO_{mismatch}$ and LO_{new} for participants (left) and the Option Model (right). (C) Proportion of correct choices on the second press (for trials 1-3 for each of the 4 branches with an incorrect first key press) for the mismatch (left) and the new (right) condition. (D) Probability of a correct first key press for the second stage of the first trial of each of the 4 branches in Blocks 5-8 for participants (left) and the Option Model (right). (E)-(H) Same as (A)-(D) for Mturk participants.

captured participants' behavior well (Fig 3.27ABD, see Table 3.2 for model parameters).

3.5.2.2 Second press reveals benefit of option composition

The results so far supported one of our predictions, $LO_{match} > LO_{mismatch}$, by showing that performance in the mismatch condition was impacted due to negative MO transfer. We next sought evidence for our second prediction, $LO_{mismatch} > LO_{new}$, where we hypothesized better performance in the mismatch condition by composing the first stage policy of MO_1 and LO_3.

In terms of performance on the first press in each trial, we did not find a significant difference between the two conditions (Fig 3.27B). However, this might be because the negative MO transfer reduced the benefit of compositionality, making it less detectable on the first press, also reflected by the small effect from the Option Model in Fig 3.27B. Positive LO transfer thus might only show a more significant effect after the first press unexpectedly failed (from negative transfer of MO_1).

Therefore, we further computed the proportion of correct choices on the second press in those trials where the first press was incorrect (Fig 3.27C). Indeed, we found that the proportion of correct choices on the second press was significantly higher in the mismatch condition than the new condition (paired t-test, $t(17) = 2.8, p = 0.012$). This result supports our second prediction, $LO_{mismatch} > LO_{new}$, revealing a benefit in the mismatch condition compared to the new condition in participants re-composing an old LO into a non-matching MO.

3.5.2.3 Mturk participants showed benefits of option composition

We collected a larger and independent sample on Mturk. Mturk participants also improved over Blocks 1-6 (Fig 3.25B) and within block (Fig 3.28), though their asymptotic performance (Blocks 5-6) was lower than the in-lab population. Specifically, we compared the average number of key presses in Blocks 5-6 in the first and second stages for both in-lab and Mturk populations. There was a main effect of stage and a marginal interaction of population and stage (2-way mixed ANOVA, stage: $F(1, 78) = 7.1, p = 0.009$; interaction: $F(1, 78) = 3.1, p = 0.08$). In particular, for the first stage, Mturk population was not significantly worse than the in-lab population (unpaired t-test, $t(78) = 0.17, p = 0.86$); but for the second stage, which was the focus of our analysis, Mturk population was significantly worse than the in-lab population (unpaired t-test, $t(76) = 3.2, p = 0.002$).

In the second stage (Fig 3.27E), there was a main effect of block on number of presses ($F(2, 120) = 17, p < 0.0001$). Specifically, the average number of key presses in LO_{new} was significantly more than LO_{match} and $LO_{mismatch}$ (paired t-test, LO_{match}: $t(60) = 4.6, p < 0.0001$; $LO_{mismatch}$: $t(60) = 3.8, p = 0.0004$). LO_{match} was not significantly different from Blocks 5-6 and $LO_{mismatch}$ (paired t-test, Blocks 5-6: $t(60) = 0.26, p = 0.8$; $LO_{mismatch}$: $t(60) = 0.8, p = 0.42$).

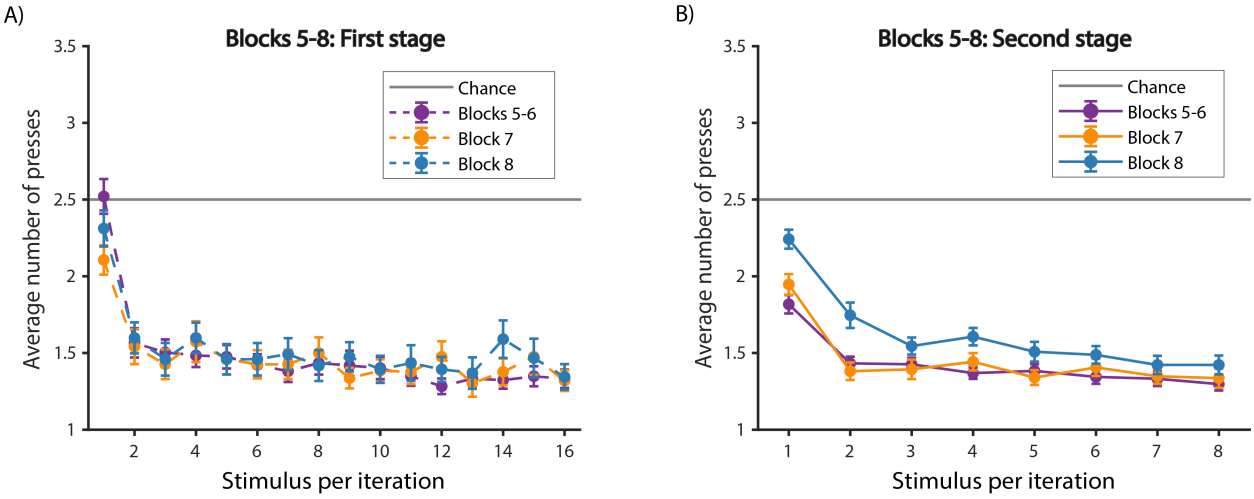

Figure 3.28: **Experiment 4 performance within Blocks 5-8 for Mturk participants.** (A) First stage. (B) Second stage.

The proportion of correct first press choices (Fig 3.27F) showed a similar pattern: there was a main effect of LO condition ($F(2, 120) = 15, p < 0.0001$) on the proportion of correct choices. In particular, the proportion of correct choice for LO_{new} was significantly lower than $LO_{mismatch}$ and LO_{match} (paired t-test, $LO_{mismatch}$: $t(60) = 4.7, p < 0.0001$; LO_{match}: $t(60) = 5.1, p < 0.0001$) in Block 7. There was no significant difference between $LO_{mismatch}$ and LO_{match} performance (paired t-test, $t(60) = 0.54, p = 0.59$). There was no difference between the mismatch condition and the new condition for second key presses (paired t-test, $t(52) = 0.08, p = 0.94$, Fig 3.27G), contrary to in-lab participants (Fig 3.27C). This difference could be attributed to MTurk participants' lower task engagement. Indeed, contrary to in lab participants, MTurk participants' performance was at chance for second key press (MTurk: paired t-test, $t(53) = 1.6, p = 0.13$; in-lab $t(17) = 3.4, p = 0.003$). Directly comparing MTurk and in-lab population for the proportion of correct second key press in both the mismatch and new conditions revealed a marginal effect of condition and a marginal interaction of population and condition (2-way mixed ANOVA, condition: $F(1, 69) = 3.3, p = 0.07$; interaction: $F(1, 69) = 3.7, p = 0.06$). This supports our interpretation that MTurk participants did not attempt to find the correct answer following an error, making the second press error analysis in this population difficult to interpret.

Finally, we looked at the probability of a correct first press in the very first trial of each of the 4 branches in the second stage (Fig 3.27H). There was a main effect of block (Friedman test, $\chi^2(2, 120) = 17, p = 0.0002$). In particular, Blocks 5-6 and $LO_{mismatch}$ were significantly above chance (sign test, both $p = 0.004$) LO_{match} was marginally above chance (sign test, $p = 0.07$); LO_{new} was significantly below chance (sign test, $p < 0.0001$).

These results can be interpreted in one of two ways. The similar performance between

LO_{match} and $LO_{mismatch}$ suggests that participants were able to efficiently re-compose the first stage of MO_1 with LO_3 in the mismatch condition in Block 7, so that they did not suffer from MO negative transfer, as did in-lab participants. Alternatively, this result might indicate a lack of MO transfer (and only positive LO transfer) in both the match and mismatch condition. The latter interpretation is supported by the fact that second stage performance in LO_{match} was lower in MTurk participants than it was for in-lab participants in all measures (unpaired t-test, number of key presses in the first 10 trials of Blocks 5-6: $t(78) = 1.8, p = 0.08$; proportion of correct choices in match condition: $t(78) = 2.4, p = 0.019$).

The Option Model could capture the negative transfer effect in LO_{new} and thus the difference between LO_{new} and $LO_{mismatch}$ (Fig 3.27EF). However, it could not fully reproduce the lack of difference between LO_{match} and $LO_{mismatch}$, since the model would first try to transfer LO_1 in the mismatch condition, resulting in worse performance for $LO_{mismatch}$.

This interpretation might suggest that the Task-Set Model explains the Mturk population better, indicating a lack of temporally extended options, and makes a specific prediction: second stage errors should not be impacted by first stage information. To test this prediction, we analyzed the specific errors participants made, as this is a specific hallmark of temporally extended option transfer vs. task-sets (Fig 3.6A). Contrary to the prediction made by the Task-Set model, but consistent with the Option Model prediction, Mturk participants did demonstrate the behavioral signature of negative option (MO) transfer in the mismatch condition (Fig 3.29): they made significantly more *option transfer* errors than *other* errors (paired t-test, $t(53) = 4.8, p < 0.0001$). While the comparison was not significant for in-lab participants (paired t-test, $t(17) = 1.5, p = 0.16$), a direct comparison between in-lab and Mturk populations did not reveal an effect of population (2-way mixed ANOVA, $F(2, 140) = 0.74, p = 0.48$). Thus, our results indicate that both MTurk participants and in-lab participants used temporally extended MOs, although MTurk participants were overall less successful at transferring them to facilitate decision making in the second stage. The results are consistent with participants re-composing low-level options into higher-level options.

3.6 Robustness of results for different parameters

We used the set of parameters from Table 3.2 for all figures in previous sections to track participants' behavioral patterns both qualitatively and quantitatively.

Here we used another set of parameters (Table 3.3) to (1) constrain parameters so that most experiments shared the same parameters while showing the qualitatively trends in participants' behavior and (2) show that the model can reproduce the same qualitative effects with a range of parameters.

In particular, we used $\alpha^1 = 0.7, \beta^1 = 4, \beta^2 = 4, m = 0.01$ for all experiments. For all in-lab experiments, we used $\alpha^2 = 0.7, f^2 = 0.001$; for all Mturk experiments, we used $\alpha^2 = 0.5, f^2 = 0.005$, which indicate slower learning rate and faster forgetting. For Experiment 1

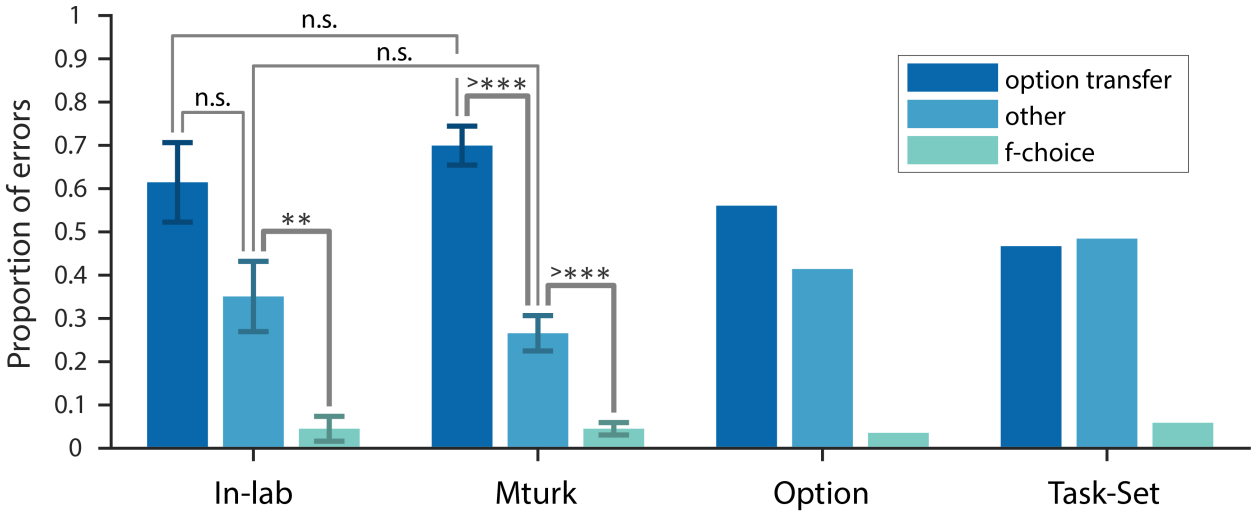

Figure 3.29: **Experiment 4 second stage errors reveal temporal options transfer and compositionality.** Error type analysis of the second stage in Block 7 for the mismatch condition for in-lab participants, Mturk participants, the Option Model and the Task-Set Model.

Exp	Sample	Model	α^1	β^1	γ^1	f^1	α^2	β^2	γ^2	f^2	m
Exp 1	In-lab	Naive	0.5	4	NA	0.0025	0.7	10	NA	0.0001	0.01
		Flat	0.5	4	NA	0.0025	0.7	10	NA	0.0001	0.01
		Task-Set	1	2	14	0.0004	0.8	3	3	0.0002	0.01
		Option	1	2	14	0.0004	0.8	3	3	0.0002	0.01
	Mturk	Option	0.8	3	100	0.01	0.6	6	5	0.004	0.01
Exp 2	In-lab	Option	0.7	3	13	0.001	0.6	4	5	0.001	0.01
Exp 3	In-lab	Option	0.7	4	100	0.01	0.8	5	15	0.001	0.01
	Mturk	Option	0.7	4	100	0.01	0.8	5	15	0.005	0.01
Exp 4	In-lab	Option	0.6	4	100	0.01	0.8	5	4	0.0002	0.01
	Mturk	Option	0.6	4	100	0.01	0.4	4	5	0.002	0.01
		Task-Set	0.6	4	100	0.01	0.4	4	5	0.002	0.01

Table 3.2: Parameters for the main text.

Exp	Sample	Model	α^1	β^1	γ^1	f^1	α^2	β^2	γ^2	f^2	m
Exp 1	In-lab	Naive	0.7	4	NA	0.001	0.7	4	NA	0.001	0.01
		Flat	0.7	4	NA	0.001	0.7	4	NA	0.001	0.01
		Task-Set	0.7	4	14	0.001	0.7	4	4	0.001	0.01
		Option	0.7	4	14	0.001	0.7	4	4	0.001	0.01
	Mturk	Option	0.7	4	100	0.01	0.5	4	4	0.005	0.01
Exp 2	In-lab	Option	0.7	4	100	0.01	0.7	4	4	0.001	0.01
Exp 3	In-lab	Option	0.7	4	100	0.01	0.7	4	20	0.001	0.01
	Mturk	Option	0.7	4	100	0.01	0.5	4	20	0.005	0.01
Exp 4	In-lab	Option	0.7	4	100	0.01	0.7	4	4	0.001	0.01
	Mturk	Option	0.7	4	100	0.01	0.5	4	4	0.005	0.01

Table 3.3: A second set of parameters that is constrained but still replicate transfer effects qualitatively.

in-lab, we used $\gamma^1 = 14, f^1 = 0.001$; for all other experiments, we used $\gamma^1 = 100, f^1 = 0.01$ to implement a lack of transfer effects in the first stage. We used $\gamma^2 = 20$ in Experiment 3 to model reduced option transfer in the second stage; for all other experiments, we used $\gamma^2 = 4$.

We recreated some of the representative analysis in the main text to demonstrate that this second set of parameters can replicate the transfer effects in human participants qualitatively well (Fig 3.30, Fig 3.31, Fig 3.32).

3.7 Discussion

Our findings provide novel and strong support for the acquisition of options in healthy human adults. Options can be thought of as choices that are more abstract than simple motor actions, but can be taken as a single choice. Using a novel two-stage protocol, we provide evidence that humans create options, and flexibly transfer and compose previously learned options. This transfer and composition ability guides exploration in novel contexts and speeds up learning when the options are appropriate, but impairs performance otherwise, as predicted by the options framework [22]. Model simulations showed that only a model including temporal hierarchy could account for all results, suggesting that human participants not only build state abstractions with one-step task-sets [116], but also temporal abstractions in the action space with multi-step options.

We developed a new model, the Option Model, to account for participants' behavior. The Option Model includes features from our previous hierarchical structure learning model [38, 36, 33] and the hierarchical reinforcement learning (HRL) options framework [158]. In our previous hierarchical structure learning model, we used non-parametric priors (CRP) over latent variables that represented the currently valid policy to create *state abstractions*: this

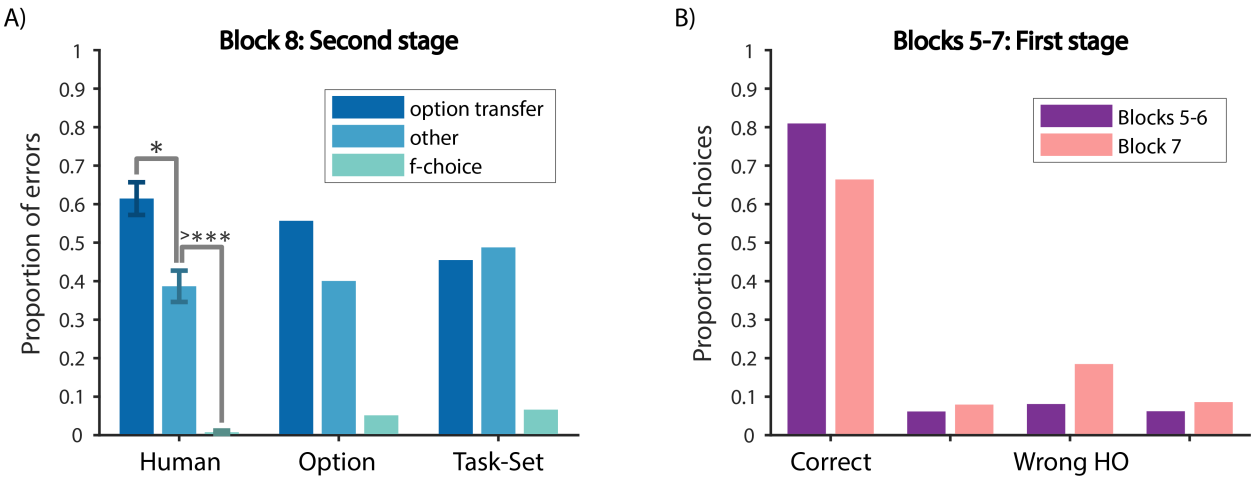

Figure 3.30: **Experiment 1 with parameters from Table 3.3.** (A) Error type analysis of the second stage in Block 8 for participants (left), the Option Model (middle) and the Task-Set Model (right). (B) Choice type analysis of the first stage in Blocks 5-7 for the Option Model.

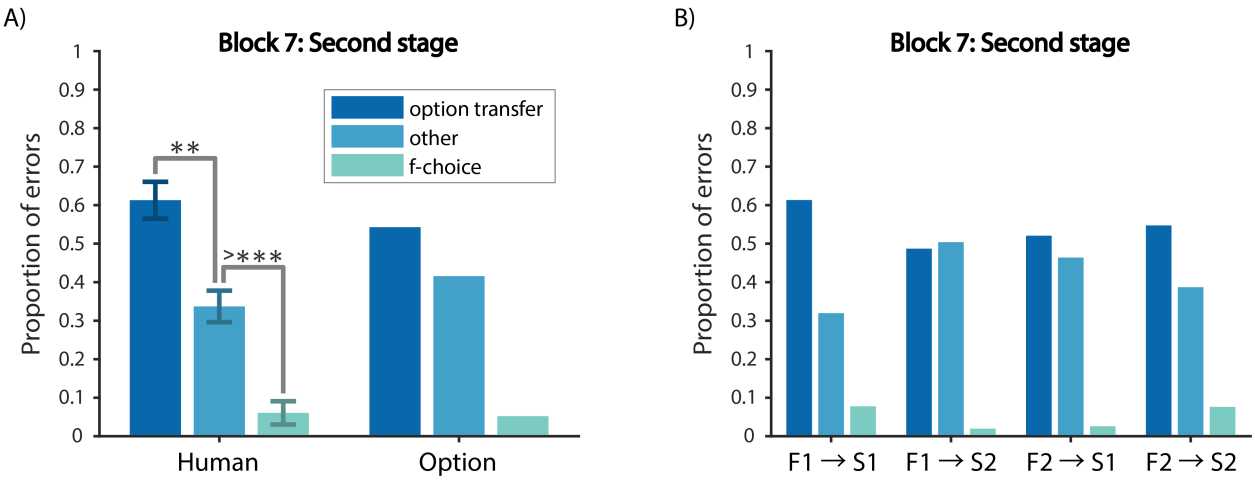

Figure 3.31: **Experiment 2 second stage choices with parameters from Table 3.3.** (A) Error type analysis of the second stage in Block 7 for participants (left) and the Option Model (right). (B) Error type analysis for each of the 4 branches in the second stage of Block 7 for the Option Model.

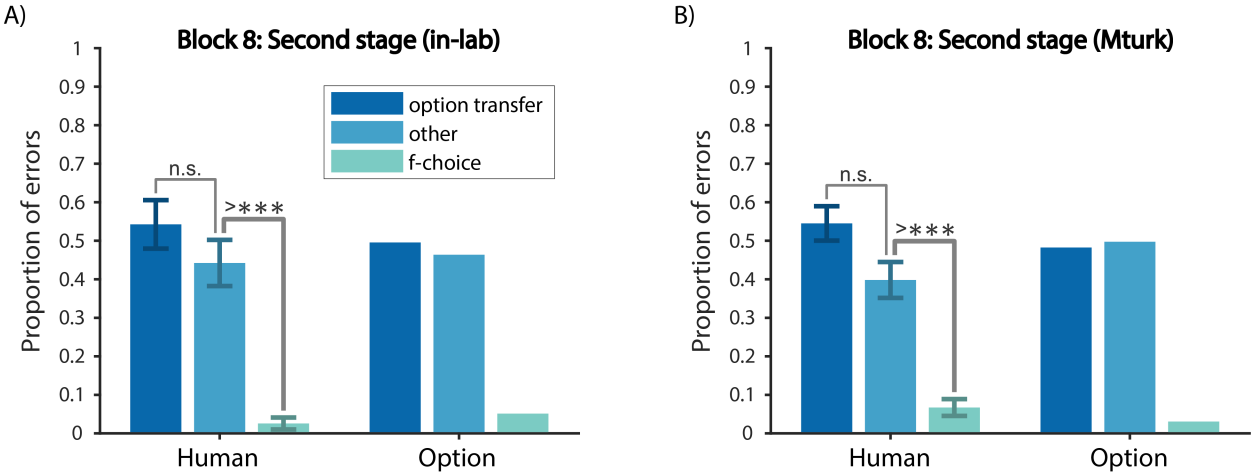

Figure 3.32: **Experiment 3 second stage choices with parameters from Table 3.3.** Error type analysis of the second stage in Block 8 for (A) in-lab participants (left) and the Option Model (right), and (B) Mturk participants (left) and the Option Model (right).

allowed the model to cluster different contexts together if the same task-set applied. This CRP prior enables the agent to identify (via Bayesian inference) novel contexts as part of an existing cluster if the cluster-defined task-set proves successful, resulting in more efficient exploration and faster learning.

On the other hand, the original formulation of the HRL options framework [158] augments the action space of traditional flat RL with *temporal abstractions* called options. Each option is characterized by an initiation set that specifies which states the option can be activated, a termination function that maps states to a probability of terminating the current option, and an option-specific policy (that leads the agent to a potentially meaningful and useful sub-goal). Multi-step options allow even more efficient transfer than task-sets, which can be thought of as simpler one-step options.

Our Option Model is inspired by the fact that task-sets and options are similar in essentials: they are policies that an agent can select as a whole, and then apply at a lower level of abstraction (applying it to make a motor choice in response to a stimulus for task-sets, or applying it across time until termination in the case of an option [34]. Thus, our model brings together state and temporal abstractions by using option-specific CRP priors to implement option-specific policies that can be flexibly selected in different contexts if they share the same environmental contingencies. Our model captures the essence of the options framework despite some subtle differences. Here, we discuss how our Option Model relates to each part of the HRL options framework.

3.7.1 Option-specific policy

The most important component of an option is the option-specific policy: what lower level-choices (either simpler options or basic actions) it constrains. In this study, we focused on the transfer of option-specific policy to test theoretical benefits of the options framework.

Theoretical work [22] suggested that useful options should facilitate exploration and speed up learning. Indeed, we observed speed up in learning through the positive transfer effects. For example, in Experiment 1, the second stage of Block 7 provided a test of positive option transfer in terms of both number of presses (Fig 3.3B) and choice types (Fig 3.7). Importantly, this positive transfer was not interfered by the negative transfer in its first stage (Fig 3.3B), suggesting that participants transferred mid-level options (MO) as a whole.

Moreover, the learning benefit was evident even in the first press (Fig 3.6B, Fig 3.10E, Fig 3.13D): participants were already significantly above chance in the first press, indicating that they could explore by immediately transferring previously learned options.

Previously learned option-specific policies also helped with option composition in the mismatch condition (Fig 3.24) of Experiment 4 (Fig 3.27). While MO_1 was usually followed by LO_1 in Blocks 1, 3, 5, in the mismatch condition, MO_1 was followed by LO_3 instead. This change indeed resulted in *option transfer* errors (Fig 3.29). However, the fact that LO_3 had been previously learned helped participants explore more efficiently. For example, once participants figured out A_2 was correct for the diamond, they would more likely explore LO_3, and thus A_4 for triangle.

The HRL options framework also suggested that non-useful options can slow down learning. Indeed, we observed negative option transfer effects in the second stage across multiple experiments in terms of number of presses (Fig 3.3B, Fig 3.10C, Fig 3.13C, Fig 3.27AE), and more importantly, error types (Fig 3.6A, Fig 3.10D, Fig 3.13D, Fig 3.16, Fig 3.29), that are consistent with the predictions of the options framework. Note that the slow down was due to negative transfer of previously learned option-specific policies. Thus testing how having a wrong sub-goal can impact learning performance is an interesting future direction.

We sought to confirm that participants were indeed learning option-specific policies, not just action sequences. Our protocol specifically used two second stage stimuli following each first stage stimulus (Fig 3.2B) to avoid this potential confound. If, for example, circle was always followed by diamond and square by triangle, participants would not need to pay attention to the actual stimulus in the second stage, and could instead plan a sequence of actions in the first stage. In contrast, here, participants could only perform well by selecting options (i.e. stimulus-dependent temporally extended policies). While pure sequence learning could not account for our results, we investigated whether it could contribute to some of its aspects. Sequence learning would predict faster reaction times for actions that often follow in a sequence [31]. Therefore, we compared the reaction time for the *sequence* and *non-sequence* error types in the second stage (Sec 3.2.2.3). We did not find significant difference between the reaction time for *sequence* and *non-sequence* error types at the beginning of blocks; we only found such difference at the end of blocks (Fig 3.8, Fig 3.15, Sec 3.2.2.3). This suggests that while the transfer effects we observe at the beginning of each block could

not be explained by pure sequence learning, participants might develop sequence learning-like expectations over time in a block, speeding up choices that came more frequently after each other.

3.7.2 Initiation set

The initiation set specifies the set of states where an option can be selected. The observable states in our tasks are the shapes shown on the screen. Therefore, at first, the initiation sets of HO and MO are first stage stimuli (e.g. circle and square, Fig 3.2B), whereas the initiation sets of LO are second stage stimuli. However, the optimal policies were also dependent on the block; thus participants needed to infer the hidden context (*state abstraction*) dictated by block. Our CRP implementation can thus be thought of as continuously adding new block contexts to the initiation set of an option throughout the task. The ability to add new contexts to the initiation sets provides our Option Model the crucial flexibility needed to achieve transfer and composition, as demonstrated by human participants. For example, if LO_3 was tied solely to the context of Block 2, where it was first learned, we would not observe the benefit of option composition in Experiment 4 in the mismatch condition.

3.7.3 Termination function

An option's termination function maps each state to the probability of terminating the current option (i.e. not using its policy anymore). How to terminate an option is closely related to the underlying theoretical question of credit assignment, which arises naturally in tasks that require hierarchical reasoning [141]: if the current policy does not generate any (pseudo-) reward for a while, should the agent continue improving the current policy or terminate it and use another policy or even something new?

With a termination function as described in the original HRL options framework, credit assignment happens in a very specific way: the policy of the currently selected option (or options if multiple nested options are selected) is updated until termination is reached. In our task, this would make behavior very inflexible. For example, when an agent entered the second stage of Block 8 in Experiment 1 (Fig 3.2B) for the first time after having correctly made a choice for the circle in the first stage, the agent would likely use LO_1 due to negative transfer of MO_1 and thus not receive reward. Because the termination function only takes state as an input, the agent would keep overwriting the LO_1 policy with LO_5 policy until termination, and thus not be able ot reuse LO_1 down the line.

Our Option Model, however, uses a more flexible form of option termination. Specifically, we use Bayesian inference (Sec 3.2.1.5), which was introduced in our previous hierarchical structure learning model [38]. At the end of each choice, the model updates the likelihood of each option being valid based on the observed reward feedback, which then determines whether the model should stop using the current option. Moreover, Q-learning only operates on the option that has the highest posterior, thus assigning credit retrospectively to the best

cause [117]. Therefore, the Option Model is more likely to create a new LO_5 and learn its policy from scratch, making it more flexible at learning and selecting options.

The crucial difference between the two is that the Option Model would create a new LO_5 and learn its policy from scratch, without overwriting the original LO_1 policy. While the Option Model can capture participants' choices well across all four experiments, the current experimental protocol was not designed specifically to test credit assignment to options, and could not distinguish between these two possibilities. This remains an important question for future research.

There is another credit assignment problem that is not fully addressed by our current protocol and modeling: choices by lower level options may affect the termination of higher level options. For example, if you get punished for boiling potatoes, should you credit this to the lower level option (boiling) or to the higher level option (making potatoes in the first place). Should you plan to cook vegetables instead, or just roast the potatoes? We have some evidence for both levels of credit assignment (e.g. in Block 7 of Experiment 2, or Block 8 in Experiment 1; Fig 3.2B), when participants were experiencing many errors in the second stage using LO_1 and LO_2. Participants might not only consider terminating or re-learning the current LO, but also naturally attribute some of the negative feedback to the choices they made in the first stage regarding MO or HO. Indeed, we observed that second stage errors potentially resulted in more *wrong* HO errors in the first stage of Experiment 2 (Fig 3.17).

In our Option Model (Sec 3.2.1.5), for simplicity, first stage choices were only determined by learning within the first stage and were not sensitive to reward feedback in the second stage. It will be important in future research to better understand interactions between option levels for credit assignment. When considered together with the termination problem, these future directions may help trace the underlying neural mechanisms for credit assignment in human learning and hierarchical decision making.

3.7.4 Possible extensions

We tested predictions of HRL options framework through positive and negative transfer of option-specific policies in the simplest possible set up of tabular representation of state and action space. Multiple aspects could be expanded on in future research to increase the generalizability of the policy in real world scenarios. First, real world policies apply to much more complex (continuous, multidimensional) state spaces. Recent work in AI expands the options framework to more realistic situations [91], where artificial agents learn how to navigate a sequence of rooms with different shapes and sizes. If each state in a room is naively paramatrized in a tabular way by (x, y) coordinates, when the agent is placed in a new room of a different shape, previously learned policy would be of not use. It is thus crucial to identify meaningful features of the state space shared by different rooms. [91] proposed learning options in a state space parametrized by distance from goals (*agent space*) to bypass this limitation.

Second, the low-level action space in real life conditions is also more complex. A good example is our flexible use of tools [1]. We can conceptualize using various tools as taking actions. Humans demonstrate great flexibility when improvising using different tools to solve the same problem or even crafting new tools. If we simply represent actions in a tabular way, after participants associated a particular tool (action) to solve a task, the policy would be of no use if this particular tool is no longer provided in the future. The key might again be figuring out meaningful dimensions of the tool (action) space that are shared in different task scenarios, such as shape and weight of the tool.

Finally, even if two problems are different in terms of both state and action space (e.g. learning to play piano vs learning to play violin [69]), knowledge of one might still help the other. Once one learned a piece on the piano, the knowledge of music theory might serve as a model to guide option transfer when learning the same piece on violin. These are important future directions for testing how humans transfer in those more real life scenarios, which might provide insight into developing more flexible and human-like AI systems with the HRL options framework.

3.7.5 Option discovery

One of the most important questions regarding options in AI is how to discover meaningful options. Discovering useful options entails learning all components of an option: initiation set, termination function, and option-specific policy that leads to a meaningful sub-goal. In this study, we designed a protocol that focused on learning option-specific policies by making all other features, including sub-goals, trivial.

Discovering options may be useful because of a key feature of our interactions with our environment. In real world scenarios, it is frequent that for a given observable state, the right choice to make depends on hidden context, task demand, or past information. This property is refered to as *non-Markovian*: the current observable information is insufficient to determine the next step. For example, when potatoes are peeled, we can use them to make either roasted potatoes or mashed potatoes. Therefore, the state *" peeled potatoes"* is a meaningful sub-goal state, and peeling potatoes is its corresponding option-specific policy.

This non-Markovian property might contribute to the hierarchical and compositional nature of human behavior. It is central to the original formulation of the options framework [158], and is also a natural objective for option discovery. In relation to our protocol, the correct action for diamond (Fig 3.2B) varies from time to time in the same block. It makes sense to create different options to capture this, and relate it to the inferred hidden cause for why the correct actions change. Indeed, we observed that the non-Markovian feature in our experiments encouraged participants to create and transfer options at multiple levels of abstractions.

We tested whether the environment needs to be non-Markovian to trigger option creation. Specifically, we designed Experiment 3 by eliminating the non-Markovian property from Experiment 1 and testing if that affects option learning and transfer (Fig 3.19). Unsurprisingly, we found weaker option transfer effects in Experiment 3; however, participants'

behavior was still not flat (Fig 3.19, Fig 3.22). Thus, our results hint at the possibility that participants create temporal options (MO), even in the absence of a need for it, echoing past results showing that humans tend to create structure unnecessarily [38, 32, 35, 175]. Furthermore, this may also show that objectives for option discovery are not limited to solving non-markovian problems. For example, [52] showed that humans could identify bottleneck states from transition statistics, reflecting graph-theoretic objectives for option discovery in humans.

3.7.6 The options framework and other learning systems

While our Option Model uses a simple form of model-free RL (Q-learning; [159]) to learn option-specific policies, the options framework is general and not limited to just Q-learning. Options can be learned or used with model-free methods [22] and model-based methods [20]. It also has strong connections to successor representations [154, 115], which might provide objectives for sub-goal discovery.

Moreover, in this study, we gave examples of potential interaction of options with the meta-learning system (Fig 3.16) and sequence learning (Sec 3.2.2.3) in human participants. How options might interact with other learning systems is an important question for future research.

3.8 Conclusion

In summary, we found compelling evidence of option learning and transfer in human participants by examining the learning dynamics of a novel two-stage experimental paradigm. Through analyzing participants' behavioral patterns and model simulations, we demonstrated the flexibility of option transfer and composition at distinct levels in humans.

To model the complex human cognition demonstrated in these experiments, we utilized the options framework to extend one-step decision making from the previous Butterfly study (see Chapter 2) to multi-step decision making. We further augment the original options framework with Chinese Restaurant Process to learn state and temporal abstractions online, thus achieving transfer and generalization. More importantly, the new augmented Option Model allows us to pose important new questions, e.g. how is credit assignment addressed in hierarchical tasks.

Humans' ability to flexibly transfer previously learned skills is crucial for learning and adaptation in complex real world scenarios. This ability is also one of the fundamental gaps that sets humans apart from current state-of-the-art AI algorithms. Therefore, our work trying to probe learning and transfer in humans might also help provide inspirations for AI algorithms to be more flexible and human-like.

Chapter 4

Conclusion

In conclusion, we have demonstrated the effectiveness of RL as a mathematical model for human learning and decision making by using it both as a quantitative tool for compressing human behavior and parametrizing individual differences (Chapter 2), and as a theoretical framework for developing new theories of complex human cognition on multi-step and hierarchical learning (Chapter 3).

We would like to highlight the key roles of applied mathematical techniques in both studies. In the first study (Chapter 2), we used *mathematical RL models* [159] not only to capture human behavior, but also to provide a quantitative and mechanistic understanding of human learning under uncertainty in the Butterfly task. For model fitting, we used *hierarchical Bayesian modeling* [73] to jointly fit all participants instead of each participant independently, which helped achieve better model fitting results compared to traditional maximum likelihood. The hierarchy introduced by hierarchical Bayes, however, resulted in intractable likelihood calculation; thus we employed *No-U-Turn sampling*, a state-of-the-art sampling technique implemented in the probabilistic programming language Stan [27, 10], to directly sample from the joint posterior of all participants' parameters and bypass the exact evaluation of the entire likelihood function. These mathematical and statistical techniques enabled us to perform robust statistical inference for all participants as a whole and make important conclusions about developmental changes in learning.

In the second study (Chapter 3), we started from the important insight of the overlap between hierarchical human learning and the theoretical benefits of *mathematical HRL models* [158]. We went a step further beyond learning to address transfer of previously learned policies. We created a novel mathematical model by augmenting HRL models with *Chinese Restaurant Process* [133] to allow transfer. This new mathematical model can effectively capture the various transfer effects that we empirically observed in hierarchical human behavior, which could not be explained by any flat RL models.

Both lines of work have challenges and opportunities ahead. For the first line of work, there have been major challenges regarding the interpretability and generalizability of RL model parameters. In particular, whether the parameter estimates from a particular model, a particular task and a particular subject sample would generalize to other models, other tasks

with different features or other populations with different demographics [126, 43, 172, 56, 109]. Despite the challenges, using RL models to compress human behavior and parametrize individual differences can be used to address important scientific questions. For example, aside from addressing developmental questions, one can also leverage the same quantitative methods to address learning and decision making changes during aging by focusing on older adults [135, 42]. There have also been promising work using similar methodologies to study the computational and neural underpinnings of psychopathologies by comparing the model parameters between patients with psychiatric disorders and healthy controls [40, 81].

Moving beyond characterizing only behavior, the mathematical RL modeling techniques described in this thesis can also help inform brain function. One fruitful and effective way is to extract trial-by-trial learning dynamics from the models and correlate them with functional brain activities to find relevant brain regions that support various components of the RL models [102, 139, 53, 113]. Moreover, new mathematical and statistical methods have been developed recently to jointly model behavioral and neural data [51, 161], potentially allowing researchers to address an even broader set of questions relating behavior to underlying brain mechanisms.

For the second line of work, the additional hierarchical structure and transfer capabilities in our model help explain the various transfer effects demonstrated by human participants, but also add challenges and complexities to model fitting. The traditional statistical model fitting techniques such as maximum likelihood and the sampling-based method used in Chapter 2 no longer suffice, since the likelihood calculation for HRL models becomes intractable (Sec 3.2.1.5). There have been ongoing efforts and early promising results to fit decision making models with hierarchy [60, 59]. There have also been active research on using deep learning methods [58, 50] to bypass explicit likelihood calculations. The ability to extract trial-by-trial learning dynamics through these novel model fitting techniques could further enable brain imaging experiments to probe neural mechanisms for hierarchical learning and transfer. Another possible future direction would be to refine our understanding of other aspects of hierarchical learning in human behavior that are not covered in Chapter 3, such as credit assignment and subgoal discovery in HRL.

Aside from deep diving in refining our current option model, there are other new directions that our research on human option learning and transfer can open the door to. For example, the transfer and generalization paradigm can be potentially extended to study individual differences in hierarchical learning and psychopathologies as well. Another possible direction is to scale up our existing option model from tabular cases to complex real world problems to help AI agents achieve human-level flexibility on learning, transfer and compositionality.

Bibliography

[1] Kelsey R Allen, Kevin A Smith, and Joshua B Tenenbaum. "The Tools Challenge: Rapid Trial-and-Error Learning in Physical Problem Solving". In: *arXiv preprint arXiv:1907.09620* (2019).

[2] John R Anderson et al. "An integrated theory of the mind." In: *Psychological review* 111.4 (2004), p. 1036.

[3] Jacob Andreas, Dan Klein, and Sergey Levine. "Modular multitask reinforcement learning with policy sketches". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 166–175.

[4] David Badre. "Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes". In: *Trends in cognitive sciences* 12.5 (2008), pp. 193–200.

[5] David Badre and Mark D'Esposito. "Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex". In: *Journal of cognitive neuroscience* 19.12 (2007), pp. 2082–2099.

[6] David Badre and Mark D'esposito. "Is the rostro-caudal axis of the frontal lobe hierarchical?" In: *Nature Reviews Neuroscience* 10.9 (2009), p. 659.

[7] David Badre and Michael J Frank. "Mechanisms of hierarchical reinforcement learning in cortico–striatal circuits 2: Evidence from fMRI". In: *Cerebral cortex* 22.3 (2011), pp. 527–536.

[8] Bernard W Balleine et al. "Hierarchical control of goal-directed action in the cortical–basal ganglia network". In: *Current Opinion in Behavioral Sciences* 5 (2015), pp. 1–7.

[9] Marjolein EA Barendse et al. "Study Protocol: Transitions in Adolescent Girls (TAG)". In: *Frontiers in psychiatry* 10 (2020), p. 1018.

[10] Beth. Baribault. *bbstanlib: A library of helper functions for Stan/MATLABStan*. en. 2019.

[11] Beth. Baribault. *matstanlib: A library of helper functions for Stan/MATLABStan*. en. 2019.

[12] Timothy E. J. Behrens et al. "Learning the value of information in an uncertain world". en. In: *Nature Neuroscience* 10.9 (2007), pp. 1214–1221. ISSN: 1546-1726. DOI: 10.1038/nn1954.

[13] Marc G Bellemare et al. "Autonomous navigation of stratospheric balloons using reinforcement learning". In: *Nature* 588.7836 (2020), pp. 77–82.

[14] Irving Biederman. "Recognition-by-components: a theory of human image understanding." In: *Psychological review* 94.2 (1987), p. 115.

[15] Johannes Bill et al. "Hierarchical structure is employed by humans during visual motion perception". In: *bioRxiv* (2019), p. 758573.

[16] Gunnar Blohm, Konrad P Kording, and Paul R Schrater. "A how-to-model guide for Neuroscience". In: *Eneuro* 7.1 (2020).

[17] Aaron M Bornstein and Kenneth A Norman. "Reinstated episodic context guides sampling-based decisions for reward". In: *Nature neuroscience* 20.7 (2017), pp. 997–1003.

[18] Wouter van den Bos et al. "Striatum–Medial Prefrontal Cortex Connectivity Predicts Developmental Changes in Reinforcement Learning". en. In: *Cerebral Cortex* 22.6 (2012), pp. 1247–1255. ISSN: 1047-3211. DOI: 10.1093/cercor/bhr198.

[19] Matthew Botvinick and David C Plaut. "Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action." In: *Psychological review* 111.2 (2004), p. 395.

[20] Matthew Botvinick and Ari Weinstein. "Model-based hierarchical reinforcement learning and human action control". In: *Phil. Trans. R. Soc. B* 369.1655 (2014), p. 20130480.

[21] Matthew M Botvinick. "Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1485 (2007), pp. 1615–1626.

[22] 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.

[23] Matthew Michael Botvinick. "Hierarchical reinforcement learning and decision making". In: *Current opinion in neurobiology* 22.6 (2012), pp. 956–962.

[24] Barbara R Braams et al. "Longitudinal changes in adolescent risk-taking: a comprehensive study of neural responses to rewards, pubertal development, and risk-taking behavior". In: *Journal of Neuroscience* 35.18 (2015), pp. 7226–7238.

[25] Julia M Rodriguez Buritica, Hauke R Heekeren, and Wouter van den Bos. "The computational basis of following advice in adolescents". In: *Journal of experimental child psychology* 180 (2019), pp. 39–54.

[26] Stephanie L Cardoos et al. "Social status strategy in early adolescent girls: Testosterone and value-based decision making". In: *Psychoneuroendocrinology* 81 (2017), pp. 14–21.

[27] Bob Carpenter et al. "Stan: A Probabilistic Programming Language". en. In: *Journal of Statistical Software* 76.1 (Jan. 2017), pp. 1–32. ISSN: 1548-7660. DOI: 10.18637/jss.v076.i01.

[28] Romain D Cazé and Matthijs AA van der Meer. "Adaptive properties of differential learning rates for positive and negative outcomes". In: *Biological cybernetics* 107.6 (2013), pp. 711–719.

[29] Yevgen Chebotar et al. "Combining model-based and model-free updates for trajectory-centric reinforcement learning". In: *arXiv preprint arXiv:1703.03078* (2017).

[30] Anastasia Christakou et al. "Neural and psychological maturation of decision-making in adolescence and young adulthood". In: *Journal of cognitive neuroscience* 25.11 (2013), pp. 1807–1823.

[31] Benjamin A Clegg, Gregory J DiGirolamo, and Steven W Keele. "Sequence learning". In: *Trends in cognitive sciences* 2.8 (1998), pp. 275–281.

[32] Anne Gabrielle Eva Collins and Michael Joshua Frank. "Motor demands constrain cognitive rule structures". In: *PLoS computational biology* 12.3 (2016), e1004785.

[33] Anne Gabrielle Eva Collins and Michael Joshua Frank. "Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning". In: *Cognition* 152 (2016), pp. 160–169.

[34] Anne GE Collins. "Learning Structures Through Reinforcement". In: *Goal-Directed Decision Making*. Elsevier, 2018, pp. 105–123.

[35] Anne GE Collins. "The cost of structure learning". In: *Journal of Cognitive Neuroscience* 29.10 (2017), pp. 1646–1655.

[36] Anne GE Collins, James F Cavanagh, and Michael J Frank. "Human EEG uncovers latent generalizable rule structure during learning". In: *Journal of Neuroscience* 34.13 (2014), pp. 4677–4685.

[37] Anne GE Collins and Jeffrey Cockburn. "Beyond dichotomies in reinforcement learning". In: *Nature Reviews Neuroscience* (2020), pp. 1–11.

[38] 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.

[39] Anne GE Collins and Michael J Frank. "How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis". In: *European Journal of Neuroscience* 35.7 (2012), pp. 1024–1035.

[40] Anne GE Collins et al. "Interactions among working memory, reinforcement learning, and effort in value-based choice: A new paradigm and selective deficits in schizophrenia". In: *Biological psychiatry* 82.6 (2017), pp. 431–439.

[41] Ronald E Dahl et al. "Importance of investing in adolescence from a developmental science perspective". In: *Nature* 554.7693 (2018), pp. 441–450.

[42] Reka Daniel, Angela Radulescu, and Yael Niv. "Intact reinforcement learning but impaired attentional control during multidimensional probabilistic learning in older adults". In: *Journal of Neuroscience* 40.5 (2020), pp. 1084–1096.

[43] Juliet Y. Davidow et al. "An Upside to Reward Sensitivity: The Hippocampus Supports Enhanced Reinforcement Learning in Adolescence". en. In: *Neuron* 92.1 (Oct. 2016), pp. 93–99. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2016.08.031.

[44] Nathaniel D Daw et al. "Model-based influences on humans' choices and striatal prediction errors". In: *Neuron* 69.6 (2011), pp. 1204–1215.

[45] Johannes H Decker et al. "Experiential reward learning outweighs instruction prior to adulthood". In: *Cognitive, Affective, & Behavioral Neuroscience* 15.2 (2015), pp. 310–320.

[46] Kristen Delevich, A Wren Thomas, and Linda Wilbrecht. "Adolescence and "late blooming" synapses of the prefrontal cortex". In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 83. Cold Spring Harbor Laboratory Press. 2018, pp. 37–43.

[47] Samantha DePasque and Adriana Galván. "Frontostriatal development and probabilistic reinforcement learning during adolescence". In: *Neurobiology of Learning and Memory* 143 (2017), pp. 1–7.

[48] Amir Dezfouli and Bernard W Balleine. "Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized". In: *PLoS computational biology* 9.12 (2013), e1003364.

[49] Amir Dezfouli and Bernard W Balleine. "Habits, action sequences and reinforcement learning". In: *European Journal of Neuroscience* 35.7 (2012), pp. 1036–1051.

[50] Amir Dezfouli et al. "Disentangled behavioural representations." In: *NeurIPS*. 2019, pp. 2251–2260.

[51] Amir Dezfouli et al. "Integrated accounts of behavioral and neuroimaging data using flexible recurrent neural network models". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

[52] 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.

[53] Carlos Diuk et al. "Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia". In: *Journal of Neuroscience* 33.13 (2013), pp. 5797–5805.

[54] Kathy T Do, Paul B Sharp, and Eva H Telzer. "Modernizing conceptions of valuation and cognitive-control deployment in adolescent risk taking". In: *Current Directions in Psychological Science* 29.1 (2020), pp. 102–109.

[55] Maria K Eckstein and Anne GE Collins. "Computational evidence for hierarchically-structured reinforcement learning in humans". In: *bioRxiv* (2019), p. 731752.

[56] Maria K Eckstein et al. "Understanding the Unique Advantage of Adolescents in Stochastic, Volatile Environments: Combining Reinforcement Learning and Bayesian Inference". In: *bioRxiv* (2020).

[57] Shiva Farashahi et al. "Feature-based learning improves adaptability without compromising precision". In: *Nature communications* 8.1 (2017), p. 1768.

[58] Alexander Fengler et al. "Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience". In: *Elife* 10 (2021), e65074.

[59] Charles Findling, Nicolas Chopin, and Etienne Koechlin. "Imprecise neural computations as a source of adaptive behaviour in volatile environments". In: *Nature Human Behaviour* 5.1 (2021), pp. 99–112.

[60] Charles Findling et al. "Computational noise in reward-guided learning drives behavioral variability in volatile environments". In: *Nature neuroscience* 22.12 (2019), pp. 2066–2077.

[61] Karin Foerde and Daphna Shohamy. "Feedback Timing Modulates Brain Systems for Learning in Humans". en. In: *Journal of Neuroscience* 31.37 (Sept. 2011), pp. 13157–13167. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.2701-11.2011.

[62] Karin Foerde and Daphna Shohamy. "The role of the basal ganglia in learning and memory: Insight from Parkinson's disease". en. In: *Neurobiology of Learning and Memory*. Memory Impairment and Disease 96.4 (Nov. 2011), pp. 624–636. ISSN: 1074-7427. DOI: 10.1016/j.nlm.2011.08.006.

[63] Erika E Forbes and Ronald E Dahl. "Pubertal development and behavior: hormonal activation of social and motivational tendencies". In: *Brain and cognition* 72.1 (2010), pp. 66–72.

[64] Roy Fox et al. "Multi-level discovery of deep options". In: *arXiv preprint arXiv:1703.08294* (2017).

[65] M. J. Frank et al. "Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning". en. In: *Proceedings of the National Academy of Sciences* 104.41 (2007), pp. 16311–16316. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0706111104.

[66] Michael J Frank and David Badre. "Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis". In: *Cerebral cortex* 22.3 (2011), pp. 509–526.

[67] Michael J. Frank, Lauren C. Seeberger, and Randall C. O'Reilly. "By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism". en. In: *Science* 306.5703 (Dec. 2004), pp. 1940–1943. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1102941.

[68] Willem E Frankenhuis and Nicole Walasek. "Modeling the evolution of sensitive periods". In: *Developmental cognitive neuroscience* 41 (2020), p. 100715.

[69] Nicholas T Franklin and Michael J Frank. "Compositional clustering in task structure learning". In: *PLoS computational biology* 14.4 (2018), e1006116.

[70] Wai-Tat Fu and John R Anderson. "From recurrent choice to skill learning: A reinforcement-learning model." In: *Journal of experimental psychology: General* 135.2 (2006), p. 184.

[71] Adriana Galvan et al. "Earlier development of the accumbens relative to orbitofrontal cortex might underlie risk-taking behavior in adolescents". In: *Journal of Neuroscience* 26.25 (2006), pp. 6885–6892.

[72] Jim Gao. "Machine learning applications for data center optimization". In: (2014).

[73] Andrew Gelman et al. *Bayesian Data Analysis*. en. Chapman and Hall/CRC, Nov. 2013. ISBN: 978-0-429-11307-9. DOI: 10.1201/b16018.

[74] Raphael T. Gerraty et al. "Dynamic Flexibility in Striatal-Cortical Circuits Supports Reinforcement Learning". en. In: *The Journal of Neuroscience* 38.10 (Mar. 2018), pp. 2442–2453. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.2084-17.2018.

[75] Jan Gläscher et al. "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4 (2010), pp. 585–595.

[76] Omer Gottesman et al. "Guidelines for reinforcement learning in healthcare". In: *Nat Med* 25.1 (2019), pp. 16–18.

[77] Shixiang Gu et al. "Continuous deep q-learning with model-based acceleration". In: *International Conference on Machine Learning*. 2016, pp. 2829–2838.

[78] Tobias U. Hauser et al. "Cognitive flexibility in adolescence: Neural and behavioral mechanisms of reward prediction error processing in adaptive decision making during development". en. In: *NeuroImage* 104 (2015), pp. 347–354. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2014.09.018.

[79] Megan M Herting et al. "The role of testosterone and estradiol in brain volume changes across adolescence: a longitudinal structural MRI study". In: *Human brain mapping* 35.11 (2014), pp. 5633–5645.

[80] Clay B Holroyd and Nick Yeung. "Motivation of extended behaviors by anterior cingulate cortex". In: *Trends in cognitive sciences* 16.2 (2012), pp. 122–128.

[81] Quentin JM Huys et al. "Advances in the computational understanding of mental illness". In: *Neuropsychopharmacology* 46.1 (2021), pp. 3–19.

[82] Amir Homayoun Javadi, Dirk HK Schmidt, and Michael N Smolka. "Adolescents adapt more slowly than adults to varying reward contingencies". In: *Journal of cognitive neuroscience* 26.12 (2014), pp. 2670–2681.

[83] Dinesh Jayaraman et al. "Time-agnostic prediction: Predicting predictable video frames". In: *arXiv preprint arXiv:1808.07784* (2018).

[84] Yiding Jiang et al. "Language as an Abstraction for Hierarchical Deep Reinforcement Learning". In: *arXiv preprint arXiv:1906.07343* (2019).

[85] Rebecca M Jones et al. "Adolescent-specific patterns of behavior and neural activity during social reinforcement learning". In: *Cognitive, Affective, & Behavioral Neuroscience* 14.2 (2014), pp. 683–697.

[86] Keno Juechems and Christopher Summerfield. "Where does value come from?" In: *Trends in cognitive sciences* 23.10 (2019), pp. 836–850.

[87] Kentaro Katahira. "How hierarchical models improve point estimates of model parameters at the individual level". en. In: *Journal of Mathematical Psychology* 73 (Aug. 2016). ISSN: 0022-2496. DOI: 10.1016/j.jmp.2016.03.007.

[88] Kentaro Katahira. "The statistical structures of reinforcement learning with asymmetric value updates". In: *Journal of Mathematical Psychology* 87 (2018), pp. 31–45.

[89] Etienne Koechlin and Thomas Jubault. "Broca's area and the hierarchical organization of human behavior". In: *Neuron* 50.6 (2006), pp. 963–974.

[90] Etienne Koechlin, Chrystele Ody, and Frédérique Kouneiher. "The architecture of cognitive control in the human prefrontal cortex". In: *Science* 302.5648 (2003), pp. 1181–1185.

[91] George Konidaris and Andrew G Barto. "Building Portable Options: Skill Transfer in Reinforcement Learning." In: *IJCAI*. Vol. 7. 2007, pp. 895–900.

[92] Wouter Kool, Samuel J Gershman, and Fiery A Cushman. "Cost-benefit arbitration between multiple reinforcement-learning systems". In: *Psychological science* 28.9 (2017), pp. 1321–1333.

[93] Konrad P Kording et al. "Appreciating the variety of goals in computational neuroscience". In: *arXiv preprint arXiv:2002.03211* (2020).

[94] Helena Chmura Kraemer et al. "How can we learn about developmental processes from cross-sectional studies, or can we?" In: *American Journal of Psychiatry* 157.2 (2000), pp. 163–171.

[95] OE Krigolson and CB Holroyd. "Evidence for hierarchical error processing in the human brain". In: *Neuroscience* 137.1 (2006), pp. 13–17.

[96] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. "Human-level concept learning through probabilistic program induction". In: *Science* 350.6266 (2015), pp. 1332–1338.

[97] Brenden M Lake et al. "Building machines that learn and think like people". In: *Behavioral and brain sciences* 40 (2017).

[98] Bart Larsen and Beatriz Luna. "Adolescence as a neurobiological critical period for the development of higher-order cognition". en. In: *Neuroscience & Biobehavioral Reviews* 94 (2018), pp. 179–195. ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2018.09.005.

[99] Tai Sing Lee and David Mumford. "Hierarchical Bayesian inference in the visual cortex". In: *JOSA A* 20.7 (2003), pp. 1434–1448.

[100] Germain Lefebvre et al. "Behavioural and neural characterization of optimistic reinforcement learning". In: *Nature Human Behaviour* 1.4 (2017), pp. 1–9.

[101] Jill Fain Lehman, John E Laird, PS Rosenbloom, et al. "A gentle introduction to Soar, an architecture for human cognition". In: *Invitation to cognitive science* 4 (1996), pp. 212–249.

[102] Yuan Chang Leong et al. "Dynamic interaction between reinforcement learning and attention in multidimensional environments". In: *Neuron* 93.2 (2017), pp. 451–463.

[103] Peng Liao et al. "Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.1 (2020), pp. 1–22.

[104] Marios C Machado, Marc G Bellemare, and Michael Bowling. "A laplacian framework for option discovery in reinforcement learning". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 2295–2304.

[105] Marlos C Machado et al. "Eigenoption discovery through the deep successor representation". In: *arXiv preprint arXiv:1710.11089* (2017).

[106] Zdeňa A Op de Macks et al. "Testosterone levels correspond with increased ventral striatum activation in response to monetary rewards in adolescents". In: *Developmental Cognitive Neuroscience* 1.4 (2011), pp. 506–516.

[107] Travis Mandel et al. "Offline policy evaluation across representations with applications to educational games." In: *AAMAS*. 2014, pp. 1077–1084.

[108] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

[109] Sarah L. Master et al. "Distentangling the systems contributing to changes in learning during adolescence". en. In: *Developmental Cognitive Neuroscience* 41 (Feb. 2020), p. 100732. ISSN: 1878-9293. DOI: 10.1016/j.dcn.2019.100732.

[110] Samuel D McDougle and Anne GE Collins. "Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning". In: *Psychonomic bulletin & review* (2020), pp. 1–20.

[111] Amy McGovern and Andrew G Barto. "Automatic discovery of subgoals in reinforcement learning using diverse density". In: (2001).

[112] Ishai Menache, Shie Mannor, and Nahum Shimkin. "Q-cut—dynamic discovery of sub-goals in reinforcement learning". In: *European Conference on Machine Learning*. Springer. 2002, pp. 295–306.

[113] Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. "Dorsal hippocampus contributes to model-based planning". In: *Nature neuroscience* 20.9 (2017), p. 1269.

[114] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518.7540 (2015), p. 529.

[115] Ida Momennejad et al. "The successor representation in human reinforcement learning". In: *Nature Human Behaviour* 1.9 (2017), p. 680.

[116] Stephen Monsell. "Task switching". In: *Trends in cognitive sciences* 7.3 (2003), pp. 134–140.

[117] Rani Moran et al. "Retrospective model-based inference guides model-free credit assignment". In: *Nature communications* 10.1 (2019), p. 750.

[118] Michael Moutoussis et al. "Change, stability, and instability in the Pavlovian guidance of behaviour from adolescence to young adulthood". In: *PLoS computational biology* 14.12 (2018), e1006679.

[119] Suraj Nair and Chelsea Finn. "Hierarchical Foresight: Self-Supervised Learning of Long-Horizon Tasks via Visual Subgoal Generation". In: *arXiv preprint arXiv:1909.05829* (2019).

[120] Shelley Nason and John E Laird. "Soar-RL: Integrating reinforcement learning with Soar". In: *Cognitive Systems Research* 6.1 (2005), pp. 51–59.

[121] Danielle J Navarro. "Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection". In: *Computational Brain & Behavior* 2.1 (2019), pp. 28–34.

[122] Eric E Nelson et al. "The social re-orientation of adolescence: a neuroscience perspective on the process and its relation to psychopathology". In: *Psychological medicine* 35.2 (2005), p. 163.

[123] Yael Niv. "Reinforcement learning in the brain". en. In: *Journal of Mathematical Psychology*. Special Issue: Dynamic Decision Making 53.3 (June 2009), pp. 139–154. ISSN: 0022-2496. DOI: 10.1016/j.jmp.2008.12.005.

[124] Yael Niv. "The primacy of behavioral research for understanding the brain". In: (2020).

[125] Yael Niv and Angela Langdon. "Reinforcement learning with Marr". In: *Current opinion in behavioral sciences* 11 (2016), pp. 67–73.

[126] Kate Nussenbaum and Catherine A. Hartley. "Reinforcement learning across development: What insights can we draw from a decade of research?" en. In: *Developmental Cognitive Neuroscience* 40 (Dec. 2019), p. 100733. ISSN: 1878-9293. DOI: 10.1016/j.dcn.2019.100733.

[127] Stefano Palminteri, Valentin Wyart, and Etienne Koechlin. "The Importance of Falsification in Computational Cognitive Modeling". en. In: *Trends in Cognitive Sciences* 21.6 (June 2017), pp. 425–433. ISSN: 1364-6613. DOI: 10.1016/j.tics.2017.03.011.

[128] Stefano Palminteri et al. "The computational development of reinforcement learning during adolescence". In: *PLoS computational biology* 12.6 (2016), e1004953.

[129] Xue Bin Peng et al. "MCP: Learning Composable Hierarchical Control with Multiplicative Compositional Policies". In: *arXiv preprint arXiv:1905.09808* (2019).

[130] JS Peper et al. "Sex steroids and brain structure in pubertal boys and girls: a mini-review of neuroimaging studies". In: *Neuroscience* 191 (2011), pp. 28–37.

[131] Anne C. Petersen et al. "A self-report measure of pubertal status: Reliability, validity, and initial norms". en. In: *Journal of Youth and Adolescence* 17.2 (Apr. 1988), pp. 117–133. ISSN: 1573-6601. DOI: 10.1007/BF01537962.

[132] David J Piekarski et al. "Does puberty mark a transition in sensitive periods for plasticity in the associative neocortex?" In: *Brain research* 1654 (2017), pp. 123–144.

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

[134] Vitchyr Pong et al. "Temporal difference models: Model-free deep rl for model-based control". In: *arXiv preprint arXiv:1802.09081* (2018).

[135] Angela Radulescu, Reka Daniel, and Yael Niv. "The effects of aging on the interaction between reinforcement learning and attention." In: *Psychology and aging* 31.7 (2016), p. 747.

[136] Anna N Rafferty et al. "Faster teaching via pomdp planning". In: *Cognitive science* 40.6 (2016), pp. 1290–1332.

[137] Aniruddh Raghu et al. "Deep reinforcement learning for sepsis treatment". In: *arXiv preprint arXiv:1711.09602* (2017).

[138] Jose JF Ribas-Fernandes et al. "A neural signature of hierarchical reinforcement learning". In: *Neuron* 71.2 (2011), pp. 370–379.

[139] 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.

[140] Lionel Rigoux et al. "Bayesian model selection for group studies—revisited". In: *Neuroimage* 84 (2014), pp. 971–985.

[141] Morteza Sarafyazd and Mehrdad Jazayeri. "Hierarchical reasoning by neural circuits in the frontal cortex". In: *Science* 364.6441 (2019), eaav8911.

[142] Marieke E van der Schaaf et al. "Distinct linear and non-linear trajectories of reward and punishment reversal learning during development: relevance for dopamine's role in adolescent decision making". In: *Developmental cognitive neuroscience* 1.4 (2011), pp. 578–590.

[143] Anna C Schapiro et al. "Neural representations of events arise from temporal community structure". In: *Nature neuroscience* 16.4 (2013), p. 486.

[144] Wolfram Schultz, Peter Dayan, and P Read Montague. "A neural substrate of prediction and reward". In: *Science* 275.5306 (1997), pp. 1593–1599.

[145] Gideon Schwarz et al. "Estimating the dimension of a model". In: *Annals of statistics* 6.2 (1978), pp. 461–464.

[146] David Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144.

[147] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), pp. 484–489.

[148] Uri Simonsohn. "Two lines: A valid alternative to the invalid testing of U-shaped relationships with quadratic regressions". In: *Advances in Methods and Practices in Psychological Science* 1.4 (2018), pp. 538–555.

[149] Özgür Şimşek and Andrew G Barto. "Using relative novelty to identify useful temporal abstractions in reinforcement learning". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004, p. 95.

[150] Satinder Singh, Richard L Lewis, and Andrew G Barto. "Where do rewards come from". In: *Proceedings of the annual conference of the cognitive science society*. Cognitive Science Society. 2009, pp. 2601–2606.

[151] Alec Solway et al. "Optimal behavioral hierarchy". In: *PLoS computational biology* 10.8 (2014), e1003779.

[152] Leah H Somerville, Rebecca M Jones, and BJ Casey. "A time of change: behavioral and neural correlates of adolescent sensitivity to appetitive and aversive environmental cues". In: *Brain and cognition* 72.1 (2010), pp. 124–133.

[153] Jeffrey M Spielberg et al. "Exciting fear in adolescence: does pubertal development alter threat processing?" In: *Developmental cognitive neuroscience* 8 (2014), pp. 86–95.

[154] Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. "The hippocampus as a predictive map". In: *Nature neuroscience* 20.11 (2017), p. 1643.

[155] Laurence Steinberg. "A social neuroscience perspective on adolescent risk-taking". In: *Developmental review* 28.1 (2008), pp. 78–106.

[156] Klaas Enno Stephan et al. "Bayesian model selection for group studies". In: *Neuroimage* 46.4 (2009), pp. 1004–1017.

[157] Richard S Sutton. "Dyna, an integrated architecture for learning, planning, and reacting". In: *ACM Sigart Bulletin* 2.4 (1991), pp. 160–163.

[158] 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.

[159] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. en. MIT Press, Oct. 2018. ISBN: 978-0-262-35270-3.

[160] Momchil Tomov et al. "Discovery of Hierarchical Representations for Efficient Planning". In: *BioRxiv* (2018), p. 499418.

[161] Brandon M Turner, Birte U Forstmann, Mark Steyvers, et al. *Joint models of neural and behavioral data*. Springer, 2019.

[162] Wouter Van Den Bos et al. "Better than expected or as bad as you thought? The neurocognitive development of probabilistic feedback processing". In: *Frontiers in human neuroscience* 3 (2009), p. 52.

[163] David C Van Essen and John HR Maunsell. "Hierarchical organization and functional streams in the visual cortex". In: *Trends in neurosciences* 6 (1983), pp. 370–375.

[164] Deena M Walker et al. "Adolescence and reward: making sense of neural and behavioral changes amid the chaos". In: *Journal of Neuroscience* 37.45 (2017), pp. 10855–10866.

[165] Jane X Wang et al. "Prefrontal cortex as a meta-reinforcement learning system". In: *Nature neuroscience* 21.6 (2018), p. 860.

[166] Sumio Watanabe. "A widely applicable Bayesian information criterion". In: *Journal of Machine Learning Research* 14.Mar (2013), pp. 867–897.

[167] CM Wessinger et al. "Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging". In: *Journal of cognitive neuroscience* 13.1 (2001), pp. 1–7.

[168] Marco A Wiering. "Multi-agent reinforcement learning for traffic light control". In: *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*. 2000, pp. 1151–1158.

[169] Robert C Wilson and Anne GE Collins. "Ten simple rules for the computational modeling of behavioral data". In: *Elife* 8 (2019), e49547.

[170] David Wingate et al. "Compositional policy priors". In: (2013).

[171] Liyu Xia and Anne Gabrielle Eva Collins. "Temporal and state abstractions for efficient learning, transfer and composition in humans". In: *bioRxiv* (2020).

[172] Liyu Xia et al. "Modeling Changes in Probabilistic Reinforcement Learning during Adolescence". In: *bioRxiv* (2020).

[173] Danfei Xu et al. "Neural task programming: Learning to generalize across hierarchical tasks". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 1–8.

[174] Danfei Xu et al. "Regression Planning Networks". In: *arXiv preprint arXiv:1909.13072* (2019).

[175] Angela J Yu and Jonathan D Cohen. "Sequential effects: superstition or rational behavior?" In: *Advances in neural information processing systems*. 2009, pp. 1873–1880.

[176] Noah Zarr and Joshua W Brown. "Hierarchical error representation in medial prefrontal cortex". In: *NeuroImage* 124 (2016), pp. 238–247.

[177] Yufan Zhao, Michael R Kosorok, and Donglin Zeng. "Reinforcement learning design for cancer clinical trials". In: *Statistics in medicine* 28.26 (2009), pp. 3294–3315.