Title
Computational Models of Learning and Hierarchy

Author
Eckstein, Maria Katharina

Publication Date
2020

Peer reviewed|Thesis/dissertation

Computational Models of Learning and Hierarchy

by

Maria K. Eckstein

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Anne G. E. Collins, Chair
Professor Silvia A. Bunge
Professor Ming Hsu
Professor Thomas L. Griffiths

Fall 2020

Computational Models of Learning and Hierarchy

Abstract

Computational Models of Learning and Hierarchy

by

Maria K. Eckstein

Doctor of Philosophy in Psychology

University of California, Berkeley

Assistant Professor Anne G. E. Collins, Chair

The aim of this thesis is to create precise computational models of how humans create and use hierarchical representations when solving complex problems. In the process, the thesis aims to understand human learning more generally, and investigates the method of computational modeling itself. The main result of the thesis is that hierarchical reinforcement learning –the layering of multiple reinforcement-learning processes at different levels of abstraction– provides a precise and comprehensive model of human behavior in complex tasks, and has the promise to explain how hierarchical representation can be created when interacting with a problem. Our investigation of human learning shows that learning proceeds differently at different ages, and suggests that different stages of life might be optimized to solve different problems. Our investigation of computational modeling reveals that even though computational models are powerful tools for compressing complex datasets into a small number of model parameters, these parameters are not generic and task-independent, as commonly believed. Instead, model parameters should be interpreted as maximally-compact behavioral measures that are fundamentally tied to task context.

Contents

Acknowledgments

I would like to thank my co-authors for their contributions to this work: Sarah L. Master, Liyu Xia, Ronald E. Dahl, Linda Wilbrecht, and Anne G.E. Collins. Sarah Master was the soul of the SLCN project (chapters 1 & 2) while she was lab manager here, and I don't know what I would have done without her. Jimmy Xia has been a great lab mate, always excited about discussions and brain storms. And Anne was the most wonderful advisor I could have wished for. I would also like to thank Silvia Bunge, who let me visit her lab for a year prior to starting grad school and advised me for another year during grad school, and who shared an incredible sense of excitement about research, and so much academic wisdom. Jack Gallant and Tom Griffiths were crucial influences throughout my PhD. Their classes have shaped my thinking more than anything else.

My family has been crucial. My brother Korbinian, physicist, one of the brightest people I have had the pleasure of meeting in my life so far; who always has an open ear, who asks the best questions, straight to the point, and who has the clearest thoughts. My parents and grandparents, none of whom have academic backgrounds, but who know enough about life to have contributed some of the most fundamental insights into what it means to get a PhD in the first place.

So many people have helped me with this work, most of them unknowingly or indirectly. The friends' dinners and adventure trips were crucial, as were so many late-night discussions, dance parties, Ironworks sessions, picnics, climbing and skiing trips, co-op dinners, etc. Naming everyone would fill pages, but here is a start, in no particular order: Ignasi and Anna, Carlos, Kate R and Kate B (damn this is fun); Kim and Dan, Rebecca and Evan, Edna, Cashmere, Ryan, Adam, Vicenc, Forrest and Annie, and everyone else who hangs out at Hilgard; Chenling, Oliver, Sachi, Brandon, Maria G, Robyn, Marc, Yev, Mathilde, Rudy, Eddie, Fadzai, Kevin, Henry and Christy, Adam, and everyone else at HiP; Deepthi and Laura, Ingeborg Treu, Sylvain, Guillaume, Xabi and Alba, Eric and Aldo, and all the people at I-House; Mostafa and Kunsel, Abhishek, and Nikhil, and everyone else at Claremont; Michael and Ori, Delphine, and everyone on Delaware; Marcus, and everyone at little mountain; Christina and Manon, Aida and Amin, David, Falk, Jocelyn, Jennifer, Stephan, and everyone in clinical, cognition, neuro, and cog-neuro; Mike, for all the climbing and skiing and everything else; Sasha, Ed, Kevin, Jess, David, Timo, Anna, and everyone else at DeepMind; Sam, Jimmy, Sarah, Aram, Haley, Lucy, Amy, Milena, Aspen, Beth, and of course Anne, and everyone else in the Collins lab; Rich, Bruno, Jack, Tom, Tania, Lucia, Nina, Bob, Steve, Celeste, Silvia, Marc, Joni, and all the other professors at Berkeley who I had the pleasure meeting; and Mia, Franca, Felix, Stani, Nadim, Georgi, and everyone else from Munich. I could not be more grateful to everyone.

Chapter 1

Introduction

This chapter motivates the study of hierarchical representations to understand complex human thought, introduces computational modeling as the main method of this thesis, and lays out a roadmap of the chapters ahead.

1.1 Seemingly Simple Abilities

Even though many things we as humans do in everyday life seem effortless to us, most of them are not simple. In the early days of artificial intelligence, researchers started compiling "MABA-MABA" lists, which stands for "Machines are better at - Men are better at" (Fitts, 1951). For example, while machines surpassed humans in the game of chess several decades ago, humans are still better at grasping and moving chess pieces. And while machines are better at storing numbers and doing complex and precise calculations on them, it takes a human to decide which calculations are meaningful, and to interpret the results.

Sometimes, observing phenomena from a different perspective helps appreciate their complexity. Even though we take our cognitive abilities for granted because they seem easy to us, they are in reality often breathtakingly complex (and still out of reach for artificial systems). For example, consider our ability to recognize structure in the world. It might seem obvious that bears have similar bone structures to moose, but might catch similar diseases as salmon, given that they are biologically similar to moose, but they eat salmon. Which kind of mental representation is necessary to make such inferences (Tenenbaum et al., 2011)? Then, consider how many different contexts we encounter every day, and with which ease we select appropriate strategies for each, reusing old strategies if useful, and otherwise creating new ones as needed. How do we represent these strategies, how do we determine which one is appropriate, and how do we learn new ones (Collins and Koechlin, 2012)? Lastly, consider how we act in situations whose complexity goes beyond our cognitive abilities, for example because they offer too many choices over too many time steps. How do we know how to break such problems into suitable sub-problems, and how do we approximate steps that are too complex to solve (Huys et al., 2015)?

1.2 Hierarchy

One answer to this question is: by employing structured, hierarchical representations. The term hierarchy has been used in many ways, and two common ones include "processing hierarchies" and "representational hierarchies". In processing hierarchies, higher levels exert control over lower levels, for example by controlling the flow of information or by setting the agenda for lower levels (E. K. Miller and Cohen, 2001; Vezhnevets et al., 2017). In representational hierarchies, higher levels form abstractions over lower levels, such that lower levels contain concrete, sensory, and fine-grained information, whereas higher levels contain general, conceptual, and integrated information (Badre, 2008; Sutton et al., 1999; Tenenbaum et al., 2011).

Both flavors of hierarchy have been invoked to explain complex cognition, oftentimes in conjunction: long-term planning problems can be solved using hierarchy because higher levels have a bird's eye view of the problem and can identify appropriate sub-goals along the way, while lower levels have a finer temporal resolution and are able to reach each sub-goal (Ribas Fernandes et al., 2011; Sutton et al., 1999; Vezhnevets et al., 2017). Cognitive flexibility and task switching can be achieved using a hierarchy over strategies: A high-level strategy is in charge of selecting one of several low-level strategies, and low-level strategies guide actual behavior. Whereas the high-level strategy is trained to identify the best low-level strategy for each context, the low-level strategies are trained to optimize behavior within each context (Collins and Frank, 2013; Vezhnevets et al., 2020). Lastly, inference in complex domains can be explained by using structured, hierarchical representations of the world around us, which are combined with new evidence to yield conclusions (Kemp and Tenenbaum, 2008).

Because of its potential to solve problems of extensive, real-world complexity, hierarchy is a key component in many artificial intelligence algorithms that attempt to overcome problems such as the combinatorial explosion of possibilities, sparseness of feedback, or extensively long planning horizons by using meta-learning, generalization, and abstraction (Duan et al., 2016; Finn et al., 2017; Sutton et al., 1999; Vezhnevets et al., 2017; Wang et al., 2016). In psychology, hierarchical representations in humans (and animals) have been the focus of many research programs (Botvinick, 2012; Chase and Simon, 1973; Diuk, Schapiro, et al., 2013; Gershman and Niv, 2010), and in neuroscience, the hierarchical organization of brain structures has guided and inspired the investigation of brain function (Badre and D'Esposito, 2009; Geddes et al., 2018; E. K. Miller and Cohen, 2001). In the spirit of interdisciplinary cognitive science, there has been a constant flow of information between these three fields (Collins, 2019; Griffiths et al., 2019; Lake et al., 2017).

1.3 Learning

Nevertheless, hierarchical representations are only useful when they capture the essence of a problem, and what this essence is is seldomly clear from the outset. Most hierarchical representations therefore arise from the interaction between agents and the problems they face, and are created over time in a learning process. The human ability to learn (and teach) has been argued to be what distinguishes humans from other species (Premack, 2007), and the ability to learn like a human

child has been argued to be what will provide the break-through in creating artificial intelligence (Turing, 1950). Unsurprisingly, the study of learning has been at forefront of psychological research since the beginning, and has remained essential throughout methodological paradigm shifts (Gopnik and Tenenbaum, 2007; Nussenbaum and Hartley, 2019; Tolman, 1948; Watson, 1913).

1.4 Computational Modeling

In order to study learning and hierarchy, many researchers have employed computational models. Computational cognitive modeling dates back to the earliest days of artificial intelligence and the birth of the cognitive sciences (Atkinson and Shiffrin, 1968; G. A. Miller, 1956; Newell et al., 1959). The basic idea of cognitive modeling is to formalize a theory about cognitive processes sufficiently to formulate it as a process model, a step-by-step algorithm or recipe of how quantitative variables are combined together to give rise to cognition. This process model can be subjected to statistical methods in order to determine how accurately it captures human behavior (Daw, 2011; Palminteri et al., 2017; Wilson and Collins, 2019). Computational modeling has many advantages: it allows –indeed forces– researchers to test very precise theories, avoiding overly broad or non-falsifiable statements. In addition, models have the inherent ability to "run forward" and make predictions outside the domain in which they were designed. Lastly, and most interestingly for the study of hierarchy, computational models make it possible to test theories that are complex, containing many moving parts with intricate relations.

Current learning research frequently employs cognitive modeling, oftentimes adapting algorithms from artificial intelligence, such as reinforcement learning and Bayesian inference (Russell and Norvig, 2009). Reinforcement learning, a concept proposed by behaviorist psychologists (Skinner, 1977; Watson, 1913) and formalized by computer scientists (Sutton and Barto, 2017), states that biological / artificial agents learn through constant interaction with their world. Agents produce actions and receive outcomes, and then adjust their actions based on the valence of the outcome, be it rewarding or punishing. In order to do this efficiently, agents consolidate the entire reward history of actions into "action values", which represent the expected valence of action outcomes, and can be stored and updated very quickly and efficiently. The framework of reinforcement learning is mathematically precise, and abstract enough to be applied to a variety of possible problems (e.g., training an animal, playing a computer game, picking a restaurant), with any kind of action (e.g., motor movement, key press, strategy), or reward (e.g., praise, points won, good food).

1.5 Aim of the Thesis

The aim of this thesis is to create precise computational models of how humans create and use hierarchical representations when solving complex problems. In the process, the thesis aims to understand human learning more generally, and investigates the method of computational modeling itself. The main result of the thesis is that hierarchical reinforcement learning –the layering

of multiple reinforcement-learning processes at different levels of abstraction– provides a precise and comprehensive model of human behavior in complex tasks, and has the promise to explain how hierarchical representation can be created when interacting with a problem. Our investigation of human learning shows that learning proceeds differently at different ages, and suggests that different stages of life might be optimized to solve different problems. Our investigation of computational modeling reveals that even though computational models are powerful tools for compressing complex datasets into a small number of model parameters, these parameters play a different role than commonly assumed, and are not as generic as often believed. We suggest a refined interpretation of learning models that potentially can help resolve discrepancies in the previous computational modeling literature.

1.6 Thesis Outline

Chapters 2 and 3 investigate human learning and computational modeling, respectively, and lay the foundation for chapters 4 and 5, which investigate the use and creation of hierarchical representations. In chapter 2, we investigate how participants of different ages (8-35 years) perform a task in which task contingencies can switch without notice, and in which not all feedback is reliable (positive feedback is reliable, but negative feedback is not). We find that adolescents aged 13-15 years performed better at this task than younger children and teenagers, but also better than older teenagers and even adults. The reason for this unique advantage was that 13-to-15-year-olds were better at overruling unreliable negative feedback. This behavior was captured by smaller learning rates for negative outcomes in a computational reinforcement learning model. When analyzed through the lens of Bayesian Inference rather than reinforcement learning, 13-to-15-year-olds showed the most accurate mental models of the task. Analyzing the parameters of the reinforcement learning and Bayesian Inference models jointly, 13-to-15-year-olds occupied a developmental sweet spot, still showing child-like time scales of learning, but already showing adult-like levels of task proficiency. In addition to these U-shaped developments, both models also revealed monotonic age changes, such that decision noise decreased with age, whereas choice persistence increased.

In chapter 3, we investigate a larger dataset that includes the task of chapter 2 as one of three learning tasks that were given to the same group of more than 300 participants aged 8-35 years. The goal of this project was to relate the results of different computational models between tasks, challenging the common assumption that computational model parameters capture the same cognitive processes across task, and that computational models of different tasks are directly comparable. We show that one model parameter, the decision noise, captured similar cognitive processes across tasks, and showed similar values across tasks for the same participant. Other model parameters, on the contrary, most notably learning rates, captured different cognitive processes across tasks, and showed different values across tasks for the same participant. For example, two learning rate parameters captured partially overlapping processes in two of the tasks; in two other tasks, they captured orthogonal processes; and a different kind of learning rate captured overlapping processes, but in the inverse way across two tasks. This shows that model parameters are not nec-

essarily comparable between studies, and that computational models are very task-specific. Due to their flexibility, computational model can capture a wide range of cognitive processes, and model parameters seem to adapt flexibly to varying demands. This highlights the need for careful validation of any possible interpretation of computational models in terms of underlying cognitive mechanisms, as they do not appear to broadly generalize across even similar tasks.

In chapter 4, we investigate whether human behavior in a complex learning task can be described using hierarchical reinforcement learning. Participants learned to make choices in several different contexts, and we hypothesized that learning would proceed according to hierarchical reinforcement learning: participants should learn a low-level strategy for each context to encode the choices, and additionally a high-level strategy that determines which low-level strategy to use in each context. We compared this model to two competitors: a flat reinforcement learning model, which used reinforcement learning but lacked hierarchy; and a hierarchical Bayesian model, which was hierarchical but used inference instead of reinforcement learning to select low-level strategies. We designed several tests within the task to examine whether participants' behavior showed signs of value learning at two levels of abstraction, and indeed, participants showed all the expected markers. Importantly, only the hierarchical reinforcement learning model, and not the flat or the Bayesian model, were able to replicate these patterns. This suggests that human behavior in a learning and generalization task usually considered complex enough to necessitate complex inference processes, can instead be captured by comparatively simple and biologically realistic computations, using hierarchical reinforcement learning.

In chapter 5, we investigate how humans create hierarchical representations of multi-step strategies, which are particularly relevant to frequent, every-day efficient hierarchical decision making. We designed a task in which participants had to discover complex action sequences to get reward. Importantly, these complex sequences were composed of simpler action sequences, and in order to learn the task, participants could learn simple action sequences and combine these into complex actions, rather than attempting to learn complex action sequences directly. Participants' behavior supported this hypothesis, confirming that participants created a hierarchical structure composed of basic actions, simple action sequences, and complex action sequences.

Chapter 2

Developmental Changes in Learning

This chapter presents a research study in which adolescents aged 13-15 years performed better than both children and adults. Two different computational models –reinforcement learning and Bayesian inference– are fitted, compared, and interpreted jointly to understand how adolescents achieved this. [1]

Abstract

During adolescence, youth venture out, explore the wider world, and are challenged to learn how to navigate novel and uncertain environments. We investigated whether adolescents are uniquely adapted to this transition, compared to younger children and adults. In a stochastic, volatile learning task with a sample of 291 participants aged 8-30, we found that adolescents 13-15 years old outperformed both younger and older participants. We developed two independent cognitive models, and used hierarchical Bayesian model fitting to assess developmental changes in underlying cognitive mechanisms. Choice parameters in both models improved monotonously. By contrast, update parameters peaked closest to optimal values in 13-15 year-olds. Combining both models using principal component analysis yielded new insights, revealing that three components contributed to the early to mid-adolescent performance peak. This research highlights early to mid-adolescence as a neurodevelopmental window that may be more optimal for behavioral adjustment in volatile and uncertain environments. It also shows how detailed insights can be gleaned by combining cognitive models.

2.1 Introduction

In mammals and other species with parental care, there is typically an adolescent stage of development in which the young are no longer supported by parental care but are not yet adult. This adolescent period can be identified in many species across the animal kingdom (Natterson-Horowitz

[1] This chapter has separately been submitted for publication, with the contributions of co-authors Sarah L. Master, Ronald E. Dahl, Linda Wilbrecht, and Anne G.E. Collins.

and Bowers, 2019) and is increasingly viewed as a critical epoch of development in which organisms explore the world, make critical decisions, and learn about important features of their environment (DePasque and Galván, 2017; Laube et al., 2020; Piekarski, Johnson, et al., 2017; Steinberg, 2005). All of these behaviors require learning and decision making that will likely have critical short and long-term impact on survival of the organism (Frankenhuis and Walasek, 2020). In humans, and likely many other species, the transition to independence almost always involves environmental changes and increased exposure to stochastic, uncertain outcomes. It is therefore possible that adolescent brains and cognitive capabilities are specifically adapted to succeed in such situations (Dahl et al., 2018; Davidow et al., 2016; Johnson and Wilbrecht, 2011; Lourenco and Casey, 2013; Sercombe, 2014).

To test this idea, we compared the behaviors of 291 participants, including 191 children and adolescents aged 8-17, and 112 adults (55 adults from the community, aged 25-30; 57 university undergraduates, aged 18-28; suppl. Fig. 2.6), on a task with volatile structure and stochastic outcomes (Fig. 2.1A, B). The goal of the task was to collect rewards, which were hidden in one of two locations (Fig. 2.1A). Which location was rewarding changed unpredictably several times ("task switch"), and the rewarded location provided rewards only 75 percent of the time (Fig. 2.1A). The task's main challenge lay in discriminating chance outcomes during stable task periods from task switches, and respond appropriately to each. It required the integration of stochastic feedback and the adaptation to a volatile environment, and thus theoretically mirrored the challenges of the adolescent period. We therefore hypothesized that adolescents would outperform both younger and older participants. Our data supported this hypothesis.

We used computational modeling to understand the cognitive processes that underlie adolescents' superior performance, as well as the strategies employed by younger children and older teenagers and adults. A variety of algorithms have been used to model human cognition, including Reinforcement learning (RL) and Bayesian inference (BI). The basic idea of RL is that choice options have "values" (their expected long-term cumulative reward). The goal of RL—maximizing long-term outcomes—can therefore be achieved by selecting options according to their values. The core of RL lies in approximating values accurately and efficiently, which can be achieved by performing small incremental updates every time an outcome is observed. This incremental procedure avoids overemphasizing any single outcome and allows RL to treat stochastic outcomes appropriately. The size of the increment captures the integration time scale: the emphasis given to recent vs. less recent outcomes. In volatile environments, RL adjusts to abrupt changes by gradually unlearning and relearning values.

RL frames our task as a *learning* problem: Participants continuously learn and adjust the value of each choice option based on trial-by-trial feedback (Fig. 2.3A, left). The same learning process occurs during stable periods and after task switches, without an explicit concept of switching: Behavioral change arises when enough updates have occurred for the values of one option to dip below the other. Basic RL algorithms are suboptimal in volatile and structured environments like ours, but can be augmented for more efficient performance (see Methods). In all cases, RL models make the fundamental assumption that humans solve challenges through continuous, value-based learning.

The most common approach in computational modeling studies is to select one type of cog-

nitive model (e.g., RL), and compare different variants of this type to find the best-fitting one, which is then interpreted as the cognitive process employed by participants, using quantitative criteria of model fit such as Bayes factors (e.g. Wagenmakers 2007), minimum description length (e.g. Grünwald 2007), or cross-validation (e.g. Browne 2000). The problem with this approach is that it cannot rule out whether a model of a different type (e.g., BI) would fit the data better altogether. The issue can be mitigated by verifying that the chosen model reproduces human behavior adequately (Palminteri et al., 2017), such that the goal of explaining behavior is achieved, and constructing additional models of a different type is unnecessary. However, a more troublesome concern is that different types of models frame behavior in terms of different cognitive processes, and one framing can be more informative, more interpretable, or summarize behavior in a more meaningful way than another. This problem is more difficult to solve because conceptual model fit is hard to quantify, and goes back to fundamental questions about the intended function of models as "scientific" (providing explanations) or "technological" (being predictive; Bernardo and Smith, 2009; Navarro, 2019, p.238). Model selection always needs to find the balance between these more qualitative (e.g., generality, explanatory adequacy) and more quantitative criteria (e.g., descriptive adequacy, complexity; Jacobs and Grainger, 1994). In short, numerical model fit (e.g., BIC, AIC, WAIC) is not the only descriptor of model quality - qualitative aspects such as model generality and explanatory power play crucial roles as well.

To address these concerns, we fitted two families of models to the current task, RL and BI. BI models combine "prior" knowledge with new observations to arrive at "posterior" conclusions about unobservable features of the environment ("hidden states"; Perfors et al., 2011; Sarkka, 2013). BI models therefore employ a "predictive model", which specifies how likely different observations are to arise from different hidden states ("likelihood"). The BI inference cycle—combining prior and likelihood to get a posterior—can continue infinitely, using each step's posterior as the prior for the subsequent step. Bayesian models deal well with stochastic outcomes because extreme likelihoods are balanced by stable priors. Environmental volatility is modeled explicitly as a change in hidden state.

BI models frame our task as an *inference* problem: Participants know that the task has two hidden states ("Left choice is correct" and "Right choice is correct"; Fig. 2.3A, right), and use trial-by-trial outcomes to determine which state is more likely. Having inferred the state, the appropriate action (left or right) can be selected. In other words, participants entertain a mental model of the task, which specifies how likely each outcome (reward, no reward) is in each hidden state, and how likely state transitions occur. In summary, whereas RL claims that participants adapt to task switches by continuously relearning choice values, BI claims that they represent state transitions explicitly, changing their behavior after detecting a switch.

We used the BI model to assess how participants' mental models developed with age. We hypothesized that adolescents' models would be better tuned for volatile and stochastic environments than children's and adults'. Because the BI model employed rational, Bayes-optimal behavior, it also allowed us to evaluate whether and how participants deviated from it: We hypothesized that adolescents would use the most accurate mental models. In addition, both RL and BI models contained parameters that controlled choice: decision noise and persistence. We expected both to decrease monotonously with age, as has been consistently observed (e.g., Master et al., 2020;

for review, see Nussenbaum and Hartley, 2019). For RL learning-rate parameters, which control integration time scales, we did not have a priori predictions because past studies differed in experimental context (Davidow et al., 2016; Master et al., 2020) and provided conflicting results (Nussenbaum and Hartley, 2019).

Model-agnostic analyses revealed that adolescents (13-15 years) outperformed younger and older participants in several measures of task performance, as predicted. We used state-of-the-art hierarchical Bayesian methods to fit RL and BI models to participant behavior, assessing age changes directly and in a statistically unbiased way (Methods; Katahira, 2016; M. D. Lee, 2011; van den Bos et al., 2017). Both models qualitatively captured participants' behavior, and choice-related parameters showed the expected age trajectories. The BI model confirmed the unique tuning of adolescents' (13-15) mental model to the task, and the RL model revealed complex developmental trajectories of learning rates. Going beyond individual models, we then used Principal Component Analysis (PCA) to expose the dimensions of largest variance in the shared parameter space. Variance between participants was captured in just four dimensions, three of which showed marked and separable developmental changes.

2.2 Results

Task

After completing a child-friendly tutorial (Methods), participants performed the following task: On each trial, two identical green boxes appeared on the screen. Participants chose one, and either received a reward (gold coin) or not (empty box; Fig. 2.1A). One box was rewarded in 75% of the trials on which it was chosen, whereas the other was never rewarded (*stochastic* aspect). After a variable number of trials, an unsignaled switch occurred, after which the opposite box was rewarding. Several unpredictable switches occurred over 120 trials (*volatility* aspect; Fig. 2.1B). Participants' goal was to collect as many gold coins as possible. More task details are provided in the Methods.

Task Behavior

Participants gradually adjusted their behavior after task switches, and on average started selecting the correct action about 3 trials after a switch, reaching asymptotic performance thereafter (Fig. 2.1C). Participants almost always repeated actions ("stayed") after receiving positive outcomes ("- +" and "+ +"), and often switched actions after receiving two negative outcomes ("- -"). Behavior was ambivalent after receiving a positive followed by a negative outcome ("+ -"), i.e., on "potential" switch trials (Fig. 2.1D; for age differences, see suppl. Fig. 2.15).

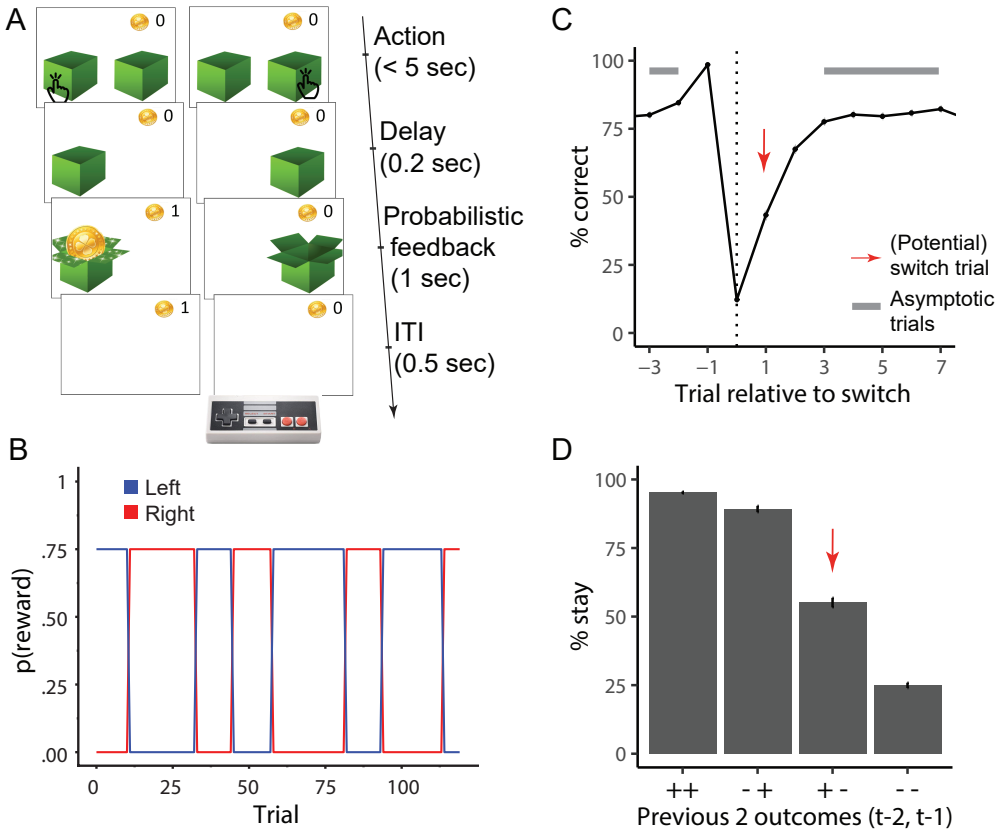


Figure 2.1: (A) Task design. On each trial, participants chose one of two boxes, using the two red buttons of the shown game controller. The chosen box either revealed a gold coin (left) or was empty (right). The probability of coin reward was 75% on the rewarded side, and 0% on the non-rewarded side. (B) The rewarded side changed multiple times, according to unpredictable task switches. (C) Average human performance and standard errors, aligned to "true" task switches (dotted line; trial 0). Switches only occurred after rewarded trials (Methods), resulting in performance of 100% on trial -1. The red arrow shows the "switch trial", grey bars show trials of asymptotic performance. (D) Average probability of repeating a previous action ("stay"), as a function of the two previous outcomes ($t-2$, $t-1$) for this action ("+": reward; "-": no reward). Error bars indicate between-participant standard errors. Red arrow highlights "potential switch trials", i.e., when a rewarded trial is followed by a non-rewarded one, which—from participants' perspective—is consistent with a task switch.

Focusing on age differences, adolescents 13-15 outperformed younger groups age 8-13 and adults (18-30) on several measures of performance (Fig. 2.2, suppl. Fig. 2.13, Fig. 2.3C-F). We tested age effects statistically with (logistic) mixed-effects regression (Methods). All measures of performance showed positive linear effects of age, indicating improved performance with age, as well as negative quadratic effects, consistent with a U-shaped relationship where adolescents 13-15

perform the task more accurately than both younger or older participants (Table 2.1).

Table 2.1: Statistics of mixed-effects regression models predicting performance measures from sex (male, female), age (years and months; "lin."), and squared age ("qua."). Overall accuracy, stay after potential (pot.) switch, and asymptotic performance were modeled using logistic regression, and z-scores are reported. Log-transformed response times on correct trials were modeled using linear regression, and t-values are reported. * $p < .05$; ** $p < .01$, *** $p < .001$.

Performance measure (Figure)	Predictor	β	z / t	p	sig.
Overall accuracy (2.2A)	Age (lin.)	0.054	3.1	0.0017	**
	Age (qua.)	-0.0014	-3.0	0.0024	**
	Sex	0.0074	0.2	0.82	
Response times (2.2B)	Age (lin.)	-0.17	-8.4	< 0.001	***
	Age (qua.)	-0.004	-7.4	< 0.001	***
	Sex	0.19	5.1	< 0.001	***
Stay after (pot.) switch (2.2C)	Age (lin.)	0.42	3.8	< 0.001	***
	Age (qua.)	-0.010	-3.5	< 0.001	***
	Sex	0.27	1.3	0.19	
Asymptotic performance (2.2D)	Age (lin.)	0.19	4.2	< 0.001	***
	Age (qua.)	-0.0048	-4.0	< 0.001	***
	Sex	0.025	0.3	0.77	

To determine the age of peak performance, we binned participants into equal-sized groups based on age (Methods; suppl. Fig. 2.13D-F; Fig. 2.3C-F). Overall task performance peaked in 13-15 year-olds (mid-adolescence), and declined steeply for both younger and older participants (Fig. 2.3C). 13-15 year-olds were also more willing to repeat previous actions after single negative outcomes, especially compared to younger children ("stay" on "(potential) switch trials"; Fig. 2.3E). This suggests that 13-15 year olds were most persistent in the face of negative feedback. 13-15 year-olds also performed best during stable task periods without switches, showing the highest accuracy on asymptotic trials, especially compared to younger participants (Fig. 2.3F).

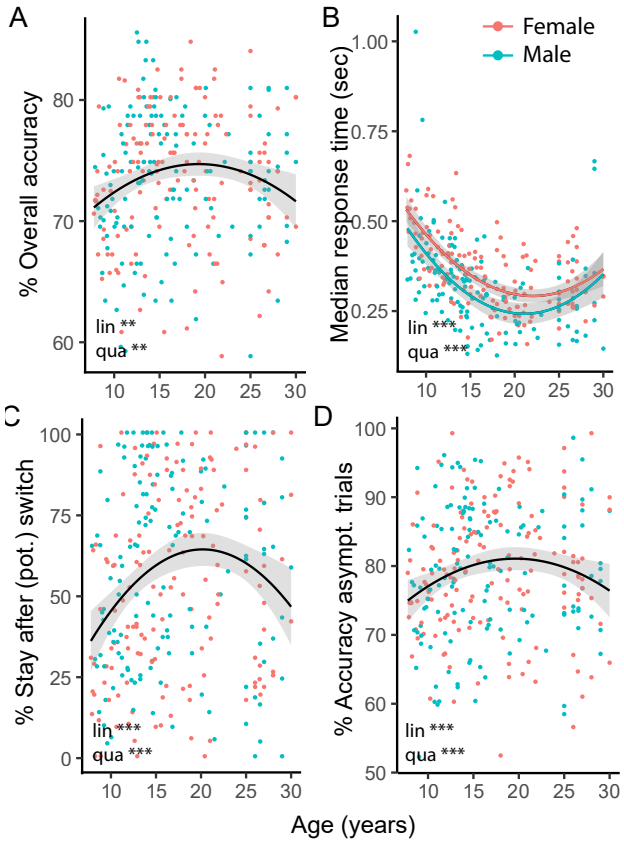


Figure 2.2: Task performance across age. Each dot shows one participant, color denotes sex. Curves show the fits of linear regression models, with shaded standard errors. "Lin.": significant effect of age on outcome; "qua.": significant effect of squared age on outcome. Stars denote p-values like before. (A) Percentage of correct choices across the entire task (120 trials). (B) Median response times on correct trials. Regression coefficients differed between males and females. (C) Fraction of stay trials after (potential, "pot.") switches (red arrows in Fig. 2.1C). (D) Accuracy on asymptotic trials (grey bars in Fig. 2.1C).

Furthermore, 13-15 year-olds adapted their choices more optimally to previous outcomes than younger or older participants. To show this, we used mixed-effects logistic regression to predict actions on trial t from predictors encoding positive or negative outcomes on trials $t-i$, for delays $1 \leq i \leq 8$ (Methods). The effects of positive outcomes were several times larger than the effects of negative outcomes (suppl. Table 2.8; Fig. 2.13B-F), in accordance with task dynamics: Positive outcomes indicated with certainty that an action was correct, justifying their strong effect on behavior, whereas negative outcomes were ambivalent as to whether a switch occurred or not, and should have smaller effects. Crucially, this pattern differed between participants of different ages, as revealed by interactions between age and previous outcomes (suppl. Fig. 2.13B, C, E, and F; suppl. Table 2.8): On trials $t-1$ and $t-2$, both positive and negative outcomes interacted with

age and squared age (all $p's < 0.014$; suppl. Table 2.8), such that the effect of positive outcomes increased with age and then slowly plateaued (suppl. Fig. 2.13C, F). For negative outcomes, the sign of the interaction was opposite for trials $t-1$ versus $t-2$ (all $p's < 0.046$; suppl. Table 2.8). This shows that the effect of negative outcomes flipped, being weakest in 13-15 year olds for trial $t-1$ (Fig. 2.13F), but strongest for trial $t-2$. In other words, 13-15 year-olds were best at ignoring single, ambivalent negative outcomes ($t-1$), and most likely to integrate long-range, meaningful negative outcomes ($t-2$), which potentially indicated task switches.

To summarize our model-agnostic results, 13-15 year-olds outperformed younger participants 8-13, older adolescents, and adults on a stochastic and volatile task, which was designed to mimic environmental challenges specific to adolescence. We next used computational modeling to investigate what cognitive processes gave rise to 13-15 year old adolescents' superior performance.

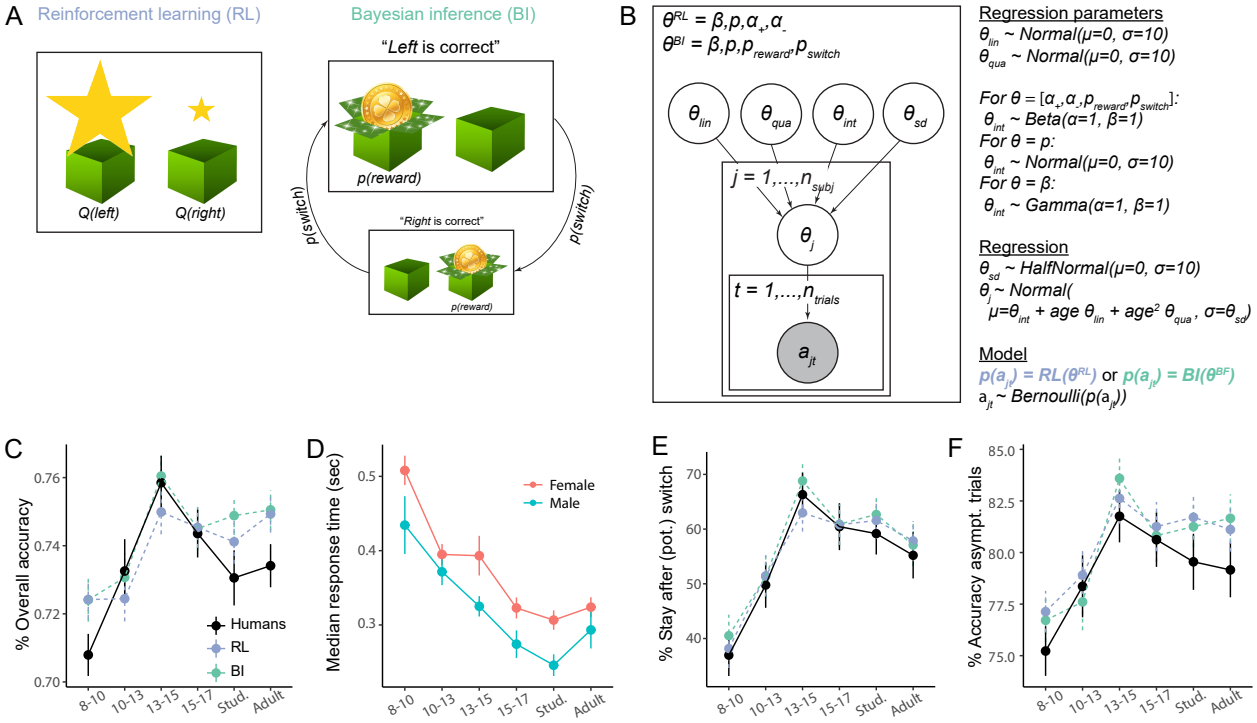


Figure 2.3: (A) Conceptual depiction of the RL and BI models. In RL (left), actions are selected based on learned values, illustrated by the size of stars ($Q(left)$, $Q(right)$). In BI (right), actions are selected based on a mental model of latent state of the task, which contains task stochasticity ($p(reward)$) and volatility ($p(switch)$). The size of each state illustrates the probability of being in that state. (B) Hierarchical Bayesian model fitting. Box on the left: Models had free parameters θ^{RL} or θ^{BI}. Individual parameters θ_j were based on group-level parameters θ_{sd}, θ_{int}, θ_{lin}, and θ_{qua} (see below). For each model (RL and BI), all parameters were simultaneously fit to the observed (shaded) sequence of actions a_{jt} of all participants j, using MCMC sampling. Right: Priors for group-level parameters were uninformative; the form of the prior differed based on parameter ranges. For each participant j, each parameter θ was sampled according to a linear regression model, based on group-wide standard deviation θ_{sd}, intercept θ_{int}, linear change with age θ_{lin}, and quadratic change with age θ_{qua}. Each model (RL or BI) provided a choice likelihood $p(a_{jt})$ for each participant j on each trial t, based on individual parameters θ_j. Action selection followed a Bernoulli distribution. See Methods for details. (C)-(F) Human behavior on the measures of Fig. 2.2, binned in age groups. (C), (E), and (F) also show simulated model behavior, verifying that models closely reproduced human behavior and age differences.

Cognitive Modeling

Models

We fitted two classes of cognitive models to the task, RL and BI. The winning RL model included four parameters: persistence p, inverse decision temperature β, and positive and negative learning rates α_+ and α_- (Methods). Notably, this model updated the values of both the chosen and unchosen action after each outcome, allowing for counterfactual learning (Boorman et al., 2011; Palminteri et al., 2016). It also allowed learning rates to differ between positive (α_+) and negative outcomes (α_-), an increasingly common idea in cognitive neuroscience (e.g., Cazé and van der Meer, 2013; Frank et al., 2004; van den Bos et al., 2012; for review, see Nussenbaum and Hartley, 2019) and AI (Dabney et al., 2020). Parameters p and β controlled the translation from RL values into choices: persistence p increased the probability of repeating choices when $p > 0$, and of alternating choices when $p < 0$; β induced decision noise (increased probability of exploratory choices) when small, and allowed for reward-maximizing choices when large. The winning BI model also had four parameters: choice-parameters p and β as in the RL model, as well as task volatility p_{switch} and reward stochasticity p_{reward}, which characterized participants' internal model of the task (Fig. 2.3A; Methods). p_{switch} ranged from stable ($p_{switch} = 0$) to volatile ($p_{switch} > 0$), and p_{reward} ranged from deterministic ($p_{reward} = 1$) to stochastic ($p_{reward} < 1$). The actual task was based on $p_{switch} = 0.05$ and $p_{reward} = 0.75$.

We fitted each model to participant data using hierarchical Bayesian fitting (Fig. 2.3B; Methods). This approach recovered individual parameters reliably (suppl. Fig. 2.14), and allowed us to estimate the effects of age on model parameters in a statistically unbiased way (Katahira, 2016; M. D. Lee, 2011; van den Bos et al., 2017). We compared different parameterizations of each model using the WAIC (Watanabe, 2013) to identify a winning RL and a winning BI model (Table 2.2). The winning RL model had the lowest score overall, revealing best quantitative fit. Nevertheless, both RL and BI models validated equally well, closely reproducing human behavior and age-related differences: Both models showed the performance peak in 13-15 year olds (Fig. 2.3C), the largest proportion of staying after (potential) switch trials (Fig. 2.3E), best asymptotic performance on non-switch trials (Fig. 2.3F), and the most efficient use of previous outcomes to adjust future actions (suppl. Fig. 2.13 D-F). Other tested models (Table 2.2) did not capture all qualitative patters (suppl. Fig. 2.16. 2.17). To conclude, despite major differences in their theoretical framework, both RL and BI captured human behavior and age differences. This finding has interesting implications, which we discuss in detail in the Discussion.

Table 2.2: WAIC model fits and standard errors for all models, based on hierarchical Bayesian fitting. Bold numbers highlight the winning model of each class. For the parameter-free BI model, the Akaike Information Criterion (AIC) was calculated precisely. WAIC differences are relative to next-best model of the same class, and include estimated standard errors of the difference as an indicator of meaningful difference. In the RL model, "α" refers to the classic RL formulation in which $\alpha_+ = \alpha_-$. "α_c" refers to the counterfactual learning rate that guides updates of unchosen actions, with $\alpha_{+c} = \alpha_{-c}$ (Methods).

	Free parameters (count)		(W)AIC	WAIC Difference
BI	–	(0)	31,959	2,668 $\pm$0
	β	(1)	$29,291 \pm 206$	868 ± 78
	β, p	(2)	$28,423 \pm 201$	$4,769 \pm 132$
	β, p, p_{reward}	(3)	$23,654 \pm 203$	51 ± 10
	β, p, p_{reward}, p_{switch}	(4)	**23,603 $\pm$ 200**	0
RL	α, β	(2)	$26,678 \pm 200$	438 ± 44
	α, β, α_c	(3)	$26,240 \pm 201$	$1,429 \pm 78$
	α, β, α_c, p	(4)	$24,811 \pm 190$	42 ± 13
	α_+, β, α_{+c}, p, α_-	(5)	$24,769 \pm 213$	$1,260 \pm 73$
	α_+, β, α_{+c}, p, α_-, α_{-c}	(6)	$23,509 \pm 211$	17 ± 10
	$\alpha_+ = \alpha_{+c}$, $\alpha_- = \alpha_{-c}$, β, p	(4)	**23,492 $\pm$ 201**	0

Age Differences in Model Parameters

All model parameters showed age effects (Fig. 2.4; suppl. Tables 2.11 and 2.12). We tested these effects statistically by modeling age explicitly in a hierarchical Bayesian model (Fig. 2.3B, suppl. Table 2.11), and also by assessing age-group differences in the posteriors of an age-less hierarchical Bayesian model (suppl. Table 2.12; Methods).

Choice-based parameters p and β were almost perfectly correlated between the winning RL and BI models, even though they were fitted independently (Spearman $\rho = 0.94$; Fig. 2.5B). This suggests that the parameters captured robust, update-independent aspects of decision making. p and β both increased monotonically with age and plateaued in older participants (Fig. 2.4A, B, E, F). This was reflected in linear and negative quadratic effects of age (suppl. Table 2.11): Persistence p increased near-linearly from age 8 until 17, and then plateaued around age 18-30 (Fig. 2.4A, E). This shows that the willingness to repeat previous actions, independent of outcomes, increased from childhood to adulthood, with steady growth during teen years.

Other parameters showed non-monotonic age trajectories. α_-, p_{reward}, and p_{switch} declined drastically from age 8 to 13-15, but then reversed their trajectory and increased again, reaching a plateau that lasted from 15-30 years (Fig. 2.4C, G-H). For α_- and p_{reward}, these changes were captured in significant pairwise differences between children (8-10) and 13-15 year-olds, as well as between 13-15 year-olds and adults (25-30; for statistics, see suppl. Table 2.12; also see Methods). For p_{switch}, age differences were captured in a significant quadratic effect of age (suppl. Table

2.11). Parameters p_{reward} and p_{switch}, reflecting participants' mental model of the task, were closest to their true values ($p_{reward} = 0.75$; $p_{switch} = 0.05$) in 13-15 year-olds. 8-10 year-old children and adults (18-30) overestimated task volatility (p_{switch}) and underestimated the reward stochasticity (p_{reward}) to a larger degree. Parameter α_- also was lowest in 13-15 year-olds, allowing them to avoid premature switching based on single negative outcomes while allowing for slow integration of outcomes and adaptive switching after multiple negative outcomes. Parameter α_+ showed a unique age trajectory with relatively stable values during childhood and adolescence (8-17), and a sudden increase in adults (18-30; Fig. 2.4D), captured in a linear effect of age (suppl. Table 2.11).

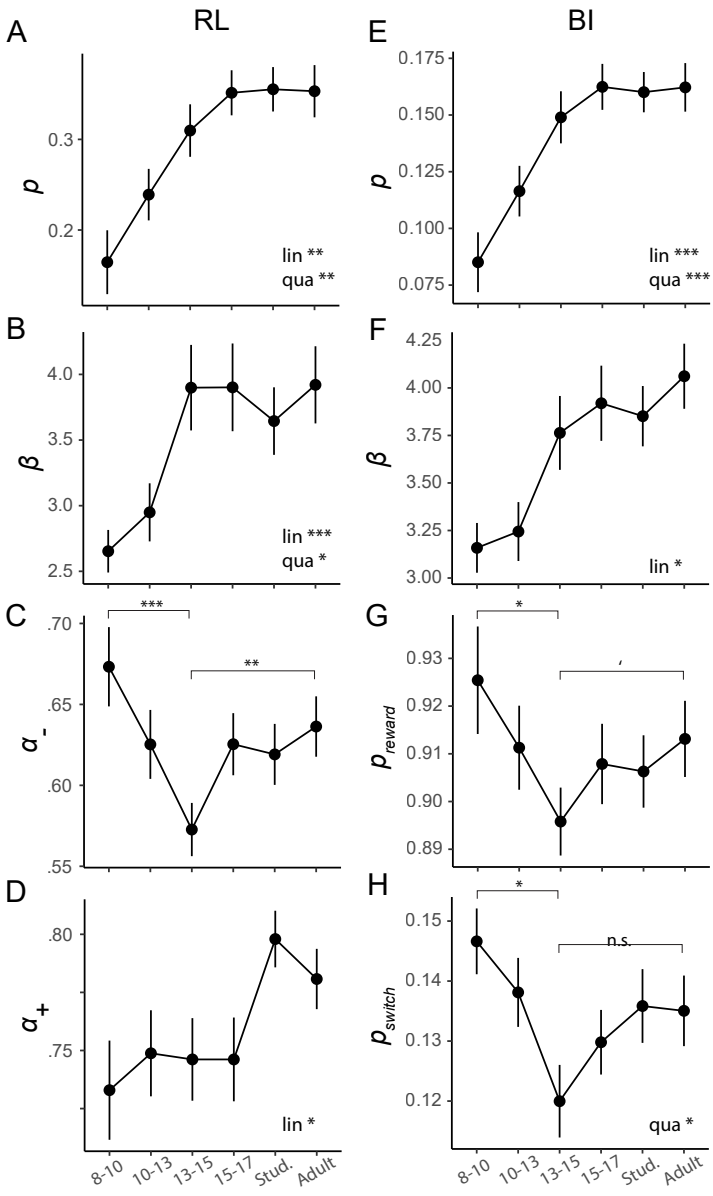


Figure 2.4: Age-related differences in model parameters for the winning RL (left column) and BI model (right). Stars indicate significant linear ("lin") and quadratic ("qua") effects of age on model parameters, obtained from the hierarchical Bayesian model, and differences between age groups, obtained from the age-less hierarchical Bayesian model (Methods; suppl. Tables 2.11 and 2.12). Means and standard errors were calculated based on individual fits from the age-less model, to avoid double-dipping (Methods). (A)-(D) RL model parameters. (E)-(H) BI model parameters. Stars indicate p-values like before.

Differences between RL and BI

Having obtained two independent sets of parameters for each participant from two computational models based on different cognitive mechanisms, we aimed to clarify how both models were related. We first asked whether each model captured different aspects of behavior, or whether both models captured the same behaviors and merely differed in form. To test this, we simulated artificial behavior from each model and assessed how well these data were captured by the opposite model. Each model was fitted worse by the opposite model than by itself (Fig. 2.5A), which reveals that each model captured unique aspects of behavior. (This difference was smaller when fitting the RL model, suggesting that it was more versatile and captured more aspects of the BI model than the other way around.)

We next asked how closely individual parameters were related between models, assessing pairwise Spearman correlations. As mentioned before, choice parameters p and β were almost perfectly correlated between models (p: $\rho = 0.97$; β: $\rho = 0.94$; Fig. 2.5B). In addition, parameter p_{reward} (BI) was strongly correlated with α_- (RL), suggesting that negative learning rate (α_-) and beliefs about task stochasticity (p_{reward}) played similar roles in the integration of negative outcomes. Parameter p_{switch} (BI) was strongly negatively correlated with β (RL), suggesting that decision noise (β) in the RL model captured aspects that were explained by beliefs about task volatility (p_{switch}) in the BI model. The only parameter that showed no large correlations with other parameters was α_+ (RL), suggesting a unique role.

Lastly, we investigated how much information each model provided about the other, using linear regression to predict each parameter from the parameters and one-way parameter interactions of the other model. Seven out of eight parameters were predicted almost perfectly (Fig. 2.5C), showing that the parameters of one model captured almost all variance in the opposite model. In other words, fitting the RL model on participants' data allowed us to nearly perfectly predict participants' BI parameters, without fitting the BI model. Note that α_+ (RL) was again an exception in that its variance was not fully captured by BI parameters. α_+ might thus account for the better fit of the RL model to human (Table 2.2) and simulated data (Fig. 2.5A), compared to the BI model.

In summary, RL and BI models captured similar aspects of behavior, as shown by large inter-model parameter correlations and amounts of explained variance; nevertheless, both models were not redundant, as evident in the fact that each was unable to perfectly fit the other.

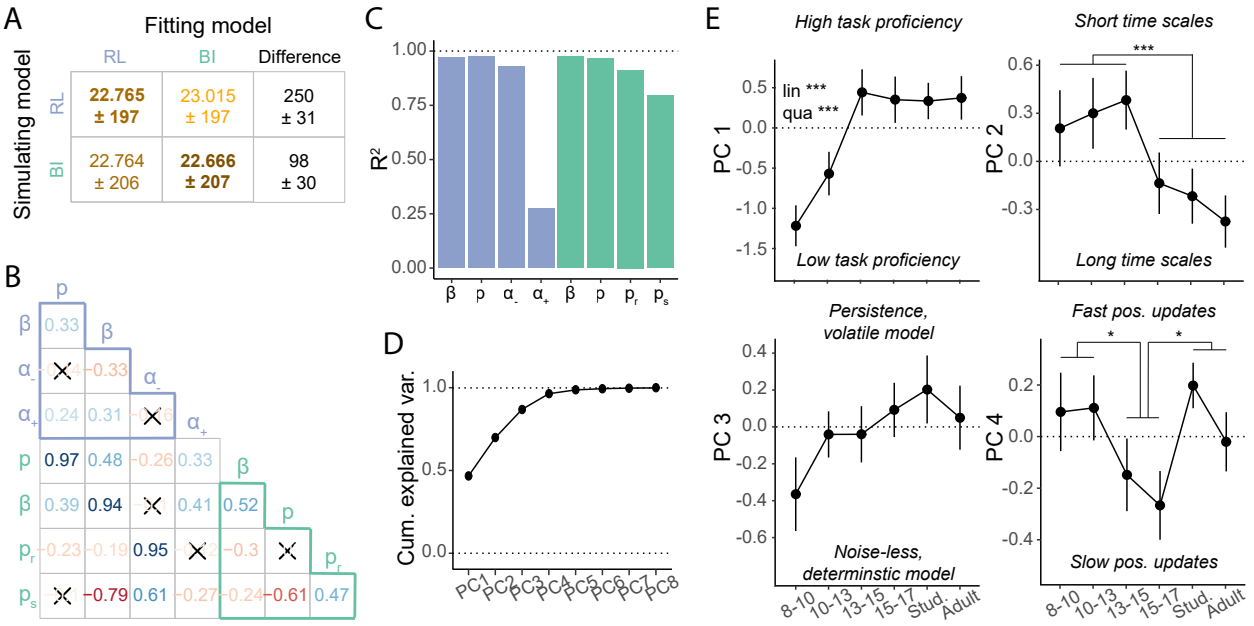


Figure 2.5: Relating RL and BI models. (A) Model recovery. WAIC scores were worse (larger; lighter colors) when recovering behavior that was simulated from one model (row) using the other model (column), than when using the same model (diagonal), revealing that the models were discriminable. The difference in fit was smaller for BI simulations (bottom row), suggesting that the RL model captured BI behavior better than the other way around (top row). (B) Spearman pairwise correlations between model parameters. Red (blue) hue indicates negative (positive) correlation, saturation indicates correlation strength. Non-significant correlations are crossed out (Bonferroni-corrected at $p = 0.00089$). Light-blue (teal) letters refer to RL (BI) model parameters. Light-blue / teal-colored triangles show correlations within each model, remaining cells show correlations between models. (C) Variance of each parameter explained by parameters and interactions of the other model ("R^2"), estimated through linear regression. All four BI parameters (green) were predicted almost perfectly by the RL parameters, and all RL parameters except for α_+ (RL) were predicted by the BI parameters. (D)-(E) Results of PCA on model parameters. (D) Cumulative variance explained by all principal components PC1-8. The first four components captured 96.5% of total parameter variance. (E) Age-related differences in PC1-4: PC1 reflected overall task proficiency and showed rapid development between ages 8-13, which were captured by linear ("lin") and quadratic ("qua") effects in a regression model. PC2 captured a step-like transition from shorter to longer updating time scales at age 15, as revealed by PC-based model simulations (Supplements). PC3 showed no significant age effects. PC4 captured the variance in α_+ and differed between adolescents 15-17 and both 8-13 year olds and adults. PC2 and PC4 were analyzed using t-tests. * $p < .05$; ** $p < .01$, *** $p < .001$.

Combining RL and BI using PCA

We next asked whether both models in conjunction provided a even better explanation of unique adolescent decision making than either model on its own, using PCA to unveil the lower-dimensional structure embedded in the shared 8-dimensional parameter space. Indeed, only four dimensions were necessary to capture 96.5% of the parameter variance (Fig. 2.5D), suggesting that individual differences between participants could be explained by variation on just these four, rather than eight different model parameters.

To identify the role of each PC, we assessed parameter loadings and gained additional insight by simulating behavior with parameter sets defined by each PC. This approach is similar to simulating behavior based on different parameter values in order to investigate the effects of parameters (e.g., simulating big versus small values of β produces high versus low task performance). We similarly simulated behavior based on small versus large values of each PC, and compared the resulting behaviors to identify the exact function of each PC (Methods). Principal component 1 (PC1), the dimension capturing the largest proportion of variance, reflected general task proficiency (suppl. Fig. 2.18A; suppl. Text). Low proficiency was caused by larger-than-average values of α_- (RL), which led to premature switching, and p_{reward} and p_{switch} (BI), reflecting overly deterministic but volatile mental models of the task (suppl. Fig. 2.18A, left). High proficiency, on the other hand, was caused by larger-than-average values of α_+ (RL), p (RL and BI), and β (RL and BI), which facilitated quick integration of positive outcomes, choice persistence, and low decision noise, respectively (suppl. Fig. 2.18A, right). PC1 was lowest in the youngest participants (8-10), but increased rapidly until age 13, at which age it reached a stable plateau that lasted throughout adulthood (Fig. 2.5E, top-left). Age differences in PC1 were characterized by linear and quadratic effects of age (Methods). Taken together, PC1 explained one side of the inverse U-shape in overall task performance (Fig. 2.2; suppl. Fig. 2.13; Fig. 2.3C-F): 13-15 year olds outperformed 8-13 year olds because younger participants had not yet mastered task proficiency.

PC1 did not explain, however, how participants aged 13-15 outperformed older participants. PC2, the dimension that captured the second-most variance after PC1, played this role by capturing the tension between updates with short versus long time scales (suppl. Fig. 2.18E). Short time scales were driven by larger-than-average values of α_+ and α_- (RL), i.e., rapid updates based on recent outcomes, which led to pronounced win-stay lose-shift behavior (suppl. Fig. 2.18B, left). In the BI model, short time scales were driven by increasingly volatile (p_{switch}) and especially deterministic (p_{reward}) mental models. Short time scales were complemented by persistence, driven by larger-than-average values of p (RL and BI). Long-term updates were the result of lower-than-average values on all parameters (suppl. Fig. 2.18B, left), allowing for gradual and slow, but precise value updates, with choice unbiased by persistence. PC2 showed a step function of age. Whereas younger participants persisted and acted on short times scales, after age 15, participants showed unbiased long-term updates (suppl. Fig. 2.5E, top-right). Differences between 13-15 year olds and older participants were therefore captured by PC2, suggesting that better performance was the result of shorter updating time scales. This is conditioned on high task proficiency (PC1), which includes slower time scales for negative outcomes, but faster time scales for positive outcomes.

PC3 captured the tension between noise-less choice (larger-than-average β) combined with a

deterministic mental model (larger-than-average p_{reward}) on one side, and persistent choice (larger-than-average p) combined with an overly volatile mental model (larger-than-average p_{switch}) on the other (suppl. Fig. 2.18C). PC3 showed no significant age-related differences (Fig. 2.5E, bottom-left). PC4 captured the unique variance of α_+ (RL), with a tension between slow and fast updates from positive outcomes (suppl. Fig. 2.18D). PC4 was lower in 13-17 year olds than both 8-13 year olds and adults (18-30; Fig. 2.5E, bottom-right), revealing that after accounting for variance in PC1-3, the remaining variance was explained by adolescents' relatively longer updating timescales for positive outcomes. In other words, positive outcomes had weaker immediate, but stronger long-lasting effects in 13-17 year olds, setting them apart from both younger and older participants.

2.3 Discussion

Across species, the adolescent transition from childhood to adulthood brings great challenges for learning and exploration. From an evolutionary perspective, these challenges may have caused the adolescent brain to evolve behavioral tendencies that promote adaptive learning in rapidly changing, uncertain environments. To test this idea, we examined choice behavior in a stochastic and volatile task adapted from rodent studies (Tai et al., 2012).

Indeed, we found that 13-15 year olds performed better than both younger (8-13) and older participants, including adults (15-30): 13-15 year olds achieved the highest overall accuracy, were most willing to wait out negative feedback (potential switch), and made the best choices during stable periods (asymptotic performance). Overall, 13-15 year olds used negative feedback most optimally to guide future choices, being least affected by proximal, but most sensitive to distal outcomes. This shows an ability to ignore ambivalent information while responding appropriately to meaningful patterns. Indeed, such inverted-U or U-shaped developmental trajectories are not unique in the development of human cognition. Evidence is accumulating that adolescents outperform adults in various domains, including probabilistic learning (Davidow et al., 2016) and reversal (van der Schaaf et al., 2011), creativity (Kleibeuker et al., 2013), and social learning (Gopnik et al., 2017). Prowess in flexibility has also been reported in studies of developing rodents (Guskjolen et al., 2017; Johnson and Wilbrecht, 2011; Simon et al., 2013).

One aspect of adolescent behavior, 13-15 year old's increased willingness to wait out negative feedback, deserves specific attention: It means that in the context of this task, 13-15 year olds were less impulsive than other age groups. This finding seems inconsistent with past research that often described the mid-adolescent period in terms of increased risk taking and higher risk of negative life outcomes. To explain why, studies of adolescent development have separated impulsivity and sensation seeking. Studies using self reports and experimental tasks showed that impulse control grows through the teen years, while sensation seeking peaks in mid to late adolescence (Albert et al., 2013; Harden and Tucker-Drob, 2011; Romer and Hennessy, 2007; Steinberg et al., 2009). The combination of not-yet-mature impulse control and high sensation seeking in mid-adolescence has been used to explain why this period is associated with higher risks (Harden and Tucker-Drob, 2011; Steinberg, 2013). Our findings do not fit into this narrative of adolescents as risk takers. There are several potential reasons for this discrepancy: (1) Our task may not tap

into sensation seeking, a process separate from impulsivity. (2) We use a behavioral task and not self-report methods, which each have different benefits and limitations. (3) Our task may elicit different learning and decision-making strategies than other tasks because it creates a stochastic and volatile environment. Individuals likely do not apply the same learning and decision rules in all contexts. This suggests that the interplay between brain development and the statistics of specific environments may be more important than previously realized (Nussenbaum and Hartley, 2019). In accordance with our findings, van den Bos and colleagues have found that adolescents displayed distinctive tolerance to ambiguity and to uncertainty during risky decision making (van den Bos and Hertwig, 2017).

To understand which cognitive and neural processes supported 13-15 year-olds' superior performance within our specific task, we employed two types of cognitive models, RL and BI. To fit human behavior, the RL model required the ability to learn from counterfactual outcomes (updating values of non-chosen actions), and to apply different learning rates to positive versus negative outcomes (learning parameters α_+ and α_-). It also required persistence, i.e., a tendency to repeat previous actions independent of outcomes, in addition to decision noise, i.e., the ability to explore non-maximizing actions (choice parameters p and β). RL models have been used extensively to shed light on neural mechanisms, and a specialized network of brain regions—including basal ganglia, cortical, and limbic regions—is thought to implement key RL computations (for reviews, see Frank and Claus, 2006; D. Lee et al., 2012; Niv, 2009; O'Doherty et al., 2015). Fitting RL models to developmental samples is thought to inform our understanding of brain development (e.g., Christakou et al., 2013; Davidow et al., 2016; Javadi et al., 2014; Master et al., 2020; for reviews, see Nussenbaum and Hartley, 2019; van den Bos et al., 2017). Using RL models, we found that choice parameters (β, p) in our study grew monotonically throughout childhood and adolescence, and only matured in late adolescence / early adulthood. This is consistent with previous developmental modeling studies (Nussenbaum and Hartley, 2019), and with a role for late-developing brain circuits in choice behavior (Giedd et al., 1999; Gogtay et al., 2004; Nussenbaum and Hartley, 2019; Sowell et al., 2003; Toga et al., 2006).

While the developmental trajectories of choice parameters have been highly consistent in the developmental modeling literature, the development of learning-rate parameters has been ambivalent and even contradictory (Nussenbaum and Hartley, 2019). One problem might be that many learning studies that are fit with RL models likely involve a variety of different learning processes, which do not only include striatal incremental learning (Yagishita et al., 2014), based on direct and/or indirect pathways (Hauser et al., 2015), but also hippocampal-based episodic memory (Bornstein and Norman, 2017; Wimmer et al., 2014), and frontal-cortical cognitive control (Badre et al., 2010; Collins and Frank, 2012; Daw et al., 2011). Differences in task contexts and task statistics likely elicit different learning strategies, and recruit different neural processes (Nussenbaum and Hartley, 2019). This potentially explains the diversity of previous findings with regard to learning parameters and limits our ability to make inferences about brain development from behavioral modeling studies. It is a future challenge to disentangle the development of multiple systems and context-based responses, for example by studying the same individuals across multiple tasks and computational models.

Another potential reason for the observed discrepancies in learning rates is that between stud-

ies, models often differ in the number and type of learning-rate parameters (e.g., positive, negative, factual, counter-factual). For example, the standard learning rate parameter α controls updates to the values of chosen actions after both positive and negative outcomes, whereas a more specialized parameter $\alpha_{c\_}$ controls updates to values of unchosen actions, but only after negative outcomes. Given the likely differences in neural substrate that underlie these different mechanisms, they likely differ in their developmental trajectory. Another reason for the discrepancies in the literature might therefore be study-by-study differences in parameterizations of computational models. In our study, learning rates from negative feedback (α_-) showed a pronounced U pattern with minimum in 13-15 year olds, whereas learning rates from positive feedback (α_+) were stable throughout childhood and adolescence, then suddenly increased in adults. These patterns likely reflect the combination of different cognitive and neural processes, which matured at different times. Indeed, the developmental trajectory of α_- was almost identical to the stochasticity parameter p_{reward} of the BI model ($\rho = 0.95$), suggesting that α_- played a role in switching behavior after negative feedback, rather than learning. As a whole, the RL model might have approximated inferential reasoning rather than performing pure incremental learning, an issue we discuss in more detail below.

To place our RL results in a broader context, we also applied Bayesian Inference (BI) models to our task data. Using BI models, we found that choice parameters β and p showed almost identical trajectories as in the RL model, strong independent support for our hypothesis that these factors increased through the second decade of life. BI model fits also provided novel results. We found that BI mental model parameters, task stochasticity p_{reward} and volatility p_{switch}, were most accurate in 13-15 year-olds. By definition, this means that 13-15 year-olds possessed the best mental model with respect to actual task statistics, whereas both younger (8-13) and older participants (15-30) demonstrated less accurate models. We had also hypothesized that children and adults would expect less volatility and stochasticity than adolescents, but only stochasticity showed this pattern. Volatility, on the other hand, appeared to be perceived as larger in younger (8-13) and older (15-30) participants compared to 13-15 year-olds. Interestingly, the BI model revealed that participants of all ages deviated markedly from Bayes-optimal behavior, employing mental models that were both too volatile and too deterministic (p_{switch}: 8-13 year-olds behaved as if they expected switches every 6.5 trials, 13-15 year-olds every 10 trials, whereas the task switched every 20 trials on average; p_{reward}: 8-13 year-olds behaved as if they expected rewards for 92% of correct responses, 13-15 year-olds for 89%, the task rewarded 75%). In summary, 13-15 year-olds exhibited mental models that were most in line with task parameters, expecting the most stochasticity and least volatility of all age groups. This is in accordance with the differentiation between "adaptation" and "settings" in Nussenbaum and Hartley, 2019, and suggests that 13-15 year-olds showed the largest ability to adjust and adapt to specific task statistics, rather than reflecting a particular, developmentally-fixed setting of any specific parameter. A similar argument about parameter optimality rather than a developmentally-specific parameter setting was made in Davidow et al., 2016.

Fitting two separate model classes led to several benefits in the understanding of the underlying cognitive processes: (1) Both models provided converging (choice parameters) and additive evidence (RL: learning parameters; BI: mental-model parameters). Converging results showed

surprisingly strong, direct replication ($\beta_{RL} \leftrightarrow \beta_{BI}$, $p_{RL} \leftrightarrow p_{BI}$), and parallelism between model parameters helped clarify the role of ambiguous parameters ($p_{reward} \rightarrow \alpha_{-}$). Independent components led to additive insights (e.g., unique parameter α_{+}), (2) Each model's conceptual framework and interpretation of the cognitive process became more distinctive in direct comparison with each other. The contrast helped sharpen claims about incremental learning (RL) versus mental-model based inferential reasoning (BI). Whereas the RL model achieved better numerical fit, the BI model provided advantages in terms of interpretability: Our concepts of interest (stochasticity, volatility) were explicitly modeled within the BI framework, potentially allowing insight into how they were processed by each participant. On the technical side, all BI parameters occupied meaningful and interpretable ranges. In the RL model, on the other hand, learning rates showed values substantially larger than 0.5. This made their interpretation difficult, as it is unlikely that high learning rates reflect the type of RL processes implemented in the brain's RL network, and suggests the model approximated some other dynamic adaptation process. This highlights the fact that numerical model fit (RL) and interpretability (BI) can sometimes be at odds. Future research is necessary to explore this topic in more depth.

A final advantage of fitting both models was the possibility to investigate patterns that go beyond model-specific parameters, using PCA on the shared parameter space. This analysis exposed a different set of factors, which differentiated 13-15 year-olds from younger participants (PC1), from older participants (PC2), or from both (PC4). PC1 reflected overall task proficiency and showed steep improvement until age 13, plateauing thereafter. This suggests that 13-15 year-olds outperformed younger participants because the younger group was too exploratory for the task, was less persistent, and possessed less accurate mental models, leading to weighing negative outcomes too much relative to positive ones. PC2 reflected participants' updating time scales and showed a step function with transition around 15 years of age. This PC suggests that 13-15 year-olds outperformed older participants because the 15-30 year-olds operated on longer time scales, i.e., were more sensitive to distant outcomes and perceived the task as less stochastic and volatile. PC4 reflected the variance in α_{+} that was not captured by previous PCs, and showed an inverse-U trajectory with minimum in 13-17 year-olds. PC4 therefore showed that 13-17 year-olds used the longest time scales when processing positive outcomes, compared to both younger (8-13) and older participants (18-30). Taken together, adolescents aged 13-15 may be at a "sweet spot" for stochastic and volatile environments because they combine mature levels of task proficiency (PC1) with youthful short updating times scales for all outcomes (PC2), but uniquely long updating time scales for positive outcomes (PC4). This combination would not be optimal in all environments, but in a stochastic and volatile environment, it led to more rewards earned. In this sense, performance in this task supports the idea that the adolescent human brain may pass through stages that have evolved to enhance success in uncertain and volatile environments.

This study shows that age played a crucial role for reward-based decision making and learning in a volatile, stochastic environment. Nevertheless, the question remains which mechanisms underlie these age effects. There is growing evidence that gonadal hormones affect inhibitory neurotransmission, spine pruning, and other variables in the prefrontal cortex of rodents (Delevich et al., 2019; Delevich et al., 2018; Drzewiecki et al., 2016; Juraska and Willing, 2017; Piekarski, Boivin, et al., 2017; Piekarski, Johnson, et al., 2017), suggesting that puberty-related changes in

brain chemistry might be the mechanism behind the observed differences. To answer this question, we investigated the trajectories of behavioral performance and model parameters over pubertal development, observing qualitatively similar patterns compared to age (suppl. Fig. 2.7, 2.8, 2.9; suppl. Tables 2.4, 2.5; for a discussion of differences, see suppl. Text). Nevertheless, pubertal measures were so highly correlated with age (suppl. Fig. 2.6) that it was difficult to interpret these findings. We therefore investigated the effects of puberty controlling for age, testing puberty effects separately within each age bin. Puberty effects in this analysis did not reach statistical significance (suppl. Fig. 2.10, 2.11, 2.12). Thus, we were unable to identify a biological mechanism underlying age besides accumulated experience over time. A related question pertains to the underlying cognitive mechanism. Future research is required to investigate whether performance in this task was related to measures of fluid intelligence, directed exploration, and impulsivity, for example.

In conclusion, we used a simple task based on volatility and stochasticity to show that adolescents outperformed adults in a task that represented the kind of learning challenge that may have ecological validity to the transitions and challenges of adolescence. In our community sample, behavior was most optimal at age 13-15. We used two models to examine the underlying cognitive processes. The results suggest that adolescent brains achieved better performance for several reasons: (1) 13-15 year-olds lay on the right spot in a monotonic trajectory between childhood and adulthood (p and β). (2) 13-15 year-olds were outliers in terms of their ability to accurately assess the volatility and stochasticity of their environment, and in terms of their integration of negative outcomes (p_{reward}, p_{switch}, and α_{-}). (3) 13-15 year-olds combined adult-like (PC1), child-like (PC2), and developmentally unique (PC4) strategies. These data suggest that multiple neural systems underlie developmental changes in brain function, at staggered time scales. Pubertal development and steroid hormones may impact a subset of these processes, yet causality is difficult to determine without manipulation or longitudinal designs (Kraemer et al., 2000).

For purposes of translation from the lab to the 'real' world, our study indicates that how youth learn and decide changes in a nonlinear fashion as they grow. This underscores the importance of youth-serving programs that are developmentally informed and avoid a one-size-fits-all approach. Finally, these data support a positive view of adolescence and the idea that the adolescent brain exhibits remarkable learning capacities that should be celebrated.

2.4 Methods

Participants

All procedures were approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley. We tested 312 participants: 191 children and adolescents (ages 8-17) and 55 adults (ages 25-30) were recruited from the community and completed a battery of computerized tasks, questionnaires, and saliva samples; 66 university undergraduate students (aged 18-50) completed the four tasks as well, but not the questionnaires or saliva sample. Community participants were prescreened for the absence of present or past psychological and neurological disorders; the undergraduate sample indicated the absence of these. Compensation for commu-

nity participants consisted in 25$ for the 1-2 hour in-lab portion of the experiment and 25$ for completing optional take-home saliva samples; undergraduate students received course credit for participation in the 1-hour study.

Exclusion Criteria Out of the 191 participants under 18, 184 completed the stochastic switching task; reasons for not completing the task included getting tired, running out of time, and technical issues. Five participants (mean age 10.0 years) were excluded because their mean accuracy was below 58% (chance: 50%), an elbow point in accuracy, which suggests that they did not pay attention to the task. This led to a sample of 179 participants under 18 (male: 96, female: 83). Two participants from the undergraduate sample were excluded because they were older than 30, leading to a sample aged 18-28; 7 were excluded because they failed to indicate their age. This led to a final sample of 57 undergraduate participants (male: 19, female: 38). All 55 adult community participants (male: 26, female: 29) completed the task and were included in the analyses, leading to a sample size of 179 participants below 18, and 291 in total (suppl. Fig. 2.6). For some analyses, we split participants into quantiles based on age. Quantiles were calculated separately within each sex.

Testing Procedure

After entering the testing room, participants under 18 years and their guardians provided informed assent and permission; participants over 18 provided informed consent. Guardians and participants over 18 filled out a demographic form. Participants were led into a quiet testing room in view of their guardians, where they used a video game controller to complete four computerized tasks. At the conclusion of the tasks, participants between 11 and 18 completed the PDS questionnaire themselves and were measured in height and weight. Participants were then compensated with $25 Amazon gift cards.

Experimental Design

The task described in this work was the last of the four tasks, a stochastic switching task. The other tasks will be or have been reported elsewhere (Master et al., 2020; Xia et al., 2020). The goal of the stochastic switching task was to collect golden coins, which were hidden in one of two green boxes. On each trial, participants decided which box to open, and task contingencies switched unpredictably throughout the task (Fig. 2.1B). Before the main task, participants completed a 3-step tutorial: A first prompt explained that one of the two boxes contained a coin (was “magical”), whereas the other one did not. Ten practice trials followed on which one box revealed a coin when selected, whereas the other was empty (deterministic tutorial). The second prompt stated that the magical box would sometimes switch sides. Participants then received eight trials on which the second box contained the coin (but not the first), followed by eight more trials on which the first box contained the coin (but not the second; switching tutorial). The third and last prompt explained that even the magical box did not always contain a coin. This prompt directly led into the main task (stochastic switching), with 120 trials.

In the main task, the correct box was rewarded in 75% of trials; the incorrect box was never rewarded. After participants reached a performance criterion (see below), it became possible for contingencies to switch (without notice), such that the previously incorrect box was now the correct one. The performance criterion was to collect 7-15 rewards, whereby the specific number was pre-randomized for each block. Any number of non-rewarded trials was allowed in-between rewarded trials. Due to this design, switches only occurred after rewarded trials. For consistency with the rodent version of the task (Tai et al., 2012), the first correct choice after a switch was always rewarded (not just in %75).

Behavioral Analyses

We assessed the effects of age on behavioral outcomes (Fig. 2.2), using (logistic) mixed-effects regression models with the package lme4 (Bates et al., 2015) in R (RCoreTeam, 2016). All models included the following set of regressors to predict outcomes of interest (e.g., overall accuracy, response times): Age, to assess the linear effect of age on the outcome; squared age, to assess the quadratic effect of age; and sex; furthermore all models specified random effects of participants, allowing participants' intercepts and slopes to vary independently. When models included additional predictors, this is noted in the main text.

We assessed the effects of previous outcomes on participants' choices (suppl. Fig. 2.13B, C, E, F) using a logistic mixed-effects regression model, which predicted actions (left, right) from previous outcomes (details below), while testing for effects of and interactions with sex, z-scored age, and z-scored quadratic age, specifying participants as mixed effects. We included one predictor for positive and one for negative outcomes at each delay i with respect to the predicted action (e.g., $i = 1$ trial ago). Outcome predictors were coded -1 for left and +1 for right choices, and 0 otherwise. Including predictors of trials $1 \leq i \leq 8$ provided the best model fit (suppl. Table 2.8). To visualize the results of this grand regression model (including all participants), we ran a separate model for each participant with the same structure, and show individual fits in suppl. Fig. 2.13B, C, E, F.

Computational Models

Reinforcement Learning (RL) Models

In RL, decisions are made based on action values, which are continuously updated based on outcomes (Sutton and Barto, 2017). A simple RL model has two parameters, learning rate α and decision temperature β. On each trial t, the value $Q_t(a)$ of action a is updated based on the observed outcome $o_t \in [0,1]$ (reward, no reward), in the following way:

$$Q_{t+1}(a) = Q_t(a) + \alpha(o_t - Q_t(a))$$

I.e., previous action values are updated in proportion to the difference between the estimated value and the actual reward, scaled by the learning rate α. The difference itself, $o_t - Q_t(a)$, is called "reward prediction error".

Over time, action values approximate the true underlying reward probabilities. Decisions are based on these values by calculating action probabilities using a softmax transform:

$$p_t(a) = \frac{exp(\beta\ Q_t(a))}{exp(\beta\ Q_t(a)) + exp(\beta\ Q_t(a_{ns}))}$$

Here, a is the selected, and a_{ns} the non-selected action.

The best-fit 4-parameter RL model was based on this 2-parameter model, with additional parameters learning rate for negative outcomes α_-, persistence p, as well as counterfactual reasoning (see below). Adding α_- allowed for separate updates of rewarded ($o_t = 1$) and non-rewarded ($o_t = 0$) trials: $Q_t(a) = Q_t(a) + \alpha_+(o_t - Q_t(a))$ iff $o_t = 1$, and $Q_t(a) = Q_t(a) + \alpha_-(o_t - Q_t(a))$ iff $o_t = 0$, with independent α_- and α_+. Choice persistence or "stickiness" p changed the value of the previously-selected action a_t on the subsequent trial, biasing toward staying ($p > 0$) or switching ($p < 0$): $Q(a_t) = Q(a_t) + p$ iff $a_t = a_{t-1}$.

Counterfactual reasoning was implemented through updates to the values of non-selected actions, using counterfactual outcomes $1 - o_t$: $Q_{t+1}(a_{ns}) = Q_t(a_{ns}) + \alpha_+((1 - o_t) - Q_t(a_{ns}))$ iff $o = 1$, and $Q_{t+1}(a_{ns}) = Q_t(a_{ns}) + \alpha_-((1 - o_t) - Q_t(a_{ns}))$ iff $o = 0$. Initially, we used separate parameters α_{+c} and α_{-n} for counterfactual updates, which were independent from α_+ and α_- for factual updates. Nevertheless, collapsing $\alpha_+ = \alpha_{+c}$ and $\alpha_- = \alpha_{-n}$ improved model fit (Table 2.2). This shows that outcomes triggered equal-sized updates to chosen and unchosen actions. Explained differently, the final model based decisions on a single value estimate—the value difference between the two available actions—, rather than on an independent value estimates for each. Chosen and unchosen actions were updated to the same degree and in opposite directions on each trial.

Action values were initialized at 0.5 for all models, reflecting equal initial values for the two actions.

Bayesian Inference (BI) Models

The BI model assumes that participants know that the task has two latent states: "Left action is correct" ($a_{left} = cor$) and "Right action is correct" ($a_{right} = cor$), where cor stands for correct (inc: incorrect). Participants assume that on each trial, the latent state switches with probability p_{switch}, and that in each state, the probability of receiving a reward for the correct action is p_{reward} (Fig. 2.3A). On each trial, participants select an action in two phases, using the Bayesian Filter algorithm (Sarkka, 2013): (1) In the *estimation phase*, participants infer the hidden state of the previous trial $t - 1$, based on the outcome o_{t-1} they received for their action a_{t-1}, using Bayes rule:

$$p(a_{t-1} = cor \mid o_{t-1}) = \frac{p(o_{t-1}|a_{t-1} = cor)\ p(a_{t-1} = cor)}{p(o_{t-1}|a_{t-1} = cor)\ p(a_{t-1} = cor) + p(o_{t-1}|a_{t-1} = inc)\ p(a_{t-1} = inc)}$$

$p(a_{t-1} = cor)$ is the prior probability that a_{t-1} was correct (on the first trial, $p(a = cor) = 0.5$ for both actions), and $p(o_{t-1}|a_{t-1})$ is the likelihood of the observed outcome o_{t-1} given action a_{t-1}. According to the mental model, likelihoods are (dropping underscripts for clarity): $p(o = 1|a = cor) = p_{reward}$, $p(o = 0|a = cor) = 1 - p_{reward}$, $p(o = 1|a = inc) = \varepsilon$, and $p(o = 0|a = cor) = 1 - \varepsilon$,

where ε is the probability of receiving a reward for an incorrect action, which was 0 in reality, but we set $\varepsilon = 0.0001$ to avoid model degeneracy. (2) In the *prediction phase*, participants integrate the possibility of state switches by propagating the inferred knowledge about the hidden state at $t-1$ forward to trial t:

$$p(a_t = cor) = (1 - p_{switch})\ p(a_{t-1} = cor) + p_{switch}\ p(a_{t-1} = inc)$$

We first assessed a parameter-free version of the BI model, truthfully setting $p_{reward} = 0.75$, and $p_{switch} = 0.05$. Lacking free parameters, this model was unable to capture individual differences and led to poor qualitative (suppl. Fig. 2.17A) and quantitative model fit (Table 2.2). The best-fit BI model had four free parameters: p_{reward} and p_{switch}, as well as the choice parameters β and p, like the winning RL model. β and p were introduced by applying a softmax to $p(a_t = cor)$ to calculate $p(a_t)$, the probability of selecting action a on trial t: $p(a_t) = \frac{1}{(1+exp(\beta(0.5-p-p(a_t=cor))))}$. When both actions had the same probability $p(a)$ and persistence $p > 0$, then staying was more likely; when $p < 0$, then switching was more likely.

Model Fitting and Comparison

We fitted parameters using hierarchical Bayesian methods (M. D. Lee, 2011; Fig. 2.3B), and found that the obtained results clearly superseded those of classical maximum-likelihood fitting in terms of parameter recovery (suppl. Fig. 2.14). Hierarchical Bayesian model fitting estimates the parameters of an entire population *data* jointly, using Bayes formula:

$$p(\theta|data) \propto p(data|\theta)p(\theta)$$

Individual parameters are embedded in a hierarchical structure, which helps resolve uncertainty at the individual level. Because we were interested in age-related differences in model parameters, we used a hierarchical structure in which parameters $\theta_j^{RL} = [p, \beta, \alpha_-, \alpha_+]$ or $\theta_j^{BI} = [p, \beta, p_{switch}, p_{reward}]$ of participant j were embedded in linear regressions:

$$\theta_j \sim Normal(\mu = \theta_{int} + age\ \theta_{lin} + age^2\ \theta_{qua},\ \sigma = \theta_{sd})$$

Each parameter θ was characterized by group-level intercept θ_{int}, slope θ_{lin}, and quadratic change with age θ_{qua}. Individual parameters θ_j were drawn from a normal distribution with standard deviation θ_{sd} around this regression line (Fig. 2.3B).

Because posteriors $p(\theta|data)$ were analytically intractable, we approximated them using Markov-Chain Monte Carlo sampling (no-U-Turn sampler), using the PyMC3 package in python (Salvatier et al., 2016). We ran 2 chains per model with 6,000 samples per chain, discarding the first 1,000 as burn-in. All models converged with small MC errors, sufficient effective sample sizes, and $\hat{R}$ close to 1 (suppl. Table 2.10). Point estimates for individual parameters θ_j were calculated as the mean over all posterior samples. For model comparison, we used the Watanabe-Akaike information criterion (WAIC), which estimates the expected out-of-sample prediction error using a bias-corrected adjustment of within-sample error (Watanabe, 2013).

To statistically test the hypothesis that parameter θ differed between age groups, we fitted a separate hierarchical Bayesian model, which did not have access to participants' age, called the "age-free" model. Instead of lying on an age-based regression line, all individual parameters were drawn from the same group-wide Normal distribution with mean θ_{mean} and standard deviation θ_{sd}. To test for differences between groups $a1$ and $a2$ without the danger of double-dipping. we assessed $\theta_{a1} < \theta_{a2}$ in each posterior sample of this model, and then calculated $p(\theta_{a1} < \theta_{a1})$ across all samples. The age-less model was also used to visualize individual parameters in suppl. Figures 2.15, 2.17, and 2.16, and to calculate group means in Fig. 2.4. Using the age-less model avoided double-dipping on age effects, which would occur if we plotted parameters across age that were fitted in an age-dependent model.

Integrating RL and BI Models

Model Recovery between RL and BI (Fig. 2.5A) We simulated one dataset per participant from each model, using parameters fitted by the age-free model. We then fitted the simulated data with both models using age-free hierarchical Bayesian fitting. We finally calculated WAIC scores and standard errors using PyMC3 (Salvatier et al., 2016).

Correlations between Model Parameters (Fig. 2.5B) We used Spearman correlation, the non-parametric version of the Pearson product-moment correlation, because parameters followed different, not necessarily normal, distributions. Results were similar when using Pearson correlation. p-values were corrected for multiple comparisons using the Bonferroni method.

Predicting Parameters from Parameters of the Other Model (Fig. 2.5C) We ran eight different regression models, predicting each parameter from the four parameters of the opposite models as well as their one-way interactions, using linear regression in R (RCoreTeam, 2016). Fig. 2.5C shows the explained variance (R^2) of each model.

Principal Component Analysis (PCA)

To extract components that covary across parameters, we ran PCA on the fitted parameters data (8 parameters per participant). PCA can be understood as a method that rotates the coordinate system to align the first axis with the dimension of largest variation in the dataset (first principle component; PC), the second axis with the dimension of second-largest variance (second PC), while being orthogonal to the first, and so on. In this way, all resulting PCs are orthogonal to each other, and explain subsequently less variance. We conducted a PCA after centering and scaling (z-scoring) the data, using the statistical programming language R (RCoreTeam, 2016).

Age Differences in PCs (Fig. 2.5E) For each PC, we ran similar regression models as for our behavioral measures of performance, predicting participants' PCs from age (linear), age (quadratic), and sex. When significant, effects were noted in Fig. 2.5E. For PC2 and PC4, we also conducted post-hoc t-tests, correcting for multiple comparison using the Bonferroni method (Table 2.3).

Table 2.3: Results of t-tests on PC2 and PC4. df: Welch-adjusted degrees of freedom.

Comparison	t	df	p	Sig.
PC2 (8-15 vs. 15-30)	3.44	266.2	< 0.001	***
PC4 (8-13 vs. 13-17)	2.28	176.8	0.047	*
PC4 (13-17 vs. 18-30)	2.49	176.6	0.028	*

2.5 Supplemental Material

Pubertal Development

Participants aged 8-17 completed the pubertal developmental scale (PDS), a questionnaire that determines pubertal status based on questions about physical development (Petersen et al., 1988). In addition, an hour after the start of the experiment and in-between tasks, participants provided a 1.8 ml saliva sample, which was analyzed for testosterone levels as a marker of pubertal development. The procedure is described in detail in Master et al., 2020. PDS scores and testosterone levels were highly correlated with age for both males and females (suppl. Fig. 2.6B), making it difficult to assess them separately. We created quantile groups for pubertal measurements similar to age: For PDS scores, we assigned all participants with score 1 to the pre-pubertal group, and divided the remaining participants into tertiles based on score, which we termed "early", "middle", and "late" puberty. Tertiles were defined separately for males and females to assure sex balance within each group (suppl. Fig. 2.6A, middle row). For testosterone levels, we created quartiles based on testosterone levels, again defining quartiles separately for males and females.

Developmental patterns were similar for pubertal development (PDS, testosterone) and age (suppl. Fig. 2.7, 2.8, 2.9). The main difference was at which time peak performance occurred: in the third quantile based on age (13-15 years), but the fourth quantile based on puberty (suppl. Fig. 2.7). Parameter trajectories also differed slightly: most notably, p and β showed more abrupt changes based on PDS, with steps between mid and late puberty. α_- and p_{reward} showed a drastic step at puberty onset (between "pre" and "early"; suppl. Fig. 2.8B). In terms of testosterone, parameters α_-, p_{reward}, and p_{switch} showed U-shaped functions similar to age, but minima occurred in the fourth rather than the third quartile (suppl. Fig. 2.8C). In terms of parameter PCs as well, trajectories were largely similar between pubertal measures and age. Slight differences included a more unique role of pre-pubertal participants, especially for PC2 in terms of PDS and PC3 for testosterone (suppl. Fig. 2.9).

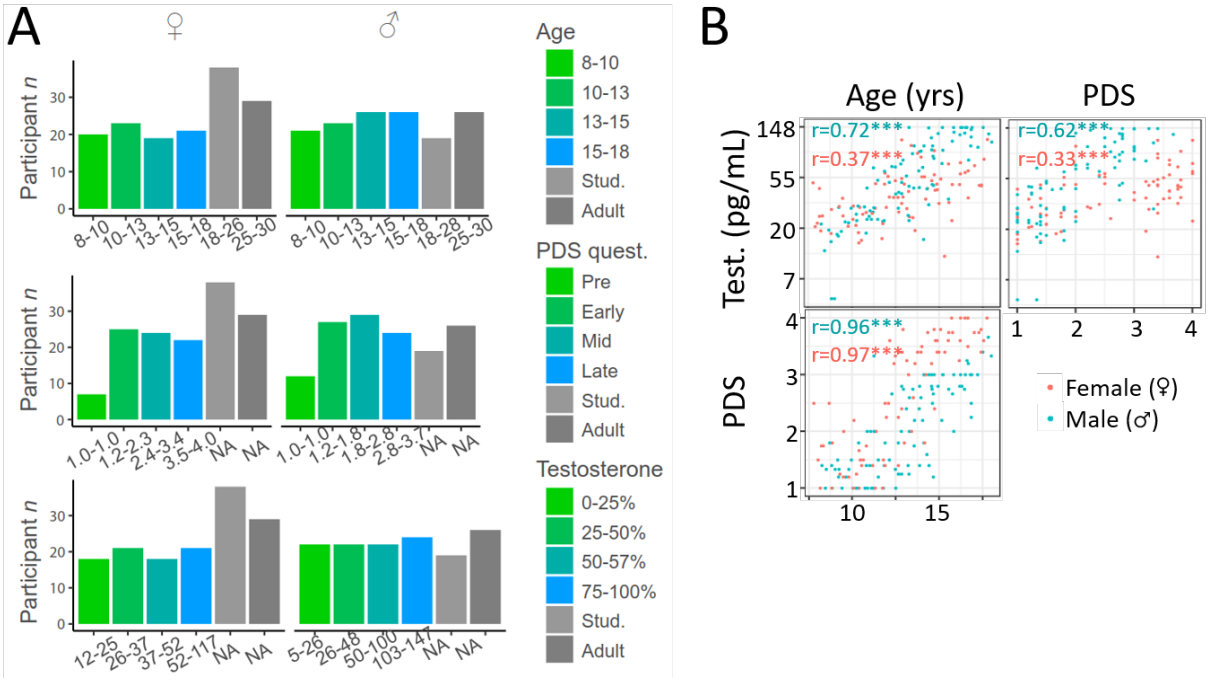


Figure 2.6: Participant sample and pubertal development. A) Number of participants in each bin, separately for each sex. Top: Age quantiles, which are the basis of Figures 2.2, 2.3, 2.4, and 2.5, and suppl. Figures 2.13, 2.15, 2.16, and 2.17. Numbers on the x-axis indicate the age ranges that went into each quantile bin, which differed slightly between males and females. The legend shows the names of the bins used throughout the paper. Middle: Bins based on the pubertal development questionnaire (PDS), which was available only for participants aged 8-17. The numbers on the x-axis show the ranges of each bins, which differed substantially between sexes. The legend shows the bin names after combining males and females. Bottom: Bins based on salivary testosterone levels, using the same conventions as above. B) Correlations between age, testosterone levels (Test.), and PDS questionnaire, for male and female participants aged 8-17. Stars refer to p-values, using the same convention as in main text figures.

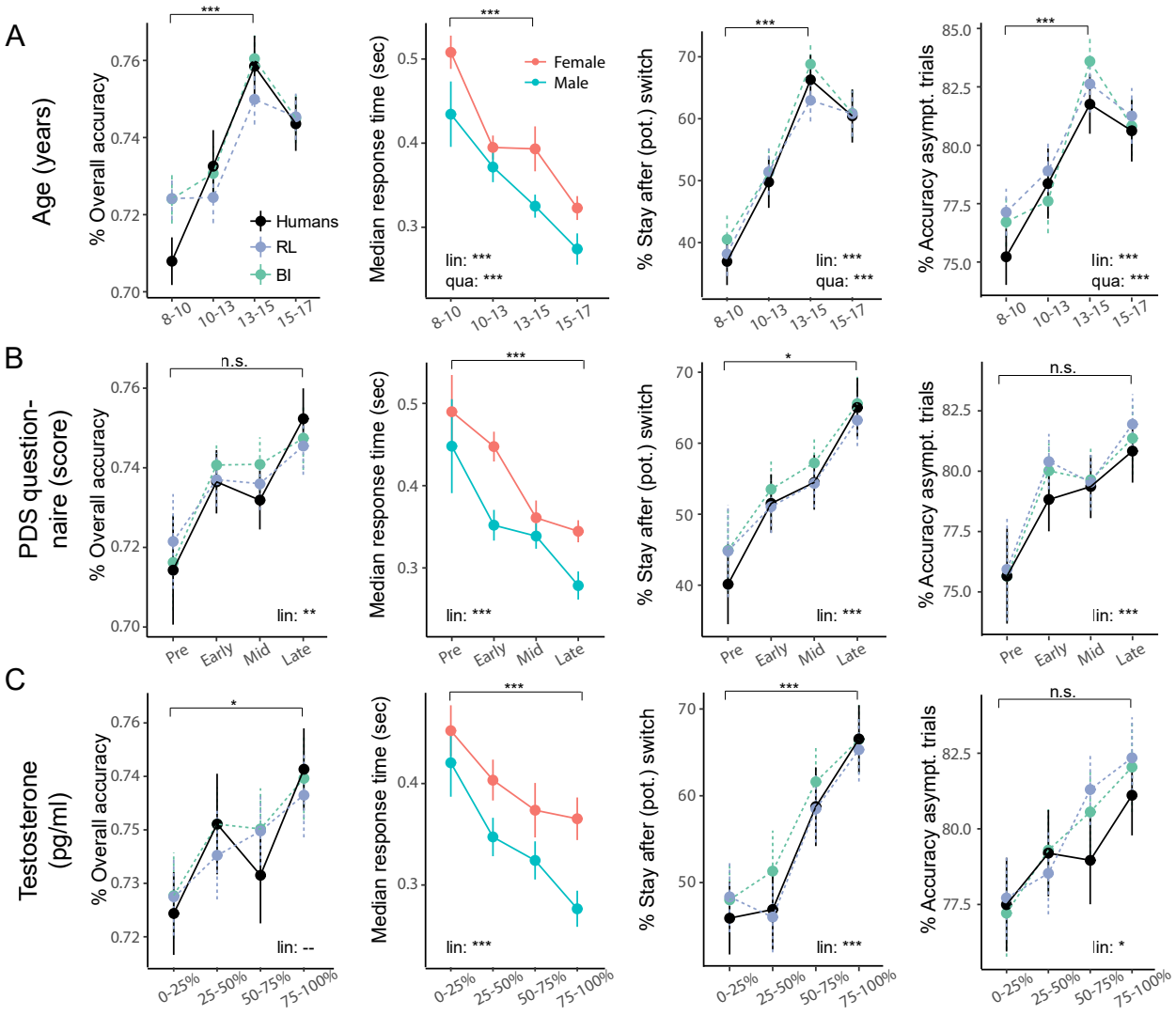


Figure 2.7: Behavior broken up by age / PDS / Testosterone bins. Significance bars and stars show the results of planned t-tests. A) Same data as in Fig. 2.3. Planned t-tests compared 8-10 year olds to 13-15 year olds. B) Same measures, but broken up by PDS bins. T-tests compared pre-pubertal to late-pubertal participants. C) Same measures, broken up by testosterone bins. T-tests compared participants in the first quartile in terms of testosterone levels to participants in the fourth quartile.

Table 2.4: Statistics of mixed-effects regression models predicting performance measures from sex (male, female) and puberty measures (PDS questionnaire / salivary testosterone). Only participants who had these measures were included in the model, restricting it to participants under the age of 18. Overall accuracy, stay after potential (pot.) switch, and asymptotic performance were modeled using logistic regression, and z-scores are reported. Log-transformed response times on correct trials were modeled using linear regression, and t-values are reported. * $p < .05$; ** $p < .01$, *** $p < .001$.

Performance measure (Figure)	Predictor	β	z / t	p	sig.
Overall accuracy (2.7B, left)	PDS	0.069	2.9	0.0038	**
	Sex	0.017	0.37	0.71	
Response times (2.7B, 2^{nd}-to-left)	PDS	-0.13	-4.9	< 0.001	***
	Sex	0.25	4.8	< 0.001	***
Stay after (pot.) switch (2.7B, 2^{nd}-to-right)	PDS	0.48	3.5	< 0.001	***
	Sex	0.76	2.9	0.0036	**
Asymptotic performance (2.7B, right)	PDS	0.25	4.2	< 0.001	***
	Sex	0.098	0.9	0.39	
Overall accuracy (2.7C, left)	Test.	< 0.0001	1.2	0.24	
	Sex	0.032	0.69	0.49	
Response times (2.7C, 2^{nd}-to-left)	Test.	-0.0034	-5.1	< 0.001	***
	Sex	0.010	1.9	0.049	*
Stay after (pot.) switch (2.7C, 2^{nd}-to-right)	Test.	0.012	3.5	< 0.001	***
	Sex	0.27	1.0	0.29	
Asymptotic performance (2.7C, right)	Test.	0.0034	2.2	0.029	*
	Sex	0.12	1.0	0.34	

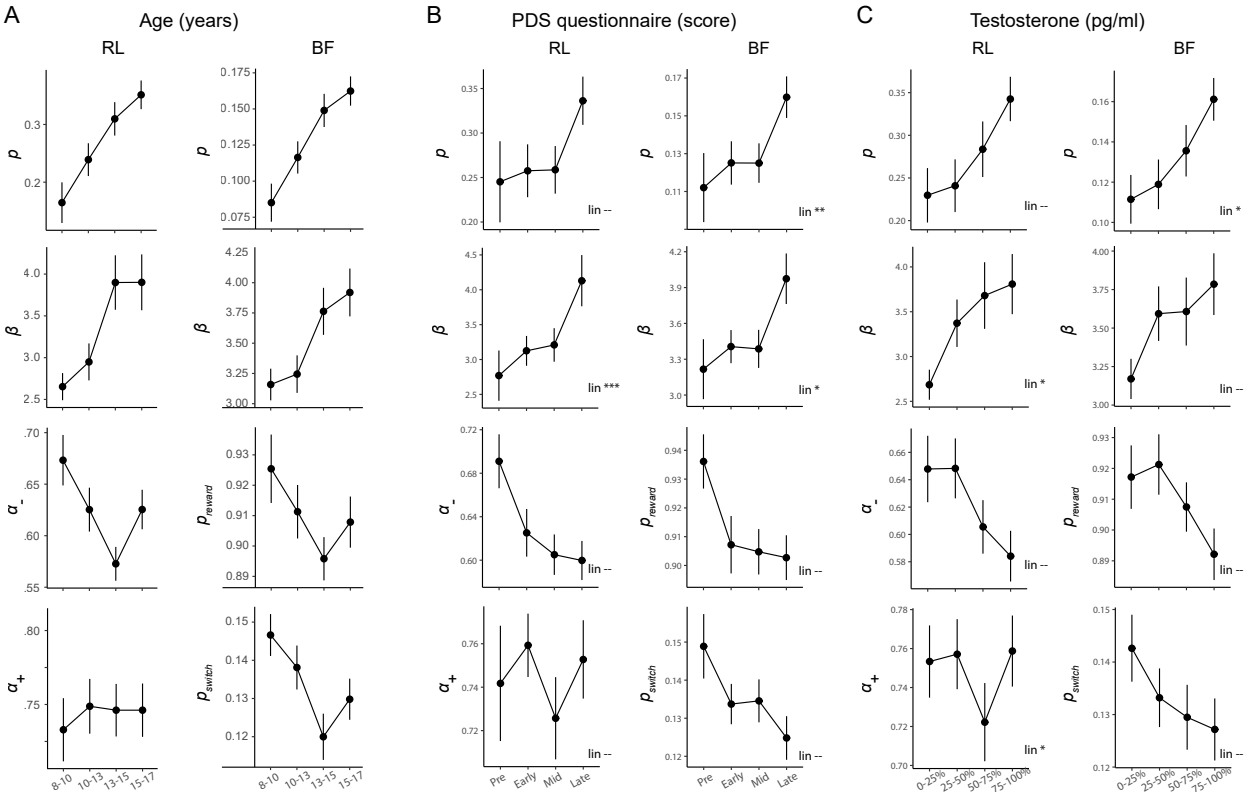


Figure 2.8: Model parameters broken up by age / PDS / Testosterone bins. A) Participants younger than 18 years of age, reproduced from Fig. 2.4. B)-C) Same data, broken up by PDS / testosterone bins. “lin.” indicates whether a linear effect of the measure of interest (PDS / testosterone) reached significance in a linear regression model.

Table 2.5: Parameter estimates and statistics from hierarchical model fitting, for pubertal predictors (PDS questionnaire, salivary testosterone), for participants under the age of 18. Significance tests against 0 for parameters whose range includes 0, NA otherwise.

Model	Parameter	$\mu \pm sd$	95% CI	p-value	sig.
PDS					
4-param. BI	p_{int}	0.11 ± 0.013	$[0.082, 0.13]$	< 0.001	***
	p_{sd}	0.089 ± 0.0085	$[0.073, 0.11]$	0	NA
	p_{lin}	0.022 ± 0.0096	$[0.0039, 0.041]$	0.0086	**
	β_{int}	3.81 ± 0.26	$[3.31, 4.34]$	0	NA
	β_{sd}	1.25 ± 0.14	$[0.98, 1.53]$	0	NA
	β_{lin}	0.31 ± 0.16	$[-0.018, 0.62]$	0.028	*
	$p_{reward\ int}$	0.88 ± 0.019	$[0.84, 0.92]$	0	NA
	$p_{reward\ sd}$	0.060 ± 0.011	$[0.038, 0.082]$	0	NA
	$p_{reward\ lin}$	$< 0.001 \pm 0.010$	$[-0.019, 0.020]$	0.48	–
	$p_{switch\ int}$	0.16 ± 0.016	$[0.13, 0.20]$	0	NA
	$p_{switch\ sd}$	0.067 ± 0.0070	$[0.053, 0.080]$	0	NA
	$p_{switch\ lin}$	-0.0098 ± 0.0099	$[-0.029, 0.0099]$	0.16	–
4-param. RL	p_{int}	0.25 ± 0.026	$[0.20, 0.30]$	< 0.001	***
	p_{sd}	0.24 ± 0.019	$[0.20, 0.28]$	0	NA
	p_{lin}	0.039 ± 0.024	$[-0.0093, 0.087]$	0.054	–
	β_{int}	3.15 ± 0.13	$[2.90, 3.41]$	0	NA
	β_{sd}	1.37 ± 0.13	$[1.12, 1.62]$	0	NA
	β_{lin}	0.41 ± 0.13	$[0.17, 0.66]$	< 0.001	***
	$\alpha_{-\ int}$	0.60 ± 0.016	$[0.56, 0.62]$	0	NA
	$\alpha_{-\ sd}$	0.16 ± 0.013	$[0.14, 0.18]$	0	NA
	$\alpha_{-\ lin}$	-0.0155 ± 0.017	$[-0.048, 0.019]$	0.18	–
	$\alpha_{+\ int}$	0.66 ± 0.028	$[0.61, 0.72]$	0	NA
	$\alpha_{+\ sd}$	0.35 ± 0.034	$[0.023, 0.15]$	0	NA
	$\alpha_{+\ lin}$	0.0085 ± 0.027	$[-0.048, 0.059]$	0.38	–
Testosterone					
4-param. BI	p_{int}	0.11 ± 0.013	$[0.081, 0.13]$	< 0.001	***
	p_{sd}	0.089 ± 0.0084	$[0.073, 0.11]$	0	NA
	p_{lin}	0.02 ± 0.010	$[0.0023, 0.040]$	0.015	*
	β_{int}	3.78 ± 0.26	$[3.29, 4.31]$	0	NA
	β_{sd}	1.28 ± 0.14	$[1.00, 1.55]$	0	NA
	β_{lin}	0.12 ± 0.17	$[-0.20, 0.45]$	0.22	–
	$p_{reward\ int}$	0.88 ± 0.019	$[0.85, 0.92]$	0	NA
	$p_{reward\ sd}$	0.056 ± 0.011	$[0.035, 0.077]$	0	NA
	$p_{reward\ lin}$	-0.0135 ± 0.010	$[-0.033, 0.0081]$	0.90	–
	$p_{switch\ int}$	0.16 ± 0.016	$[0.13, 0.19]$	0	NA
	$p_{switch\ sd}$	0.067 ± 0.0069	$[0.054, 0.081]$	0	NA
	$p_{switch\ lin}$	-0.0082 ± 0.010	$[-0.029, 0.012]$	0.22	–
4-param. RL	p_{int}	0.24 ± 0.025	$[0.20, 0.29]$	< 0.001	***
	p_{sd}	0.24 ± 0.0195	$[0.20, 0.28]$	0	NA
	p_{lin}	0.038 ± 0.025	$[-0.0091, 0.190]$	0.066	–
	β_{int}	3.16 ± 0.14	$[2.89, 3.43]$	0	NA
	β_{sd}	1.42 ± 0.13	$[1.17, 1.69]$	0	NA
	β_{lin}	0.28 ± 0.13	$[0.037, 0.54]$	0.013	*
	$\alpha_{-\ int}$	0.60 ± 0.017	$[0.55, 0.62]$	0	NA
	$\alpha_{-\ sd}$	0.16 ± 0.013	$[0.13, 0.18]$	0	NA
	$\alpha_{-\ lin}$	-0.035 ± 0.018	$[-0.070, -0.0016]$	0.24	–
	$\alpha_{+\ int}$	0.66 ± 0.028	$[0.61, 0.72]$	0	NA
	$\alpha_{+\ sd}$	0.10 ± 0.030	$[0.045, 0.16]$	0	NA
	$\alpha_{+\ lin}$	-0.017 ± 0.026	$[-0.066, 0.036]$	0.015	*

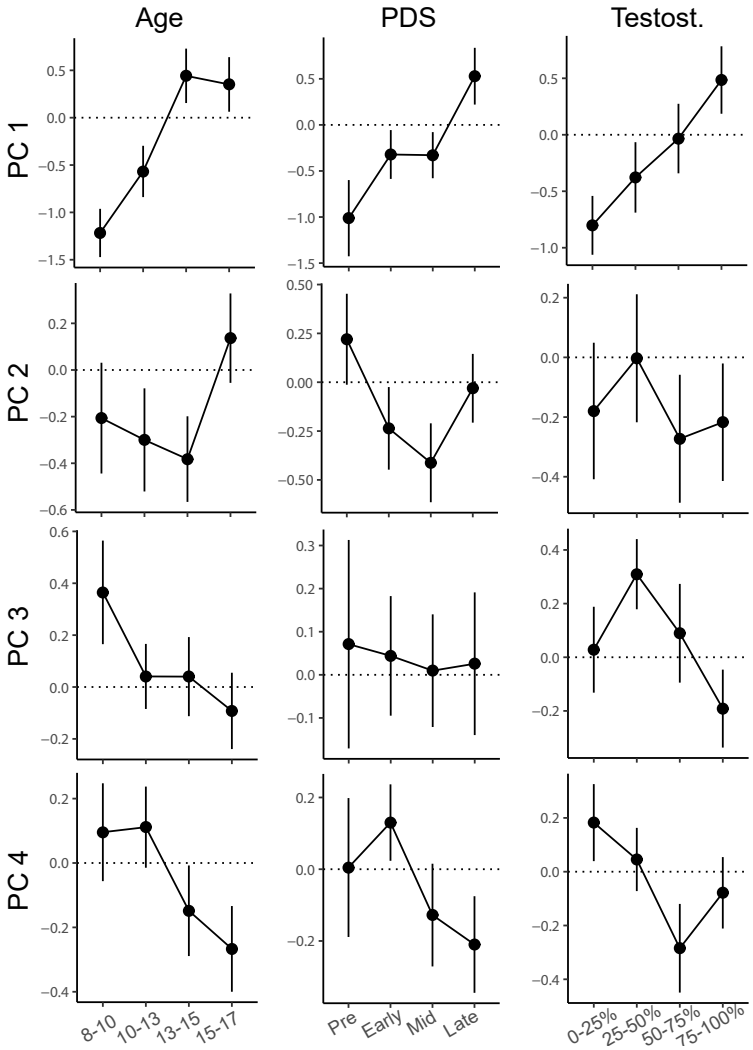


Figure 2.9: Model parameter PCs broken up by age / PDS / Testosterone bins. Left row: Participants younger than 18 years of age, reproduced from Fig. 2.5. Middle (right) row: same data, but broken up by PDS (testosterone) bins.

Effects of Puberty After Controlling for Age

We next sought to control for age and examine the effect of puberty alone. To this end, we investigated the continuous effects of puberty within each age bin, to eliminate confounds with age (Master et al., 2020). In concordance with the finding that behavior peaked in the third age bin (13-15 years), but in the fourth PDS bin (75-100th percentile), all measures of behavior increased qualitatively with respect to PDS in the third and fourth age bins (suppl. Fig. 2.10A, right-most column). Nevertheless, this pattern is difficult to interpret because pubertal status was heavily con-

founded with sex in the fourth age bin, such that girls scored higher on the PDS questionnaire than boys of the same age, in concordance with typical age differences in pubertal maturation. Within the age bins that contained participants across the entire range of pubertal status (10-13, 13-15, and 15-17 years), few significant effects of PDS (suppl. Fig. 2.10A) or salivary testosterone levels (suppl. Fig. 2.10B) were observed, possibly including some that occurred by chance. In our data, stay after (pot.) switch trials showed a qualitative decrease with PDS score in 10-13 year olds, was constant in 13-15 year olds, and showed a qualitative increase in 15-17 year olds. This could indicate a weak U-shaped effect or simply experimental noise.

In the case of fit model parameters, pubertal development did not show significant positive relationships with choice parameters p and β, which we might predict if pubertal development was a driving mechanism in growth for these parameters between ages 8-18 (suppl. Table 2.6; suppl. Fig. 2.11, 2.12). In terms of learning parameters, pubertal development also did not show significant negative relationships with α_- and α_+ (RL), or p_{reward} and p_{switch} (BI), which we might predict if pubertal onset was driving the decrease of these parameters between ages 8-15. If anything, we saw the opposite pattern in males: α_-, p_{reward}, and p_{switch} showed a qualitatively positive relationship with PDS scores (suppl. Fig. 2.11) and testosterone (suppl. Fig. 2.12) in the 10-13 year old age group, and a qualitatively negative relationship with PDS in the 13-15 year old age group. Overwhelmingly, these relationships were not statistically significant.

Trend relationships found within the 13-15 year-old group included a marginal effect of PDS on α_+ (β=0.075, p=0.092), a marginal effect of sex on p_{switch} in the testosterone model (β=0.047, p=0.078), and a significant interaction between sex and testosterone on p_{switch} (β=0.00097, p=0.015; suppl. Table 2.6). Note that these statistical tests were not corrected for multiple comparisons, making it possible that these results were observed by chance, and should thus be interpreted carefully. The cross-sectional design of our experiment may limit our ability to detect pubertal effects (Kraemer et al., 2000). It is possible that experiments with greater power, longitudinal studies, and studies of hormone manipulation may further inform these largely negative results.

Table 2.6: Statistics of regression models testing effects of puberty within the age bin 13-15 years. This bin was chosen because it contained participants across the full range of pubertal development.

Outcome	Predictor	β	p	Sig.
Testosterone				
p (RL)	Test.	-0.00096	0.57	
	Sex	0.062	0.65	
	Interaction	0.0011	0.58	
β (RL)	Test.	-0.022	0.23	
	Sex	1.86	0.22	
	Interaction	0.034	0.13	
α_-	Test.	-0.00033	0.69	
	Sex	0.047	0.48	
	Interaction	0.0014	0.16	
α_+	Test.	-0.00074	0.47	
	Sex	0.0026	0.97	
	Interaction	0.00055	0.65	
p (BF)	Test.	-0.00052	0.43	
	Sex	0.045	0.40	
	Interaction	0.00083	0.30	
β (BF)	Test.	-0.018	0.12	
	Sex	1.12	0.21	
	Interaction	0.021	0.12	
p_{reward}	Test.	-0.00038	0.31	
	Sex	0.0012	0.97	
	Interaction	0.00027	0.54	
p_{switch}	Test.	0.00053	0.10	
	Sex	0.047	0.078	’
	Interaction	0.00097	0.015	*
PDS				
p (RL)	PDS	0.0044	0.95	
	Sex	0.18	0.52	
	Interaction	0.079	0.43	
β (RL)	PDS	0.87	0.30	
	Sex	2.37	0.45	
	Interaction	0.67	0.55	
α_-	PDS	-0.024	0.52	
	Sex	0.071	0.61	
	Interaction	0.063	0.21	
α_+	PDS	0.075	0.092	’
	Sex	0.21	0.21	
	Interaction	0.051	0.39	
p (BF)	PDS	0.011	0.69	
	Sex	0.084	0.45	
	Interaction	0.032	0.43	
β (BF)	PDS	0.62	0.21	
	Sex	1.96	0.30	
	Interaction	0.64	0.34	
p_{reward}	PDS	-0.0080	0.63	
	Sex	0.023	0.72	
	Interaction	0.022	0.33	
p_{switch}	PDS	-0.010	0.51	
	Sex	0.010	0.86	
	Interaction	0.0057	0.82	

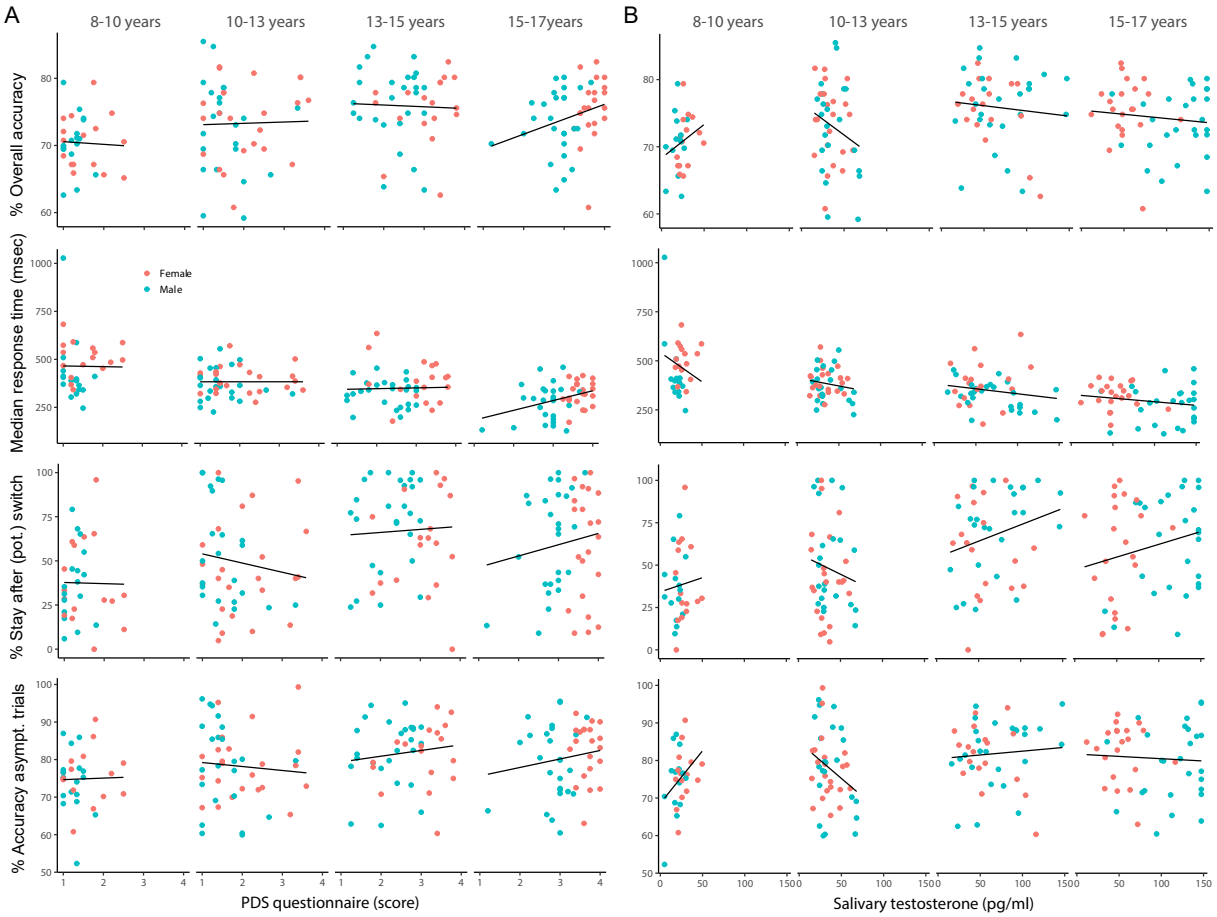
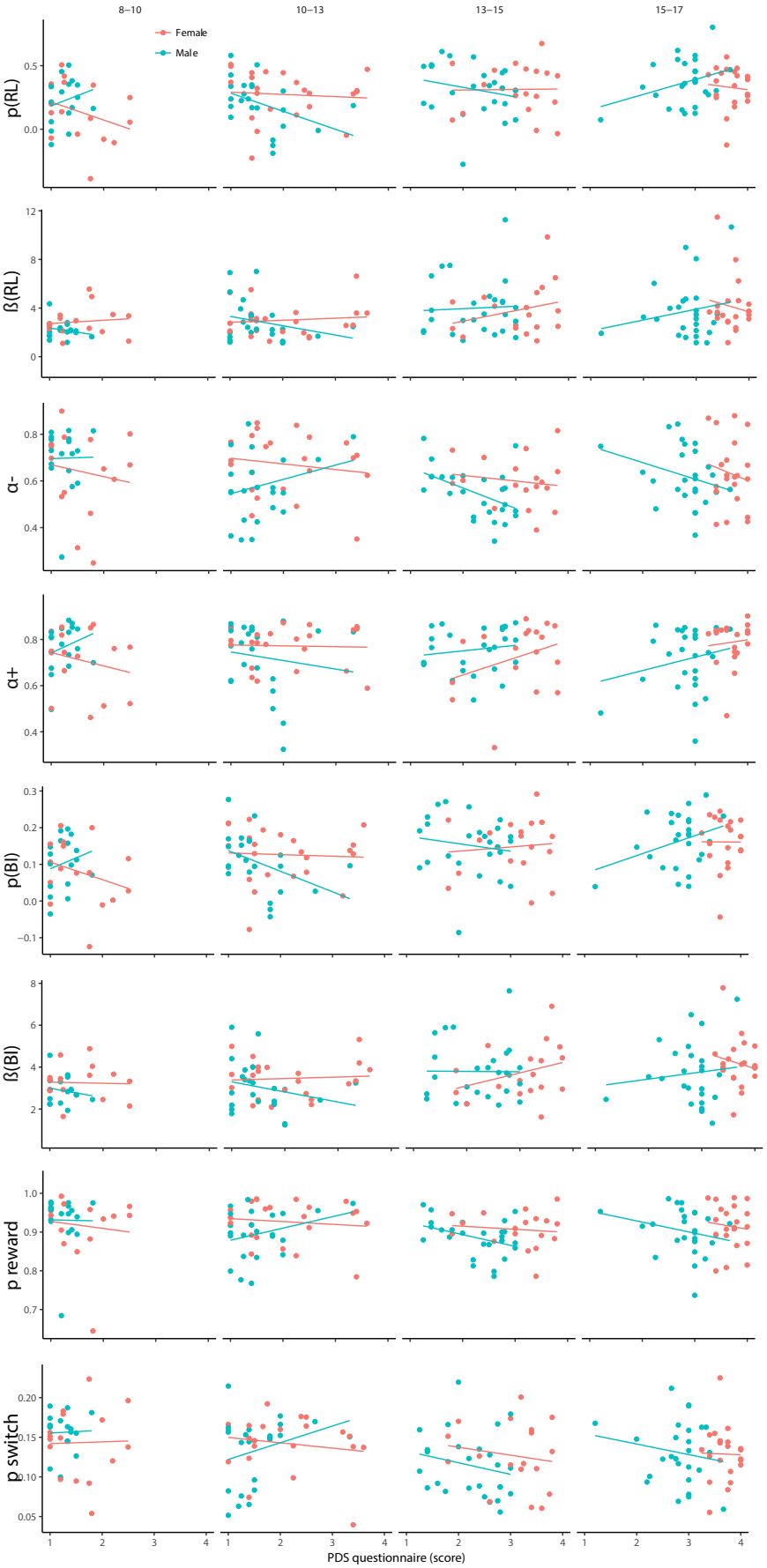
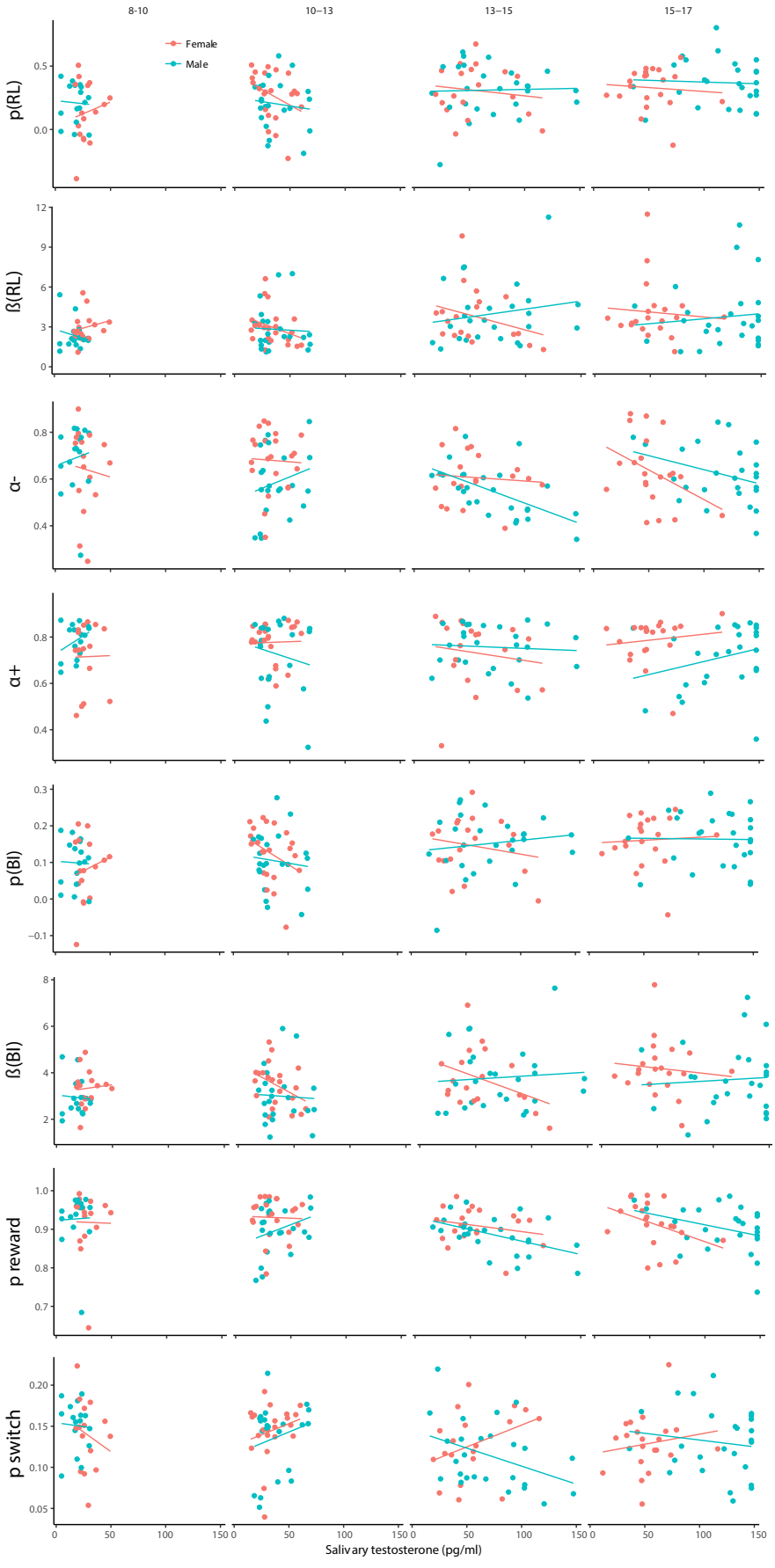


Figure 2.10: Effects of pubertal status on performance, controlling for age. Each column shows one age group, colors denote sex. Pubertal status was determined by (A) PDS questionnaire, or (B) salivary testosterone.


8–10
10–13
13–15
15–17
Female
Male
p(RL)
ß(RL)
α-
α+
p(BI)
ß(BI)
p reward
p switch
PDS questionnaire (score)


8-10
10–13
13–15
15–17
Female
Male
p(RL)
ß(RL)
α-
α+
p(BI)
ß(BI)
p reward
p switch
Salivary testosterone (pg/ml)

Additional Behavioral Analyses

The youngest children showed the lowest overall and asymptotic accuracy (Fig. 2.3C, F) and were the most likely to switch after a single negative outcome (Fig. 2.3E, suppl. Fig. 2.15B, middle). This explains why they were also fastest at switching (suppl. Fig. 2.13A, D; suppl. Table 2.7). Response times were the only performance measure in which 13-15 year olds were outperformed by another age group, university undergraduates (age 18-28; Fig. 2.2B, 2.3D)). Potential reasons for undergraduates' faster responses include greater familiarity with lab-based psychological experiments, more experience with computers, and increased motivation to finish the task quickly.

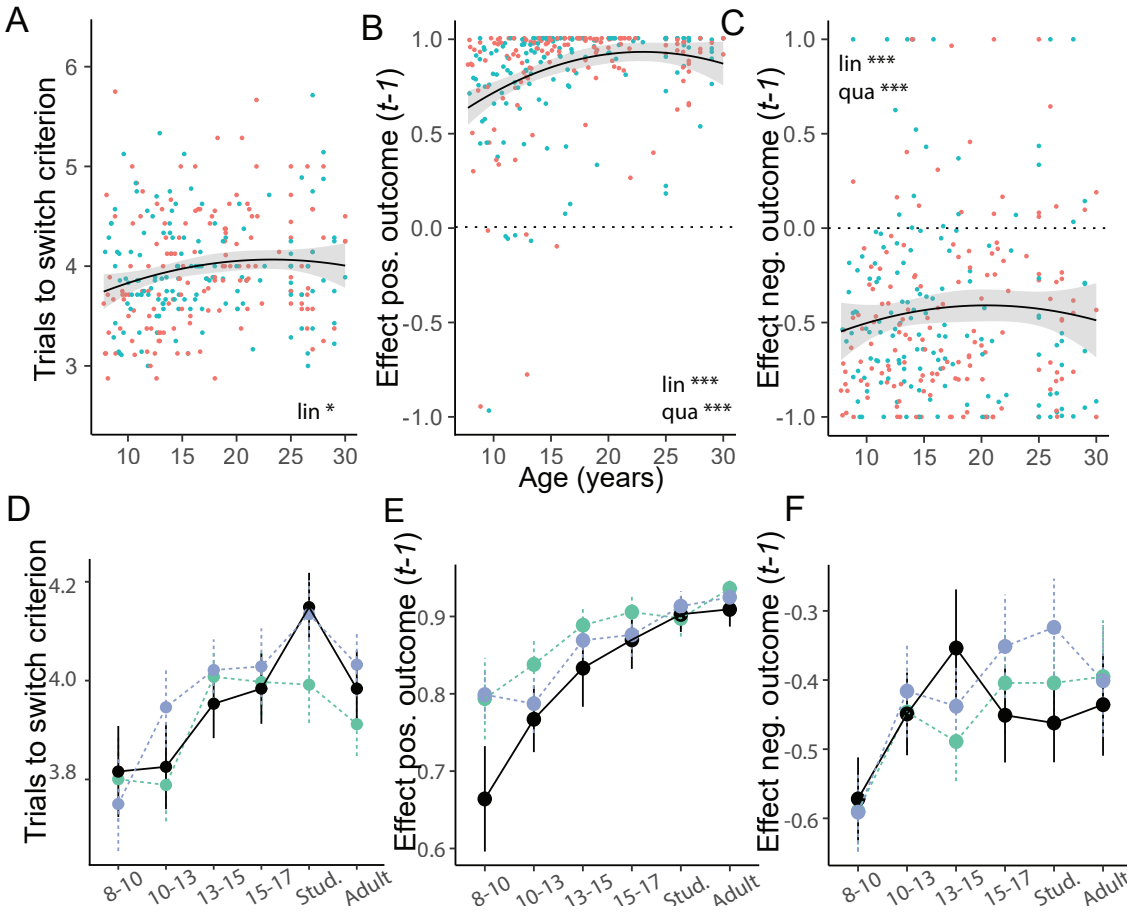


Figure 2.13: Human behavior (A-C) and model validation (D-F) for additional behavioral measures. (A, D): Number of trials after task switch until participants reached performance criterion (2 correct responses). (B-F): Effect of previous negative (B, E) and positive (C, F) outcomes on choices. "$t-1$": Outcome occurred 1 trial before choice, i.e., delay $i = 1$. Regression weights were tanh transformed for visualization.

Table 2.7: Statistics of mixed-effects regression model predicting switch criterion from sex (male, female), age (years and months; "lin."), and squared age ("qua."). * $p < .05$; ** $p < .01$, *** $p < .001$.

Behavioral measure (Figure)	Predictor	β	t	p	sig.
Switch criterion (2.13A)	Age (lin.)	0.067	2.0	0.048	*
	Age (qua.)	-0.0014	-1.6	0.11	
	Trial	0.0059	10.0	< 0.001	***
	Sex	0.0022	0.04	0.97	

Statistics for Regression Models

We conducted regression models predicting future choice from past choice and outcomes. The full statistics of these models are shown in Table 2.8.

Meta-Priors for Hierarchical Bayesian Models

When specifying the hierarchical Bayesian models (Fig. 2.3B and age-less version) for parameter fitting, we chose uninformative priors, as detailed in Table 2.9.

Statistics for Hierarchical Bayesian Models

We verified convergence of the Hierarchical Bayesian Model (Fig. 2.3B) using the Markov-Chain error, effective sample size (n), and the R-hat statistic ($\hat{R}$), using the functions provided by the PyMC3 toolbox (suppl. Table 2.10; Salvatier et al., 2016).

Table 2.10: Statistics for hierarchical Bayesian models. We report the average and the range (min and max over all model parameters) for the two winning models.

Model		**MC error**	**Effective** n	$\hat{R}$
4-param. RL	mean	< 0.001	$2,517$	1.001
	range	$[< 0.001; 0.002]$	$[155; 4,261]$	$[1.000; 1.015]$
4-param. BI	mean	0.002	816	1.001
	range	$[< 0.001; 0.01]$	$[281; 1,576]$	$[1.000; 1.004]$

Assessing Model Identifiability using Generate and Recover

All model fits are relative. In other words, when model A fits data better than model B, there is no guarantee that model A fits the data "well". Both models could fit the data poorly, with model B being even worse than model A. To ensure that our models fit well, we validated our parameter

Table 2.8: Logistic mixed-effect regression, predicting future actions from past actions and outcomes (methods). The number of predictors ($i \leq 8$) was chosen as to provide the best model fit: $AIC_{i\leq 3}$: 31.046; $AIC_{i\leq 4}$: 31.013; $AIC_{i\leq 5}$: 31.001; $AIC_{i\leq 6}$: 30.981; $AIC_{i\leq 7}$: 30.963; $AIC_{i\leq 8}$**: 30.962**; $AIC_{i\leq 9}$: 30.966; $AIC_{i\leq 10}$: 30.964.

Predictor	**delay** i	β	z	p	**Sig.**
Intercept		-0.01	-0.74	0.46	
Main effects					
Age (lin.)		-0.13	-1.40	0.16	
Age (qua.)		0.12	1.30	0.19	
Pos. outcome	1	2.19	68.09	< 0.001	***
	2	0.84	27.36	< 0.001	***
	3	0.24	7.87	< 0.001	***
	4	0.13	4.30	< 0.001	***
	5	-0.017	-0.54	0.58725	
	6	-0.017	-0.56	0.57548	
	7	-0.0035	-0.12	0.90613	
	8	-0.077	-2.77	0.0057	**
Neg. outcome	1	-0.73	-37.09	< 0.001	***
	2	-0.24	-10.64	< 0.001	***
	3	0.0055	0.22	0.82278	
	4	0.13	5.39	< 0.001	***
	5	0.12	4.87	< 0.001	***
	6	0.12	4.73	< 0.001	***
	7	0.13	5.32	< 0.001	***
	8	0.016	0.71	0.47857	
Interaction age (lin.)					
Pos. outcome	1	0.90	4.50	< 0.001	***
	2	0.84	4.19	< 0.001	***
	3	0.50	2.52	0.012	*
	4	-0.069	-0.35	0.73	
	5	0.088	0.44	0.66	
	6	-0.38	-1.94	0.052	
	7	-0.18	-0.94	0.35	
	8	-0.27	-1.49	0.14	
Neg. outcome	1	0.67	5.27	< 0.001	***
	2	-0.37	-2.48	0.013	*
	3	0.16	1.03	0.30	
	4	-0.089	-0.55	0.58	
	5	0.012	0.07	0.94	
	6	0.066	0.41	0.68	
	7	0.011	0.07	0.94	
	8	-0.068	-0.47	0.63	
Interaction age (qua.)					
Pos. outcome	1	-0.64	-3.14	0.0017	**
	2	-0.89	-4.41	< 0.001	***
	3	-0.38	-1.90	0.057	
	4	0.0020	0.01	0.99	
	5	-0.066	-0.33	0.74	
	6	0.36	1.80	0.072	
	7	0.15	0.75	0.456	
	8	0.29	1.62	0.11	
Neg. outcome	1	-0.56	-4.34	< 0.001	***
	2	0.30	2.00	0.046	*
	3	-0.16	-0.97	0.33	
	4	0.092	0.57	0.57	

Table 2.9: Hyper-priors and priors used in hierarchical Bayesian model fitting. In the age-based model, individuals' parameters were drawn from a Normal distribution around a parameter-specific, age-specific mean θ_m, with parameter-specific standard deviation θ_{sd} (top row of the table; see Fig. 2.3B for details). In the age-free model, individuals' parameters were drawn from parameter-specific group-level prior distributions (subsequent rows in the table). The shapes of these distributions were based on allowed parameter ranges (e.g., Gamma distribution for parameters with range [0, ∞], Beta distribution for parameters with range [0, 1]). The same prior distribution was used for all individuals, i.e., no age information was present in the age-free model. The distributions of individuals' parameters were themselves parameterized by prior parameters. In the age-based model, prior parameter θ_{sd} was distributed according to a HalfNormal (Normal, truncated at 0; middle section of the table), and parameterized by hyper-parameter $sd = 10$ to allow for a wide, non-informative shape (bottom section). Group-level prior θ_m was defined as an age-based regression function, parameterized by θ_{int}, θ_{lin}, and θ_{qua} for each parameter θ (middle section). The prior on the intercept θ_{int} of each parameter in the age-based model (middle section) had the same shape as the group-level prior distribution in the age-free model (top section), and was parameterized by the same hyper-priors (bottom section). In the age-less model, prior parameters parameterized the distributions of individual model parameters (middle section).

Level	Parameter	Distribution / Value	Explanation
Indiv. param.			
Age-based			
	θ	Normal($\mu = \theta_m$, $\sigma = \theta_{sd}$)	See Fig. 2.3B
Age-less			
	β	Gamma($\alpha = a_\beta$, $\beta = b_\beta$)	Range $[0, \infty[$
	p	Normal($\mu = m_p$, $\sigma = sd_p$)	Wide Normal
	α_+	Beta($\alpha = a_{\alpha+}$, $\beta = b_{\alpha+}$)	
	α_-	Beta($\alpha = a_{\alpha-}$, $\beta = b_{\alpha-}$)	
	p_{reward}	Beta($\alpha = a_{\ reward}$, $\beta = b_{\ reward}$)	
	p_{switch}	Beta($\alpha = a_{\ switch}$, $\beta = b_{\ switch}$)	
Prior			
Age-based			
	θ_{sd}	HalfNormal($\mu = m$, $= sd$)	Truncated Normal, $[0, \infty[$
	θ_m	$\theta_{int} + \theta_{lin}\ age + \theta_{qua}\ age^2$	Age-based regression
	$\theta_{int}, \theta = \beta$	Gamma($\alpha = a$, $\beta = b$)	
	$\theta_{int}, \theta = p$	Normal($\mu = m$, $\sigma = sd$)	
	$\theta_{int}, \theta \in [\alpha_+, \alpha_-, p_{reward}, p_{switch}]$	Beta($\alpha = a$, $\beta = b$)	Uniform, range [0, 1]
	θ_{lin}, θ_{qua}	Normal($\mu = m$, $\sigma = sd$)	
Age-less			
	a_β, b_β, $a_{\alpha+}$, $b_{\alpha+}$, $a_{\alpha-}$, $b_{\alpha-}$, $a_{p\ reward}$, $b_{p\ reward}$, $a_{p\ switch}$, $b_{p\ switch}$	Gamma($\alpha = a$, $\beta = b$)	
	m_p	Normal($\mu = m$, $\sigma = sd$)	
	sd_p	HalfNormal($\mu = m$, $= sd$)	
Hyper-prior			
	a	1	
	b	1	
	m	0	
	sd	10	

fitting and model comparison method by first simulating and then recovering parameters from each model (Wilson and Collins, 2019). An identifiable model will recover the simulated parameters well during fitting, whereas an unidentifiable model will not. We also compared the results of maximum likelihood and hierarchical Bayesian model fitting using this procedure. Both BF and RL model parameters were recovered well when using hierarchical Bayesian model fitting (age-free model), but recovery was much worse when using maximum likelihood (suppl. Fig. 2.14A), a well-known fact (Katahira, 2016). Hierarchical Bayesian model fitting also led to more consistent estimates of parameters β and p between both models (suppl. Fig. 2.14B), showing that this method was especially suited in our case. These results lend credence to the superior fit that can be achieved using Hierarchical Bayesian methods, and to the precision with which model parameter can be estimated.

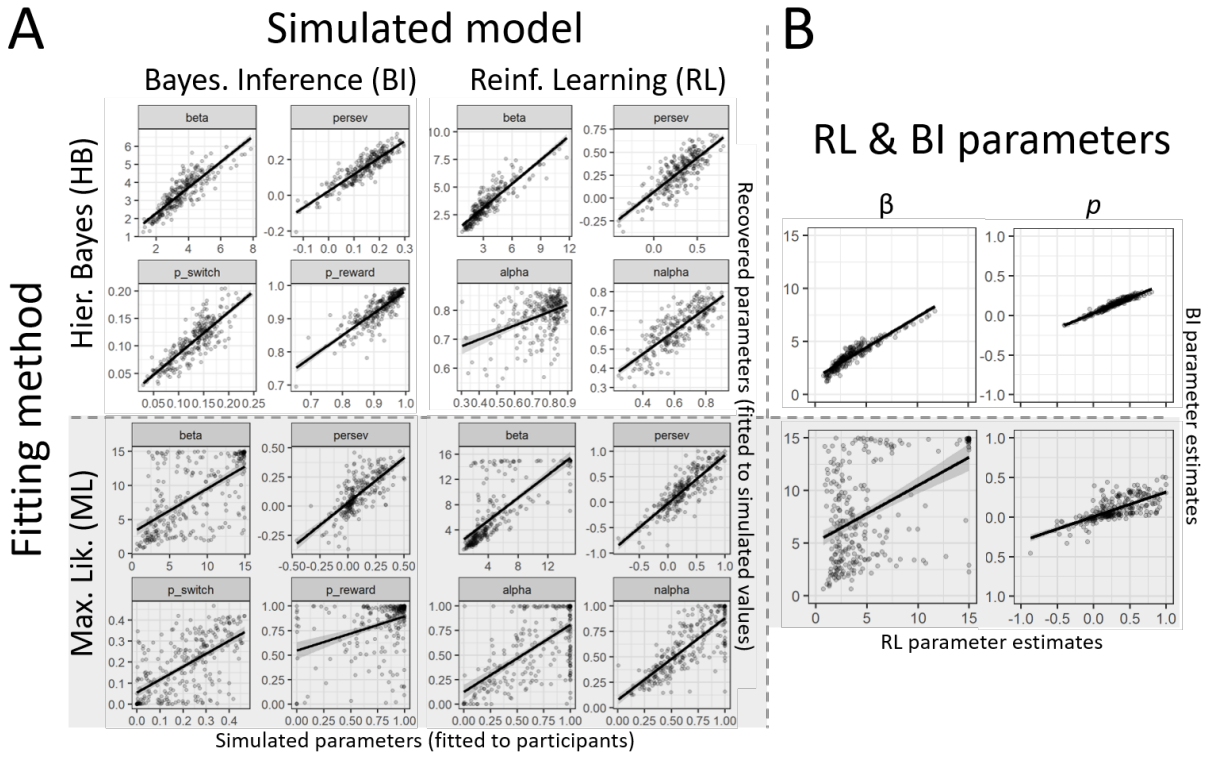


Figure 2.14: Model validation using hierarchical Bayesian model fitting (top, unshaded), as well as classical Maximum likelihood fitting (bottom, shaded). The results of hierarchical Bayesian fitting are presented in the main text. A) Simulate-and-recover procedure. The x-axes of all graphs show the parameter values of simulated datasets; the y-axes show the recovered parameters obtained by fitting these datasets using the same models. Recovered parameters should be as close to the simulated ones as possible, i.e., lie on the identity line. Black lines and shaded areas indicate best-fit regression lines. The left half presents simulate-and-recover results for the BI model, the right for the RL model. The top half shows the results of hierarchical Bayesian model fitting (our method), the bottom of the standard maximum likelihood method. This figure shows the well-established finding that hierarchical Bayesian model fitting outperformed maximum likelihood. B) Consistency in the estimation of parameters β and p. Human data was fit using RL and BI models to compare the estimates of β (left row) and p (right row) between models. When both (independent) models lead to the same estimates, dots lie on the identity line. This was indeed the case for hierarchical Bayesian fitting (top row), but not for maximum likelihood fitting (bottom row).

Qualitative Fit of RL and BI Models

To test the qualitative fit of our models, we simulated behavior using fitted parameters (from the age-free model; Methods), and checked whether the simulated behavior was able to reproduce

the patterns of interest in the human data (Palminteri et al., 2017). Indeed, both the winning RL and BI models captured human learning curves, as well as sex and age differences, very closely (suppl. Fig. 2.15). Simpler, non-winning models, on the other hand, failed to capture human characteristics (suppl. Fig. 2.17, 2.16).

Raw fitted parameters, obtained from the age-free model (Methods; suppl. Fig. 2.17, 2.16), show that age differences are evident even when age slopes were not part of the fitting model, i.e., when individual parameters were not biased by age effects at the group level. To evaluate age effects in a statistically sound way, we used a hierarchical Bayesian model that explicitly modeled age effects (Fig. 2.3B). Significant effects (suppl. Table 2.11) are shown as lines in suppl. Figures 2.17 and 2.16. To asses effects of age groups, we tested differences in posterior samples of the age-free model (Methods). Statistics are shown in suppl. Table 2.12.

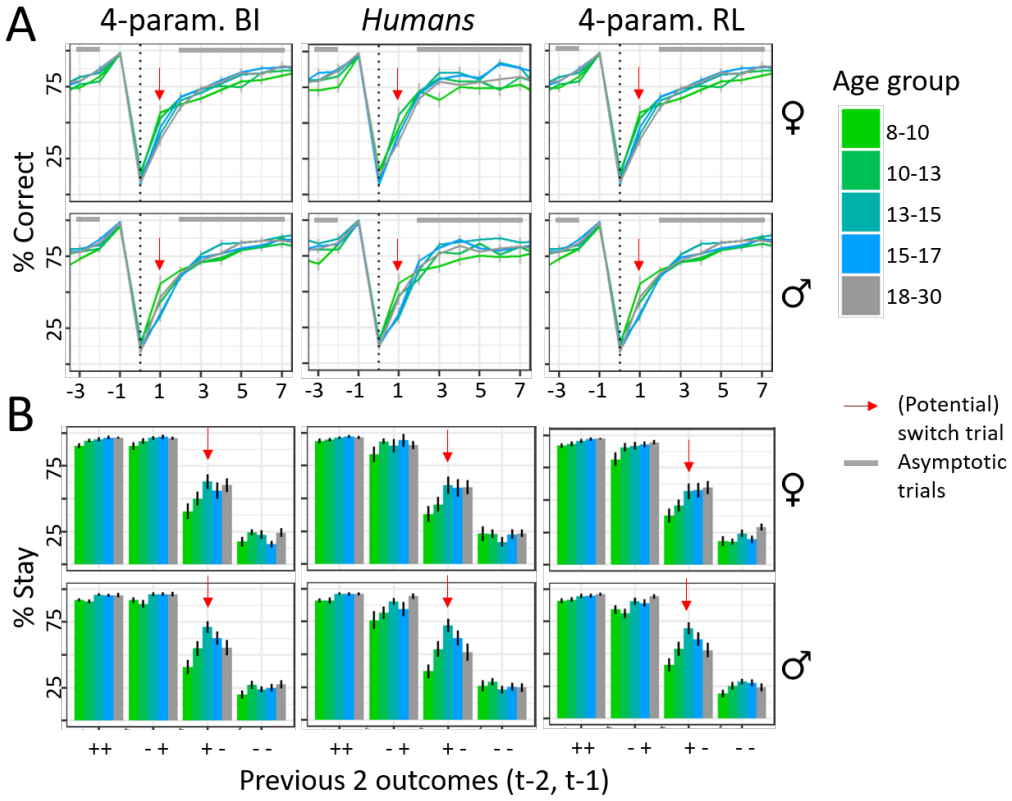


Figure 2.15: Human and model behavior, showing that models closely reproduced human patterns, A) Behavior in response to switch trials. Colors refer to age groups, red arrows show switch trials, grey bars trials of asymptotic performance. Both models captured quicker switching on switch trials in younger (light green) compared to older participants (blue and grey), and best performance on asymptotic trials in adolescents (green-blue). B) Stay probability in response to outcomes 2 trials back. Both RL and BI replicated human behavior and age differences, including linear increase in staying after positive outcomes ("+ +" and "- +"), and the inverse-U shape on potential switch trials (red arrow; "+ -" condition). Qualitative (non-significant) sex differences were also captured.

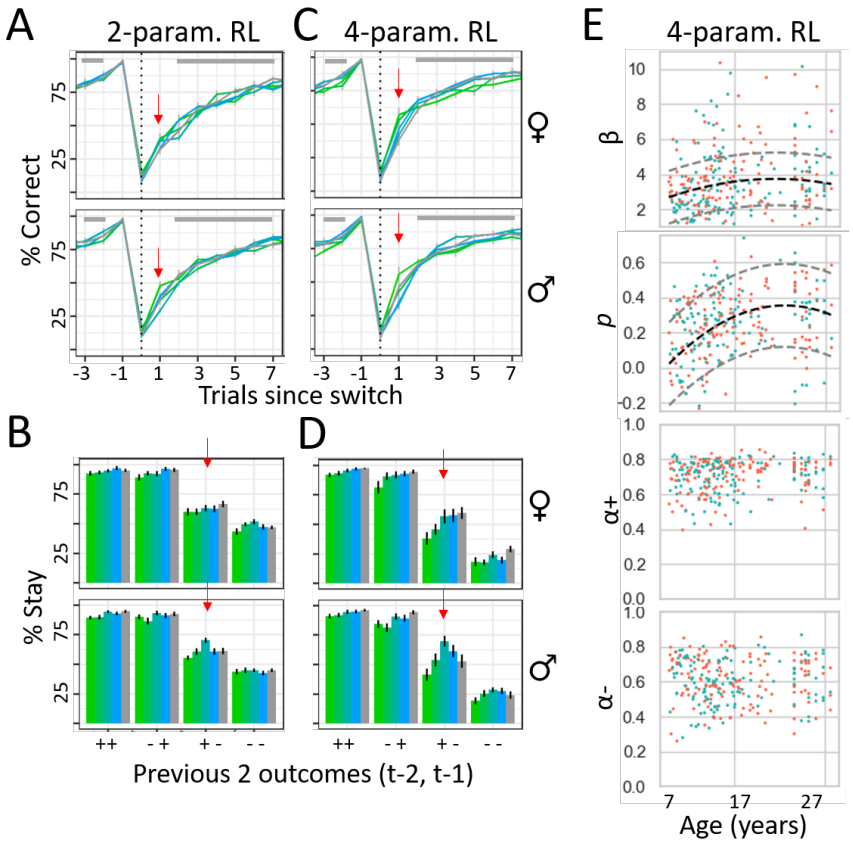


Figure 2.16: Qualitative fit of different versions of the RL model. Model behavior is shown in the same way as human behavior in suppl. Fig. 2.15. A-B) Behavior of simulations from the basic, 2-parameter version, with free parameters α and β. Lacking counter-factual updating and the ability to differentiate positive and negative outcomes, the model was unable to capture the shape of human learning curves and age differences. C-D) Behavior of simulations from the winning, 4-parameter version of the RL model, in which free parameters β, p, α_+, and α_- were fitted to participants using hierarchical Bayesian model fitting. To avoid double-dipping into age differences when visualizing the model, we fitted the model *without* access to participants' age (Methods). E) Fitted parameters of each individual, based on the same model. Dashed lines show age differences when significant (Table 2.10), based on the model with access to participants' age (Fig. 2.3B). This is the same data as summarized in Fig. 2.4A-D. Colors denote age groups, red arrow (potential) switch trials, and grey bars asymptotic trials, as in suppl. Fig. 2.15.

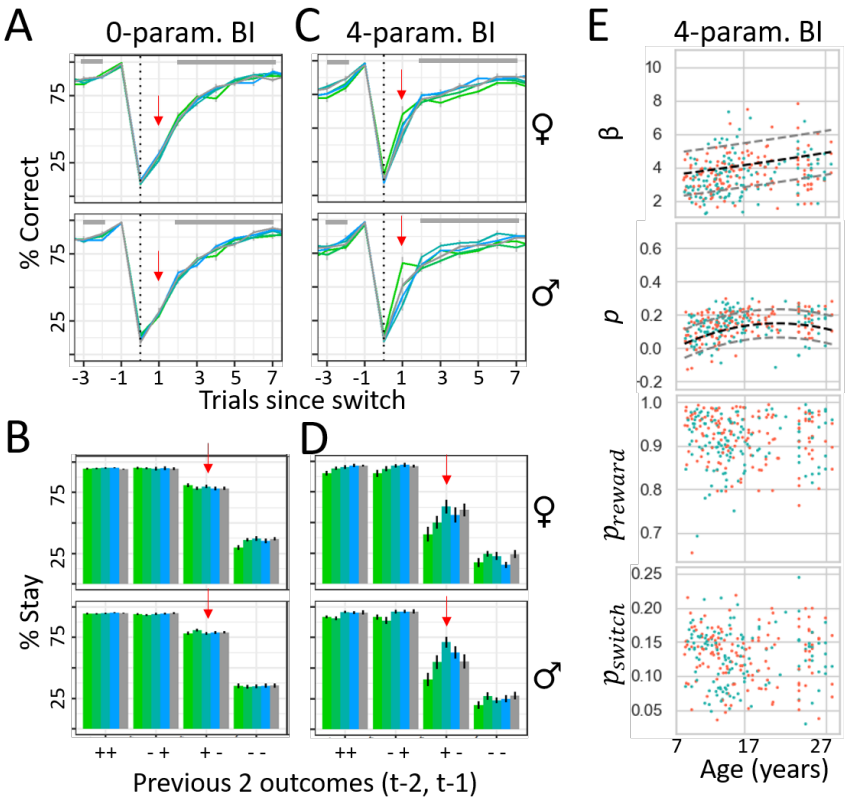


Figure 2.17: Qualitative fit of different versions of the BI model. Model behavior is shown in the same way as human behavior in suppl. Fig. 2.15. A-B) Behavior of simulations from the basic, 0-parameter version, in which truthfully $p_{reward} = 0.75$ and $p_{switch} = 0.05$. Lacking free parameters, the model predicted the same behavior for all participants, and was unable to capture age differences. C-D) Behavior of simulations from the winning, 4-parameter version of the BI model, in which free parameters β, p, p_{reward}, and p_{switch} were fitted to participants using hierarchical Bayesian model fitting. To avoid double-dipping into age differences when visualizing the model, we fitted the model *without* access to participants' age (Methods). E) Fitted parameters of each individual, based on the same model. Dashed lines show age differences when significant (Table 2.10), based on the model with access to participants' age (Fig. 2.3B). This is the same data as summarized in Fig. 2.4E-H.

Table 2.11: Parameter estimates and statistics from hierarchical model fitting. Significance tests against 0 for parameters whose ranges include 0, NA otherwise.

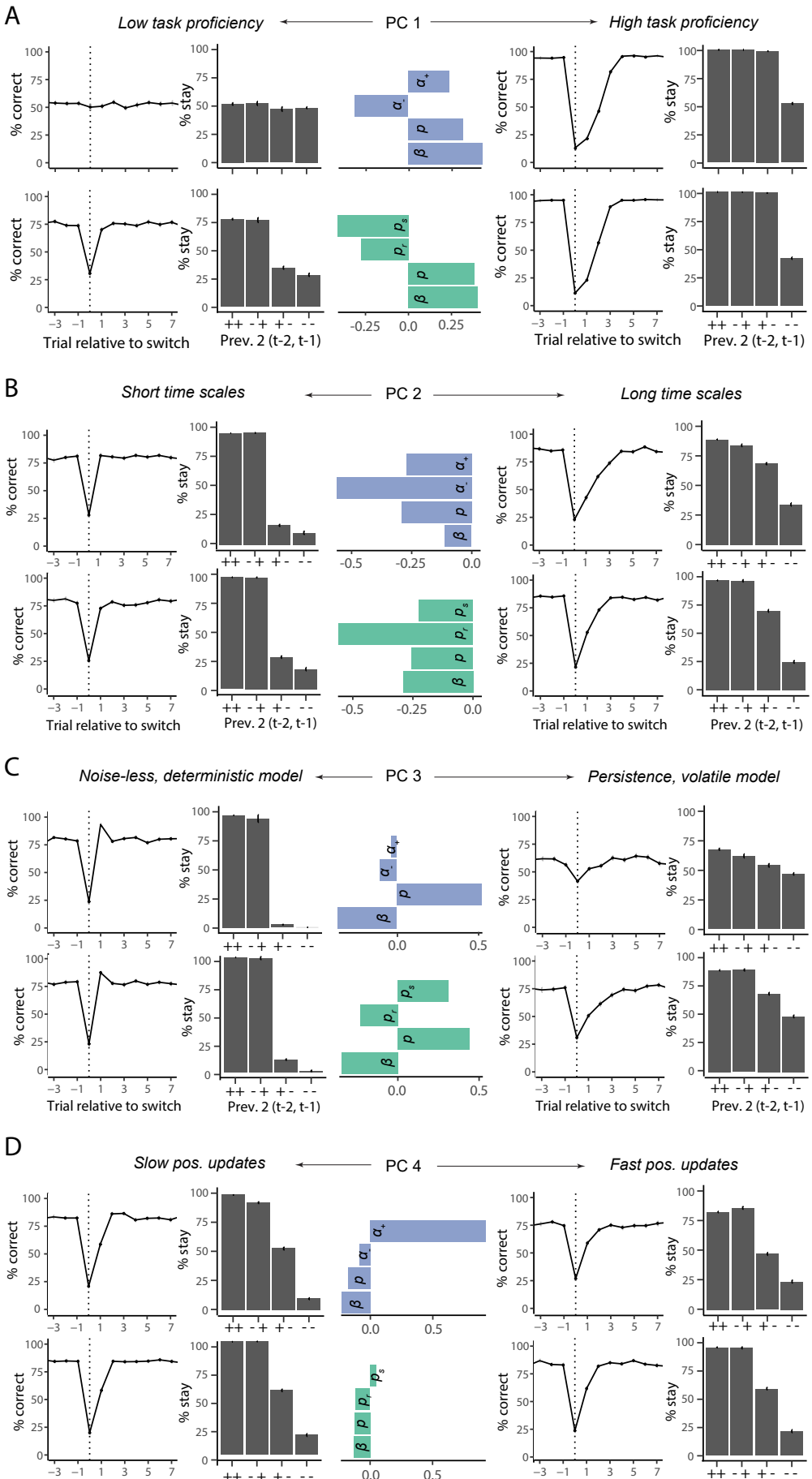
Model	Parameter	$\mu \pm sd$	95% CI	p-value	sig.
4-param. RL	p_{int}	0.34 ± 0.027	$[0.29, 0.39]$	< 0.001	***
	p_{sd}	0.24 ± 0.015	$[0.21, 0.26]$	0	NA
	p_{lin}	0.11 ± 0.020	$[0.075, 0.15]$	< 0.01	**
	p_{qua}	-0.050 ± 0.020	$[-0.089, -0.012]$	0.0051	**
	β_{int}	3.48 ± 0.15	$[3.18, 3.79]$	0	NA
	β_{sd}	1.48 ± 0.10	$[1.29, 1.69]$	0	NA
	β_{lin}	0.36 ± 0.11	$[0.14, 0.57]$	< 0.001	***
	β_{qua}	-0.22 ± 0.11	$[-0.42, -0.015]$	0.020	*
	$\alpha_{- \; int}$	0.60 ± 0.018	$[0.56, 0.63]$	0	NA
	$\alpha_{- \; sd}$	0.16 ± 0.0093	$[0.14, 0.18]$	0	NA
	$\alpha_{- \; lin}$	0.011 ± 0.015	$[-0.017, 0.040]$	0.77	
	$\alpha_{- \; qua}$	0.013 ± 0.014	$[-0.013, 0.040]$	0.84	
	$\alpha_{+ \; int}$	0.73 ± 0.034	$[0.66, 0.79]$	0	NA
	$\alpha_{+ \; sd}$	0.081 ± 0.021	$[0.042, 0.12]$	0	NA
	$\alpha_{+ \; lin}$	0.055 ± 0.024	$[0.0045, 0.10]$	0.015	*
	$\alpha_{+ \; qua}$	-0.015 ± 0.021	$[-0.055, 0.027]$	0.25	
4-param. BI	p_{int}	0.13 ± 0.013	$[0.11, 0.16]$	< 0.001	***
	p_{sd}	0.081 ± 0.0061	$[0.069, 0.093]$	0	NA
	p_{lin}	0.04 ± 0.008	$[0.023, 0.054]$	< 0.001	***
	p_{qua}	-0.02 ± 0.007	$[-0.038, -0.010]$	< 0.001	***
	β_{int}	4.27 ± 0.27	$[3.76, 4.83]$	0	NA
	β_{sd}	1.39 ± 0.12	$[1.16, 1.64]$	0	NA
	β_{lin}	0.39 ± 0.17	$[0.054, 0.72]$	0.011	*
	β_{qua}	$< 0.001 \pm 0.16$	$[-0.32, 0.30]$	0.49	
	$p_{reward \; int}$	0.87 ± 0.016	$[0.84, 0.91]$	0	NA
	$p_{reward \; sd}$	0.064 ± 0.0087	$[0.046, 0.081]$	0	NA
	$p_{reward \; lin}$	0.0045 ± 0.0096	$[-0.014, 0.024]$	0.68	
	$p_{reward \; qua}$	-0.0017 ± 0.0085	$[-0.018, 0.015]$	0.43	
	$p_{switch \; int}$	0.16 ± 0.014	$[0.14, 0.19]$	0	NA
	$p_{switch \; sd}$	0.071 ± 0.0053	$[0.062, 0.083]$	0	NA
	$p_{switch \; lin}$	-0.0066 ± 0.0095	$[-0.025, 0.012]$	0.24	
	$p_{switch \; qua}$	0.014 ± 0.0082	$[-0.0013, 0.030]$	0.042	*

Table 2.12: Parameter differences between specific age groups. p-values were obtained by assessing means for each parameter for three age groups (8-10, 13-15, and 18-30) and show in how many MCMC samples the group mean of 8-10 year olds (18-30 year olds) was smaller than the group mean of 13-15 year-olds.

Parameter	Compared groups	p-value	sig.
α_-	8-10 vs 13-15	0	***
	13-15 vs Adult	0.0045	**
p_{reward}	8-10 vs 13-15	0.019	*
	13-15 vs Adult	0.078	'
p_{switch}	8-10 vs 13-15	0.023	*
	13-15 vs Adult	0.13	

Using Model Simulations to Elucidate the Role of each PC

We simulated data from our computational models based on the obtained principal components (PCs) in order to visualize the role of each PC. It is common practice to simulate data based on small or large values of a parameter (e.g., small or large decision noise β) to assess the role of this parameter for model behavior (e.g., better or worse performance). We similarly simulated data based on small or large values of each PC to clarify the precise of each PC: We calculated two sets of parameters for each PC, one that represented high levels of this PC ("plus"), and one that represented low values ("minus"). Low levels were determined by subtracting 4 times the inverse-z-scored factor loading of a PC (suppl. Fig. 2.18, center) from the population mean of each parameter, low levels were determined by adding it. (For PC2 of the BI model, we added and subtracted 2 times the factor loading instead, to ensure $p_{reward} < 1$.) We then simulated behavior based on the resulting parameters to assess the effect of low versus high values of each PC (suppl. Fig. 2.18).


A
Low task proficiency
PC 1
High task proficiency
% correct
% stay
Trial relative to switch
Prev. 2 (t-2, t-1)
B
Short time scales
PC 2
Long time scales
C
Noise-less, deterministic model
PC 3
Persistence, volatile model
D
Slow pos. updates
PC 4
Fast pos. updates

Chapter 3

What can we Learn from Computational Modeling?

This chapter proposes a refined interpretation of computational model parameters. We show that some parameters are consistent across tasks (e.g., decision noise), but others are not (e.g., learning rates), and suggest that model parameters can capture different cognitive processes depending on the task in which they were measured. We recommend that model parameters always be interpreted with regard to task context. [1]

Abstract

Computational cognitive modeling has revolutionized the cognitive and brain sciences in many ways over the past decades. Computational studies create cognitive models that distill entire datasets of individuals' behavior into a small number of model parameters, which explain behavior succinctly, but without losing the ability to simulate original behavioral patterns in their full complexity. Parameters are often interpreted to reflect intrinsic individual characteristics that lay the foundation for more complex cognition, and to correspond to isolatable elements of brain function that underlies real-world behavior. To examine this assumption, we invited 291 participants between the ages of 8-30 years to complete three classic learning tasks in a single session, and fitted high-quality reinforcement learning (RL) models to each task. When we compared model parameters of the same individuals between tasks, some parameters (e.g., decision temperature β), but not others (e.g., learning rate α). were comparable between similar tasks. When we compared parameters across dissimilar tasks, no parameters were comparable, suggesting a lack of parameter generalizability when task context was not taken into account. Further analyses suggested the same parameters captured different cognitive processes in different tasks, suggesting a lack of parameter interpretability. Together, these results question basic assumptions of computational neuroscience –that model parameters are generalizable and interpretable across tasks– and highlight that indi-

[1] Research in this chapter was conducted together with co-authors Sarah L. Master, Liyu Xia, Ronald E. Dahl, Linda Wilbrecht, and Anne G.E. Collins.

vidual characteristics depend on specific contexts. Future work in computational neuroscience and the science of learning needs to take into account both short and also possibly long terms effects of subject/agent context.

3.1 Introduction

In the last decades, the cognitive neurosciences have made major strides in computational modeling, and demonstrated how central reinforcement learning (RL) may be in human behavior. RL models were first used to study relatively simple cognitive processes, including stimulus-outcome and stimulus-response learning (Gläscher et al., 2009; O'Doherty et al., 2004; Schultz et al., 1997), which were originally studied by behaviorists (Skinner, 1977; Watson, 1913). Strikingly, RL models have since also successfully explained far-reaching, goal-directed behavior with rich temporal structure (Daw et al., 2011; Momennejad et al., 2017; Ribas Fernandes et al., 2011). RL is even thought to lie at the heart of complex, abstract thinking that requires mental models with hierarchical structure (Botvinick, 2012; Collins and Koechlin, 2012; Eckstein and Collins, 2018), even during infancy (Werchan et al., 2015, 2016).

The current boom of RL research in machine learning and artificial intelligence (AI) has provided cognitive researchers with powerful algorithms and a strong mathematical foundation for the computational modeling of cognition, and many ideas have been transferred from AI to the cognitive sciences, and vice versa (e.g., Daw et al., 2011; Hassabis et al., 2017; Momennejad et al., 2017; Wang et al., 2018, for reviews, see Collins, 2019; Griffiths et al., 2019; Lake et al., 2017). The cognitive sciences have integrated many ideas from AI into models of human and animal cognition, including model-free and model-based algorithms, hierarchies over time scales, state spaces, and learning itself (i.e., meta-learning), temporal-difference algorithms, successor representation, etc. These RL models were fitted to human behavior across a wide range of tasks, including classic and operant conditioning, learning, decision making, problem solving, etc. In this sense, RL has emerged as a potentially unifying model of human cognition, explaining both basic cognitive processes and sophisticated problem solving, based on a compelling theoretical foundation, and with the promise to elucidate brain function: An extensive number of studies has provided evidence that a specialized network of brain regions, including the basal ganglia and prefrontal cortex, implements computations similar to those of RL algorithms, guiding choices based on action values, and updating action values based on reward prediction errors (for reviews, see Frank and Claus, 2006; Glimcher, 2011; D. Lee et al., 2012; Niv, 2009; O'Doherty et al., 2015; for a focus on development, see Bolenz et al., 2017; Nussenbaum and Hartley, 2019; van den Bos et al., 2017).

Despite this stunning progress, the precise meaning of RL models and their components (e.g., model parameters, reward prediction errors) often remains elusive. In the standard modeling approach, entire datasets of individuals' behavior are condensed into a small number of model parameters, using model fitting. The assumption is that computational models carve cognition at its joints, dissecting participants' mental processes into a small number of meaningful components, such as learning and decision making, and that the individually-fitted model parameters fully characterize the cognitive process using a few components relevant to brain and mind. Summariz-

ing complex behaviors using computational models is thought to provide a rigorous and succinct explanation of the behavior, while retaining the ability to reproduce the original behavior in its complexity. In other words, models are expected to divide the cognitive process in a way that isolates its fundamental elements and explains their intricate interplay, and to provide the few free parameters that thereby differentiate individuals from each other. Much of the appeal of cognitive modeling rests on the assumption that model parameters thereby reflect stable, interpretable, and generalizable individual characteristics: **Generalizability**, for our purposes, means that an individual's parameters measured in one context provide information that goes beyond that specific context. In other words, measuring an individual's parameters on one task is expected to provide information about their behavior in a (related but) different task, or –optimally– in the real world. **Interpretability** means that parameters reflect fundamental elements of cognition or neural processing that are intrinsic to a subject and stable across tasks, computational models, and real-world behavior. In other words, parameters are expected to reflect the same cognitive or neural variables when measured using different tasks, and these variables are expected to be intrinsic to participants. This study will focus on the question whether parameters generalize between tasks and capture the same cognitive processes across tasks.

Even though rarely stated explicitly, parameter generalizability and interpretability are at the heart of most computational (neuro)science research. The belief in parameter generalizability, for example, is evident in efforts to determine the fixed means and distributions of model parameters in a human population, with the goal of identifying generic parameter priors that can be used to jump start parameter estimation in future studies (Gershman, 2016). Using these empirical priors improved model fitting in similar tasks, and the hope is that "empirical priors can be potentially applied to a wide range of models and tasks that share similar parameterizations." Indeed, "the priors are robust across parameterizations, suggesting that they are fairly transferable". This conviction that parameters are stable individual traits that generalize across models and tasks is shared by many, if not most researchers in the field. Another example is the approach of computational neuropsychology, whose goal is an "understanding of the neural processes underlying decision-making in the normal and abnormal brain" (Niv, 2009), or more concretely, "how the brain learns to select actions to maximize future reward" (O'Doherty et al., 2015). "Fundamental for understanding brain function is to determine what computations are performed in neuronal populations that support a particular cognitive process", and the hope is that computational models succeed at "parcellating the computational mechanisms underlying cognition" (Hauser et al., 2019). A main focus of computational neuropsychology has been to locate specific RL computations within the brain, and one of its major successes is the association of reward-prediction errors with the midbrain-dopamine system (Schultz et al., 1997; Watabe-Uchida et al., 2017; for reviews, see Frank and Claus, 2006; Glimcher, 2011; D. Lee et al., 2012; Niv, 2009; O'Doherty et al., 2015). A specific network of brain regions, forming loops between the basal ganglia and the cortex, with inputs from midbrain dopamine, is assumed to implement neural processes that resemble RL computations (Frank and Claus, 2006; Glimcher, 2011; D. Lee et al., 2012; Niv, 2009; O'Doherty et al., 2015). Within this system, phasic dopamine signaling has been shown to relate to learning based on reward prediction errors (Frank et al., 2004; Steinberg, 2013), and the differential roles carried out by positive and negative learning rates in RL models are often ascribed to the characteristics

of striatum D1 and D2 dopamine receptors, expressed in the direct and indirect pathways, respectively (Collins and Frank, 2014; Tai et al., 2012; Verharen et al., 2019; for review, see Cox and Witten, 2019). These findings have thus linked model components to cognitive and brain processes in a specific task context. However, often implicit in this approach is the interpretability and generalizability of RL models: Generic model variables (e.g., reward-prediction error; positive learning rate) are taken to map precisely onto specific neural substrates (e.g., basal ganglia; D1 receptors), independent of the task in which they were measured.

Two more domains have focused on individual differences: computational psychiatry and developmental psychology. In computational psychiatry, "models are particularly useful as tools for measuring hidden variables and processes that are difficult or impossible to measure directly" (Huys et al., 2016), and they "show great promise in mapping latent decision-making processes onto dissociable neural substrates and clinical phenotypes" (Brown et al., 2020). Measuring hidden variables and mapping onto neural substrates and clinical phenotypes is what we define as interpretability, and the standard approach of comparing model parameters between studies relies on their between-task generalizability. Computational psychiatry has advanced our understanding of mental illnesses most notably including depression, schizophrenia, and Parkinson's (for reviews, see Adams et al., 2016; Hauser et al., 2019; Huys et al., 2016). However, results of different studies are often inconclusive, and many questions remain. In computational developmental research, the hope of modeling is to identify a small number of neurally-interpretable variables whose interplay can explain complex, and often non-linear trajectories of cognitive development: "Models can help to illuminate developmental change in cognitive processes or neural representations that are otherwise difficult to tease apart" (Nussenbaum and Hartley, 2019). Model parameters are often interpreted as characteristics of individuals that are short-time stable (weeks or months), but change gradually to explain development (years). Nevertheless, developmental studies so far have not been able to identify consistent age trajectories of model parameters. On the contrary, depending on the study, learning rate parameters have been found to increase (e.g., Davidow et al., 2016; Master et al., 2020), decrease (e.g., Decker et al., 2015), show U-shaped trajectories (Eckstein et al., 2020), or stay stable over a given age range (e.g., Palminteri et al., 2016; for a comprehensive review focused on these differences, see Nussenbaum and Hartley, 2019; for other reviews, see Bolenz et al., 2017; van den Bos and Hertwig, 2017). This apparent inconsistency contradicts the predominant view that generic RL parameters reflect what Nussenbaum and Hartley, 2019, call "static learning biases", i.e., that the same parameter reflects the same learning bias across studies, being unaffected by task demands. Similar inconsistencies in parameters have become apparent in the non-developmental modeling literature: Fitted parameter values often differ widely across studies, even when participant samples and tasks are comparable, suggesting that different tasks elicit different parameter values. Parameters also do not seem to be stable over time: Within the same task, participants exhibited different learning rates depending on whether the current context was stable or volatile (Behrens et al., 2007), and learning rates have even been shown to continuously adapt to task statistics (Cazé and van der Meer, 2013; Daw et al., 2006), suggesting that individuals' parameter values, as specified in simple RL models, are not fixed. This lack in parameter consistency, especially between differing tasks or task contexts, could be the consequence of a lack in parameter generalizability and interpretability.

The goal of the current project was to systematically investigate this possibility, assessing parameter generalizability and interpretability using three different tasks within the same participants, and employing state-of-the-art RL modeling. We hope to explain why previous research has often led to contradictory results, and offer an updated interpretation of model parameters that can resolve these inconsistencies. To achieve this, we asked 291 participants between the ages of 8 and 30 years to perform three different learning tasks. We fitted separate RL models to each task, conducting extensive model comparison and validation. The wide age range of participants led to a wide range in fitted parameter values, which allowed us to precisely characterize similarities and differences of parameters between tasks. The within-participant design allowed us to test directly whether the same participants showed the same parameters across tasks (generalizability), and the combination of multiple tasks allowed us to assess whether the same parameters captured the same cognitive processes (interpretability).

Before showing our results, we briefly introduce the RL models we used to fit human data (see section Computational Models for details). RL explains how agents (e.g., human, animal, artificial) adapt their behavior to their environment in order to maximize rewards and minimize punishment (Sutton and Barto, 2017). In a nutshell, agents learn a policy $\pi(a|s)$ that determines which action a to take in each state s of the world. In our models, this policy is based on the values of each action in each state $Q(a|s)$ (Fig. 3.1A). Agents learn values by paying attention to the outcomes of their actions at each time step t –desired, positive outcomes are called "rewards" r, and undesired, negative outcomes are called negative rewards. One simple learning method is to average past value estimates with new outcomes, so that over time, value estimates reflect the true reward contingencies: $Q_{t+1}(a|s) = Q_t(a|s) + \alpha \times (r_t - Q_t(a|s))$. How much a learner weighs past estimates compared to new outcomes is determined by parameter α, the learning rate. Small learning rates favor past experience and lead to stable learning over long time horizons, while large learning rates favor new outcomes and allow for faster and more flexible changes focusing on shorter time horizons. Different learning rates α_+ and α_- were used to distinguish learning from positive and negative rewards (e.g., Eckstein et al., 2020; Frank et al., 2004; Palminteri et al., 2016; van den Bos et al., 2012) Policy choices were made by translating action values $Q(a|s)$ into action probabilities $p(a|s)$ (see Fig. 3.1A and section Computational Models). How deterministically versus noisily this translation is executed was determined by exploration parameters β, also called inverse decision temperature, and/or ε, also called decision noise. Small decision temperatures $\frac{1}{\beta}$ favor the selection of only the highest-valued actions, enabling exploitation, whereas large decision temperatures select actions of all values equally likely, enabling exploration. Other model parameters included Forgetting, a decay of action values over time, and Persistence, a tendency to repeat the same action independent of outcomes (see section Computational Models for details). This way of constructing RL models is fairly standard and commonly-used in psychology and neuroscience.

In the following section, we will detail our experimental procedures and then answer two questions. Part I: Generalizability - Are parameters consistent within individuals? Part II: Interpretability - Do parameters reflect the same cognitive processes across tasks, and what are these processes? To foreshadow our results, Part I revealed that generalization differed between parameters: Decision noise parameters generalized well, especially between similar tasks, whereas positive learn-

ing rates showed some discrepancies, especially between dissimilar tasks. Negative learning rates failed to generalize in every way. Part II revealed that the cognitive processes captured by decision noise parameters were more consistent across tasks than those captured by learning rates. Learning rates captured overlapping, orthogonal, or even opposite processes, depending on the tasks. Nevertheless, even though cognitive processes differed, both decision noise parameters and learning rates captured consistent behavioral patterns across tasks. This suggests that computational models are consistent in a different way than commonly assumed: Rather than capturing the same cognitive process in each task, parameters might capture the same behavioral patterns, and behavioral patterns heavily depend on task demands. Just like we would not compare accuracy between a perception and a language task, we might not be able to compare learning rates between a stochastic and a deterministic task.

3.2 Results

Study Design

Our sample of 291 participants was balanced between females and males across the age range (8-30 years), and all ages were similarly represented (Fig. 3.1B, left). To reduce noise, we excluded participants based on performance criteria specific to each task (see section Participant Sample). Due to worse performance, more younger than older participants were excluded, which is an important caveat for the interpretation of age effects (Fig. 3.1B, right). Participants completed four computerized tasks, questionnaires, as well as a saliva sample during the 1-2 hour lab visit (Fig. 3.1C). Our tasks –called "Butterfly" (BF), "Probabilistic-Switching" (PS), and "Reinforcement learning and Working memory" (RL-WM)– were all classic reinforcement learning tasks: Participants made choices and received binary feedback in the form of point/win or no point/lose.

The tasks varied on several common experimental dimensions, including feedback stochasticity, number of available actions, memory demands (number of stimuli to learn about), and environmental volatility (Fig. 3.1D). For example, in two tasks (PS and BF), negative feedback was stochastic, such that most but not all incorrect actions led to negative outcomes, whereas in the third (RL-WM), negative feedback was deterministic, such that every incorrect action led to a negative outcome. A different set of two tasks (PS and RL-WM) provided diagnostic positive feedback, such that every positive outcome indicated a correct action, whereas in the third (BF), positive feedback was non-diagnostic, such that positive outcomes could indicate both correct and incorrect actions. Two tasks (BF and RL-WM) had larger memory demands, presenting several different stimuli/states for which correct actions had to be learned, whereas the third (PS) only presented a single state. More similarities and differences are summarized in Fig. 3.1D, and each task is described in more detail below, and in section Task Design. Overall, the BF task shared more similarities with both PS and RL-WM than either of these shared with each other. This allowed us to investigate whether task similarity played a role in parameter generalizability and interpretability.

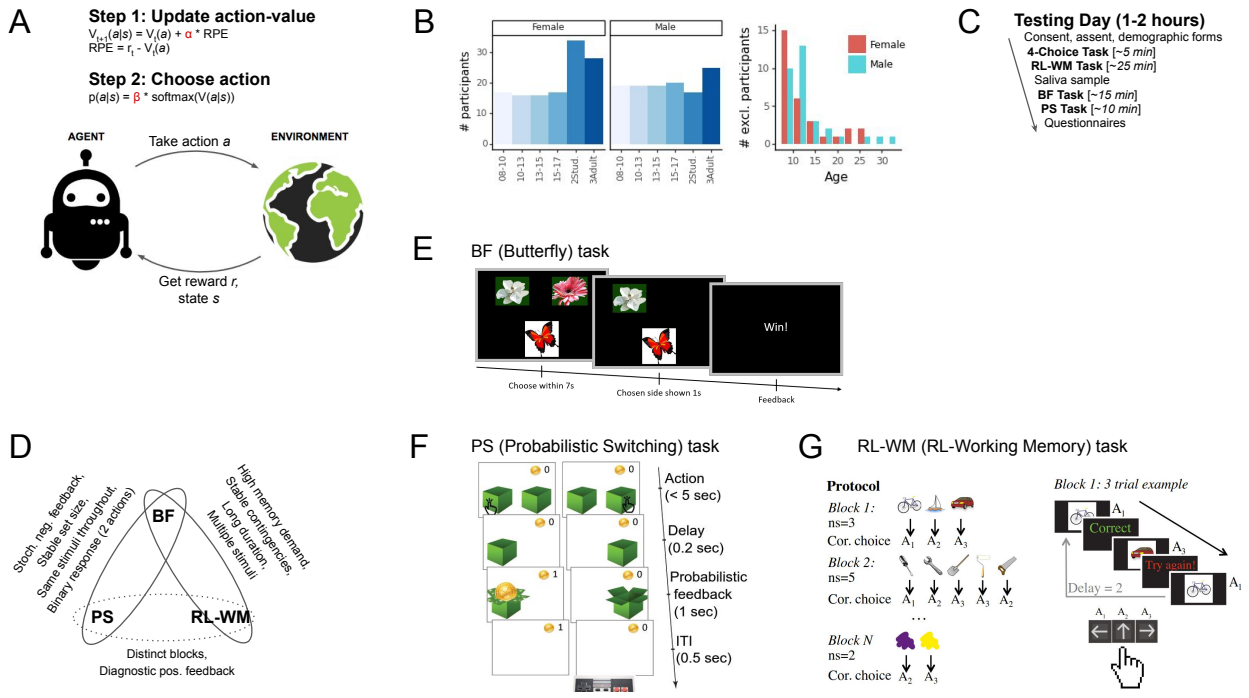


Figure 3.1: (A) Visual summary of the RL algorithm. An agent interacts with an environment by taking actions a in state s and receiving rewards r in return. The agent learns values $Q(s|a)$ for its actions based on the reward prediction error (RPE), and uses these values to calculate probabilities $p(a|s)$ for action selection. (B) Participant sample. Left: Final participant sample, broken up by age group (quartiles within each sex). Right: Histogram of age and sex of excluded participants. (C) Experimental procedure. Participants took part in a single 60-120 minute lab visit, during which they completed four experimental tasks, a saliva sample, and questionnaires. One task was excluded because too many participants failed to reach criteria. (D) The three remaining tasks shared similarities with each other, but also differed in important ways. Similarities between each pair of tasks are shown on the edges connecting the tasks. (E) Procedure of the Butterfly (BF) task. Participants saw one of four butterflies on each trial, and selected one of two flowers in response. Each butterfly had a stable preference for a specific flower throughout the task, but rewards were delivered stochastically (70% for correct responses, 30% for incorrect). (G) Procedure of the probabilistic switching (PS) task. Participants saw two boxes on each trial and selected one with the goal of finding gold coins. At each point in time, one box was correct and had a high (75%) probability of delivering a coin, whereas the other was incorrect (0%). At unpredictable intervals, the correct box switched sides. (G) Procedure of the "RL-WM" task. Participants saw one stimulus at a time and selected one of three responses. All correct responses and no incorrect responses were rewarded. Stimuli were presented in blocks containing 2-5 different stimuli. The number of stimuli in a block ("ns") is called set size . The task was designed to disentangle set-size sensitive working memory processes from set-size insensitive reinforcement learning processes.

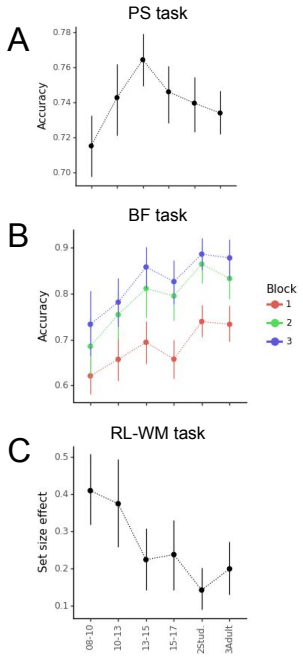


Figure 3.2: (A) Performance in the PS task increased markedly from early childhood (8-10 years) to mid-adolescence (13-15), and then decreased in late adolescence (15-17) and adulthood (18-30) (for details, refer to Eckstein et al., 2020). (B) Performance in the BF task increased with age and plateaued in early adulthood, as captured in decreases in decision temperature $\frac{1}{\beta}$ and increases in learning rate α (Xia et al., 2020). (C) The RL-WM task showed that the effect of set size on performance (regression coefficient) decreased with age, which was captured by increases in RL learning rate, but stable WM limitations (Master et al., 2020).

Each task was first analyzed independently, and detailed results have been published or submitted elsewhere (Eckstein et al., 2020; Master et al., 2020; Xia et al., 2020). We summarize here the key results. In the BF task, participants saw one of four butterflies on each trial, and aimed to pick the one of two flowers that was preferred by this butterfly. Each butterfly had a stable preference for one flower, and participants received a stochastic reward (80% probability) when they chose this flower. Nevertheless, sometimes the butterfly liked the opposite flower, and participants got a reward with 20% when they chose the opposite flower (Fig. 3.1E; see section Testing Procedure for details). The BF task has been used previously to investigate the role reward sensitivity and its interplay with episodic memory, shedding light on the neural substrate of these processes, notably the striatum and hippocampus, and revealing a unique role of adolescence in stochastic learning (Davidow et al., 2016). In our sample, performance on the BF task increased with age through the early-twenties and then stabilized (Xia et al., 2020; Fig. 3.2B). Using hierarchical Bayesian methods to fit RL models, we showed that this performance increase was driven by increasing positive

learning rate α_+ and decreasing decision noise $\frac{1}{\beta}$. Forgetting rates decreased very slightly with age, and negative learning rate α_- was 0, suggesting that participants ignored negative outcomes (see Fig. 3.3A and 3.3C for model parameters).

In the PS task, participants saw two boxes and selected one on each trial, with the goal of collecting gold coins. For some period of time, one box was correct and led to a stochastic reward (75% probability), while the other was unrewarded (0% probability). Then, the contingencies switched unpredictably and unsignaled, and the opposite box became the correct one. A 120-trials session contained 2-7 switches (Fig. 3.1F; see section Testing Procedure). The PS task was adapted from the rodent literature, where it has been used to show a causal link between stimulation of striatal spiny projection neurons and subsequent choices (Tai et al., 2012). Probabilistic switching tasks are also very common in the human literature (e.g., Cools et al., 2002; Cools et al., 2009; Dickstein et al., 2010; Peterson et al., 2009; Swainson et al., 2000; van der Schaaf et al., 2011; Waltz and Gold, 2007; for reviews, see Izquierdo et al., 2017; Lourenco and Casey, 2013) In our study, we found that human youth age 13-15 years markedly outperformed younger youth (8-12), older youth (16-17), and even young adults (18-30) on the PS task, suggesting that adolescent brains might be specifically adapted to perform well in stochastic and volatile environments. Computational modeling, using hierarchical Bayesian fitting, revealed that some model parameters (e.g,. decision temperature $\frac{1}{\beta}$, Persistence) increased monotonically from childhood to adulthood, whereas others (e.g., negative learning rate α_-, Bayesian inference parameters) showed pronounced U-shapes with peaks in 13-15 year-olds, similar to performance. Blending RL and a Bayesian inference models using principle component analysis (PCA) revealed that adolescents operated at a sweet spot that combined mature levels of task performance with child-like, short time scales of learning, and provided an explanation for adolescents' superior performance (Eckstein et al., 2020).

The RL-WM task was designed to dissociate the effects of RL and working memory, and has been used in diverse samples of adult participants (Collins, 2018; Collins, Albrecht, et al., 2017; Collins, Brown, et al., 2014; Collins, Ciullo, et al., 2017; Collins and Frank, 2012, 2017; McDougle and Collins, 2019). In the RL-WM task, participants see one stimulus at a time (e.g., bee) and choose one of three actions in response (left, up, right; Fig. 3.1G, right). Feedback is deterministic, i.e., reliably identifies each action as correct or incorrect. The goal of the RL-WM task is to learn the correct response for each stimulus. The key feature of the task is that stimuli appear in independent blocks of different sizes, ranging from 2-5 stimuli (e.g., the bee could be embedded in a block containing just 1 other animal, or up to 4 other animals). As set sizes increase, participants have been shown to shift the balance between using their capacity-limited, but reliable working memory system, to using their unlimited, but slower RL system (Collins and Frank, 2012). The RL-WM task estimates both memory systems, RL as well as working memory. The current study was the first to use the RL-WM task in youth. We found that participants aged 8-12 learned slower than participants aged 13-17, and were more sensitive to set size (Fig. 3.2C). Computational modeling revealed that developmental changes in RL were more protracted than changes in working memory: RL learning rate α_+ increased until age 18, whereas WM parameters showed weaker and more subtle changes early in adolescence (Master et al., 2020).

When fitting computational models to each task, we carefully verified that they captured participant behavior satisfactorily (see section Computational Models): For each task, we compared a large number of competing models, based on different parameters and cognitive mechanisms, and selected the best using model fit; we used hierarchical Bayesian methods for model fitting and comparison when possible, obtaining state-of-the-art parameter estimates for each individual (M. D. Lee, 2011). Crucially, we validated all models by simulating synthetic behavior based on the best model and human-fitted parameters to ensure that each model accurately reproduced human behavior and age differences in each task, and that parameters were identifiable (Wilson and Collins, 2019; refer to individual publications for details).

Part I: Parameter Generalizability

With the data of each task stemming from the same participants, we were now able to investigate how the computational models were related, and whether parameters reflected individual characteristics that were generalizable and interpretable across tasks. To investigate parameter generalizability, we first tested whether different tasks showed similar parameter values and whether parameters showed similar age trajectories across tasks.

Absolute Parameter Values

On the contrary, participants showed markedly different parameter values across tasks (Fig. 3.3A). In RL-WM, learning rates α_+ and α_- were close to 0, the lowest allowed value (mean α_+: 0.07; α_-: 0.03), whereas in PS, they were closer to 1, the highest allowed value (mean α_+: 0.77; α_-: 0.62). In BF, they were in the low-intermediate range (mean α_+: 0.22; α_- was best fitted at 0). To test these differences statistically, we conducted repeated-measures analyses of variance (ANOVA), conducting a separate model for each parameter, and predicting parameter values from task identity (PS, BF, or RL-WM). When the ANOVA showed a significant effect of task, we followed up with post-hoc repeated-measures t-tests, using the Bonferroni correction, to test differences between each pair of tasks.

For α_+, the ANOVA showed a significant effect of task on parameter values ($F(2) = 2.018$, $p < 0.001$), and t-tests revealed significant differences between all pairs of tasks (PS vs BF: $t(246) = 66$, $p < 0.001$; BF vs RL-WM: $t(246) = 12$, $p < 0.001$; RL-WM vs PS: $t(246) = 51$, $p < 0.001$). For α_-, both ANOVA ($F(1) = 2.357$, $p < 0.001$) and follow-up t-test were significant (RL-WM vs PS: $t(246) = 49$, $p < 0.001$). Suppl. Table 3.4 shows similar results when these analyses were controlled for age, using mixed-effects regression. These results confirm that the three learning tasks produced significantly different learning rate estimates for the same participants. Whereas RL-WM and BF estimated low learning rates, suggesting slow but consistent learning on long time scales, PS estimated high learning rates, suggesting fast learning and quick changes, with more weight on the most recent outcomes than on trial history, and a focus on short time scales. Interestingly, these patterns echo task differences: Whereas RL-WM and BF presented stable environments in which perfect performance can be obtained through slow, but consistent

updating, PS presented a volatile environment with frequent switches that required quick updating and increased flexibility.

For noise parameters ($\frac{1}{\beta}$ and ε), an ANOVA revealed a significant main effect of task ($F(2) = 830$, $p < 0.001$), and t-tests revealed that all pairwise differences were significant (PS vs BF: $t(246) = 25$, $p < 0.001$; BF vs RL-WM: $t(246) = 35$, $p < 0.001$; RL-WM vs PS: $t(246) = 32$, $p < 0.001$). Whereas PS suggested intermediate decision noise ($\frac{1}{\beta} = 0.33$, corresponding to 3.7-fold multiplication of action value differences), BF ($\frac{1}{\beta} = 0.095$, 10.6-fold multiplication of value differences) and RL-WM ($\varepsilon = 0.025$, corresponding to random choice on just 2.5% of all trials) suggested lower decision noise. One potential reason for differences in decision noise is the interdependence between learning rate and decision noise in typical RL models. According to this view, the observed differences in decision noise are epiphenomenon of differences in learning rates. Even though our model parameters were identifiable (see primary papers), making this explanation less likely, significant correlations were still present between learning rates and decision noise in some tasks (see suppl. Fig. 3.6). Another explanation are differing task demands: Volatile tasks like PS might require more exploration (decision noise) than stable tasks like BF and RL-WM, because they necessitate discovering a new and different correct response after task switches. Similarly, a task with deterministic feedback (RL-WM) necessitates less exploration than BF.

For the *Forgetting* parameter, task differences were significant in the ANOVA model ($F(1) = 161$, $p < 0.001$) and follow-up t-test (RL-WM vs BF: $t(246) = 49$, $p < 0.001$), revealing that more forgetting occurred in RL-WM than in BF. This might reflect the fact that participants were able to rely more on short-term working memory in RLWM, where associations were deterministic. Unlike previous research, these differences in parameters cannot be explained by differences in participant samples, testing procedures, or research labs, nor by a lack in modeling quality (Palminteri et al., 2017; Wilson and Collins, 2019). These results prove that the same participants can show different model parameters in different tasks.

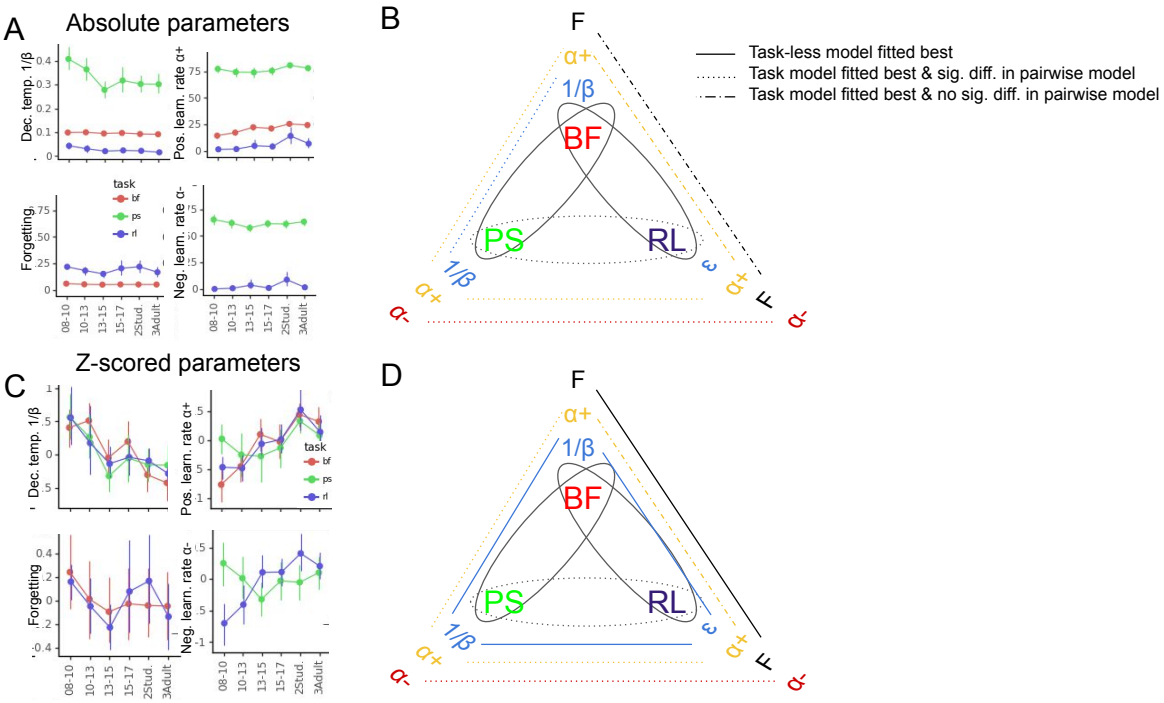


Figure 3.3: (A) Fitted parameters of each task (red: BF, green: PS, blue: RL-WM), plotted over participant age, averaged within quartile-based age groups. Dots indicate means, error bars specify the confidence level (0-1) for interval estimation of the population mean. Parameters differed substantially between tasks, with some parameters (e.g,. α_+, α_-) occupying opposite ends of the allowed spectrum of values. (B) and (C) Same as (A), but for within-task z-scored parameters, and within-task z-scored measures of behavior, respectively. (B) Age trajectories were consistent for decision-noise parameters, but not learning-rate parameters. (C) Some behavioral measures were consistent, while others were not.

Parameter Age Trajectories

Table 3.1: Statistics of mixed-effects regression models predicting z-scored parameter values from task (BF, PS, RL-WM), age, and squared age (months). The task-less grand model is reported when it had the best model fit ($\frac{1}{\beta}$ and Forgetting). Otherwise, pairwise follow-up models are shown (α_+ and α_-), whose p-values are corrected for multiple comparison using the Bonferroni correction. * $p < .05$; ** $p < .01$, *** $p < .001$.

Parameter	Tasks	Predictor	β	p (Bonf.)	sig.
α_+	PS & BF	Task (main effect)	2.75	0.003	**
		Task * linear age (interaction)	-0.28	0.006	**
		Task * quadratic age (interaction)	0.006	0.024	*
	PS & RL-WM	Task (main effect)	1.67	0.174	
		Task * linear age (interaction)	0.17	0.279	
		Task * quadratic age (interaction)	0.004	0.45	
	BF & RL-WM	Task (main effect)	1.08	0.51	
		Task * linear age (interaction)	-0.14	0.60	
		Task * quadratic age (interaction)	0.003	0.75	
$\frac{1}{\beta}/\varepsilon$	—	Intercept	1.86	< 0.001	***
		Age (linear)	-0.17	0.003	**
		Age (quadratic)	0.004	< 0.001	***
α_-	PS & RL-WM	Task (main effect)	4.15	< 0.001	***
		Task * linear age (interaction)	0.43	< 0.001	***
		Task * quadratic age (interaction)	-0.010	< 0.001	***
Forgetting	—	Intercept	0.37	0.44	
		Age (linear)	-0.034	0.53	
		Age (quadratic)	0.001	0.63	

Comparing absolute parameter values between tasks, as we just did, has important shortcomings. For example, if varying task demands lead to major differences in absolute values (as our results suggest), these differences could overshadow aspects of the parameters that were similar between tasks, such as the shapes of parameters' age trajectories, independent of absolute values. Indeed, a recent review compared not absolute parameter values, but parameter age trajectories between tasks when assessing the consistency of the developmental modeling literature to date (Nussenbaum and Hartley, 2019). We identified parameters' age trajectories by z-scoring each parameter within each task, such that means and variances were equated (Fig. 3.3B). Z-scored parameters (age trajectories; Fig. 3.3B) showed more consistent patterns than absolute values (Fig. 3.3A).

We tested for task differences in age trajectories using mixed-effects regression models that predicted z-scored parameter values from two age predictors (linear age and squared age), and with and without task as a predictor. When the model including task fit better than the model without task, then taking task differences into account increased the amount of explained variance beyond the increased complexity of adding task as a predictor, suggesting significant task differences. When this was the case, we conducted post-hoc comparisons between each pair of

tasks to determine which age trajectories differed, using regression models that predicted z-scored parameters from age (linear and squared) and task (focusing on two tasks at a time).

For α_+, the regression including task showed a slightly lower AIC score, indicating better model fit ($AIC_{with\ task} = 2,042$, $AIC_{without\ task} = 2,044$). Follow-up pairwise comparisons showed that z-scored values differed between PS and the other two tasks (RL-WM marginally; Fig. 3.3B; Table 3.1), even though overall age trajectories were qualitatively similar (linear effect of age: BF $\beta = 0.33$, $p < 0.001$, RL-WM $\beta = 0.28$, $p < 0.001$, PS $\beta = 0.052$, $p < 0.001$; quadratic effect: BF $\beta = -0.007$, $p < 0.001$, RL-WM $\beta = -0.006$, $p < 0.001$, PS $\beta = -0.001$, $p < 0.001$). Taken together, differences in α_+ age trajectories were more nuanced (Fig. 3.3B) than differences in absolute values (Fig. 3.3A), with similar trajectories for BF and RL-WM, but a differing trajectory for PS.

For α_-, including task as a predictor improved model fit ($AIC_{with\ task} = 1,373$, $AIC_{without\ task} = 1,395$), and z-scored parameter values differed significantly between PS and RL-WM in the follow-up regression (Table 3.1): In RL-WM, α_- showed a linear increase and inverse-U-shaped curvature (linear effect of age: $\beta = 0.32$, $p < 0.001$; quadratic: $\beta = -0.07$, $p < 0.001$). In PS, α_- showed the inverse pattern, a linear decrease and U-shaped curvature (linear: $\beta = -0.11$, $p < 0.001$; quadratic: $\beta = 0.003$, $p < 0.001$). Taken together, differences in α_- age trajectories (Fig. 3.3B) were as striking as differences in absolute values (Fig. 3.3A). RL-WM α_- increased monotonically, whereas PS α_- showed a U-shape with lowest point in 13-to-15-year-olds, two qualitatively different trajectories. Note that in BF, the best fitting model included a fixed $\alpha_- = 0$, again differing from the other two tasks.

For noise parameters, on the other hand, adding task as a predictor did not improve model fit ($AIC_{with\ task} = 2,054$, $AIC_{without\ task} = 2,044$), suggesting that age trajectories did not differ between tasks. The winning grand regression model revealed a linear decrease in decision noise across tasks, with a quadratic modulation that reflected slowing of the change with age (Fig. 3.3B; Table 3.1). This confirms that differences in scale (mean and variance; Fig. 3.3A) obscured similar age trajectories (Fig. 3.3B): a monotonic decrease over age, consistent with the previous literature (Nussenbaum and Hartley, 2019).

For parameter Forgetting, adding task (BF vs. RL-WM) as a predictor did not improve regression fit ($AIC_{with\ task} = 1,411$, $AIC_{without\ task} = 1,406$). The grand model did not reveal significant age effects (Fig. 3.3B; Table 3.1). Note that in PS, forgetting did not improve model fit, or equivalently, all participants had *forgetting*$= 0$, differing from the other two tasks. Together, age trajectories, which signify the conservation of participants' parameter values relative to each other, were similar for decision noise parameters and for positive learning rates in BF and RL-WM, but different for positive learning rates in PS, and strikingly different for negative learning rates (RL-WM and PS).

Predicting Age Trajectories

Table 3.2: Statistics of regression models predicting each z-scored parameter of one task (BF, PS, RL-WM) from the corresponding parameter of each of the other tasks. Because statistics were identical when predicting task A's parameter from task B's parameter and when doing the inverse, only one test is reported below.

Parameter	Tasks	β	p	sig.
α_+	BF, RL-WM	0.23	< 0.001	***
	BF, PS	0.13	0.035	*
	RL-WM, PS	-0.073	0.25	
$\frac{1}{\beta}$, ε	BF, RL-WM	0.19	0.0022	**
	BF, PS	0.28	< 0.001	***
	RL-WM, PS	0.039	0.54	
α_-	RL-WM, PS	-0.12	0.058	$
Forgetting	BF, RL-WM	0.097	0.13	

While parameter differences, assessed in sections Absolute Parameter Values and Parameter Age Trajectories, can reveal a lack of similarity, they cannot reveal its presence. We next tested directly whether parameters generalized between tasks, assessing how well age trajectories in one task predicted age trajectories of the same parameter in the other tasks.

For α_-, RL-WM and PS showed a marginal *negative* relationship (Table 3.2), suggesting that predicting this parameter in one task from the other would lead to *below*-chance predictions. For Forgetting, BF and RL-WM were not predictive of each other (Table 3.2). For both α_+ and noise parameters ($\frac{1}{\beta}$ and ε), z-scored parameters in BF predicted z-scored parameters in PS and RL-WM, and z-scored parameters in PS and RL-WM predicted z-scored parameters in BF. Nevertheless, z-scored parameters in PS and RL-WM did not predict each other (Table 3.2). This shows that predicting age trajectories across participants was only possible when task BF was involved. A potential explanation for this finding is that the BF task was more similar to both RL-WM and PS than they were to each other (Fig. 3.1D; see also section Relating Parameters and Behaviors Using Principal Component Analysis). In other words, similarity in task characteristics might be a determining factor in parameter generalization (see section Conclusion for a more detailed discussion of this exploratory result).

Summary Part I

In summary, Part I revealed that (1) participants showed strikingly different absolute values of noise parameters ($\frac{1}{\beta}$ and ε) and learning rates (α_+, α_-) across three learning tasks. Intriguingly, parameter values were in tune with task demands. (2) After equating parameter means and variances, age trajectories of noise parameters lacked differences, age trajectories of positive learning rates α_+ differed between one task and the two others, and age trajectories of negative learning rates α_- showed large qualitative differences. This suggests that differences in absolute parameter

values and parameter age trajectories are relatively independent from each other. (3) Age trajectories in noise parameters and α_+ were predictable from one task to another, as long as either the predicting or the predicted task was BF, the task that shared most similarity with each of the other tasks. This suggests that task similarity played a role in parameter consistency.

Part II: Parameter Interpretability

Part II investigates parameter interpretability, i.e., whether the same parameters captured the same cognitive processes in each task. This question is crucial for the interpretation of parameters and cognitive models. For example, when the same parameters capture different cognitive processes depending on the task, findings from different tasks cannot be compared easily, and without a mapping between parameters and cognitive processes, it is more difficult to relate findings to previous work.

Relating Parameters and Behaviors Using Principal Component Analysis

Principal component analysis (PCA) is a statistical tool that decomposes the variance of a dataset into so-called "principal components" (PCs). PCs are linear combinations of a dataset's original features (e.g., response times, accuracy, learning rate), and explain the same variance in the dataset as these original features. The advantage of PCs compared to original features is that they are orthogonal to each other and therefore capture independent aspects of the data. In addition, subsequent PCs explain subsequently less variance, such that the top PCs explain the bulk of a dataset's variance and are able to reconstruct the entire dataset, up to small details and random noise. The goal is then to understand what the top PCs capture, and "factor loadings" contain this information, the original features' weights on each PC (see section Principal Component Analysis (PCA) for details).

Having focused on individual parameters in Part I, Part II integrates different parameters as well as behavioral measures. First, we conducted a PCA on all 54 features of our dataset (39 behavioral and 15 model parameters). For more information about each behavioral feature, see section Understanding Parameters Based On Behaviors The first principal component (PC0), capturing the largest proportion of variance in the dataset (25.1%; Fig. 3.4A), reflected task performance, broadly defined: Behaviors that indicated poor task performance loaded negatively (e.g., number of missed trials, response time, response time variability), whereas behaviors that indicated good task performance loaded positively (e.g., mean accuracy, win-stay choices; Fig. 3.4B, top row). This shows that the largest source of variation in our dataset were individual differences in task performance.

To facilitate the interpretation of subsequent PCs, we first equated the role of each feature with respect to task performance, by flipping the signs of all features that played a negative role for task performance, as indicated by a negative weight on PC0. This step ensured that the directions of factor loadings on PC1 and PC2 were interpretable in the same way for all features, irrespective of their role for task performance in PC0. Thus preprocessed, loadings showed that the next two principal components (PC1: 8.9% explained variance; PC2: 6.2%) encoded task contrasts: PC1

contrasted PS to RL-WM, with positive factor loadings on PS features, and negative ones on the corresponding RL-WM features, while BF features had near-zero loadings (Fig. 3.4B, middle row). PC2 contrasted BF to PS and RL-WM, with positive loadings on BF features and negative ones on the corresponding RL-WM and BF features (Fig. 3.4B, bottom row). This shows that, after accounting for individual differences in performance (PC0), the next-most variance in our dataset arose from differences between tasks. PS and RL-WM showed the greatest differences, followed by differences between BF and both other tasks. This pattern is in accordance with similarities in terms of task features (Fig. 3.1D; note that missed trials and response times did not follow the task contrasts in PC1-PC2, suggesting that these features did not differentiate tasks.)

Aiming to elucidate the roles of parameters, we next assessed parameters' factor loadings. In PC0, all noise and Forgetting parameters loaded negatively and all α_+'s loaded positively, suggesting that across tasks, noise and Forgetting parameters affected performance negatively, whereas α_+ affected performance positively (Fig. 3.4B, top row), in accordance with these parameters' general roles in RL models (Sutton and Barto, 2017).

α_-, on the other hand, loaded positively on RL-WM but negatively on PS, suggesting that increased learning from negative feedback improved performance in RL-WM, but reduced performance in PS (Fig. 3.4B, top row). Like in Part I, this opposing pattern is in accordance with tasks demands: In RL-WM, negative feedback was diagnostic, such that an optimal strategy would use a maximum negative learning rate α_- to learn from every feedback. In PS, on the other hand, negative feedback was non-diagnostic, such that an optimal strategy needs to integrate several outcomes over a longer time horizon, using a lower negative learning rate. This result confirms patterns in Part I that suggested that α_- played very different roles in RL-WM compared to PS, reflecting each task's unique demands.

On PC1 and PC2, noise parameters, α_+, and α_- encoded the task contrasts described above, loading positively on PS but negatively on RL-WM (PC1), and positively on BF but negatively on RL-WM and PS (PC2; Fig. 3.4B, middle and bottom rows). This shows that these parameters differed sufficiently between tasks to be discriminable (as opposed to, e.g., response times and the number of missed trials, which did not show task contrasts). It also reveals that parameters contained sufficient task-specific characteristics to make it possible to associate them with the correct task. This degree of dissociation would not be expected if parameters captured the same cognitive mechanisms in each task.

Taken together, (1) The largest amount of variance arose from individual differences in task performance, broadly defined. (2) the contrast between PS and RL-WM explained more variance than the contrast between BF and both other tasks, confirming the BF task was more similar to both PS and RL-WM than these were to each other, in accordance with similarities in task characteristics (Fig. 3.1D). (3) Decision noise and Forgetting parameters were negatively associated with task performance, whereas learning rate α_+ was associated positively. Learning rate α_- showed opposite signs in both task, positive in RL-WM but negative in PS, in accordance with which setting was optimal for each. (4) Decision noise parameters and learning rates contained enough task-specific variance to be identified with each task, suggesting that the same parameters captured different aspects of cognition across tasks.

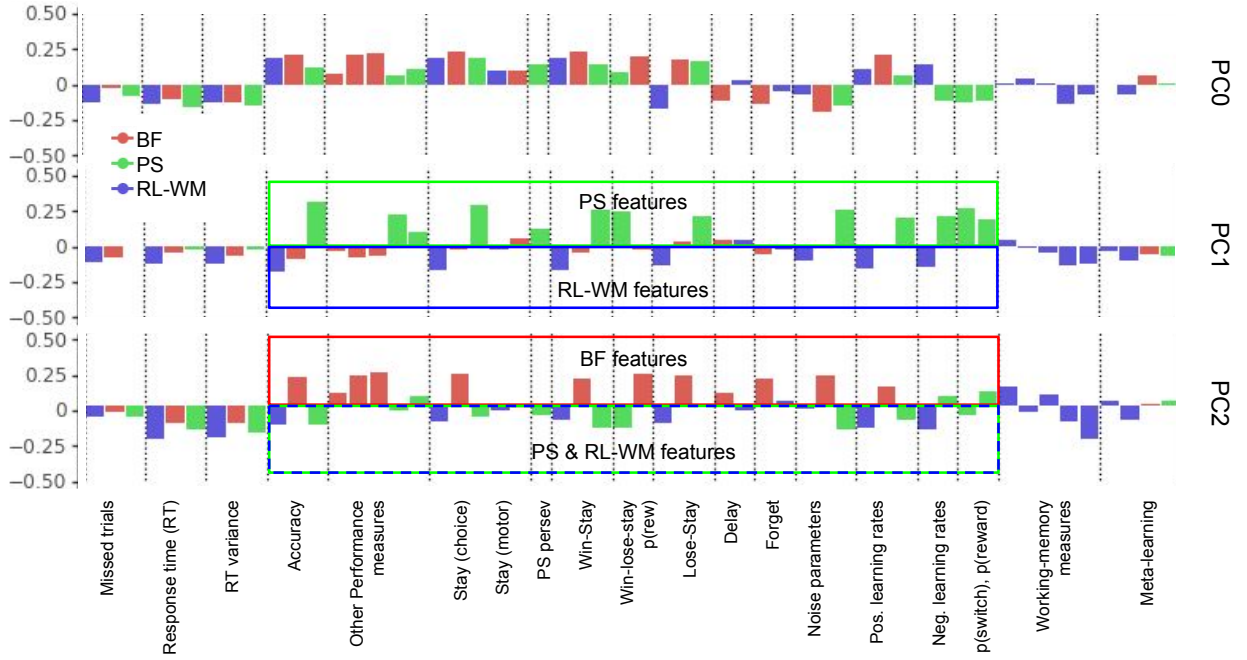


Figure 3.4: PCA of all 54 features of the dataset (39 behavioral; 15 model parameters). Factor loadings (y-axis) of each feature (x-axis) for the first three PCs (rows) of our dataset. PC0 (top row) captured broadly-defined task performance. PC1 and PC2 were re-oriented according to each feature's role for task performance to make loading directions comparable between features with opposing effects on performance. PC1 encoded a contrast between PS and RL-WM, and PC2 between BF and both other tasks.

Do Parameters Capture Overlapping Cognitive Processes?

Differences between the three learning tasks (Fig. 3.1D) likely led to the recruitment of different cognitive processes. The BF task was the most classic RL task, requiring the gradual integration of stochastic feedback over time in order to learn the correct responses for multiple stimuli (Fig. 3.1E). Similarly, the PS task required the gradual integration of stochastic negative feedback, but it not require gradual integration of positive feedback because it was diagnostic. The PS task's main challenge was to infer when a switch occurred, introducing a component of inference that was not central in the BF task (Fig. 3.1F; Eckstein et al., 2020). In the RL-WM task, each block was structurally similar to the BF task and likely relied on similar RL processes. Nevertheless, the RL-WM task also motivated the use of working memory because (1) feedback was deterministic, i.e., perfectly diagnostic, and (2) some blocks had very small set sizes, which favors the use of working memory (Fig. 3.1G; Master et al., 2020). A commonality between all tasks was the need for choice randomness or exploration to find the correct responses. Taken together, all three tasks likely required the integration of feedback over time as well as exploration, but only BF relied on

these processes principally, whereas PS and RL-WM likely also relied on inference and working memory, respectively.

To interpret model parameters in the same way in each task, they need to represent the same cognitive mechanism in each (e.g., α_+ could capture integration of positive feedback over time in each task). An alternative hypothesis is that parameters represented different processes depending on the task (e.g., α_+ could capture integration over time in BF, but reasoning/inference in PS). To test these possibilities, we first investigated whether models of different tasks captured similar variance, suggesting overlapping cognitive mechanisms. Specifically, we probed how much of each parameter's variance was explained by the parameters of a different model, using regression. When significant amounts of variance were explained, we next asked which parameters played the largest role in explaining that variance. If parameters reflected similar cognitive mechanisms across tasks, then corresponding parameters should show the largest regression coefficients (e.g., PS $\frac{1}{\beta}$ could show a significant coefficient when predicting BF $\frac{1}{\beta}$), but if parameters reflected different mechanisms, then different parameters should show significant weights (e.g., RL-WM ρ could show a significant coefficient when predicting BF α_+).

Because models contained several predicting features, regular regression models led to overfitting and were not interpretable. To avoid overfitting, we therefore used repeated, k-fold cross-validated Ridge regression (see section Ridge Regression). This method provides unbiased estimates for explained variance R^2 and regression coefficients w, and allows for statistical comparison based on bootstrapping. Indeed, for each parameter, similar variance was explained when using fitted parameters for prediction as when using raw behavioral features, confirming that computational models compressed behavior with minimal information loss (compare Fig. 3.5A and suppl. Fig. 3.8).

We found that of all parameters, most variance was predicted for two BF parameters: α_+ (using PS parameters for prediction: mean $R^2 = 18.1\%$, $sd = 0.3\%$, $p = 0$; RL-WM parameters: mean $R^2 = 12.8\%$, $sd = 0.3\%$, $p = 0$) and $\frac{1}{\beta}$ (PS: mean $R^2 = 10.2\%$, $sd = 0.2\%$, $p = 0$; RL-WM: mean $R^2 = 6.5\%$, $sd = 0.2\%$, $p = 0$; Fig. 3.5A, left panel). Mean R^2, standard deviations sd, and p-values representing $p(R^2 < 0)$, are based on 1,000 bootstrapping iterations (see section Ridge Regression for details). The fact that of all parameters, BF α_+ and BF $\frac{1}{\beta}$ were predicted best is in accordance with the BF task being most similar to each of the other two tasks, with more differences between these two (Fig. 3.1D). It is also in accordance with roles of BF $\frac{1}{\beta}$ and BF α_+ in cognitive processes that are shared between tasks, likely including exploration and integration of feedback over time, respectively. Combining PS and RL-WM parameters in a single regression model explained more variance than each task did on its own (α_+: mean $R^2 = 24.3\%$, $sd = 0.2\%$, $p = 0$; $\frac{1}{\beta}$: mean $R^2 = 14.6\%$, $sd = 0.2\%$, $p = 0$; Fig. 3.5A, left panel), indicating that the PS and RL-WM models captured partly non-overlapping cognitive processes: If both tasks captured the exact same processes, combining them would not explain more variance than using just one.

We next focused on noise parameters, examining whether they captured the same cognitive processes across tasks, by analyzing regression weights. When predicting BF $\frac{1}{\beta}$, both PS $\frac{1}{\beta}$ and RL-WM ε showed significant regression coefficients, revealing consistency (PS: $w = 0.14$, $p = 0.031$; RL-WM: $w = 0.12$, $p = 0.038$; Fig. 3.5B, top). Nevertheless, the inverse was not true, and BF $\frac{1}{\beta}$

did not show significant weights when predicting PS $\frac{1}{\beta}$ or RL-WM ε (PS $\frac{1}{\beta}$: $w = 0.09$, $p = 0.27$; RL-WM ε: $w = 0.04$, $p = 0.63$; Fig. 3.5B, bottom left and bottom right). The first result suggests that PS and RL-WM noise parameters captured cognitive processes that were also captured by BF $\frac{1}{\beta}$, and the second suggests that PS and RL-WM noise parameters also captured cognitive processes that were not captured by BF $\frac{1}{\beta}$. In addition, when predicting BF $\frac{1}{\beta}$, significant regression coefficients were obtained by PS parameters Persistence ($w = -0.19$, $p = 0.0029$) and α_- ($w = 0.14$, $p = 0.032$), and by RL-WM parameters α_- ($w = -0.18$, $p = 0.045$) and working-memory weight ρ ($w = -0.19$, $p = 0.023$; Fig. 3.5B, top). This shows that the matching between noise parameters was not perfect between tasks, and suggests that non-noise parameters PS Persistence, PS α_-, RL-WM α_-, and RL-WM ρ captured aspects of noise, and/or that BF $\frac{1}{\beta}$ captured non-noise processes.

We next focused on learning rates and found that several parameters, including learning-rate and non-learning rate parameters, captured overlapping variance. BF α_+ predicted RL-WM parameters α_+ ($w = 0.20$, $p = 0.013$) and α_- ($w = 0.24$, $p = 0.0022$), revealing consistency within learning-rate parameters. In the inverse model, however, BF α_+ was predicted by both RL-WM α_- ($w = 0.22$, $p = 0.011$) and working-memory weight ρ ($w = 0.16$, $p = 0.050$) and working-memory capacity κ ($w = 0.15$, $p = 0.020$; Fig. 3.5B, top right). In other words, the variance captured by RL-WM's RL parameters (α_+, α_-) was only captured by BF α_+, but the variance captured by BF α_+ was captured by both RL-WM's RL (α_-) and working-memory parameters (ρ, κ). This suggests that RL-WM's RL parameters captured only RL processes, while BF's α_+ captured both RL and working-memory processes, in accordance with previous research (Collins and Frank, 2012).

BF α_+ was furthermore predicted by most PS parameters ($\frac{1}{\beta}$: $w = -0.19$, $p = 0.0026$; α_-: $w = -0.21$, $p < 0.001$; Persistence: $w = 0.23$, $p < 0.001$), with the notable exception of its direct counterpart PS α_+ ($w = 0.0042$, $p = 0.94$, n.s.; Fig. 3.5B, top left). Similarly in the inverse models, PS α_- was predicted by BF α_+ ($w = -0.25$, $p = 0.0018$), showing a notable negative relationship, and PS α_+ was not predicted by any BF parameter ($\frac{1}{\beta}$: $w = -0.077$, $p = 0.37$; α_+: $w = 0.058$, $p = 0.48$; Forgetting: $w = 0.015$, $p = 0.82$). Taken together, this suggests that the processes reflected by BF α_+ were captured by an interplay between multiple PS parameters, rather than any single one. Notably, PS α_+ shared no overlap with BF α_+, being completely orthogonal, and PS α_- was negatively related to BF α_+.

Highlighting further differences, BF α_- was 0 for all participants, as opposed to PS α_- and RL-WM α_-, which showed pronounced age trajectories (Fig. 3.3A). Variance in PS α_- was thereby unexplained by RL-WM parameters (mean $R^2 = -0.0205$, $sd = 0.0024$, $p = 1$), suggesting that there was no overlap in the cognitive processes captured by PS α_- and the entire RL-WM model, including RL-WM's own α_-. Even though PS α_- and RL-WM α_- were similar in that they both significantly predicted BF α_+, they exhibited opposite signs (PS: $w = -0.19$, $p < 0.001$; RL-WM: $w = 0.22$, $p = 0.011$; similar pattern for BF $\frac{1}{\beta}$, PS: $w = 0.13$, $p = 0.023$; RL-WM: $w = -0.18$, $p = 0.045$). This confirms that α_- played opposing roles in PS compared to RL-WM, as suggested by previous analyses. The unique role of PS α_+ was highlighted by the fact that it was impossible to predict significant variance in this parameter using any other parameters (BF: $R^2 = -0.0080$, $sd = 0.0037$, $p = 1$; RL-WM: $R^2 = -0.019$, $sd = 0.0027$, $p = 1$; Fig. 3.5A, middle). This shows

that PS α_+ reflected unique cognitive processes, which were not captured by RL-WM or BF, potentially inference.

We next focused on parameters that were not shared between tasks, notably RL-WM's working-memory parameters. No variance in these parameters was explained by any other task (working-memory weight ρ: PS mean $R^2 = -0.017$, $sd = 0.0028$, $p = 1$; BF mean $R^2 = -0.014$, $sd = 0.0031$, $p = 1$; capacity κ: PS mean $R^2 = -0.015$, $sd = 0.0015$, $p = 1$; BF mean $R^2 = -0.020$, $sd = 0.0015$, $p = 1$; Forgetting: PS mean $R^2 = -0.017$, $sd = 0.0024$, $p = 1$; BF mean $R^2 = -0.010$, $sd = 0.0015$, $p = 1$). Significant variance, on the other hand, was explained in RL-WM's RL parameters (α_+: PS mean $R^2 = 0.014$, $sd = 0.0033$, $p = 0$; BF mean $R^2 = 0.020$, $sd = 0.0020$, $p = 0$; α_-: PS mean $R^2 = 0.057$, $sd = 0.0036$, $p = 0$; BF mean $R^2 = 0.076$, $sd = 0.0045$, $p = 0$). This suggests that the RL-WM model's working-memory parameters captured largely unique cognitive processes, while its RL parameters captured overlapping processes, and supports the notion that the RL-WM model successfully disentangled working-memory and reinforcement-learning processes.

Summarizing this section, noise parameters showed many similarities across tasks, learning-rate parameters showed fewer similarities, and working-memory parameters appeared relatively unique. PS and RL-WM noise parameters captured much of the cognitive processes captured by BF $\frac{1}{\beta}$, but they also captured additional processes beyond BF $\frac{1}{\beta}$, and of the processes shared with BF $\frac{1}{\beta}$, both captured different aspects. This reveals substantial overlap in the noise parameters between tasks, but also highlights important differences. Learning-rate parameters showed some consistencies and many differences between tasks, and revealed substantial overlap with non-learning rate parameters, most notably RL-WM's working-memory parameters. BF α_+ shared variance with both RL-WM's RL (α_+ and α_-) and working-memory parameters (ρ, κ), and also with all PS parameters *except* its counterpart PS α_+. PS α_+ was entirely orthogonal to all other model parameters, and α_- showed inverse roles in RL-WM compared to PS. In sum, learning-rate parameters did not capture unitary variance across tasks, but showed marked differences.

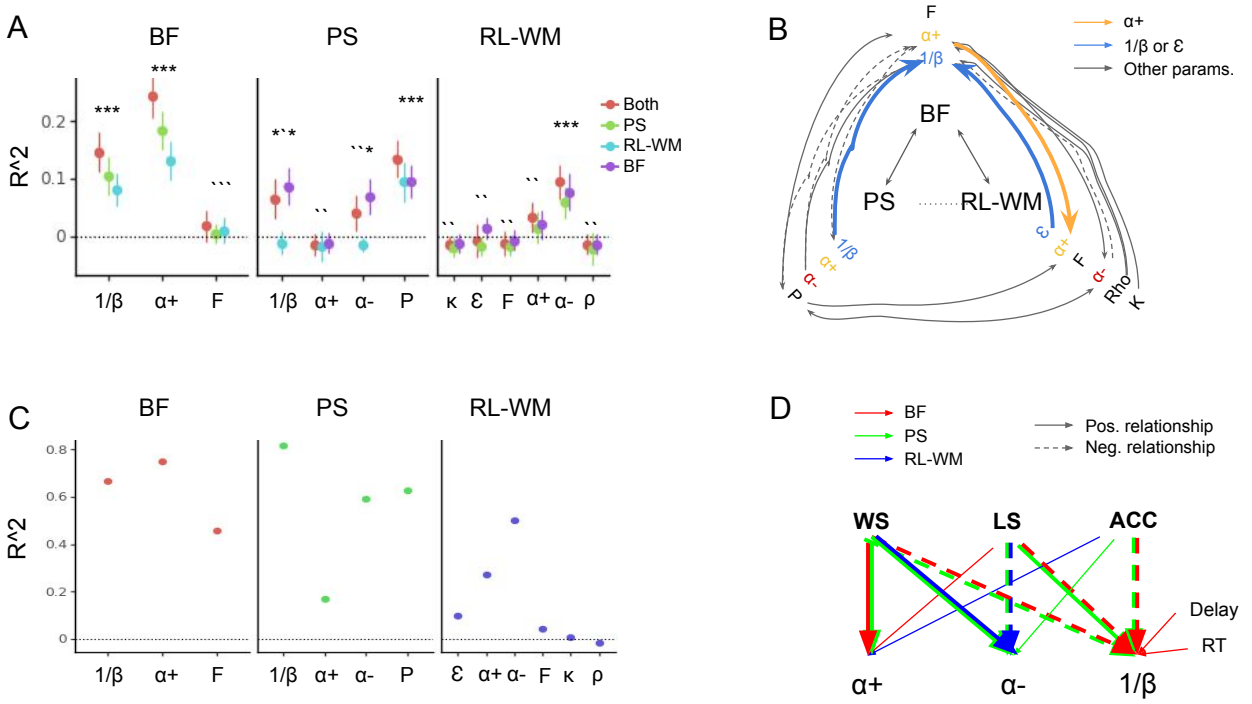


Figure 3.5: (A) Percentage of explained variance (R^2) when predicting each model parameter (x-axis) from the parameters of the other models. Each parameter was predicted using three different regression models: One containing the parameters of both other tasks (red), and two containing the parameters of just one other task (green: PS parameters; blue: RL-WM parameters; purple: BF parameters). (B) Summary of regression coefficients of the models in part A. Each arrow denotes a significant regression coefficient ($p < 0.05$). Arrow direction indicates direction of prediction (e.g., an arrow pointing from a PS parameter to a BF parameter denotes a significant coefficient when predicting the BF parameter using the PS model). Colored fat arrows show the only cases in which corresponding parameters predict each other, i.e., in which predicting and predicted parameters are identical. Dashed lines indicate negative and full lines positive coefficients. (C) Explained variance (R^2) when predicting model parameters by behavioral features of all three tasks. (D) Summary of regression coefficients of the models in part B. Each arrow signifies a significant regression coefficient ($p < 0.05$) when predicting a model parameter (bottom row) from behaviors (top row) of all three tasks. All significant within-task relationships are shown; arrow color indicates the task of the behavior and parameter it connects. Fat arrows indicate that arrows are duplicated across tasks, i.e., task-based consistency.

Understanding Parameters Based On Behaviors

Our final set of analyses focuses on why parameters showed different developmental trajectories (Fig. 3.3A and B) and captured different cognitive processes (Fig. 3.5A and B) depending on the

task. One possibility is that computational modeling failed at isolating specific cognitive processes, for example due to model misspecification (Nussenbaum and Hartley, 2019). In this interpretation, model parameters did not capture the meaningful patterns intended by modelers, and the promise of RL fell short of its high expectations (Sutton and Barto, 2017). Another explanation is that parameters captured intended patterns of behavior, but these patterns relied on different cognitive processes depending on the task. For example, learning rates might always capture the behavioral response to feedback, but what this response looks like relies on different cognitive mechanisms in different tasks (e.g., depending on whether feedback is diagnostic or not).

To investigate these possibilities, we assessed the relationships between model parameters and participant behavior. Using regularized Ridge regression like above, we predicted each model parameter from five selected behavioral features of each task (features are detailed below). When two parameters capture identical cognitive processes, they will be predicted by the exact same behavioral features in this analysis (e.g., all noise parameters might be predicted by BF accuracy). We will call this case "absolute consistency". When two parameters capture orthogonal processes, they will be predicted by different, non-overlapping features (e.g., Forget might be predicted by accuracy in one task, but by Delay in another). We will call this "absolute inconsistency". Crucially, there also is a third case: When two parameters capture the same behavioral patterns in each task, but these patterns differ between tasks, then the parameters will be predicted by corresponding features in both tasks (e.g., each learning rate would be predicted by accuracy in its own task). We will call this case "task-based consistency", which means that parameters capture the same behavioral patterns in each task, and differences between parameters arise from differences in behavioral patterns between tasks.

Table 3.3: Statistics of mixed-effects regression models predicting z-scored behavioral features from task (BF, PS, RL-WM), age, and squared age (months). The task-less grand model is reported when it had the best model fit (win-stay, Delay). Otherwise, pairwise follow-up models are shown (RT, ACC, lose-stay), with p-values corrected for multiple comparison using the Bonferroni correction. * $p < .05$; ** $p < .01$, *** $p < .001$.

Parameter	Tasks	Predictor	β	p (Bonf.)	sig.
RT	PS & BF	Task (main effect)	2.15	0.006	**
		Linear age	-0.23	0.003	**
		Task * linear age (interaction)	-0.25	0.003	**
		Task * quadratic age (interaction)	0.007	0.003	**
	PS & RL-WM	Task (main effect)	-0.76	0.69	
		Linear age	-0.48	< 0.001	***
		Task * linear age (interaction)	0.10	0.45	
		Task * quadratic age (interaction)	-0.003	0.288	
	BF & RL-WM	Task (main effect)	1.40	0.63	
		Linear age	-0.23	0.003	**
		Task * linear age (interaction)	-0.15	0.084	
		Task * quadratic age (interaction)	0.004	0.129	
ACC	PS & BF	Task (main effect)	1.27	0.36	
		Linear age	0.27	< 0.001	***
		Task * linear age (interaction)	-0.10	0.87	
		Task * quadratic age (interaction)	0.001	1	
	PS & RL-WM	Task (main effect)	-2.15	0.033	*
		Linear age	0.17	0.036	*
		Task * linear age (interaction)	0.20	0.118	
		Task * quadratic age (interaction)	-0.004	0.33	
	BF & RL-WM	Task (main effect)	-0.88	0.60	
		Linear age	0.27	< 0.001	***
		Task * linear age (interaction)	0.10	0.57	
		Task * quadratic age (interaction)	-0.003	0.57	
WS	—	Intercept	-3.05	< 0.001	***
		Age (linear)	0.31	< 0.001	***
		Age (quadratic)	-0.007	< 0.001	***
LS	PS & BF	Task (main effect)	-0.90	0.42	
		Linear age	0.075	0.87	
		Task * linear age (interaction)	0.12	0.42	
		Task * quadratic age (interaction)	-0.004	0.29	
	PS & RL-WM	Task (main effect)	4.84	< 0.001	***
		Linear age	0.20	0.015	*
		Task * linear age (interaction)	-0.51	< 0.001	***
		Task * quadratic age (interaction)	0.012	< 0.001	***
	BF & RL-WM	Task (main effect)	3.94	< 0.001	***
		Linear age	0.075	0.54	
		Task * linear age (interaction)	-0.39	< 0.001	***
		Task * quadratic age (interaction)	0.008	0.003	**
Delay	—	Intercept	0.95	0.035	*
		Age (linear)	-0.09	0.07	
		Age (quadratic)	0.002	0.14	

Age Trajectories of Behavioral Features Before presenting the results of the regression analysis, we discuss the behavioral features and their age trajectories, using similar regression models as in section Parameter Age Trajectories. Response times, reflecting choice fluidity and task engagement, sped up with age in all tasks, whereby age trajectories differed significantly between PS and BF in pairwise follow-up models (grand model $AIC_{with\ task} = 1.868$, $AIC_{no\ task} = 1.871$; for detailed statistics, see Table 3.3; Fig. 3.3C). Accuracy, reflecting subjective ease and task engagement, showed a significant increase with age, and no significant pairwise differences in age trajectories after correcting for multiple comparisons, despite the better fit of the model including task compared to the model without task ($AIC_{with\ task} = 2.015$, $AIC_{no\ task} = 2.024$; Fig. 3.3C; Table 3.3). Win-stay (WS) behavior reflects participants' tendency to repeat rewarded actions; specifically, it is the proportion of rewarded actions that are repeated on the next trial relative to all rewarded actions. Similarly, lose-stay (LS) behavior reflects participants' tendency to repeat non-rewarded actions. Win-stay behavior increased with age, without task differences ($AIC_{with\ task} = 1.961$, $AIC_{no\ task} = 1.959$; Fig. 3.3C; Table 3.3). Lose-stay behavior showed marked task differences ($AIC_{with\ task} = 2.075$, $AIC_{no\ task} = 2.109$), with inverse trajectories in RL-WM compared to the other tasks: In RL-WM, lose-stay behavior decreased monotonically until mid-adolescence (linear effect of age: $w = -0.31$, $p < 0.001$; quadratic effect: $w = 0.007$, $p < 0.001$), whereas in BF, it increased slightly (linear effect of age: $w = 0.075$, $p < 0.001$; quadratic effect: $w = -0.001$, $p < 0.001$). In PS, lose-stay behavior showed an inverse-U trajectory (linear effect: $w = 0.20$, $p < 0.001$; quadratic: $w = -0.005$, $p < 0.001$; Fig. 3.3C). Lastly, the Delay pattern measured the decrease in accuracy with increasing delay between presentations of the same stimulus (only present for RL-WM and BF because only these presented several stimuli; see section **??**). Delay did not show significant age changes (Table 3.3), and did not differ between tasks ($AIC_{with\ task} = 1.405$, $AIC_{no\ task} = 1.402$).

Regression Results We investigated the consistency in behavioral features captured by noise parameters and learning rates. Showing a high degree of task-based consistency, both BF $\frac{1}{\beta}$ and PS $\frac{1}{\beta}$ were predicted by task accuracy (BF $\frac{1}{\beta}$: $w = -0.19$, $p = 0.0076$; PS $\frac{1}{\beta}$: $w = -0.36$, $p < 0.001$; Fig. 3.5D) and win-stay behavior of the respective task (BF $\frac{1}{\beta}$: $w = -0.30$, $p < 0.001$; PS $\frac{1}{\beta}$: $w = -0.58$, $p < 0.001$), and the relationship with loose-stay behavior depended on the adaptivity of loose-stay behavior for the task (BF; PS). Showing absolute consistency, both were also predicted by BF win-stay behavior (PS $\frac{1}{\beta}$: $w = -0.12$, $p = 0.032$). RL-WM ε was not predicted by any behavioral feature, neither within the RL-WM task (all $|w|$'s < 0.16, all p's > 0.14), nor from other tasks (all $|w|$'s < 0.023, all p's > 0.85). Taken together, BF and PS decision noise parameters showed strong task-based consistency some absolute consistency, whereby reduced decision noise was related to improved task accuracy and increased repetition of rewarded choices, as expected. Noise in RL-WM was not captured by behavioral features.

Both PS α_+ and BF α_+ were predicted by win-stay behavior in the respective task (PS α_+: $w = 0.27$, $p < 0.001$; BF α_+: $w = 0.74$, $p < 0.001$; Fig. 3.5D), revealing task-based consistency. RL-WM α_+ was predicted by RL-WM task accuracy ($w = 0.24$, $p = 0.033$). Furthermore, both PS α_- and RL-WM α_- were negatively predicted by lose-stay behavior of the respective task (PS

α_-: $w = -0.71$, $p < 0.001$; BF α_-: $w = -0.41$, $p < 0.001$), and positively predicted by win-stay behavior (PS α_-: $w = 0.29$, $p < 0.001$; RL-WM α_-: $w = 0.16$, $p = 0.009$), revealing strong task-based consistency. PS α_- was also negatively predicted by PS task accuracy ($w = -0.28$, $p < 0.001$). In summary, learning-rate parameters showed high levels of task-based consistency, being related to win-stay behavior, lose-stay behavior, and task accuracy, as intended.

Taken together, these results suggest that both noise parameters and learning-rate parameters were quite consistent across tasks in terms of which behavioral patterns they captured. Differences in developmental time courses (Fig. 3.3A and B) and cognitive processes (Fig. 3.5A and B) occurred in conjunction with stark differences in behavior between tasks (Fig. 3.3C).

3.3 Conclusion

The past decades have seen a surge in computational modeling, and many studies have fitted model parameters to individual participants in the hope of distilling complex behaviors into a small number of meaningful individual characteristics. Parameters are usually interpreted to be both generalizable –making predictions about individuals that extend to other tasks and real-life situations– and interpretable –identifying the fundamental elements of cognitive and/or neural processing that are at the core of laboratory tasks and real-world problem solving. Though rarely stated explicitly, these assumptions underlie conclusions that have been drawn across all major fields of computational psychology and neuroscience. For example, associating model parameters with neural substrates (e.g., learning rates with dopamine receptors), identifying parameters that differ between healthy participants and those with psychiatric conditions (e.g., "blunted" positive learning rates in depression), determining humans' generic parameter settings (e.g., learning rates from negative outcomes; relationships between learning from negative versus positive outcomes), or assessing the development of parameters (e.g., learning rates across age), all rely on the assumption that model parameters capture the same (neuro)cognitive processes across studies and across paradigms, i.e., that model parameters are generalizable and interpretable in a generic, task-independent sense. Our results challenge this assumption, and show that a more nuanced understanding of model parameters is required to help resolve major discrepancies in the literature (e.g., Adams et al., 2016; Hauser et al., 2019; Nussenbaum and Hartley, 2019).

In a within-participant design and using state-of-the-art model fitting techniques –avoiding potential discrepancies caused by technical issues–, RL learning rates still showed wide discrepancies between tasks, revealing a lack of parameter generalizability. Discrepancies were evident in both absolute values and age trajectories, and, revealing a lack of parameter interpretability, learning rates also captured orthogonal or even opposite cognitive processes across tasks. Decision noise parameters, on the other hand, showed consistent age trajectories and cognitive processes. Both this consistency of decision noise parameters and the inconsistency of learning rates confirm patterns that have started to emerge in the literature (Nussenbaum and Hartley, 2019). Interestingly, both decision noise and learning rates were consistent across tasks in terms of the behavioral patterns they captured. Taken together, these findings clearly demonstrate that fitted values of learning rates should *not* be assumed to generalize between tasks, or even to capture similar cognitive pro-

cesses across studies, while these assumptions can likely be made for decision noise parameters. The lack of generalizability in learning rates, however, does not suggest a failure of cognitive modeling, but the need for an adapted interpretation of model parameters, given that parameters were consistent across tasks in terms of behavioral patterns. Instead of equating model parameters to (neuro)cognitive processes, we suggest they should be interpreted as maximally-compact behavioral summaries: In the same way behaviors differ between tasks due to task differences, so do model parameters.

Results Summary

Part I of this study focused on parameter generalizability, testing whether the same participants showed the same parameter values in different tasks. The evidence was positive for decision noise parameters ($\frac{1}{\beta}$ and ε), and negative for negative learning rates (α_-), with intermediate generalizability for positive learning rates (α_+): Decision noise parameters showed different absolute values across tasks, but they displayed the same age trajectories, and these age trajectories could be predicted from one task to another, as long as the two tasks were sufficiently similar. This demonstrates that a participant's relative decision noise, compared to other participants, was stable and generalized between tasks.

Learning rates, on the other hand, differed strikingly between tasks, especially in terms of absolute values. Interestingly, the observed differences were in accordance with task demands: In the volatile PS task, which required rapid switching and adaption, participants showed large positive and negative learning rates, whereas in the stable BF and RL-WM tasks, which required continuous incremental learning, participants showed small learning rates. In other words, participants' learning rates were of the general magnitude that was most appropriate for each task. Parameter adaptivity was also evident in the age trajectories of negative learning rates: A monotonic increase occurred in the RL-WM task, which provided diagnostic negative feedback, but a quadratically-modulated decrease occurred in the PS task, which had non-diagnostic negative feedback. In other words, both tasks showed a general increase in parameter optimality, a tendency that has been observed before (Nussenbaum and Hartley, 2019). Nevertheless, the shapes of the developmental trajectories differed profoundly between tasks (monotonic increase vs U-shape), highlighting the lack of generalization of learning rates between tasks. Positive learning rates revealed intermediate generalizability between decision noise and negative learning rates, showing qualitatively similar trajectories, but the inability to predict trajectories between tasks.

Part II examined parameter interpretability, i.e., whether parameters captured the same cognitive processes across tasks. All decision noise parameters showed a negative relationship with task performance, and all positive learning rates showed a positive relationship, whereas negative learning rates showed opposite relationships depending on the task. This highlights that with respect to their roles for performance, decision noise parameters and positive learning rates could be interpreted consistently across tasks, while negative learning rates could not. Regularized regression showed that variance in BF noise parameters was captured by both PS and RL-WM noise parameters, and that both captured slightly different aspects. Indeed, the BF task was the most prototypical RL task of the three, suggesting that parameters in more similar tasks potentially capture more sim-

ilar cognitive processes. Negative learning rates, on the other hand, showed large discrepancies: PS and RL-WM negative learning rates predicted the same parameter, but inversely, suggesting opposing cognitive processes; and RL-WM negative learning rate –indeed, the entire RL-WM model– failed to predict PS negative learning rate, suggesting a lack in overlap. Negative learning rates therefore could not be interpreted in terms of the same cognitive processes across these tasks. For positive learning rates, BF learning rates predicted RL-WM learning rates, revealing shared processes; in the inverse model, RL-WM working-memory and RL parameters predicted BF learning rates, suggesting that BF learning rates captured both RL and working-memory processes. (No variance in RL-WM working-memory parameters was captured by PS or BF parameters, suggesting that they successfully isolated working-memory processes.) PS positive learning rates were not predicted by any other parameter, suggesting an orthogonal, independent cognitive process, but PS Persistence was related to other the tasks' learning rates. In sum, learning-related cognitive processes were captured by a different interplay of positive and negative learning rates, working-memory, and other parameters (e.g., Persistence) in each task; learning rates therefore should not be interpreted as the entirety of learning processes captured by a cognitive model.

Lastly, the relation between model parameters and task behavior was strikingly consistent across tasks for both decision noise and learning rates. When decision noise decreased, accuracy and win-stay behavior increased, and loose-stay behavior increased when adaptive (PS) and decreased when not (BF). Similarly, when positive learning rates increased, win-stay behavior increased, and when negative learning rates increased, loose-stay behavior decreased. This shows that even learning rates, sometimes capturing widely different cognitive processes across tasks, consistently summarized the same behavioral patterns in each task: Decision noise was related to task accuracy and adaptive stay behavior, and positive and negative learning rates captured win-stay and loose-stay behavior, respectively. In sum, our recommendation for future research is to shift the interpretation of model parameters from a notion of task-independent measures of (neuro)cognitive processing, to a notion of maximally-compact behavioral summaries with the ability to highlight task differences. Specific task characteristics need to be the foundation for interpreting model parameters. For example, stochastic feedback might elicit incremental updating, whereas volatile tasks might elicit state inference, and many stimuli might elicit working-memory processing. The use of different cognitive processes results in differences in win-stay and loose-stay behavior, which will be reflected in differences between model parameters.

Potential Neural Substrates

Can computational modeling still inform brain science? We answer this question in the positive. Assuming that model parameters capture meaningful patterns, they are just as likely to relate to brain processes as any other meaningful behavioral pattern that has been used in fMRI studies, and our results provide no argument against using computational modeling in brain science. The only caveat is that parameter-associated brain areas are expected to generalize between tasks that rely on similar cognitive processes, and not as a general rule. More specifically, the generalizability of decision noise parameters across tasks is compatible with a single underlying neural system. The cognitive process likely relates to broad, non-specific motivation, task engagement, and overall

attention, and is potentially based in the evolutionary-old basal ganglia system, which plays a fundamental role for motivated behavior (Linda citations?). Another possibility is that decision noise is modulated by cortical maturation (with 'sparsification' of representations due to spine pruning, and maturation of inhibition and neuromodulation; Linda citations).

The lack of generalizability of learning rate parameters is compatible with different neural interpretations: Similar to decision noise, learning rates might reflect the same neural system in each task, most likely the midbrain-dopamine system (Niv, 2009; O'Doherty et al., 2015). Task differences might be caused by modulation of this system by environmental factors, including uncertainty (Gershman, 2017a; Gershman and Uchida, 2019; Starkweather et al., 2018), potentially through development (Lin et al., 2020). An alternative interpretation is that unlike decision noise, learning rates are not tied to a specific neural system, but capture different systems depending on task demands. For example, learning rates might reflect midbrain-dopamine activity when incremental, stochastic learning is required (Bayer and Glimcher, 2005), but hippocampal processing when episodic memory is required (Davidow et al., 2016), and yet another set of brain structures for working memory (Collins, Ciullo, et al., 2017). Future research is necessary to discriminate between these alternatives, using brain imaging in tasks with differing demands.

Reconciling Discrepancies in the Previous Literature

Discrepancies in the literature can point to a replication crisis and the existence of fundamental problems, but our results show that this does not need to be the case for cognitive modeling. Discrepancies in the computational modeling literature have been ascribed to a range of methodological issues, including model misspecification, lack of model comparison and validation, and inappropriate fitting methods (Daw, 2011; M. D. Lee, 2011; Nussenbaum and Hartley, 2019; Palminteri et al., 2017; Wilson and Collins, 2019), as well as the lack of parameter reliability and validity, even though this seems to be a smaller issue when model fitting is done appropriately (Brown et al., 2020). Our results show that this is not the only possible explanation, and that discrepancies can arise –and are even expected– when different studies use tasks that recruit different cognitive processes and elicit different behaviors.

We recommend that previous modeling studies be reinterpreted in this light to reconcile discrepancies. Nussenbaum and Hartley, 2019, for example, suggested a reinterpretation based on parameter optimality. In this view, instead of treating model parameters as stable individual characteristics (e.g., 10-year olds have a learning rate of 20%; 12-year olds of 40%), they should be interpreted in terms of their optimality for a specific task (e.g., adults' parameters are more optimal than children's). In this interpretation, parameter differences arise from differences in the ability to identify optimal parameters, or in the flexibility of adjusting parameters to task demands. Reinterpretation needs to specifically focus on task characteristics. Parameter commonalities and differences need to be mapped onto task factors including volatility (Behrens et al., 2007), uncertainty (Gershman and Uchida, 2019), feedback diagnosticity (Eckstein et al., 2020), response type (go no-go, binary, several options), and working-memory load (Collins and Frank, 2012). Future research will be crucial to fully understand the effects of all these factors, and investigate their interplay. A larger number of tasks, systematically varying several axes of interest, is needed

to verify our exploratory hypotheses about the mapping between model parameters and cognitive processes (e.g., stochastic feedback ↔ incremental updating). This research is also necessary to determine the extent of parameter consistency and inconsistency between tasks. In our study, noise parameters were more consistent than learning rates, and positive learning rates were more consistent that negative ones. Even though these results are consistent with patterns in the literature (Nussenbaum and Hartley, 2019), a larger pool of tasks is needed to understand their full extent. Furthermore, our results suggested that parameters generalized better between similar tasks, and a larger number of tasks will be able to confirm or disprove this pattern. Future research is also needed to investigate the relationships between cognitive constructs measured using computational models, and traditional cognitive constructs from psychology, including cristalline and fluid intelligence (Wechsler and Matarazzo, 1972), risk taking (Gullone et al., 2000), and memory span (Conway et al., 2005). Understanding the interplay between these constructs is necessary to connect computational modeling to previous research. Lastly, future research should systematically investigate model similarity. While our results suggest that parameters generalized more between similar tasks, this could also be a function of model similarity because similar tasks were modeled using similar models. De-correlating cognitive models from experimental tasks will improve our understanding. Relatedly, previous research has shown that not all individual differences can be captured in parameter differences, and that participants might instead employ entirely different computational models (e.g., Palminteri et al., 2016). Future research is needed to understand the relationship between computational models, model parameters, and task demands.

Why Parameters Capture Different Processes in Different Tasks: Parameters are Specific not Generic

What features of RL models lead to their capturing different cognitive processes in different tasks? RL is a general framework that has been applied to a variety of tasks in the cognitive literature, spanning from simple conditioning paradigms all the way to complex, goal-directed, temporally-extended, and hierarchical decision making. Nevertheless, the computational models used in each case are strikingly similar. For example, the same learning rate parameter is used when a) slowly acquiring a preference for one option over another through hundreds of repetitions, based on unreliable and noisy feedback; b) quickly recognizing whether underlying task contingencies have switched, requiring an abrupt switch in response patterns; and c) deciding which general strategy to try out in a new context, based on experience in other contexts. Furthermore, the same learning rates also account for choice (i.e., which action is selected in each trial) and meta-learning (i.e., slow improvement at a specific task over time), an issue that a small number of studies has addressed using learning rates at different levels of hierarchy (Botvinick, 2012; Eckstein and Collins, 2018; Ribas Fernandes et al., 2011; Wang et al., 2018). Nevertheless, often generic RL models with a single learning rate are used, which are unable to reflect these differences. In sum, because RL models are so compact, the same parameters need to capture different cognitive processes when used in different domains, a fact that is facilitated by their flexibility. In many ways, model parameters are more similar to task-specific measures (e.g., accuracy reflects different things in a learning

task versus language test), than to stable individual characteristics (e.g., intelligence, which can be measured using different tasks). This means that it is increasingly important to understand what parameters measure in each task; relating model parameters to other cognitive constructs and real-world behavior might be a task-by-task endeavor. Fundamentally, model parameters seem to be task-specific rather than task-independent and generic.

Outlook

This discovery reflects a larger pattern of realization in psychology that we cannot assume that we can assess individuals' function outside of a subjective context. IQ tests for example are not neutral assessments of an individual's function and scores are mediated by the familiarity with the context of the test questions and even testing itself. Task skin and parameters may operate in a similar way and shift the system settings within an individual. Knowledge about these processes could be valuable to develop individualized learning plans. This may be valuable as we better appreciate and seek to accommodate diversity in cognitive function.

3.4 Methods

Participant Sample

Sample Overview

All procedures were approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley. We tested 312 participants: 191 children and adolescents (ages 8-17) and 55 adults (ages 25-30) were recruited from the community and completed a battery of computerized tasks, questionnaires, and saliva samples; 66 university undergraduate students (aged 18-50) completed the four tasks as well, but not the questionnaires or saliva sample. Community participants of all ages were prescreened for the absence of present or past psychological and neurological disorders; the undergraduate sample indicated the absence of these. Compensation for community participants consisted in $25 for the 1-2 hour in-lab portion of the experiment and $25 for completing optional take-home saliva samples; undergraduate students received course credit for participation in the 1-hour study.

Participant Exclusion

Two participants from the undergraduate sample were excluded because they were older than 30, and 7 were excluded because they failed to indicate their age. This led to a sample of 191 community participants under 18, 57 undergraduate participants between the ages of 18-28, and 55 community participants between the ages of 25-30. Of the 191 participants under 18, 184 completed the PS task, and 187 completed the BF and RL-WM task. Reasons for not completing a task included getting tired, running out of time, and technical issues. All 57 undergraduate participants completed the PS task, 55 completed the BF task, and 55 completed the RL-WM task.

All 55 community adults completed the PS and BF task, and 45 completed RL-WM. Appropriate exclusion criteria were implemented separately for each task to exclude participants who failed to pay attention and who performed critically worse than the remaining sample (for PS, see Eckstein et al., 2020; BF: Xia et al., 2020; RL-WM: Master et al., 2020). Based on these criteria, 5 participants under the age of 18 were excluded from the PS task, 10 from the BF task, and none from the RL-WM task. One more community adult participant was excluded from the BF task, but no adult undergraduates or community participants were excluded for PS or RL-WM.

The performance criterion led to the exclusion of the majority of our developmental sample in the fourth task of our study, which was modeled after a rodent task and used in humans for the first time (Johnson and Wilbrecht, 2011). We therefore excluded this task from the current analysis. For some analyses, we split participants into quantiles based on age. Quantiles were calculated separately within each sex.

Testing Procedure

After entering the testing room, participants under 18 years and their guardians provided informed assent and permission; participants over 18 provided informed consent. Guardians and participants over 18 filled out a demographic form. Participants were led into a quiet testing room in view of their guardians, where they used a video game controller to complete four computerized tasks, in the order shown in Fig. 3.1C. At the conclusion of the tasks, participants between 11 and 18 completed the PDS questionnaire (Petersen et al., 1988) and were measured in height and weight. Participants were then compensated with $25 Amazon gift cards.

Task Design

Probabilistic Switching (PS)

The goal of the task was to collect golden coins, which were hidden in two green boxes. The task could be in one of two states: “Left box is correct” or “Right box is correct”. In the former, selecting the left box led to reward in 75% of trials, while selecting the right box never led to a reward (0%). Several times throughout the task, and unpredictably, task contingencies changed without notice (after participants had reached a performance criterion indicating they had learned the current state), and the task switched states, which led to a reversal of the task contingencies. Participants completed 120 trials of this task (2-9 reversals), which took approximately 5-15 minutes. For more information about additional task details, the employed tutorial, exact instructions, and switching rules; as well as a full analysis of this data set, please refer to Eckstein et al., 2020.

Butterfly (BF)

The goal of the task was to collect as many points as possible, by guessing correctly which of two flowers was associated with each of four butterflies. Correct guesses were rewarded with 70% probability, and incorrect guesses with 30%. The task contained 120 trials (30 for each butterfly)

that were split into 4 equal-sized blocks, and took between 10-20 minutes to complete. More detailed information about methods and results in this data set can be found in Xia et al., 2020.

Reinforcement Learning-Working Memory (RL-WM)

The goal of the task was to collect as many points as possible by pressing the correct key for each stimulus. Pressing the correct key always led to reward deterministically, and the correct key for a stimulus never changed. Stimuli appeared in blocks that varied in the number of different stimuli, with stimulus set sizes ranging from 2-5. In each block, each stimulus was presented 12-14 times, for a total of 13 * set size trials per block. Three blocks were presented for set sizes 2-3, and 2 blocks were presented for set sizes 4-5, for a total of 10 blocks. The task took between 14-25 minutes to complete. For more details, as well as a full analysis of this data set, please refer to Master et al., 2020.

Pubertal Measures

We administered the pubertal development scale (Petersen et al., 1988) and collected saliva samples to investigate the role of pubertal maturation on learning and decision making. Pubertal analyses are not the focus of the current study and will be or have reported elsewhere (e.g., Master et al., 2020; Xia et al., 2020). For details about how we assessed pubertal measures, refer to Master et al., 2020.

Computational Models

We fitted a separate RL model to each task, using state-of-the-art methods for model construction, fitting, and validation (Palminteri et al., 2017; Wilson and Collins, 2019). The PS and BF models were fitted using hierarchical Bayesian methods with Markov-Chain Monte-Carlos sampling, which is an improved method compared to maximum likelihood that leads to better parameter recovery, amongst other advantages (Gelman et al., 2013; Katahira, 2016; Watanabe, 2013). The RL-WM model was fitted using the classic non-hierarchical maximum-likelihood method because model parameter K is discrete, which renders hierarchical sampling less tractable. In all cases, we verified that the model parameters were recoverable by the selected model-fitting procedure, and that the models were identifiable. Details of model-fitting procedures can be found in Eckstein et al., 2020; Master et al., 2020; Xia et al., 2020

For the PS task, we fitted two separate models, one based on RL, and the other based on Bayesian Inference (Eckstein et al., 2020). For the other two tasks, we fitted a single, RL-based model. As explained in Introduction, RL proceeds in two steps: value-learning and action selection. During learning, action values are updated based on new outcomes r:

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

$Q_t(a|s)$ indicates the value of action a in state s on trial t, for example the estimated reward probability of selecting the red flower (a) in response to the purple butterfly (s) on the BF task. The

learning rate $0 < \alpha < 1$ determines the weight of new information compared to old value estimates, and the differences between the estimated value and actual received reward is called the "reward prediction error" $(r_t - Q_t(a|s))$. For action selection, learned action values $Q(a|s)$ are translated into action probabilities $p(a|s)$, using the softmax function:

$$p(a_i|s) = \frac{exp(\beta\, Q(a_i|s))}{\sum_{a_j \in A} exp(\beta\, Q(a_j|s))}$$

where A refers to the set of all available actions (PS and BF have two actions, RL-WM has three), and a_i and a_j to individual actions within the set. The free parameter $0 < \beta$ is the inverse decision temperature, or exploration: higher values of beta lead to more deterministic selection of the higher-valued action.

In PS and BF, positive and negative learning rates are differentiated in the following way:

$$Q_{t+1}(a|s) = Q_t(a|s) + \alpha_+(r_t - Q_t(a|s) \iff r_t = 1$$

$$Q_{t+1}(a|s) = Q_t(a|s) + \alpha_-(r_t - Q_t(a|s) \iff r_t = 0$$

(In BF, the best model only treated α_+ as a free parameter, and α_- was set to 0 for all participants.) In RL-WM, α_- is a function of α_+, such that $\alpha_- = b * \alpha_+$, where b is the neglect bias parameter that determines how much negative feedback is neglected compared to positive feedback. Throughout the paper, we report α_-.

In PS, an additional free parameter p captured *choice Persistence* (also called "sticky choice" or "choice perseverance"), which biased choices toward staying ($p > 0$) or switching ($p < 0$) on the subsequent trial. p worked by creating modified action values $Q'(a|s)$, which were submitted to the softmax function instead of $Q(a|s)$:

$$Q'_t(a|s) = Q_t(a|s) + p \iff a_t = a_{t-1}$$

$$Q'_t(a|s) = Q_t(a|s) \iff a_t \neq a_{t-1}$$

In addition, the PS model included counter-factual learning parameters α_{C+} and α_{C-}, which added counter-factual updates based on the inverse outcome and affecting the non-chosen action. For example, after receiving a positive outcome ($r = 1$) for choosing left (a), counter-factual updating would lead to an "imaginary" negative outcome ($\bar{r} = 0$) for choosing right ($\bar{a}$).

$$Q_{t+1}(\bar{a}|s) = Q_t(\bar{a}|s) + \alpha_{C+}(\bar{r} - Q_t(\bar{a}|s)) \iff r = 1$$

$$Q_{t+1}(\bar{a}|s) = Q_t(\bar{a}|s) + \alpha_{C-}(\bar{r} - Q_t(\bar{a}|s)) \iff r = 0$$

$\bar{a}$ indicates the non-chosen action, and $\bar{r}$ indicates the inverse of the received outcome, $\bar{r} = 1 - r$. The best model fits were achieved with $\alpha_{C+} = \alpha_+$ and $\alpha_{C-} = \alpha_-$, so counter-factual learning rates are not reported in this paper.

In BF, the best fitting model included a forgetting mechanism, which was implemented as a decay in Q-values applied to all action values of the three stimuli (butterflies) that were not shown on the current trial:

$$Q_{t+1}(a|s) = (1 - f) * Q_{t+1}(a|s) + f * 0.5.$$

The free parameter $0 < f < 1$ reflects the individual tendency to forget.

In addition to an RL module, the RL-WM model also included a working-memory module with perfect recall of recent outcomes, but subject to forgetting and capacity limitations. Perfect recall was modeled as an RL process on working-memory weights $W(a|s)$ with learning rate $\alpha_{WM+} = 1$. On trials with positive outcomes ($r = 1$), the model reduces to:

$$W_{t+1}(a|s) = r_t$$

On trials with negative outcomes ($r = 0$), multiplying $\alpha_{WM+} = 1$ with the neglect bias b leads to potentially less-than perfect memory:

$$W_{t+1}(a|s) = W_t(a|s) + b * (r_t - W_t(a|s))$$

Working-memory weights $W(a|s)$ were transformed into action policies $p_{WM}(a|s)$ in a similar way as RL weights $Q(a|s)$ were transformed into action probabilities $p_{RL}(a|s)$, using a softmax transform, but combined with undirected noise:

$$p(a_i|s) = (1-\varepsilon) * \frac{exp(\beta\ Q(a_i|s))}{\sum_{a_j \in a} exp(\beta\ Q(a_j|s))} + \varepsilon * \frac{1}{|a|}$$

$|a| = 3$ is the number of available actions, and $\frac{1}{|a|}$ is the uniform policy over these actions. Forgetting is implemented as a decay in working-memory weights $W(a|s)$ (but not RL Q-values):

$$W_{t+1}(a|s)_{t+1} = (1-f) * W_t(a|s)_t + f * \frac{1}{3}$$

Capacity limitations of working memory were modeled as an adjustment in the weight w of $p_{WM}(a|s)$ compared to $p_{RL}(a|s)$ in the final calculation of action probabilities $p(a|s)$:

$$w = \rho * (min(1, \frac{K}{ns}))$$

$$p(a|s) = w * p_{WM}(a|s) + (1-w) * p_{RL}(a|s)$$

The free parameter ρ is the individual weight of working memory compared to RL, ns stands for a block's set size, and K captures individual differences in working-memory capacity.

In addition to the RL model, we also fit a Bayesian Inference model to the PS task. This model contained a mental model of the task, which was based on two states "Left is correct" and "Right is correct", and used Bayesian inference to infer the current state based on recent outcomes. The free parameters of this model were the parameters of the task (switch probability on each trial p_{switch}, and probability of reward for a correct choice p_{reward}), in addition to choice parameters Persistence p and inverse decision temperature β. Detailed information about this model is provided in Eckstein et al., 2020. For additional details on any of these models, as well as detailed model comparison and validation, the reader is referred to the original publications (RL-WM: Master et al., 2020; BF: Xia et al., 2020; PS: Eckstein et al., 2020).

Ridge Regression

In section "Do Parameters Capture Overlapping Cognitive Processes?", we use regularized, cross-validated Ridge regression to determine whether parameters captured overlapping variance, which would point to capturing similar cognitive processes. Ridge regression is used to avoid problems caused by overfitting, regularizing regression weight parameters w based on their L2-norm. Regular regression works by identifying a vector of regression weights w that minimizes the linear least squares $||y - wX||_2^2$. $||a||_2^2 = \sqrt{\sum_{a_i \in x} a_i^2}$ is the L2-norm of a vector, vector y represents the outcome variable (in our case, a parameter fitted to each participant), matrix X represents the predictor variables (in our case, parameters of a different task fitted to each participant), and vector w represents the weight assigned to each feature in X (in our case, the weight assigned to each predicting parameter).

When datasets are small compared to the number of predictors in a regression model, overfitting can lead to "exploding" regressions weights w. Ridge regression avoids this issue by not only minimizing the linear least squares like regular regression, but also the L2 norm of weights w, minimizing $||y - wX||_2^2 + \alpha * ||w||_2^2$. Parameter α is a meta-parameter of Ridge regression and needs to be chosen by the experimenter. To avoid bias in the selection of α, we used a repeated cross-validation procedure. At each iteration of the procedure, we split the dataset into a predetermined number s of equal-sized splits, fitted Ridge regression to each of the splits independently using a different value for α, and then determined the best value of α based on cross-validation between the splits, using the amount of explained variance, R^2, as the selection criterion. Based on a coarse pre-analysis, we determined the search space $\alpha \in$ [0, 10, 30, 50, 100, 300, 500, 1,000, 3,000, 5,000, 10,000, 100,000, 1,000,000]. To avoid biases based on the random assignment of participants into data splits, this procedure was repeated $n = 100$ times for each value of α. To avoid biases caused by s, the entire process was repeated for $s \in [2,3,4,5,6,7,8]$. The final value of s was selected based on model fit (explained variance R^2).

This process was conducted separately for each model, i.e., each combination of an outcome parameter (e.g., PS α_+) and a predicting task (BF or RL-WM). Meta-parameters s and α were allowed to differ (and differed) between models (see supplements). The final values of R^2 (Fig. 3.5A) and the final regression weights w (Fig. 3.5B) were determined by running 1,000 Ridge regression models based on the best meta-parameters identified using this procedure, and calculating means and standard deviations sd of R^2 and w over the repetitions. Statistical tests were conducted by assessing the proportion p of repetitions in which a null hypothesis (e.g., $R^2 < 0$) was accepted. Values of $p < 0.05$ were deemed significant.

Principal Component Analysis (PCA)

PCA performs a "change of basis": Instead of describing the dataset using the original features (in our case, 54 behaviors and model parameters), it creates new features –called Principal Components (PCs)– that are linear combinations of the original features and capture the same variance, but are orthogonal to each other. PCs are created by eigendecomposition of the covariance matrix of the dataset: the eigenvector with the largest eigenvalue shows the direction in the dataset in

which most variance occurs, and represents the first PC. Eigenvectors with subsequently smaller eigenvalues form subsequent PCs.

PCs can be interpreted by evaluating their "factor loadings", the linear weight of each original feature on the PC. PCA is related to Factor analysis, and often used for dimensionality reduction. In this case, only a small number of PCs is retained whereas the majority is discarded, in an effort to retain most variance with a reduced number of features.

Our dataset consisted of 54 features, 39 behavioral measures and 15 model parameters. For an explanation of the model parameters, refer to section "Computational Models" above, or the original publications of each model (RL-WM: Master et al., 2020; BF: Xia et al., 2020; PS: Eckstein et al., 2020). For each task, behavioral measures include: number of missed trials, average response times, response time variability (standard deviation of response times), accuracy (overall percentage of correct trials), win-stay strategy (percentage of trials in which a rewarded choice was repeated), and loose-stay tendency (percentage of trials in which a non-rewarded choice was repeated). For PS, we additionally included win-loose-stay tendencies, which is the proportion of trials in which participants stay after a winning trial that is followed by a losing trial. This is an important measure for this task because the optimal strategy required staying after single losses.

We also included behavioral persistence measures in all tasks. In BF and RL-WM, these included a measure of action repetition (percentage of trials in which the previous key was pressed again, irrespective of the stimulus and feedback) and choice repetition (percentage of trials in which the action was repeated that was previously selected for the same stimulus, irrespective of feedback). In PS, both measures are identical because it does not have different stimuli, so we only included one of them.

Last, we included task-specific measures of performance. In BF, the average accuracy for the first three presentations of each stimulus, reflected early learning speed; and the asymptote, intercept, and slope of the learning progress in a regression model predicting performance (for details about these measures, see Xia et al., 2020). In PS, the number of reversals (because reversal was contingent on performance; and the average trial-to-criterion after a switch (number of trials until 2 correct choices have been made after a task switch) offered additional measures of performance. In BF and RL-WM, we also included a model-independent measure of forgetting. In BF, this was the effect of delay on performance in the regression model mentioned above. In RL-WM, this was the effect of delay in a similar regression model, which also included set size, the number of previous correct choices, and the number of previous incorrect choices, whose effects were also included. Lastly for RL-WM, we included the slope of accuracy and response times over set sizes, as a measure of the effect of set size on performance. For PS, we also included the difference between early (first third of trials) and late (last third) performance as a measure of learning.

3.5 Supplemental Material

Table 3.4: Statistics of mixed-effects regression models predicting parameter values from task (BF, PS, RL-WM), age, and squared age (months). Only effects including task are reported. * $p < .05$; ** $p < .01$, *** $p < .001$.

Parameter	Tasks	Predictor	β	p	sig.
α_+	PS & BF	Task (main effect)	0.79	< 0.001	***
		Task * linear age (interaction)	-0.025	0.009	**
		Task * quadratic age (interaction)	0.001	0.021	*
	PS & RL-WM	Task (main effect)	0.84	< 0.001	***
		Task * linear age (interaction)	-0.012	0.41	
		Task * quadratic age (interaction)	< 0.001	0.55	
	BF & RL-WM	Task (main effect)	0.048	0.70	
		Task * linear age (interaction)	-0.12	0.37	
		Task * quadratic age (interaction)	< 0.001	0.36	
$\frac{1}{\beta}$	PS & BF	Task (main effect)	0.49	< 0.001	***
		Task * linear age (interaction)	-0.026	0.046	*
		Task * quadratic age (interaction)	0.001	< 0.001	***
α_-	PS & RL-WM	Task (main effect)	11.70	< 0.001	***
		Task * linear age (interaction)	0.58	< 0.001	***
		Task * quadratic age (interaction)	-0.013	< 0.001	***
Forgetting	PS & RL-WM	Task (main effect)	0.10	0.36	
		Task * linear age (interaction)	0.005	0.70	
		Task * quadratic age (interaction)	< 0.001	0.67	

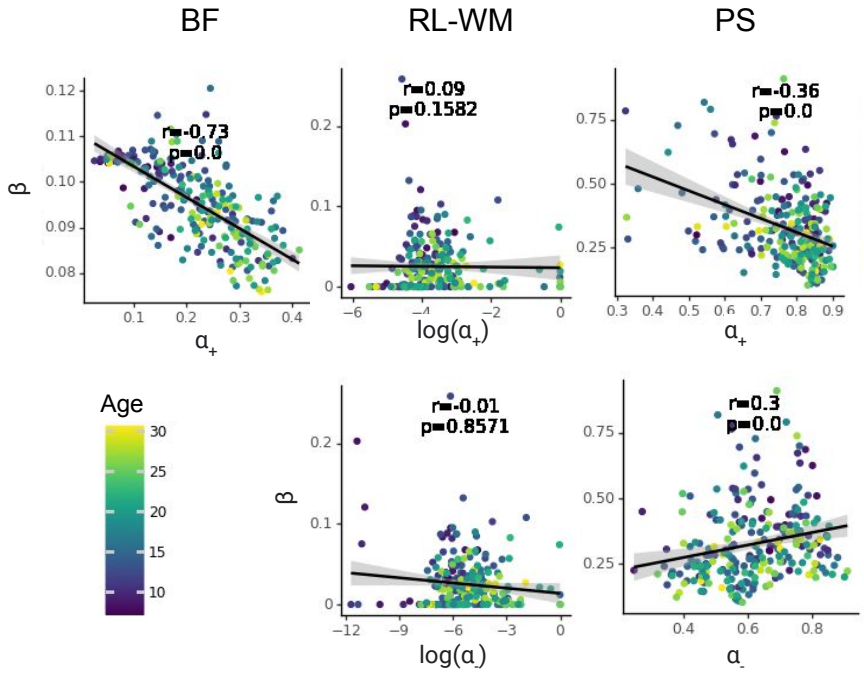


Figure 3.6: Scatter plots and Spearman correlations. Each dot is one participant, colors indicate age.

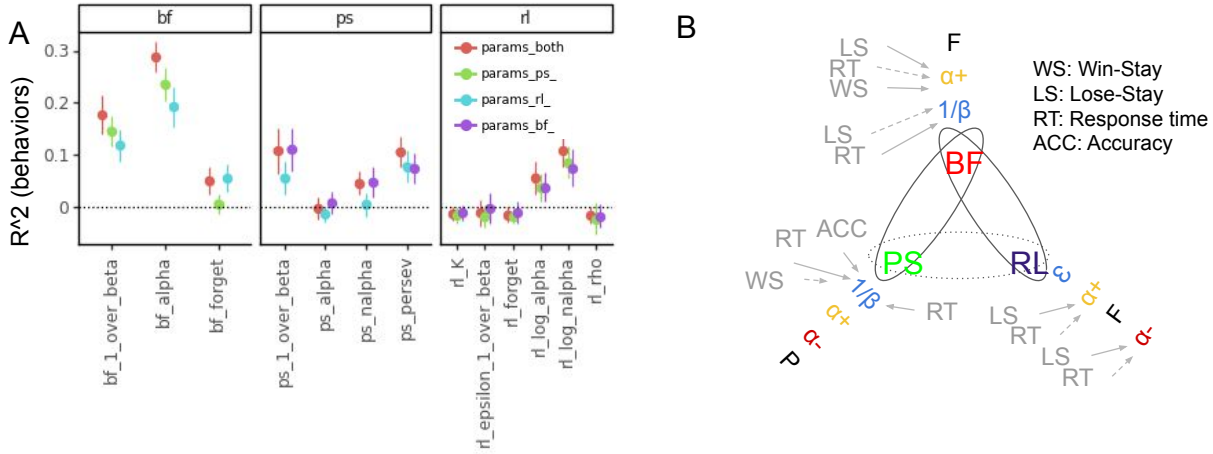


Figure 3.7: Same as Fig. 3.3A and B, but using behavioral features as predictors.

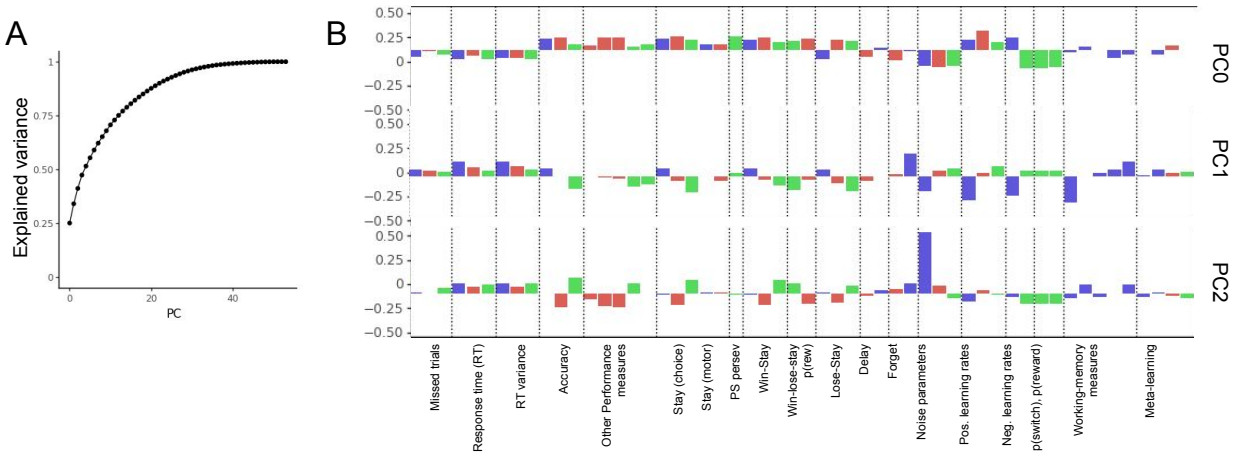


Figure 3.8: Additional PCA results. (A) Cumulative explained variance of all PCs in the analyses presented in section Relating Parameters and Behaviors Using Principal Component Analysis. (B) Alternative results, conducting separate PCAs based on parameters only, or behaviors only, shown in the same figure.

Chapter 4

Hierarchically-Structured Reinforcement Learning in Humans

This chapter presents a computational model that captures hierarchical learning and choice in humans. The model executes reinforcement learning processes at two levels of abstraction, inspired by the hierarchical organization of the brain's reinforcement learning system. [1]

Abstract

Humans have the fascinating ability to achieve goals in a complex and constantly changing world, still surpassing modern machine learning algorithms in terms of flexibility and learning speed. It is generally accepted that a crucial factor for this ability is the use of abstract, hierarchical representations, which employ structure in the environment to guide learning and decision making. Nevertheless, how we create and use these hierarchical representations is poorly understood. This study presents evidence that human behavior can be characterized as hierarchical reinforcement learning (RL). We designed an experiment to test specific predictions of hierarchical RL using a series of subtasks in the realm of context-based learning, and observed several behavioral markers of hierarchical RL, such as asymmetric switch costs between changes in higher-level versus lower-level features, faster learning in higher-valued compared to lower-valued contexts, and preference for higher-valued compared to lower-valued contexts. We replicated these results across three independent samples. We simulated three models: a classic RL, a hierarchical RL, and a hierarchical Bayesian model, and compared their behavior to human results. While the flat RL model captured some aspects of participants' sensitivity to outcome values, and the hierarchical Bayesian model some markers of transfer, only hierarchical RL accounted for all patterns observed in human behavior. This work shows that hierarchical RL, a biologically-inspired and computationally simple algorithm, can capture human behavior in complex, hierarchical environments, and opens the avenue for future research in this field.

[1] This chapter has been published separately (Eckstein and Collins, 2018), co-authored with my thesis advisor Anne G.E. Collins, and with the contributions of Lucy Whitmore and Sarah L. Master to data collection.

4.1 Introduction

Research in the cognitive sciences has long highlighted the importance of hierarchical representations for intelligent behavior, in domains including perception (T. S. Lee & Mumford, 2003), learning and decision making (Botvinick et al., 2009; Botvinick et al., 2015), planning and problem solving (Chase & Simon, 1973), cognitive control (E. K. Miller & Cohen, 2001), and creativity (Collins & Koechlin, 2012), among many others (Griffiths et al., 2019; Tenenbaum et al., 2011). The common thread across all these domains is the insight that hierarchical representations—i.e., the simultaneous representation of information at different levels of abstraction—allow humans to behave adaptively and flexibly in complex, high-dimensional, and ever-changing environments. Exhaustive non-hierarchical (*flat*) representations, in contrast, are insufficient to achieve human-like behaviors.

To illustrate, consider the following situation. Mornings in your office, your colleagues are working silently, or quietly discussing work-related topics. After work, they are laughing and chatting loudly at their favorite bar. In this example, a context change induced a drastic change in behavior, despite the same interaction partners (i.e., "stimuli"). Hierarchical theories of cognition capture this behavior by positing that we learn strategies hierarchically, activating different behavioral strategies (or "task-sets") in different contexts. Although hierarchical representations can incur additional cognitive cost (Collins, 2017), they provide a range of advantages compared to exhaustive flat representations: Once a task-set has been selected (e.g., office), attention can be focused on a subset of environmental features (e.g., just the interaction partner) (Frank & Badre, 2012; Leong et al., 2017; Niv et al., 2015; Wilson & Niv, 2012). When new contexts are encountered (e.g., new workplace, new bar), entire task-sets can be reused, allowing for generalization (Collins & Koechlin, 2012; Donoso et al., 2014; Taatgen, 2013). Old skills are not catastrophically forgotten (Flesch et al., 2018). In addition, hierarchical representations deal elegantly with incomplete information, for example when contexts are unobservable (Collins & Frank, 2013; Collins & Koechlin, 2012). All these advantages are evident in the current study.

Although we know that hierarchical representations are essential for flexible behavior, how humans create these representations and how they learn to use them is still poorly understood. Here, we hypothesize that learning and using hierarchical representations can be explained under a hierarchical reinforcement learning (RL) framework, in which simple RL computations are combined to simultaneously operate at different levels of abstraction. RL theory (Sutton & Barto, 2017) formalizes how to adjust behavior based on feedback in order to maximize rewards. Standard RL algorithms estimate how much reward to expect when selecting actions in response to stimuli, and use these "action-value" estimates to select actions. Old action-values are updated in proportion to the "reward prediction error", the discrepancy between action-values and received reward, to produce increasingly accurate estimates. Such "flat" RL algorithms operate over unstructured, exhaustive representations (SI Appendix, Fig. 3A), converge to optimal behavior, are computationally inexpensive, and have led to recent breakthroughs in artificial intelligence (AI) (Sutton & Barto, 2017).

Broad evidence suggests that the brain implements computations similar to RL: Dopamine neurons generate reward prediction errors (Bayer & Glimcher, 2005; Schultz et al., 1997), and a

wide-spread network of frontal cortical regions (D. Lee et al., 2012) and basal ganglia (Abler et al., 2006; Tai et al., 2012) represents action values. Specific brain circuits thereby form "RL loops" (G. Alexander et al., 1986; Collins & Frank, 2013), in which learning is implemented through the continuous updating of action values (Niv et al., 2015; Schultz, 2013). In this sense, estimating action-values via RL is an algorithm of special interest to cognition: There is strong evidence that the brain implements a simple mechanism to perform the necessary computations. Nevertheless, RL algorithms have important shortcomings: They suffer from the curse of dimensionality (an exponential drop in learning speed with increasing problem complexity); they lack flexibility for behavioral change; and they cannot easily generalize or transfer old knowledge to new situations. Hierarchical RL (Konidaris, 2019) mitigates these shortcomings by nesting RL processes at different levels of temporal (Botvinick, 2012; Momennejad et al., 2017; Ribas Fernandes et al., 2011) or state abstraction (Farashahi et al., 2017; Leong et al., 2017).

Recent research has provided support for a plausible implementation of hierarchical RL in the brain: The neural circuit that implements RL is multiplexed, such that distinct RL loops operate at different levels of abstraction along the rostro-caudal axis (G. Alexander et al., 1986; W. H. Alexander & Brown, 2015; Badre, 2008; Badre & D'Esposito, 2009; Badre & Frank, 2012; Balleine et al., 2015; Frank & Badre, 2012; Haruno & Kawato, 2006; Koechlin, 2016). Consistent with this architecture, recent studies have shown signatures of RL values and reward prediction errors at different levels of abstraction in the human brain (Diuk, Tsai, et al., 2013; Ribas Fernandes et al., 2011). However, previous studies did not provide evidence that neural signatures of hierarchical value support learning and generalizing hierarchically structured behavior. Thus, it remains unknown whether hierarchical RL indeed supports hierarchical behavior. The goal of this study is to fill this gap. We investigate hierarchical RL in a novel paradigm that promotes the creation and reuse of hierarchical structure. We provide a fully-fledged computational model that accounts for behavior across a variety of relevant situations: context-dependent learning, context switches, generalization to new contexts, partially-observable problems, and choices at different levels of abstraction. To our knowledge, this is the first study that tests all predictions of hierarchical RL in a single paradigm. Because hierarchical RL makes specific behavioral predictions in each situation, we are able to test the model *qualitatively* against human behavior (Palminteri et al., 2017). We then compare our hierarchical RL model *quantitatively* to the two most relevant competing models, flat RL and hierarchical Bayes. The former employs RL, but without hierarchical structure. The latter assumes that high-level decisions are based on Bayesian inference of task-set reliability, rather than RL using task-set values (Donoso et al., 2014).

In the following, we first introduce our hierarchical RL model and experimental paradigm. We then test whether humans show qualitative behaviors that are predicted by the hierarchical RL model, as well as the two competing models. We first show evidence for hierarchical representations in humans, as predicted by both hierarchical RL and hierarchical Bayes, but not flat RL. We employ multiple independent analyses, including switch cost measures and positive and negative transfer. We then provide evidence for human hierarchical value learning, which is only consistent with the hierarchical RL model. We next provide quantitative support for these qualitative results, and show that model comparison supports the hierarchical RL model over flat RL and hierarchical Bayes. The majority of results replicates across three independent participant samples.

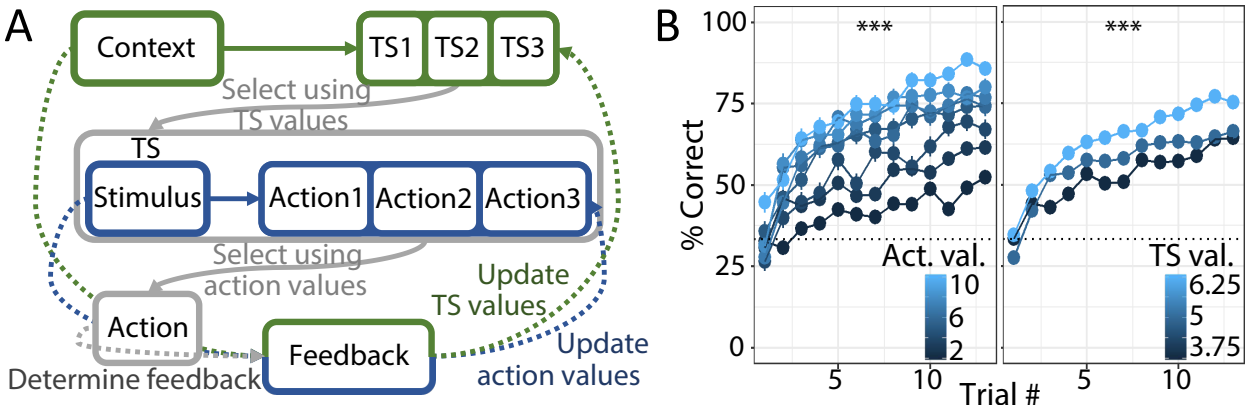


Figure 4.1: A) Schematic of the hierarchical RL model. A high-level RL loop (green) selects a task-set TS in response to the observed context, using TS values. The chosen task-set provides action-values, based on which the low-level RL loop (blue) selects an action in response to the observed stimulus. Task-set and action-values are both learned based on action feedback. B) Human learning curves during the initial learning phase, averaged over blocks. Colors denote underlying action-values (left) and task-set values (right), respectively. Stars show that *both* affect performance (main text), consistent with hierarchical RL. *** indicates $p < 0.001$.

4.2 Results

Computational Models

Our hierarchical RL model is composed of two hierarchically-structured RL processes. The high-level process manages behavior at the abstract level by acquiring a "policy over policies": It learns which task-set to choose in each context, using "task-set values" (the estimated expected reward of selecting a task-set in a given context). The low-level process acquires these task-sets: it learns which actions to choose in response to each stimulus by estimating "action values" (the estimated expected reward of selecting an action for a given stimulus, within a specific task-set; Fig. 4.1A).

At the beginning of the task, task-sets and actions are picked randomly, but over time, trial-and-error learning leads to the formation of meaningful task-sets, which represent policies that are specialized for particular contexts. Trial-and-error learning also underlies the policy over task-sets that determines which task-set is selected in each context. Thus, our hierarchical RL model is based on two nested processes, which create an interplay between learning stimulus-action associations (low level) and context-task-set associations (high level). SI Appendix, (Fig. 4) shows a step-by-step visualisation of this model.

Formally, to select an action a in response to stimulus s in context c, hierarchical RL goes through a two-step process: (1) It selects a task-set TS based task-set values in the current context, $Q(TS|c)$, using $p(TS|c) = \frac{exp(Q(TS|c))}{\sum_{TS_i} exp(\beta_{TS}\ Q(TS_i|c))}$. The inverse temperature β_{TS} captures task-set choice stochasticity (Fig. 4.1A). The chosen task-set TS provides a set of action-values $Q(a|s,TS)$,

which are used to (2) select an action a, according to $p(a|s,TS) = \frac{exp(Q(a|s,TS))}{\sum_{a_i} exp(\beta_a\ Q(a_i|s,TS))}$, where β_a captures action choice stochasticity (Fig. 4.1A; for trial-by-trial behavior, see SI Appendix, Fig. 4B). After executing action a on trial t, feedback r_t reflects the continuous amount of reward received, which guides learning at both levels of abstraction, i.e., to update the values of the selected task-set and action: $Q_{t+1}(TS|c) = Q_t(TS|c) + \alpha_{TS}\ (r_t - Q_t(TS|c))\ Q_{t+1}(a|s,TS) = Q_t(a|s,TS) + \alpha_a\ (r_t - Q_t(a|s,TS))$ α_{TS} and α_a are learning rates at the levels of task-sets and actions (Fig. 4.1A; SI Appendix, Fig. 4C).

The flat RL model uses the same mechanism for value learning and action selection, but lacks hierarchical structure: It treats each combination of context and stimulus as a unique state (methods). The hierarchical Bayesian model creates a task-set structure like hierarchical RL, but selects task-sets according to their inferred reliability, rather than task-set values (methods).

Task Design

We designed a task in which participants learned to select the correct actions for different stimuli (Fig. 4.2A). The mapping between stimuli and actions varied across three contexts, creating three distinct task-sets (Fig. 4.2B). Each context appeared in three blocks of 52 trials, for a total of 9 blocks. Contexts differed in average rewards, allowing us to test for RL values at the level of task-sets. After an initial-learning phase of this task (Fig. 4.2A), participants completed four test phases (Fig. 4.2C) to hone in on specific predictions of hierarchical RL. Detailed information about the task is provided in Fig. 4.2, the methods, and SI Appendix.

Learning Curves and Effects of Reward

As expected, participants' performance increased within a block, showing adaptation to context changes (Fig. 4.1B). We also verified that participants were sensitive to continuous differences in reward magnitudes (tape length). RL predicts better performance for larger rewards because these lead to larger action-values, which make correct actions more distinguishable from incorrect ones (see suppl. Fig. 4B for details). Participants indeed showed better performance for high-reward stimuli (Fig. 4.1B, left). This effect was predicted by both hierarchical and flat RL. Hierarchical RL additionally predicts better performance for high-valued contexts: Larger rewards create larger reward-prediction errors at the task-set level, which allow for better discrimination between correct and incorrect task-sets, and lead to better task-set selection and performance (see SI Appendix, Fig. 4A for details). As predicted, participants also showed an effect of task-set values on performance (Fig. 4.1B, right).

To quantify both effects, we conducted a mixed-effects logistic regression model predicting trialwise accuracy from action-values, task-set values, and their interaction (fixed effects), specifying participants, trial, and block as random effects. We approximated action-values as average stimulus-action rewards, and task-set values as average context-task-set rewards, as shown in Fig. 4.2B. The model revealed significant effects of both action-values, $\beta = 0.38$, $p < 0.001$, and task-set values, $\beta = 0.20$, $p < 0.001$, on performance (for complete statistics and results in other sam-

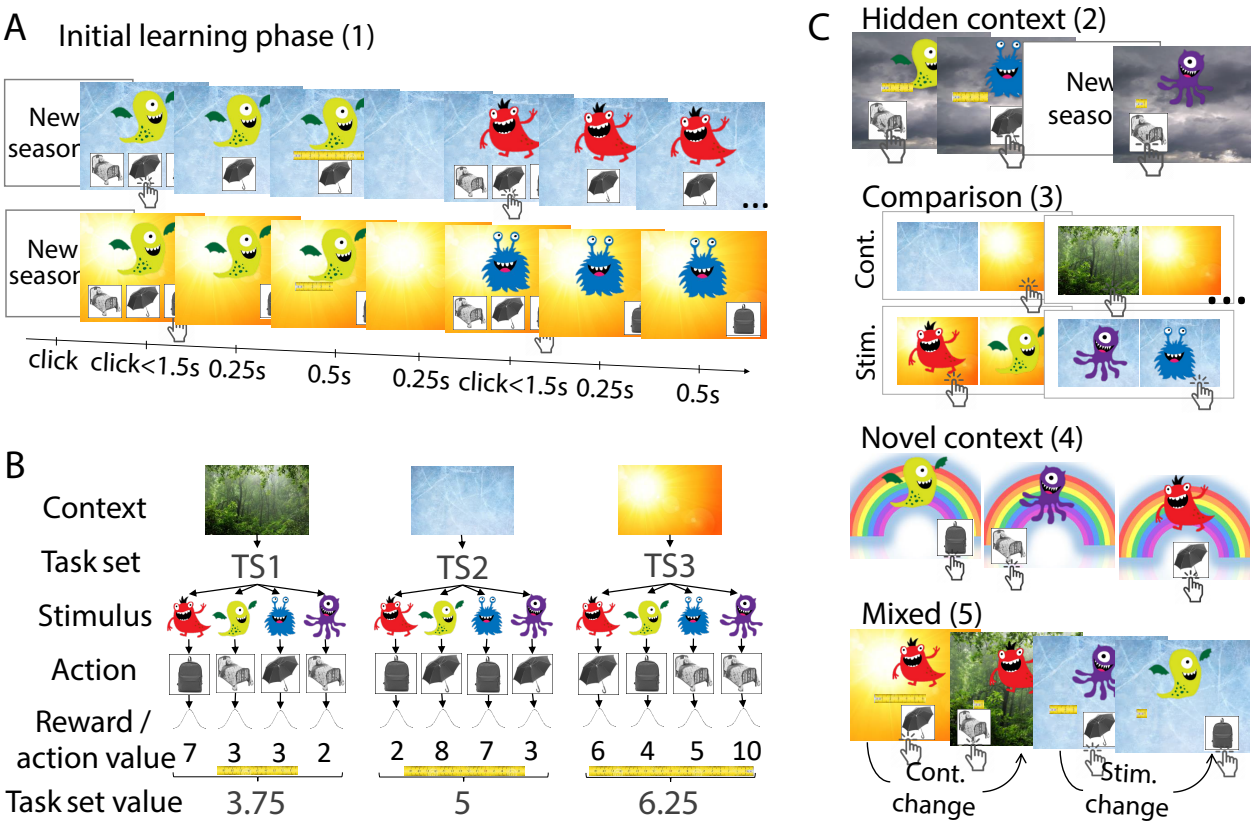


Figure 4.2: Task design. A) In the initial-learning phase, participants saw one of four stimuli (aliens) in one of three contexts (seasons), and had to find the correct action (item) through trial and error. Each context had a different mapping between stimuli and correct actions, and contexts were presented blockwise. Feedback indicated correctness deterministically, but different context-stimulus-action combinations lead to different rewards (with Gaussian noise). B) Example mapping between stimuli and actions for each context, defining three task-set TS1-3. Average rewards (*task-set values*) differed between contexts. All actions and stimuli had equal average rewards. C) Additional test phases. The hidden-context phase, presented after initial learning, was identical except that contexts were unobservable (season hidden by clouds). This allowed us to test whether participants reactivated previously-learned task-sets. In the comparison phase, participants saw either two contexts ("Cont.") or two stimuli ("Stim.") on each trial, and selected their preferred one. We used subjective preferences to assess task-set values (contexts) and action-values (stimuli). The novel-context phase was similar to initial learning, but had a new context and no feedback, to test how participants generalized previous knowledge to new situations. The final mixed phase was similar to initial learning, but not blocked, i.e., both stimuli and contexts could change on every trial, to test for asymmetric switch costs. All test phases were separated by "refresher blocks" similar to initial learning, to alleviate carry-over effects and forgetting.

ples, see SI Appendix, table 1). This provides initial evidence that human choices were sensitive to RL values at two levels of abstraction—actions and task-sets—, as predicted by hierarchical RL.

Hierarchical Representation

We tested participants' abstractions in more detail using three independent analyses: switch costs in the mixed phase of the task, reactivation of task-sets in the hidden-context phase, and task-set selection errors during initial learning.

Asymmetric Switch Costs

Asymmetric switch costs can be evidence for hierarchical representations because changes across trials are more challenging at higher than lower levels of abstraction (Collins, Cavanagh, et al., 2014; Monsell, 2003). For example, switching contexts is more cognitively costly than switching stimuli within a context. To test for such asymmetries in our paradigm, we compared trials on which a different stimulus was presented than on the previous trial (but the same context) to those on which a different context was presented (but the same stimulus), using the mixed phase (Fig. 4.2C). As expected, participants responded significantly slower after context switches than after stimulus switches, $t(25) = 3.47,\ p = 0.002$. This was not due to participants' initial surprise about the interleaved presentation of contexts in the mixed phase, as the result held throughout the phase (see SI Appendix). Asymmetric switch costs therefore suggest that participants created hierarchical representations, nesting stimuli within contexts, as predicted by hierarchical RL and hierarchical Bayes.

Reactivating Task-Sets

Did representing the task hierarchically benefit performance, e.g., did it support positive transfer? In the hidden-context phase of our task, contexts were not observable, such that participants could either relearn old stimulus-action mappings from scratch (no transfer), or reactivate previous task-sets, with the correct mappings already in place (transfer). By enabling reactivation of old task-sets, hierarchy has been shown previously to enable better performance and faster learning (Collins & Koechlin, 2012; Donoso et al., 2014; Koechlin, 2016).

If participants reactivated task-sets, we expect a specific pattern of performance in the hidden-context phase, specifically on the first few trials after a context switch, before any stimulus is repeated: Because every trial provides feedback about the appropriateness of the chosen task-set, task-set selection should become more accurate on each trial, and consequently, accuracy should improve. If, on the other hand, participants did not use task-sets and instead re-learned stimulus-response associations from scratch, as predicted by flat RL, performance can only increase after a stimulus is repeated. Because no stimuli are repeated until the 5^{th} trial in our task, the first four trials provide the perfect testing ground to pitch these two predictions against each other, as illustrated in Fig. 4.3A: Hierarchical RL simulations show increasing performance, whereas flat

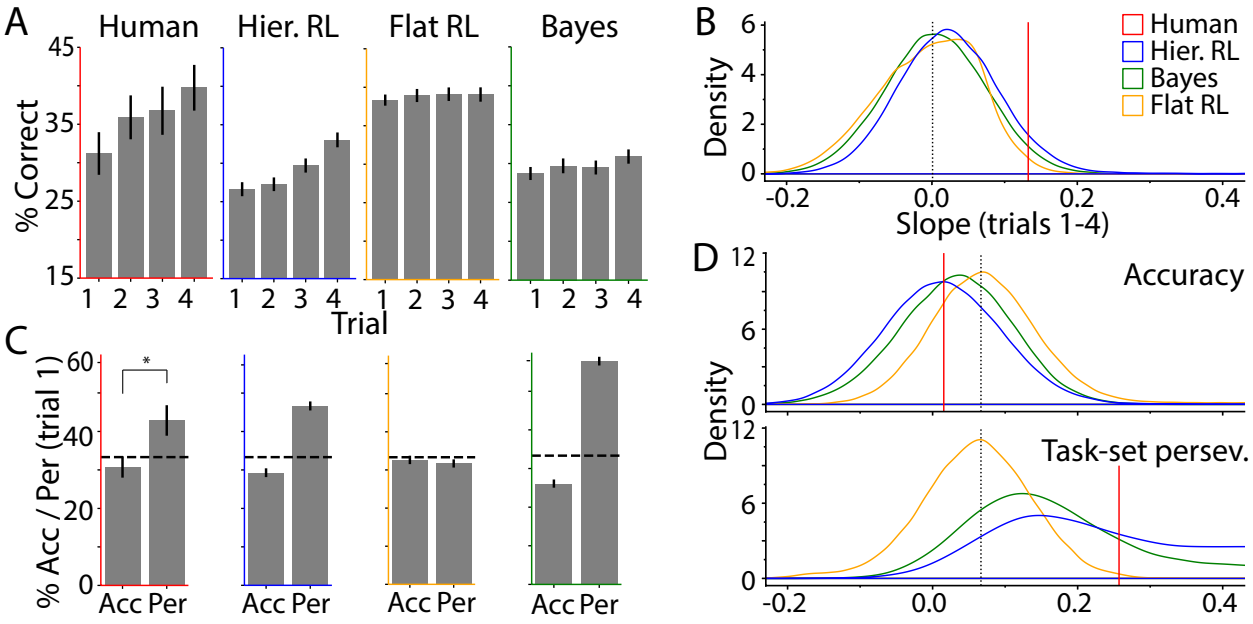


Figure 4.3: A)-B) Participants reactivated task-sets in the hidden-context phase. A) Human performance (left, red) increased over the first four trials following a context switch, even though different stimuli were presented on each trial. The "best" (methods) simulation based on the hierarchical RL model showed qualitatively similar behavior (blue). The effect was absent in the flat RL model (orange), and present but weaker in the Bayesian hierarchical model (green). B) Slopes of the performance increase in part A), as densities over 50,000 simulations per model, with parameters sampled uniformly at random. These densities approximate marginal model likelihoods for the calculation of Bayes factors. The densities of hierarchical RL and hierarchical Bayes were shifted toward larger slopes, making human-like performance slopes more likely. Dotted line indicates chance. C)-D) Task-set perseveration errors in the initial-learning phase. C) Percent correct trials ("Acc") and percent task-set perseveration errors ("Per") on the first trial after a context switch. Humans (left, red): Star denotes significance in repeated-measures t-test. Models: Hierarchical RL and hierarchical Bayes, but not flat RL, qualitatively reproduced human behavior. D) Accuracy and task-set perseveration errors for all simulations, as densities.

RL simulations show no change (simulation details in methods and section "Modeling Behavioral Patterns Jointly").

Human behavior qualitatively matched the predictions of hierarchical RL: Performance increased steadily over the first four trials after a context switch (Fig. 4.3A), evident in the significant correlation between item position (1-4) and performance, $r = 0.19$, $p = 0.048$. This shows that participants recalled previously-learned stimulus-action mappings rather than relearning them, a signature of task-set transfer.

We next assessed quantitatively which of our three candidate models captured this behavior best. We compared the models using Bayes Factors (BF), which we estimated using a method related to Approximate Bayesian Computation (ABC; see methods and SI Appendix; (Sunnaker et al., 2013)). Our method involved simulating synthetic data from each model and estimating the likelihood of human behavior under the simulated data, as illustrated in Fig. 4.3B. Hierarchical RL surpassed both flat RL, $BF = 5.12$, and also hierarchical Bayes, $BF = 1.96$, in model comparison (SI Appendix, table 3). This confirms the qualitative result, showing that human performance in the hidden-context phase was better captured by hierarchical than flat models.

Task-set Perseveration Errors

We showed that hierarchy allowed for positive transfer, enabling participants to reactivate old task-sets. However, hierarchy can also lead to negative transfer: When participants select the wrong task-set, the "correct" action according to that task-set is likely to be incorrect in the current context. We call such errors "task-set selection errors", and focus on a specific subtype, *task-set perseveration errors*. Here, actions are chosen that would have been correct in the previous context, but are incorrect in the current one. Contrary to flat RL, hierarchical models predict task-set perseveration (methods and example in SI Appendix, Fig. 4A), reflected in high proportions of task-set perseveration errors and low initial accuracy (Fig. 4.3C and D).

We tested this prediction on the first trial after each context switch during initial learning, and found that participants were more likely to make task-set perseveration errors than to select correct actions, $t(25) = 2.1$, $p = 0.046$, in accordance with hierarchical model simulations (Fig. 4.3D). Task-set perseveration persisted several trials into the new block, as evident in a logistic regression predicting task-set perseveration errors from trial index ($\beta = -6.83\%$, $z = -9.31$, $p < 0.001$), task-set values ($\beta = -2.43\%$, $z = -1.00$, $p < 0.001$), and action-values ($\beta = -14.03\%$, $z = -8.45$, $p < 0.001$), controlling for block, and specifying random effects of participants.

In summary, the presence of task-set perseveration errors in humans is qualitative evidence for hierarchical processing. Quantitative model comparison supports this conclusion, showing that hierarchical models fit human error patterns better than flat RL (hierarchical vs flat RL: $BF = 14.99$; hierarchical Bayes vs flat RL: $BF = 10.32$; hierarchical RL vs Bayes $BF = 1.40$).

RL Values at Different Levels of Abstraction

Our results so far focused on hierarchical representations in general, showing that participants created, reactivated, and transferred task-sets. We now test predictions that are unique to hierarchical

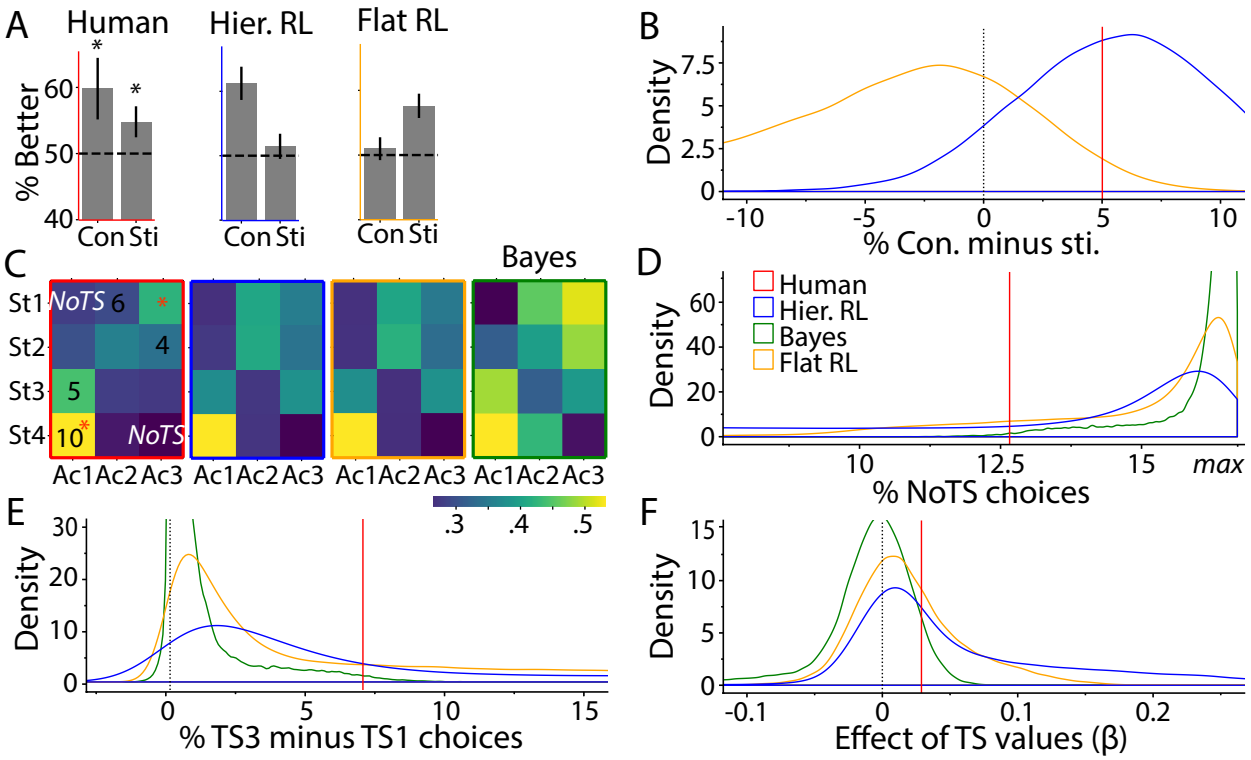


Figure 4.4: Effects of task-set values on behavior. A)-B) Comparison phase. A) Humans (red) performed better for contexts ("Con") than stimuli ("Sti"; "% Better": percentage choosing higher-valued alternative). Stars indicate significant difference from chance (dashed line). The hierarchical RL simulation showed the same qualitative pattern, whereas flat RL showed the opposite. B) Difference between context and stimulus condition, as simulation-based densities. C)-E) Novel-context phase. C) Raw action frequencies. "Ac1-3": actions; "St1-3": stimuli. Humans (red frame): Overlaid numbers show actions-values in TS3, the highest-valued task-set, which was chosen frequently. Red stars indicate actions that were correct in multiple task-sets, also selected frequently. "NoTS" indicates actions that were incorrect in all task-sets, selected rarely. Models: Hierarchical (blue frame) and flat RL (orange frame) were qualitatively similar to humans, hierarchical Bayes (green frame) made different predictions. D) NoTS choices in all simulations. E) Difference between percentage of actions consistent with TS3 and TS1. F) Initial-learning phase. Regression weights predicting performance from task-set values, showing that values at both levels affected performance more in the RL models.

RL, assessing whether participants acquired RL values at the level of task-sets as well as actions.

Task-Set Values Affect Subjective Preference

A classic approach to assess RL values in humans is to investigate subjective preferences (Jocham et al., 2011). To investigate whether participants acquired values at both levels, we thus used a comparison phase, where participants selected their preferred out of two items on each trial. Items were either two contexts or two stimuli—testing task-set and action values, respectively (Fig. 4.2C).

The hierarchical RL model selected contexts based on the task-set values acquired during initial learning, and showed a strong preference for high-valued over low-valued contexts (SI Appendix, Fig. 7A). The flat RL model selected contexts based on average action-values in this context, and showed a much weaker preference (SI Appendix, Fig. 7A). The hierarchical Bayesian model did not track values over contexts and was thus not simulated in this phase. As predicted by hierarchical RL, participants preferred high-valued over low-valued contexts, $t(25) = 2.56$, $p = 0.017$, indicating RL values at the level of contexts. Quantitative model comparison (Fig. 4.4B) strongly favored hierarchical over flat RL, $BF = 1171.65$. For completeness, we also confirmed participants' RL values at the level of stimuli, as predicted by both flat and hierarchical RL, and evident in the preference for high-valued over low-valued stimuli, $t(25) = 2.11$, $p = 0.045$. In conclusion, participants' preferences were best accounted for by the hierarchical RL model.

We next investigated a different model prediction in the comparison phase: The hierarchical RL model takes two steps to retrieve action-values, but only one to retrieve task-set values. This suggests stimulus selection should be slower and noisier than context selection. Flat RL, on the contrary, takes one step to retrieve action-values, but multiple steps to calculate context-values, suggesting the inverse pattern. Humans showed the patterns predicted by hierarchical RL: RTs were numerically slower and performance was significantly worse for contexts than for stimuli (mixed-effects regression, RTs: $\beta = 148.21$, $t(25) = 1.63$, $p = 0.12$, Acc.: $\beta = 0.28$, $z = 2.0$, $p = 0.048$; Fig. 4.4B). Though the effect on RTs did not reach significance here, it was strongly significant in the replication (see SI Appendix, table 1). Quantitative model comparison strongly favored hierarchical over flat RL in terms of accuracy, $BF = 39.64$.

Task-set Values Affect Performance

As explained above, human initial learning was affected by both action-values and task-set values (Fig. 4.1B), in accordance with hierarchical RL. To compare our models in this regard, we calculated the effects of task-set values on performance, using a simplified regression model (see SI Appendix). Supporting our qualitative findings, the hierarchical RL model provided a better fit than value-less hierarchical Bayes, $BF = 6.62$, and crucially, than flat RL, $BF = 1.49$ (Fig. 4.4F).

Task-set Values Affect Generalization

We showed above that participants preferred high-valued over low-valued contexts (SI Appendix, Fig. 7A). We now test whether participants showed similar task-set preferences in the novel-

context phase, that is, when generalizing old knowledge to a new context. For simulations, our hierarchical RL model applied its highest-valued task-set throughout the novel-context phase. The hierarchical Bayes model applied its most reliable task-set. The flat RL model chose actions based on average values (methods).

We labeled each action in the novel-context phase as one of the following: correct in task-set TS3, TS2, TS1, both TS3 and TS1, both TS2 and TS1, or not correct in any task-set (NoTS). Despite the lack of feedback, human participants showed consistent preferences for certain stimulus-action combinations over others (Fig. 4.4C; see SI Appendix, Fig. 2 for heatmaps of task-set values). They chose NoTS actions less often than other actions, controlling for the frequency of each category, $t(25) = 2.24$, $p = 0.034$. Mappings shared between multiple task-sets (TS2 and TS1; TS3 and TS1) were more frequent than mappings that only occurred in one task-set (TS1, TS2, TS3), controlling for chance level, $t(25) = 2.83$, $p = 0.0091$. This confirms that participants reused old task-sets for new contexts, in accordance with our findings in the hidden-context phase, and prior literature (Collins & Frank, 2013). Quantitative model comparison confirmed that the number of NoTS choices was captured better by hierarchical RL than by flat RL, $BF = 1.78$, or hierarchical Bayes, $BF = 45.60$.

Highlighting the role of task-set values, hierarchical RL predicted more actions from the highest-valued TS3 than from the lowest-valued TS1, and a greater difference between the two than flat RL or hierarchical Bayes (Fig. 4.4E). Humans showed the same pattern, selecting more TS3 than TS1 actions, $t(25) = 2.58, p = 0.016$. Bayes Factors confirmed that this difference was captured better by hierarchical RL than flat RL, $BF = 1.59$, or hierarchical Bayes, $BF = 32.01$. Taken together, our hierarchical RL model captured both the reuse of old task-sets in new contexts, and the preference for high-valued over low-valued task-sets.

Modeling Behavioral Patterns Jointly

Human behavior followed predictions of hierarchical RL qualitatively, and Bayes Factors confirmed quantitatively that this model fit better than the competing ones. However, we treated each behavioral measure independently. We next sought to confirm that it was possible to obtain all behavioral results simultaneously based on a single set of parameters. To this end, we chose one "best" set of parameters for each model (methods), and showed the behavior of this simulation side-by-side with humans, for each behavioral measure. As expected, neither flat RL nor hierarchical Bayes, replicated all qualitative patterns in Figs. 4.3A, 4.3C, 4.4A, and 4.4C. But importantly, a single set of parameters could capture all qualitative patterns in the hierarchical model. Note that because parameters were not obtained through model fitting, behavior can deviate quantitatively from human data.

4.3 Discussion

The goal of the current study was to assess whether human flexible behavior could be explained by hierarchical reinforcement learning (RL), i.e., the concurrent use of RL at different levels of

abstraction (Botvinick et al., 2009; Diuk, Schapiro, et al., 2013). We proposed a hierarchical RL model that acquires low-level strategies—or "task-sets"—using RL, and also learns to choose between these task-sets using RL. We contrasted this model with a flat RL model to highlight the unique contribution of hierarchy, and to a hierarchical Bayesian model to highlight the contribution of a hierarchical value representation.

Our hierarchical RL model predicted unique patterns of behavior in a variety of situations. To assess whether humans employed hierarchical RL, we designed a context-based learning task in which multiple subtasks tested these predictions. Indeed, participants' behavior followed the predictions in all subtasks. The first prediction was that participants would create hierarchical representations. Several independent results supported this claim, including asymmetric switch costs, task-set perseveration errors, and task-set reactivation. These results could not be accounted for by the flat RL model, but were also compatible with the hierarchical Bayesian model.

To address the unique predictions of hierarchical RL, we sought evidence of hierarchical values. Hierarchical RL predicts value-based (1) context preferences, (2) performance differences between contexts, and (3) generalization in new contexts. Human behavior showed the predicted patterns: (1) When asked to pick their preferred contexts, participants selected higher-valued ones more often. This suggests that they had formed abstract task-set values, in addition to low-level action-values. Participants also performed better when choosing between high-level contexts than low-level stimuli, in accordance with the "blessing of abstraction" (Gershman, 2017b; Goodman et al., 2011; Kemp et al., 2007). (2) Task-set values affected performance, with better performance of higher-valued task-sets. This shows that hierarchical representation can explain performance differences between contexts. (3) When faced with a new context, participants reused previous task-sets, preferring higher-valued over lower-valued ones. This suggests that task-set values guided generalization of old knowledge to new situations.

In summary, human behavior showed all qualitative patterns predicted by hierarchical RL. To quantify the differences with hierarchical Bayes and flat RL, we conducted formal model comparison using Bayes Factors. Because marginal model likelihoods were intractable, we approximated them using simulations, similar to (M. D. Lee, 2011; Steingroever et al., 2016; Sunnaker et al., 2013). Bayes Factors instantiate an implicit Occam's razor that accounts for differences in model complexity, such as the larger number of parameters in the hierarchical models compared to flat RL, differences in the functional form of each model, and differences in parameter spaces (MacKay, 1992; Steingroever et al., 2016). In this way, Bayes Factors implement a more comprehensive tradeoff between parsimony and goodness-of-fit than traditional methods.

In our paradigm, Bayes Factors showed that hierarchical RL and hierarchical Bayes captured behavioral aspects of *hierarchy* better than flat RL (e.g., task-set reactivation, task-set perseveration), whereas flat RL and hierarchical RL captured *value-based* aspects better (e.g., value-based generalization, effects of values on performance). Furthermore, hierarchical RL uniquely captured the influence of two sets of values on behavior. Overall, Bayes Factors favored the hierarchical RL model over flat RL and hierarchical Bayes. Based on this quantitative confirmation, we next asked whether all results could be jointly observed when simulating the hierarchical RL model with a single set of parameters, to confirm that different parameters were not responsible for different behaviors. We used simulation summary statistics to identify a "best" set of parameters for each

model. Only the hierarchical RL simulation qualitatively replicated all human behaviors, but not flat RL or hierarchical Bayes. This shows that seemingly different behaviors, including trial-and-error learning (initial-learning phase), "inference" of missing information (hidden-context phase), subjective preferences (comparison phase), and generalization (novel-context phase), can all be explained in the same overarching hierarchical RL framework.

Note that we have not explored the full space of possible models. In particular, it would be possible to construct a hierarchical Bayesian model that tracks task-set and action-values rather than their reliability, but uses Bayesian inference rather than RL to perform updates. This model might capture the behavioral patterns we observed here. Indeed, our results show evidence for humans' ability to track values at multiple levels of hierarchy in support of generalizable behavior, but do not speak directly to the exact update process. However, we favor the hierarchical RL formulation of such updates because it is inspired by a rich literature on brain circuits that makes its implementation plausible, and because it is algorithmically simple, with the ability to account for complex cognitive processes.

Many computational models have addressed cognitive hierarchy. How are they related to our model? One important class of hierarchical models is purely Bayesian (Solway et al., 2014; Tenenbaum et al., 2011; Tomov et al., 2019). These models aim to explain, on a computational level of analysis (Marr, 1982), the fundamental purpose of hierarchy for cognitive agents. Our model, on the other hand, is algorithmic, like many pure-RL models: It aims to describe dynamically which cognitive steps humans take when they make decisions in complex environments. Our model is also inspired by the structure of human neural learning circuits (G. Alexander et al., 1986; W. H. Alexander & Brown, 2015; Badre, 2008), thereby extending to the implementational level of analysis.

Some models of hierarchical cognition are method hybrids: Some combine Bayesian inference at the abstract level with RL at the lower level (Collins & Koechlin, 2012; Frank & Badre, 2012). Other, resource-rational models, combine Bayesian principles of rationality with cognitive constrains (Lieder & Griffiths, 2019). Frank and Badre (Badre & Frank, 2012; Frank & Badre, 2012) proposed a hybrid model that uses Bayesian inference to arbitrate between multiple types of hierarchy and flat RL. In general, hybrid models assume a role for Bayesian inference at higher levels of hierarchy, contrary to our hierarchical RL model. This is an important difference: Hierarchical RL mimics a form of inference (for example, identifying the latent task-set at the beginning of a block; SI Appendix, results 2.1), but cannot do it optimally. It is an important direction for future research to identify whether human behavior is suboptimal in the same way.

Computational models at different levels of analysis (Marr, 1982) are not mutually exclusive. Bayesian inference offers a perspective based on optimality, but it is often intractable and approximations are computationally expensive. RL, on the other hand, uses values to approximate expectations instead of calculating them exactly. Because of its relative computational simplicity, and because it is biologically well supported, RL has often been used as an algorithmic and implementational model. Recent research showed that a neural network implementing hierarchical RL approximated the results of Bayesian inference (Collins & Frank, 2013). In other words, hierarchical RL might allow for optimal behavior using simpler computations.

Hierarchical RL was initially proposed in AI (Konidaris, 2019; Vezhnevets et al., 2017). A

number of AI algorithms has recently been used to model human cognition as well (Momennejad et al., 2017; Ribas Fernandes et al., 2011; Sutton et al., 1999; Wang et al., 2018), showcasing how intertwined the two fields have become (Collins, 2019; Lake et al., 2017; Sutton & Barto, 2017). Nevertheless, most hierarchical RL algorithms in AI focus on hierarchy over the time scale of choices (*temporal abstraction*, e.g., breaking up long-term goals into sub-goals). Our hierarchical model, in contrast, focuses on *choice abstraction* (i.e., allowing choice at the level of task-sets and motor actions), a still rare approach in AI (Vezhnevets et al., 2020).

To conclude, classic RL has been a powerful model for simple decision making in animals and humans, but it cannot explain hallmarks of intelligence like flexible behavioral change, continual learning, generalization, and inference of missing information. Recent advances in AI have proposed hierarchical RL as a solution to a number of such shortcomings, and we found that human behavior showed many signs of hierarchical RL, which were captured better by our hierarchical RL model than competing ones.

There is no debate that achieving goals and receiving punishment are some of the most fundamental motivators that shape our learning and decision making. Nevertheless, almost all decisions humans face pose more complex problems than what can be achieved by flat RL. Structured hierarchical representations have long been proposed as a solution to this problem, and our hierarchical RL model uses only simple RL computations, known to be implemented in our brains, to solve complex problems that have traditionally been tackled with intractable Bayesian inference. This research aims to model complex behaviors using neurally plausible algorithms, and provides a step toward modeling human-level, everyday-life intelligence.

4.4 Methods

Participants

We tested three independent groups of participants, with approval from UC Berkeley's institutional review board. All were university students, gave written informed consent, and received course credit for participation.

The pilot sample had 51 participants (26 women; mean age$\pm$sd: 22.1 ± 1.5), 3 of whom were excluded due to past or present psychological or neurological disorders. Due to a technical error, data were not recorded in the comparison phase for this sample. The second and main sample had 31 participants (22 women; mean age$\pm$sd: 20.9 ± 2.1), 4 of whom were excluded due to disorders, and one of whom was excluded because average performance in the initial-learning phase was below 35% (chance is 33%). We added the mixed testing phase for this sample. The third sample had 32 participants (15 women; mean age$\pm$sd $= 20.8 \pm 5.0$), 2 of whom were excluded due to disorders. Five participants did not complete the experiment and were excluded when data was missing. The task was minimally adapted for EEG data collection. All statistical tests were conducted in all samples (SI Appendix, table 1 and Fig. 1), and the SI Appendix discusses sample differences in detail.

Task Design

Participants first received instructions and underwent the initial-learning phase of the task. The purpose of initial learning was for participants to acquire distinct task-sets, i.e., specific stimulus-action mappings for each context. We also used the initial-learning phase to test for the effects of action-values and task-set values on performance, and to assess errors types predicted by hierarchical RL.

In the beginning, participants were instructed to "feed aliens to help them grow as much as possible". A tutorial with instructed trials followed, then participants practiced a **simplified task** without contexts: On each trial, participants saw one of four stimuli and selected one of three actions by pressing J, K, or L on the keyboard (Fig. 4.2A). Feedback was given in form of a measuring tape whose length indicated the amount of reward. Correct actions produced consistent long (mean=5.0) and incorrect actions short tapes (mean=1.0, Fig. 4.2). When no action was selected, participants were reminded to respond faster next time, and the trial was counted as missed. Participants received 10 training trials per stimulus (40 total), with a maximum response time of 3,000 msec. Order was pseudo-randomized such that each stimulus appeared once in four trials, and the same stimulus never appeared twice in a row.

The **initial-learning phase** had the same structure as training, but stimuli were presented in one of three contexts, each with a unique mapping between stimuli and actions (Fig. 4.2B). The context remained the same for a block of 52 trials. At the end of a block, a context change was explicitly signaled, before the next block began with a new context. Participants went through 9 blocks (3 per context) for a total of 468 trials. Participants needed to respond within 1.5s, then received reward. Rewards varied between 2-10 for correct actions (Fig. 4.2B); rewards for incorrect actions remained 1. We chose these numbers to maximize differences between contexts, while controlling for differences between stimuli and actions. The **hidden-context phase** was identical to initial learning and participants knew they would encounter the same contexts as before, but this time, they were "hidden" (Fig. 4.2C). There were 9 blocks with 10 trials per stimulus per block (360 total). Context switches were signaled. The purpose of the **comparison phase** was to assess participants' subjective preferences for contexts and stimuli, as estimates of their task-set and action-values. Participants were shown two contexts (context condition), or two stimuli in the same context (stimulus condition), and selected their preferred one (Fig. 4.2C). Participants saw each of three pairs of contexts 5 times, and each of 18 pairs of stimuli 3 times, for a total of $15 + 198 = 213$ trials. Participants had 3 sec to respond.

The purpose of the **novel-context phase** was to probe generalization, specifically the reuse of old task-sets in a new context. This phase was identical to the initial-learning phase, except that it introduced a new context in extinction, i.e., without feedback (Fig. 4.2C). Participants received 3 trials per stimulus (12 total). The purpose of the final **mixed phase** was to probe switch costs, assessing whether switching contexts was more costly than switching stimuli, indicating hierarchical representation. The mixed phase was identical to the initial-learning phase, except that contexts as well as stimuli could change on every trial. Participants received 3 blocks of 84 trials (252 total), each with 7 repetitions per stimulus-context combination. To alleviate carry-over effects and forgetting between test phases, we interleaved them with **refresher blocks**, shorter

120-trial versions of the initial-learning phase. More details on task design are provided in the SI Appendix,.

Computational Models

We will address in turn how each model behaves in each phase. During **initial learning**, the flat RL model implemented classic model-free (*delta-rule*) RL (Sutton & Barto, 2017): It treated every combination of a context and a stimulus as a unique state, and learned one RL value for each state and action, as visualized in SI Appendix, Fig. 3A. Using main text notations, values were updated based on $Q_{t+1}(a|s,c) = Q_t(a|s,c) + \alpha\ (r - Q_t(a|s,c))$, and actions were selected based on $p(a|s,c) = \frac{exp(Q(a|s,c))}{\sum_{a_i} exp(\beta\ Q(a_i|s,c))}$.

The flat RL model acquired 36 action-values, based on three parameters α, β, and f, whereas the hierarchical RL model acquired 9 task-set-values and 36 action-values (45 total), with six free parameters α_a, α_{TS}, β_a, β_{TS}, f_a, and f_{TS} (equations in main text). SI Appendix, (Fig. 3) visualizes the difference between both models, and SI Appendix, Fig. 4 explains hierarchical RL behavior trial-by-trial. The forgetting parameters $f \in [f_a, f_{TS}]$ captured value decay in both models: $Q_{t+1} = (1-f)\ Q_t + f\ Q_{init}$.

The hierarchical Bayes model also learned task-sets, but acquired their action-values based on correct-incorrect rather than continuous feedback: $Q_{t+1}(a|s,TS) = Q_t(a|s,TS) + \alpha\ (correct - Q_t(a|s,TS))$. The main difference to hierarchical RL was the selection of task-sets: The Bayesian model chose task-sets based on estimated reliability rather than task-set values, using Bayes theorem to obtain task-set reliabilities: $p_{t+1}(TS|c) = \frac{p(r|s,TS,a)p_t(TS|c)}{p(r|s,a)}$, with $p(r|s,TS,a) = Q(a|s,TS)$. Another difference was that hierarchical RL updated $Q(TS|c)$ only for the chosen task-set, whereas hierarchical Bayes kept $p(TS|c)$ up-to-date at all times for all task-sets (Collins & Koechlin, 2012; Donoso et al., 2014).

Q-values for both models were initialized at the expected reward of chance performance, $Q_{init} = 1.67$. The subsequent testing phases started from the Q-values obtained at the end of initial learning.

In the **hidden-context phase**, contexts were not shown, such that models could not directly reuse acquired values that depended on contexts (flat RL: $Q(a|c,s)$; hierarchical RL: $Q(TS|c)$; Bayes: $p(TS|c)$). All models instead initialized these values at Q_{init} after each context switch, and then relearned them using the same update equations as before. For flat RL, this resulted in learning an entire new policy $Q(a|c,s)$. For hierarchical models, only high-level information ($Q(TS|c)$ for RL, $p(TS|c)$ for Bayes) had to be relearned, but not action values $Q(a|s,TS)$. This ability to transfer learned values is one of the main advantages of hierarchy.

For the **comparison phase**, we only simulated RL models because the Bayesian model does not provide values at the level of contexts. To select between two stimuli, RL models first computed the "state value" (Sutton & Barto, 2017) of each, based on action-values: $V(c,s) = max_a\ Q(a|c,s)$ (flat RL) and $V(c,s) = \max_a\ Q(a|s,TS)\ p(TS|c)$, where $p(TS|c) = softmax(Q(TS|c))$ (hierarchical RL). Models then selected one stimulus based on a softmax over the two state values. To select between contexts, the hierarchical model repeated the same computation for task-set values:

$V(c) = max_{TS}\ Q(TS|c)$. The flat model, lacking task-set values, used averages over action-values to estimate context preferences on-the-fly: $V(c) = mean_s\ V(c,s)$.

In the **novel-context phase**, models were faced with a context for which they had not learned values. Flat RL used averages over previous action-values to choose: $Q(a|c_{new},s) = mean_c\ Q(a|c,s)$. Hierarchical RL [Bayes] applied the previously highest-valued [most reliable] task-set: $Q(TS|c_{new}) = max_c\ Q(TS|c)$ [$p(TS|c_{new}) = max_c\ p(TS|c)$].

Model Comparison

The Bayes Factor BF quantifies the support for one model M_1 over another model M_2 by assessing the ratio between their marginal likelihoods, $BF = \frac{p(data|M_1)}{p(data|M_2)}$. $BF > 1$ provides evidence for M_1. Marginal model likelihoods represent the probability of the data under the model, marginalizing over model parameters θ: $p(data|M) = \int p(data|M,\theta)\ p(\theta)\ d\theta$.

For each model, we simulated datasets by drawing model parameters θ uniformly at random. Due to uniform sampling, $p(\theta)$ is equal for all θ, such that the empirical distribution over simulations approximates the marginal likelihood. To obtain Bayes Factors, we computed the same summary statistics s_m as for humans for each individual simulation (e.g., performance slope in hidden-context phase). We estimated model densities $\hat{s}_m$ based on a large number of simulations. We obtained marginal model likelihoods as the probability of the human summary statistic s_h under the model, $p(s_h|\hat{s}_m)$. Bayes Factors are given by $BF = \frac{p(s_h|\hat{s}_{m1})}{p(s_h|\hat{s}_{m2})}$.

We drew parameters uniformly at random in a range allowing as broad coverage of possible behavior as possible: $0 < \alpha_a, \alpha_{TS}, f_a, f_{TS} < 1$ and $1 < \beta_a, \beta_{TS} < 20$. Each synthetic dataset consisted of 26 agents simulated on the exact same inputs received by the 26 participants, such that the noise in the synthetic statistics was identical to the one in the human dataset. We simulated 50,000 datasets for each model to assure convergence of the density estimates.

We presented one example datasets for each model in the bar graphs of figures 4.3A, 4.3C, 4.4A. These datasets were obtained by first selecting all of the 50,000 model simulations that fell within a certain range of human behavior for *all* summary statistics (50%-150% for flat and hierarchical RL; 10%-190% for hierarchical Bayes). We then simulated one new dataset per model based on the median parameter values of the selected models. The supplementary methods provide a detailed discussion of our model comparison method and selection of the example datasets.

4.5 Data Availability

All data for this study will be made available for researchers only through the NIMH NDA data base. Analysis and modeling code is available on github: https://github.com/MariaEckstein/TaskSets.

4.6 Supplementary Methods

Participants

We tested our paradigm in three independent samples of participants, replicating most of our major findings. The three versions of the task differed in a few ways: The first sample did not receive the mixed test, and an error in data collection led to the loss of the comparison test data. The second sample received the same task as the first, expect for the addition of the mixed phase. The largest changes—though overall still minor—were necessary for the third sample to enable EEG data collection. The changes concerned mostly timing parameters. Inter-trial intervals were drawn uniformly between 500 and 1.000 milliseconds, in 50 millisecond increments (fixed at 250 milliseconds in previous versions). Intervals before feedback presentation were drawn uniformly between 400 and 800 milliseconds. Testing took place under different conditions, notably in an EEG lab with dimmed light and using a different computer and monitor. Lastly, experimental sessions lasted for 2 hours to accommodate for setting up EEG electrodes on participants' scalps. We chose sample 2 to present in the main text because it is the first sample that includes data from all phases.

Learning curves were qualitatively similar across the samples (suppl. Fig. 4.5). Overall performance differed slightly, albeit non-significantly, and the large majority of statistical tests replicated across samples (suppl. table 4.6). In other words, the results reported in the main text were mostly robust to small changes in task design.

The most notable difference concerned overall task performance of the three samples. EEG participants performed slightly, albeit non-significantly, better, showing qualitatively steeper learning curves (suppl. Fig. 4.5), numerically better performance and fewer initial selection errors in the initial-learning phase, larger performance increases in the hidden-context phase, and better overall performance in the comparison phase (supple. table 4.6). It is unclear why performance seemed slightly better in the EEG sample than in the other samples. The most likely reasons include increased attention due to the more involved EEG procedure and changes in timing parameters that slowed the task down slightly.

Table 4.6 also suggests that task-set values might have had slightly larger effects in the EEG sample than in the other two: In the mixed phase, the effect size of RT switch costs was numerically twice as large, as were the effect size of correlation in the novel-context phase, and differences between stimulus and context condition in the comparison phase, for both accuracy and response times.

Based on these exploratory findings, future research could explore links between task performance and task-set structure.

One test in specific differed between samples: "More initial selection errors than accurate trials" was statistically significant in the main (second) sample, but not in the first and third (EEG). The most likely explanation were overall performance differences. It is only possible to conduct more mistakes than correct actions when many mistakes are committed. We do not think that this difference in outcomes between samples invalidates any of our claims about hierarchical processing and RL.

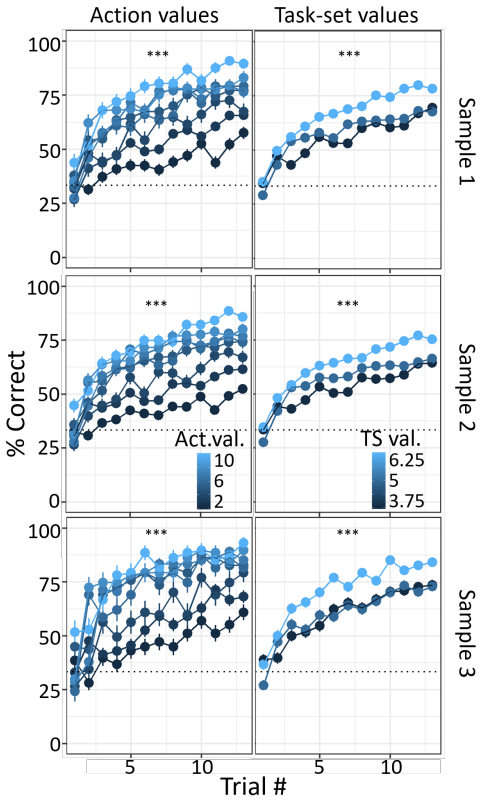


Figure 4.5: Learning curves of all three samples.

Statistical tests in the three samples.

	Sample 1 No comparison / mixed	**Sample 2** Main sample	**Sample 3** Adapted for EEG
Final sample size (excluded, number and reason)	48 (3 disorder)	26 (4 dis., 1 chance perform.)	30 (2 dis., up to 5 stop early)
Mean accuracy during initial learning (sd)	59.8% (9.3%)	55.8% (9.3%)	63.2% (11.5%)
Hierarchical Representation			
Initial-learning phase			
Effect action-val. on perf.	$\beta = 0.28 (std = 0.04), z = 8.02, p < 0.001$	$\beta = 0.38 (std = 0.05), z = 7.65, p < 0.001$	$\beta = 0.40 (std = 0.05), z = 8.34, p < 0.001$

Effect task-set val. on perf.	$\beta = 0.11 (std = 0.04), z = 3.05, p = 0.002$	$\beta = 0.20 (std = 0.05), z = 4.00, p < 0.001$	$\beta = 0.16 (std = 0.05), z = 3.31, p < 0.001$
Interaction between both	$\beta = -0.02 (std = 0.01), z = -2.80, p = 0.005$	$\beta = -0.04 (std = 0.01), z = -3.70, p < 0.001$	$\beta = -0.03 (std = 0.01), z = -3.49, p < 0.001$
Mixed phase			
Asymmetric RT switch costs	NA	$t(25) = 3.47, p = 0.002, d = 0.30$	$t(24) = 4.46, p < 0.001, d = 0.65$
Hidden-context phase			
Reactivating task-sets	$r = 0.12, t = 1.72, p = 0.087$	$r = 0.19, t = 2.0, p = 0.048$	$r = 0.34, t = 3.92, p < 0.001$
Initial-learning phase			
Task-set perseverance errors $>$ accurate actions	$t(47) = 0.73, p = 0.47, d = 0.19$	$t(25) = 2.1, p = 0.046, d = 0.71$	$t(29) = 0.44, p = 0.67, d = 0.15$
RL values at different levels of abstraction			
Comparison phase			
Stimulus accuracy > 0.5	NA	$t(25) = 2.11, p = 0.045, d = 0.83$	$t(28) = 2.58, p = 0.009, d = 0.96$
Context accuracy > 0.5	NA	$t(25) = 2.56, p = 0.017, d = 1.00$	$t(28) = 5.61, p < 0.001, d = 2.08$
Context $>$ stimulus acc.	NA	$\beta = 0.28$ $(std = 0.14)$, $z = 1.98$, $p = 0.048$	$\beta = 0.60$ $(std = 0.11)$, $z = 5.43$, $p < 0.001$
Context $<$ stimulus RT	NA	$\beta = 148.21$ $(se = 91.14)$, $t(25) = 1.63$, $p = 0.12$	$\beta = 384.90$ $(se = 55.90)$, $t(27) = 6.89$, $p < 0.001$
Novel-context phase			
TS3 $>$ TS1	$t(47) = 3.81, p < 0.001, d = 0.96$	$t(25) = 2.58, p = 0.016, d = 0.87$	$t(26) = 2.04, p = 0.052, d = 0.71$
TS2 $>$ TS1	$t(47) = 3.19, p = 0.003, d = 0.37$	$t(25) = 1.93, p = 0.065, d = 0.59$	$t(26) = 1.18, p = 0.25, d = 0.38$
TS3and1 $>$ TS2and1	$t(47) = 3.14, p = 0.003, d = 0.59$	$t(25) = 1.37, p = 0.18, d = 0.30$	$t(26) = 3.82, p < 0.001, d = 1.07$

Task Design

Additional Task Information, Including Randomization

The association between task-sets and contexts was randomized between participants; for example, the winter context might be associated with the highest-valued TS3 for participant 1, but with the lowest-valued TS1 for participant 2. Similarly, the association between stimuli and alien characters, and between actions and items was randomized. The position of actions, i.e., the keyboard key associated with it, were not randomized between trails. In other word, the bed, umbrella,

and backpack always appeared in the same position on the screen within one participant, but they differed between participants.

We randomized the mapping between contexts (e.g., winter) and their role (e.g., highest-valued task-set) to avoid systematic biases in participant response, e.g., due to the semantics of different objects. We also conducted basic analyses to confirm that specific objects did not lead to systematic response biases.

Trial order was randomized in the following way for the initial-learning phase and hidden-context phase: Each phase consisted of several blocks. Each block included a single context (season), but all stimuli (aliens). The order of context blocks within each phase was pseudo-randomized such that each context block appeared once within a macro-block of three blocks, and the same context never appeared twice in a row. Context changes were signaled explicitly, even though changes in the context were visually very salient. This was done to avoid mistakes based on uncertainty about the current context. Both the initial-learning phase and the hidden-context phase consisted of three macro-blocks (nine blocks total).

Stimulus order within each block was pseudo-randomized in a similar way: A mini-block consisted of the four stimuli randomized in order, and mini-blocks were combined such that the same stimulus never appeared twice in a row. This randomization ensured that each stimulus was presented equally often in each position across a block. Thirteen mini-blocks formed one block for the initial-learning phase, and ten for the hidden-context phase, for a total of 9 (blocks) * 13 (mini-blocks) * 4 (stimuli) = 468 trials in the initial-learning phase, and 9 * 10 * 4 = 360 trials in the hidden-context phase.

The novel-context phase only contained a single block because it introduced a single new context (rainbow). This block consisted of 3 mini-blocks, with stimuli randomized as before, for a total of 3 (mini-blocks) * 4 (stimuli) = 12 trials. No feedback was given. The low number of trials was chosen to limit the risk of participants disengaging in the absence of feedback.

The mixed phase was structured slightly differently. It consisted of mini-blocks of 12 trials (one for each combination of stimuli [4] and contexts [3]). Trial order was randomized within each mini-block. Seven mini-blocks were combined into one block, and self-paced breaks separated three blocks in total, for a total of 3 (blocks) * 7 (mini-blocks) * 12 (items per mini-block) = 252 trials. The correct mappings between contexts, stimuli, and actions were the same in the mixed phase as before during initial learning and in the hidden-context phase, and participants received the same kind of feedback as before. The only difference was that contexts were no longer presented blockwise, and both stimuli and contexts were allowed to switch on every trial.

The comparison phase reused the same objects as before, but presented participants with a different task: Instead of selecting an action for a given stimulus (and context), participants saw two different stimuli (and a context) on the screen and had to select their preferred one, via button press ("stimulus condition"). The context was always the same for two stimuli that were presented together, to facilitate the task for participants as well as choice analysis.

We counted a trial as correct in this test when participants chose the stimulus that had led to larger reward during initial learning (larger action-value). For example, presented with the red and the purple alien in the rainy season, a correct choice would be to pick the red alien because it had led to a reward of 7, and not the purple alien, which had led to a reward of 2 (Fig. 2B).

The context condition was similar to the stimulus condition, except that participants saw just two contexts without any stimuli, and selected their preferred one. We counted an action as correct in this condition when participants selected the context with the larger average reward (task-set value), as shown in 1B). We used the stimulus condition of this phase to test for the formation of action-values in our participants, and the context condition to test for task-set values.

Trial order in the comparison phase was randomized using a block structure like before. Each block in the context condition consisted of all pairs of two contexts, i.e., 2 choose 3 [contexts] = three trials. Each block in the stimulus condition consisted of 2 choose 2 [stimuli] = six trials for each context, i.e., 6 (pairs) * 3 (contexts) = 18 trials in total. Trials were randomized within each block, and blocks were combined such that the same trial did not repeat twice in a row. The context condition had five blocks (5 * 3 = 15 trials total) and the stimulus condition 3 (3 * 18 = 54 trials total).

In addition to the context and stimulus condition just described, we also tested an item condition (presenting each pair of two items together), a "pure" stimulus condition (only stimuli, without contexts), and a "mixed" stimulus condition (two different stimuli, with two different contexts). These conditions were not of interest and presented after the context and stimulus condition.

Task Timing

We limited response times to 1.5 seconds. Multiple considerations went into this decision. First, this is a usual task timing for most reinforcement learning experiments, and keeping the timing similar allows for comparison between them. Another, more pragmatic, reason was to keep the experiment within 60 minutes to limit participant fatigue, while ensuring a sufficient number of trials in each phase. Last, we aimed to motivate participants to employ reinforcement learning rather than cognitive control or effortful strategizing, which require more time.

Details about Task-Sets

The task-sets were constructed such that there was only one correct action for each stimulus in each context, but the same action could be correct for multiple stimuli. E.g., the bed was the only correct action for the yellow alien in the rainy context in the example shown in 2B. Selecting the bed for the yellow alien in this context led to a measuring tape of length 3. The backpack and the umbrella were both incorrect, and selecting either of these led to a tape of length 1. In the same context, the bed was also the correct response for the purple alien, for which the reward was a tape of length 2.

To obtain action-values and task-set values in Fig. 2B, we assessed the average rewards for a correct response. For example, choosing the backpack for the red alien in the rainy context was rewarded with measuring tapes of length 7$\pm$noise, averaging out to an expected reward of 7. Similarly for task-set values, we averaged the rewards of correct actions in each task-set (Fig. 2B). When we indicate "higher-valued" task-sets, we refer to task-sets that have larger task-set values thus calculated.

The heatmaps in suppl. Fig. 4.6 show the three task-sets visually. The information is identical to what is presented in 2B. Task-set values are shown side-by-side with human raw action frequencies in the novel-context phase (replicated from Fig. 4C).

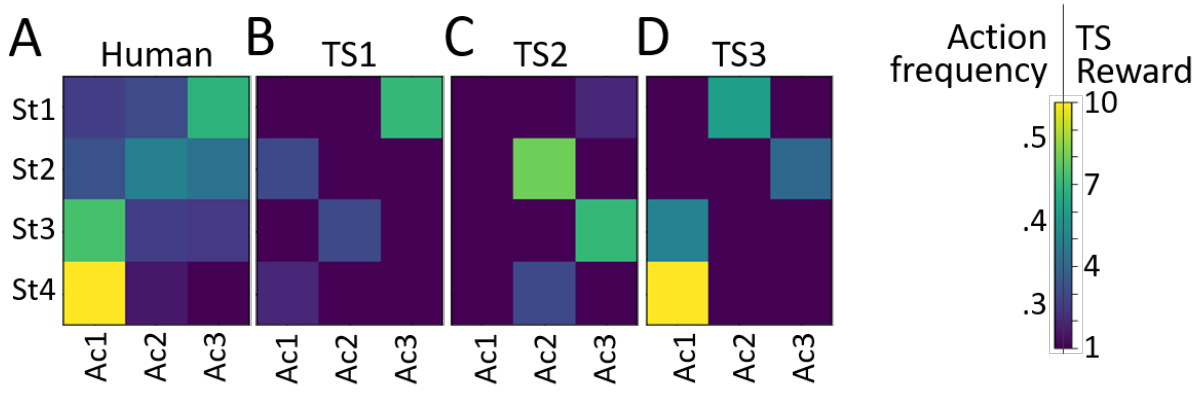


Figure 4.6: A) Raw human action probabilities in the novel-context phase. B)-D) Action-values of the three task-sets.

Regression Models in Humans and Simulations

To assess the effects of action-values and task-set values on performance, we calculated regression models predicting performance from both, and assessed their respective regression weights. To analyze human data (Fig. 1B), we ran a mixed-effects regression model that controlled for a range of other factors as well (e.g., block, item position), as explained in the main text. To estimate model likelihoods, we had to get a similar measure from our computational models. Nevertheless, running the mixed-effects model we used for humans in all 150,000 model simulations was infeasible due to the computational demands. We therefore ran a simpler regression model in simulations, predicting performance from just action-values and task-set values. We present the regression weights of task-set values obtained from these models in the model distributions (Fig. 4F). For model comparison only, we re-ran the simplified regression model on human data as well; the results of the simplified model are shown in Fig. 4F (red line), and were used to calculate the model likelihoods for this measure.

Computational Models

Values in Flat and Hierarchical RL

The hierarchical RL model had nine task-set values, which specify the value of each task-set (3) for each context (3). The number of three task-sets was chosen to accommodate for the three

contexts, and is consistent with previous research on how many task-sets humans entertain in parallel (Donoso et al., 2014).

In addition, the hierarchical RL model contains 3 (task-sets) * 4 (stimuli) * 3 (actions) = 36 action-values, specifying the value of each action for each stimulus, in each task-set. The flat RL model only contains 3 (contexts) * 4 (stimuli) * 3 (actions) = 36 values in total.

Visualization of Flat and Hierarchical RL

Flat RL learned independent action-values for each context-stimulus-action combination, which can be visualized in a single "flat" value table (suppl. Fig. 4.7, left). Hierarchical RL learned values for each context-task-set combination and for each task-set-stimulus-action combination, therefore requiring two separate value tables (suppl. Fig. 4.7, right).

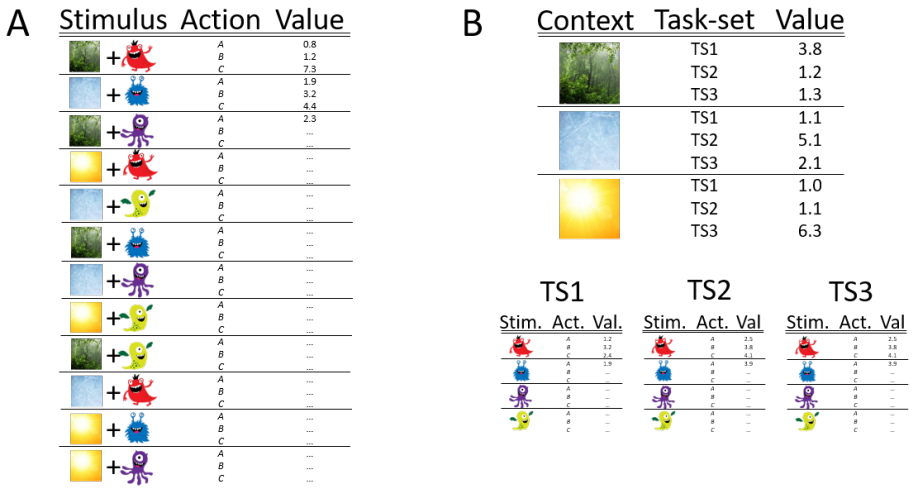


Figure 4.7: Visualization of learning in flat and hierarchical RL. A) Flat RL learns an "exhaustive" value table, treating each combination of context and stimulus independently from each other. B) Hierarchical RL learns distinct "task-sets", which contain stimulus-action values or "low-level values" (bottom). Task-sets are associated with contexts using a table of task-set or "high-level" values (top).

Trial-By-Trial Behavior of a Hierarchical RL Agent

To shed light on the processes that underlie trial-by-trial choices, we zoomed in on the behavior of a hierarchical RL agent in the initial-learning phase (suppl. Fig. 4.8). Agent behavior showed several interesting patterns, for example better performance in high-valued than low-valued contexts, just like humans. Unlike in human participants, we were able to assess the RL process directly in the simulated agent, allowing us to investigate the precise dynamics that gave rise to this pattern.

The most striking behavioral differences between contexts arose for task-set selection: While the highest-valued context was in place (suppl. Fig. 4.8A, red), the agent selected the same (red) task-set throughout. During the lowest-valued context (green), on the other hand, the agent circled through all three task-sets inconsistently. An intermediate amount of task-set switching occurred in the middle-valued context (blue). Consistent task-set selection therefore went hand-in-hand with better performance. The reason for this was that consistent task-set selection allowed the agent to use feedback maximally efficiently: Consistent task-set selection means that every feedback is used to update the action-values of the same task-set, which therefore quickly turns into an optimal strategy for this context (suppl. Fig. 4.8B, action-values inside the red box). In contrast, inconsistent task-set selection was related to poor performance because it led to suboptimal use of feedback: Action-value updates were applied to all task-sets, which impeded the creation of a single optimized task-set (green box), and perturbed action-values of other, already-optimized, task-sets.

Differences in task-set switching ultimately arose from differences in reward sizes between contexts (Fig. 2B). Mechanistically, large rewards quickly led to large task-set values and consistent task-set selection. Consistent task-set selection allowed for more efficient use of feedback and the formation of optimized action-values. Optimized action-values, in turn, enabled correct action selection and led to rewards, which further increased task-set values, etc. Small rewards had the opposite effect, leading to small task-set values, frequent task-set switching (suppl. Fig. 4.8D), suboptimal action-values, lack of rewards, etc.

The performance differences exemplified by this agent are a general feature of hierarchical RL. Hierarchical RL also showed the other behaviors observed in humans, such as task-set perseveration errors (suppl. Fig. 4.8A), quick task-set reactivation after context switches (in initial learning and hidden-context phase), and effects of reward size on performance (suppl. Fig. 4.8).

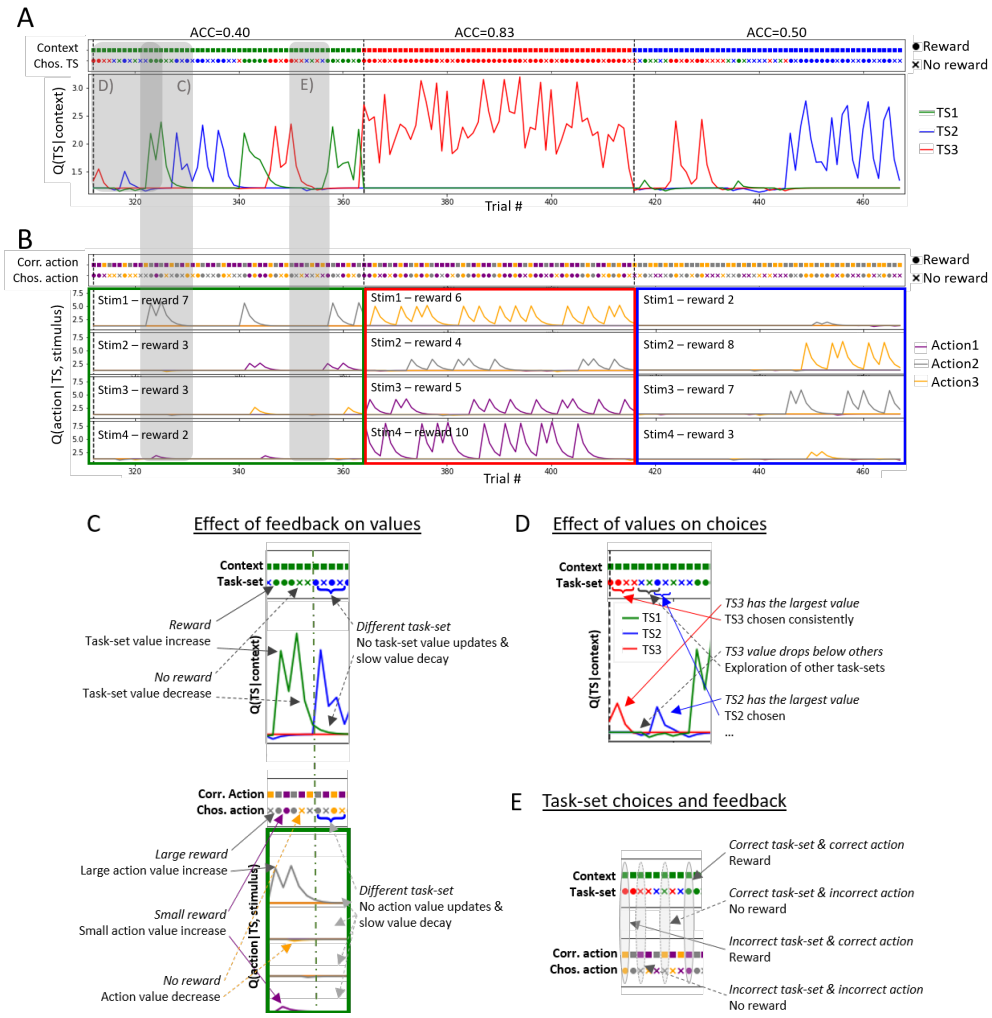


Figure 4.8: Trial-by-trial behavior a hierarchical RL agent from the population shown in Figures 3 and 4. A) RL at task-set level. Top: Sequence of contexts and agent's selected task-sets. There was no imposed mapping between contexts and task-sets; identical colors highlight which task-sets became specialized for each context during training. Bottom: Evolution of task-set values. B) RL at action level. Top: Sequence of stimuli (represented by correct actions) and agent's selected actions. Bottom: Action-values over time. Only the specialized task-set is shown for each context, indicted by box color. Task-sets contain action-values for all four stimuli (Stim1, Stim2, etc.). C) Top (bottom): Task-set (action) values were affected by reward (increase), no reward (decrease), and non-selection (slow decay, i.e., forgetting). D) Probability of task-set selection was based on relative task-set values. Agents circled through task-sets when no task-set was optimized and rewards were rare. E) Feedback reflects the validity of action selection, not of task-set selection. Task-set values arise from an indirect process mediated through action-values.

Model Comparison

Model Comparison is Relative The goal of model comparison is to test which one of two (or more) models explains a given dataset better. As such, model comparison is always relative, and does not provide an "absolute" measure of model fit. The quality of model comparison depends on the models included in the comparison, and on the supporting analyses that are performed externally to model comparison. We compared our hierarchical RL model to a flat RL and a hierarchical Bayesian model to test both of its unique aspects, aiming to provide as meaningful a model comparison as possible. We also supported our modeling analyses with qualitative behavioral analyses, and show that human behavior exhibited specific patterns that were only predicted by hierarchical RL, but not the other models.

Model Comparison Method Established approaches for model fitting and comparison such as maximum likelihood and sampling-based hierarchical Bayesian methods (Daw, 2011; M. D. Lee, 2011; Wilson & Collins, 2019) were not applicable in our case, because the model likelihood was intractable. Approximate Bayesian Computation (ABC) and other likelihood-free methods (Sunnaker et al., 2013; Turner & Sederberg, 2014; Turner & Van Zandt, 2012) were also infeasible in our setting. We therefore chose an alternative method for model comparison, namely simulation-based Bayes Factors, similar to (M. D. Lee, 2011; Steingroever et al., 2016; Sunnaker et al., 2013).

We took great care to ensure that our results were not based on our chosen hyper-parameters. Specifically, we explored different ranges of β parameters, the only parameters whose range was not given naturally. Changing these ranges had slight effects on the distributions of the models, but did not affect Bayes Factors in a meaningful way, and therefore did not affect our conclusions. We also explored different numbers of simulations per model, and found that this number had no noticeable effects beyond a reasonably high number of a few thousands. We also explored different values of the range parameter when selecting human-like simulations (suppl. Fig. 4.9A). The results were highly consistent for different ranges.

An overview of the results of model comparison for all measures is provided in table 4.3.

Table 4.2: Bayes factors. Numbers greater than 1 support the model mentioned first (usually hierarchical RL), numbers smaller than 1 support the model mentioned second.

	HRL vs flat	HRL vs Bayes	Bayes vs flat
Hierarchical representation			
Hidden-context phase, slope (Fig. 3B)	5.12	1.96	2.61
Initial-learning phase, accuracy 1^{st} trial (Fig. 3c)	2.88	1.31	2.21
Initial-learning phase, task-set perseverance 1^{st} trial (Fig. 3c)	14.49	1.40	10.32
RL values at two levels of abstraction			
Comparison phase, stimulus accuracy (suppl. Fig. 4.11B)	0.48	NA	NA
Comparison phase, context accuracy (suppl. Fig. 4.11A)	1171.65	NA	NA
Comparison phase, cont. minus stim. acc. (Fig. 4B)	39.64	NA	NA
Initial learning, regr. TS values on perf. (Fig. 4F)	1.49	6.62	0.22
Novel-context phase, frequency NoTS choices (Fig. 4D)	1.78	45.60	25.55
Novel-context phase, TS3 minus TS1 chocies (Fig. 4E)	1.59	32.01	20.14

Selection of Example Model Simulations

We presented one example simulation from each model in the bar graphs of figures 3A, 3C, and 4A. We obtained these simulation results in two steps: We first defined performance criteria around human behavior (see below) and selected a small subset of the 50,000 simulations that we had created for each model that matched these criteria. We then calculated the median parameter values across the selected simulations for each model, which are shown in table 4.3. We used the obtained parameters to create a new simulation for each model. The re-simulation of an independent dataset avoids problems of double-dipping and biased selection that would arise if an already-simulated dataset was presented based on its preferable performance.

Table 4.3: Median parameters of all models selected for similarity to human behavior. These parameter values were used to simulate a single new dataset to show side-by-side with humans

	α_a	β_a	f_a	α_{TS}	β_{TS}	f_{TS}
Hierarchical RL	0.49	9.74	0.47	0.29	13.77	0.17
Flat RL	0.49	10.11	0.31	NA	NA	NA
Hierarchical Bayes	0.79	13.28	0.21	NA	12.08	0.44

We defined the following criteria to select model subsets. We took note of human behavior for each of our summary measures (e.g., amount of NoTS choices in novel-context phase, accuracy in the context condition of the comparison phase, etc.). We then calculated a range around human performance for each summary measure, and selected all simulated datasets that fell within the ranges *of all measures simultaneously*. Fig. 4.9A shows how many simulations were selected from each model based on the range around human performance.

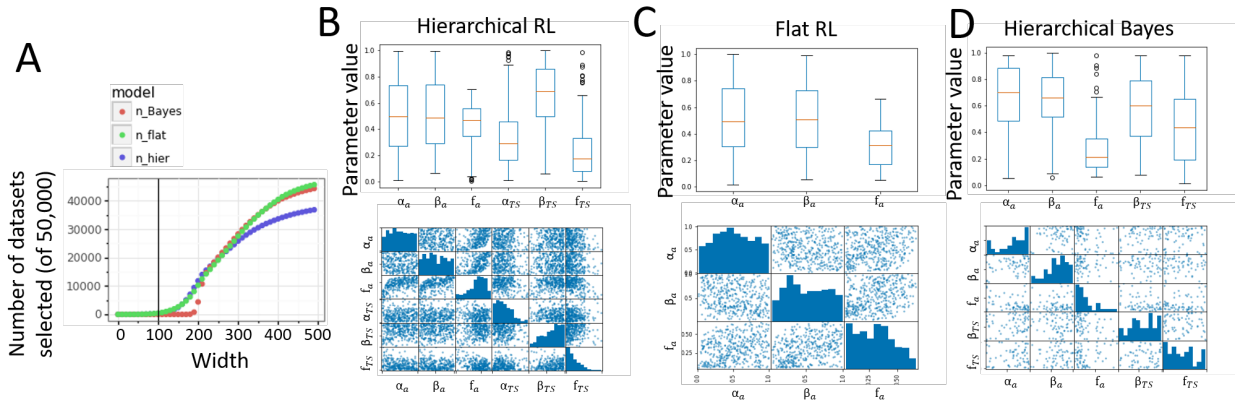


Figure 4.9: A) Number of selected simulations for each model, dependent on the chosen range around human performance. B)-D) Box plots and correlation matrices for the parameters of the selected datasets. Parameter values of β_a and $beta_{TS}$ are scaled down by a factor of 20 for easier comparison. All other parameters naturally lie in the range between 0 and 1. B) Hierarchical RL, C) Flat RL, D) Hierarchical Bayes.

We chose a window width of 100 around human measurement to select hierarchical and flat RL simulations, and width 180 to choose hierarchical Bayesian simulations. At width 100 (50%-150% of human performance), 314 / 50,000 (0.63%) hierarchical RL simulations were selected, and 434 / 50,000 (0.87%) flat RL ones (0 hierarchical Bayes datasets). At width 180 (10%-190% of human performance), 65 (0.13%) hierarchical Bayes simulations were selected.

We next inspected the parameters of the selected simulations in each model. Parameters were no longer distributed uniformly (suppl. Fig. 4.9B-D) like in the source distribution of the initial

50,000 simulations per model. This shows that for each model, human-like behavior was more likely under some parameters than others. For example, human-like performance in the hierarchical RL model was more likely to arise with low forgetting of task-set values - this can be interpreted as relating to participants' ability to reuse task-sets, which would not be the case if task-set values were fully forgotten.

Simulating new datasets in this way avoids problems of double-dipping when presenting example model behavior. But despite its conceptual similarity to ABC rejection sampling (Sunnaker et al., 2013), this method is not a parameter fitting procedure, i.e., the parameters in table 4.3 should not be interpreted as "best fits to the human data". One aspect that is lost by this procedure, for example, is the strong interdependence between parameters, for example relationships between β_a and f_a in hierarchical RL (suppl. Fig. 4.9B-D, bottom). The parameters obtained in this way were a basic estimate of reasonable parameter settings at the group level, and were merely meant to highlight the qualitative patterns that can arise from each model.

4.7 Supplementary Analyses

Additional Analysis of Initial Learning

Most of our analyses in the main paper followed the same logic: We made a behavioral prediction based on the hierarchical RL model, revealed the predicted pattern in human behavior, and concluded that hierarchical RL captured human behavior in qualitative terms. Here, we report an analysis that tests a prediction based on the hierarchical Bayesian model, to provides an inverse test.

Some problems can only be solved by Bayesian inference, but not through RL (in its basic form). For example, a Bayesian model is able to switch to correct behavior immediately after receiving a single diagnostic feedback, whereas RL needs to try out many actions to achieve this (in the right set-up). We therefore tested for markers of immediate switching in human data, as evidence for the hierarchical Bayesian model, and against hierarchical RL.

Nevertheless, we found no evidence for such behavior in our task. To test this, we first selected diagnostic and non-diagnostic trials in the hidden-context phase, and then compared participants' performance in the trials immediately following these trials. We defined diagnostic trials as trials in which feedback (correct vs. incorrect) indicated the correct task-set with certainty, such as receiving a reward after selecting the umbrella for the blue alien, which is only correct in TS1. Non-diagnostic trials were defined as trials in which the was not the case, e.g., receiving a reward after selecting the backpack for the red alien is correct in both TS1 and TS2 (even though reward amounts differ; see below). Most correct trials were diagnostic, whereas all incorrect trials were non-diagnostic in this sense (e.g., not receiving a reward after selecting the umbrella for the green aliens still leaves both TS1 and TS3 as possible candidates). We therefore restricted our analysis to correct trials only.

We found no difference in performance between dignostic and non-diagnostic trials (hidden-context phase, performance on trials immediately following diagnostic trials: 68.0%; perfor-

mance on trials immediately following non-diagnostic trials: 66.6%; difference between the two in repeated-measures t-test: t(25)=0.60, p=0.56). This shows that we were unable to find the behavior predicted by the hierarchical Bayesian model in our task, consistent with our hypothesis that human cognition employs hierarchical RL rather than Bayes.

Nevertheless, our task was not designed to test this prediction specifically, and the test just described had one potential confound, so we defer from drawing definite conclusions from it. The potential confound is that all correct feedback in our task indicates the correct task-set with certainty, not just the trials we termed diagnostic above: Even though some stimulus-action mappings are shared between task-sets, their rewards always differ. For example, action 3 is correct in both TS1 and TS2 for alien 1 (Fig. 2B). But because the reward is 7 for TS1 but 2 for TS2, correct feedback indicates with certainty which one is in place.

Thus, an agent with a perfect model of the task could, in theory, know with certainty which TS to select after a single reward. Whether our discrimination of diagnostic and non-diagnostic trials is valid therefore depends on whether humans have such a perfect model of our task. To conclude, our results suggest that participants were not able to quickly switch to a correct task-set after a single diagnostic feedback, despite the fact that the structure of the task could allow such. Instead, their learning process shows a slower trajectory, consistent with our hierarchical RL model rather than perfect inference.

Additional Analyses of Mixed Phase

Basic Behavior

Average performance in the mixed phase was 50.3% (sd=20.0%), as compared to 55.8% (sd=9.3%) during initial learning (chance=33.3%). The numerically lower performance might reflect the increased difficulty of the task when both stimuli and contexts were allowed to change on every trial. Nevertheless, differences between action-values and task-set values persisted. As expected, learning within blocks was not evident (suppl. Fig. 4.10A).

Switch Costs

Blocking contexts during initial learning might induce expectations in participants that contexts are necessarily blocked. The slower RTs after context switches than stimulus switches in the mixed phase could be a result of a violation of participants' expectation thus formed, rather than an index of hierarchical structure. We took several measures to alleviate this concern. First, participants were told explicitly at the beginning of the mixed phase that contexts would change quickly and unpredictably:

"You will next encounter the *chaotic season*. In the chaotic season, the weather changes very quickly. It can be rainy one day, and then sunny the next."

Nevertheless, expectations might persist implicitly. To investigate whether this was the case, we compared the RT effect in the first and second halves of the mixed phase. Our reasoning was that expectations about trial order should fade away quickly once participants realize that contexts

are presented in random order. Therefore, RT effects should diminish over time. If the RT effects were caused by participants' hierarchical representation, on the other hand, the RT effect should persist.

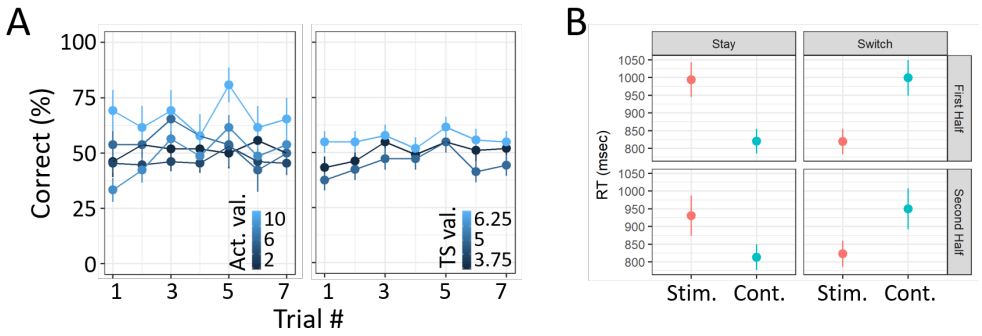


Figure 4.10: Mixed phase behavior. A) Performance broken up by action-values ("Act. val.", left panel) and task-set values ("TS val.", right panel). B) Response times on correct trials, in the mixed phase. The left panel shows stay trials, i.e., when the same stimulus ("Stim.", red) or context ("Cont.", blue) is repeated. The right panel shows switch trials. The top and bottom panel show the similarity between the first and second half of the mixed phase.

We split the mixed phase into two halves of equal size, and tested for the RT effects in both halves separately. The RT effect was present in both, with no significant difference between them (suppl. Fig. 4.10B); all trials: t(25)=3.5, p=0.002; only first half: t(25)=2.5, p=0.02; only second half: t(25)=2.7, p=0.01; difference between first and second half: t(25)=0.9, p=0.40). These results suggest that expectations formed by blocked context presentation were unlikely a complete explanation of the RT effects. The more likely explanation was the hierarchical representation of stimuli within contexts.

Additional Analysis of Comparison Phase

In the main text, we only showed the performance difference between stimulus and context condition, but not raw performance in each individually. Suppl. Fig. 4.11 and table 4.2 provide this information. As shown in the figure, the hierarchical RL model was likely to obtain better performance in the context condition (Fig. 3A), but worse performance in the stimulus condition, compared to flat RL. Bayes factors therefore favored hierarchical RL over flat RL in the context condition, $BF = 1.31$, but flat RL over hierarchical RL in the stimulus condition, $BF = 0.79$.

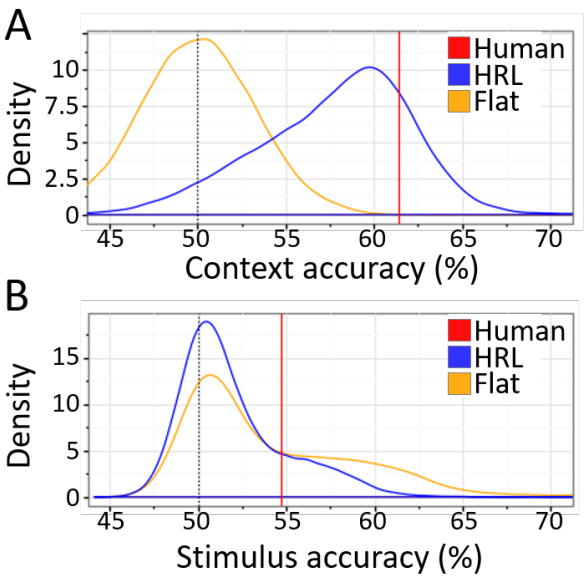


Figure 4.11: A)-B) Distribution over model behavior in the comparison phase. A) Accuracy for the context condition, i.e., % of trials in which the higher-valued context was chosen. B) Same for stimulus condition.

Relationship Between Model Parameters and Hierarchical Behavior

The behavioral predictions differed qualitatively between our tested models: E.g., the hierarchical RL model predicted that performance would improve during the first four trial of the hidden-context phase, whereas the flat RL stated that performance could not yet improve. Despite this qualitative difference in prediction—the presence versus absence of an effect—, the observable behavioral pattern—i.e., the slope over performance in the first four trials—lay on a continuum. In other words, even some flat RL simulations showed positive slopes, solely due to noise. And many hierarchical RL simulations did not show positive slopes because their underlying parameters prohibited learning.

Bayes Factors take these factors into account when comparing models quantitatively. In this section, we aim to understand how each model produced different markers of hierarchical behavior, expecting systematic links between certain parameters and model behaviors for model that provides a mechanism to explain the result (e.g., positive slopes in the hidden-context phase for hierarchical RL), and the lack thereof when the model cannot.

In the flat model, the forget parameter f influenced the frequency of task-set perseveration errors (suppl. Fig. 4.13, "Task-set perseveration errors"), as well as overall accuracy (suppl. Fig. 4.13, "Accuracy trial 1"), as expected. Nevertheless, the levels of task-set perseveration errors never went above, and accuracy never dropped below chance, such that these behaviors did not provide evidence for systematicity, and hierarchy. Extreme values of f also influenced

performance differences between context and stimulus condition in the comparison phase (suppl. Fig. 4.13, "Context minus stimulus"), which could be interpreted as a sign of hierarchy, whereby larger learning rates were related to bigger spreads in this measure.

In the hierarchical model, the forget parameter f_a played a similar role for "Accuracy trial 1" and "Context minus stimulus", but its role for "Task-set perseveration errors" was reversed. f_a also influenced "Task-set reactivation" (suppl. Fig. 4.12. Other model parameters showed additional relationships with behavioral markers in this model, such that, for example, very small high-level learning rates α_{TS} were associated with more task-set perseveration errors and smaller effects of task-set values on performance ("Effect of task-set values"), and high-level beta β_{TS} was associated with increased task-set reactivation, reduced task-set perseveration errors, and increased effect of task-set values on performance. This shows that in the hierarchical model, behavioral markers of hierarchical behavior arose from a complex interplay between model parameters.

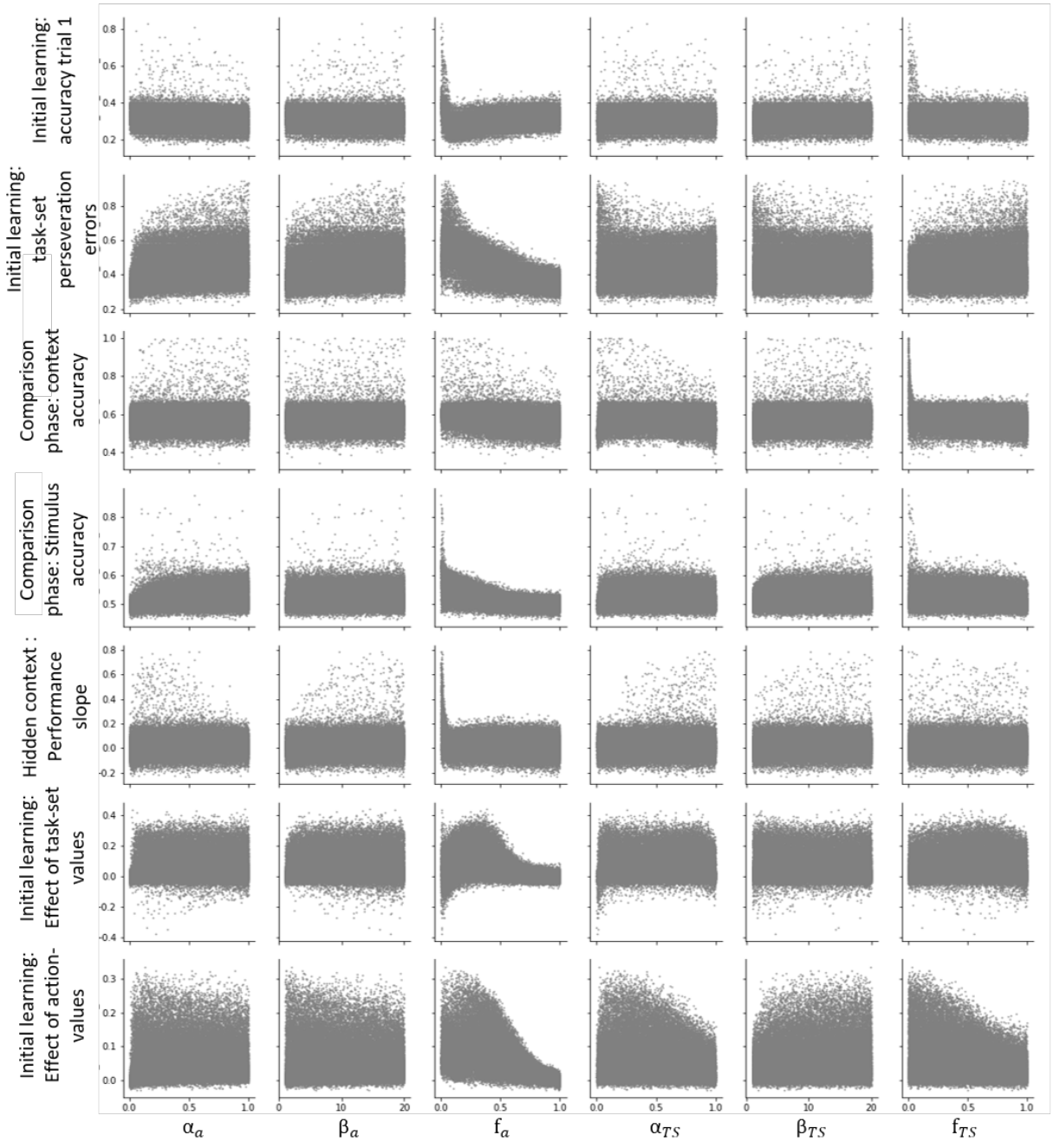


Figure 4.12: Relationship between hierarchical model parameters and behavioral markers across 50,000 simulations.

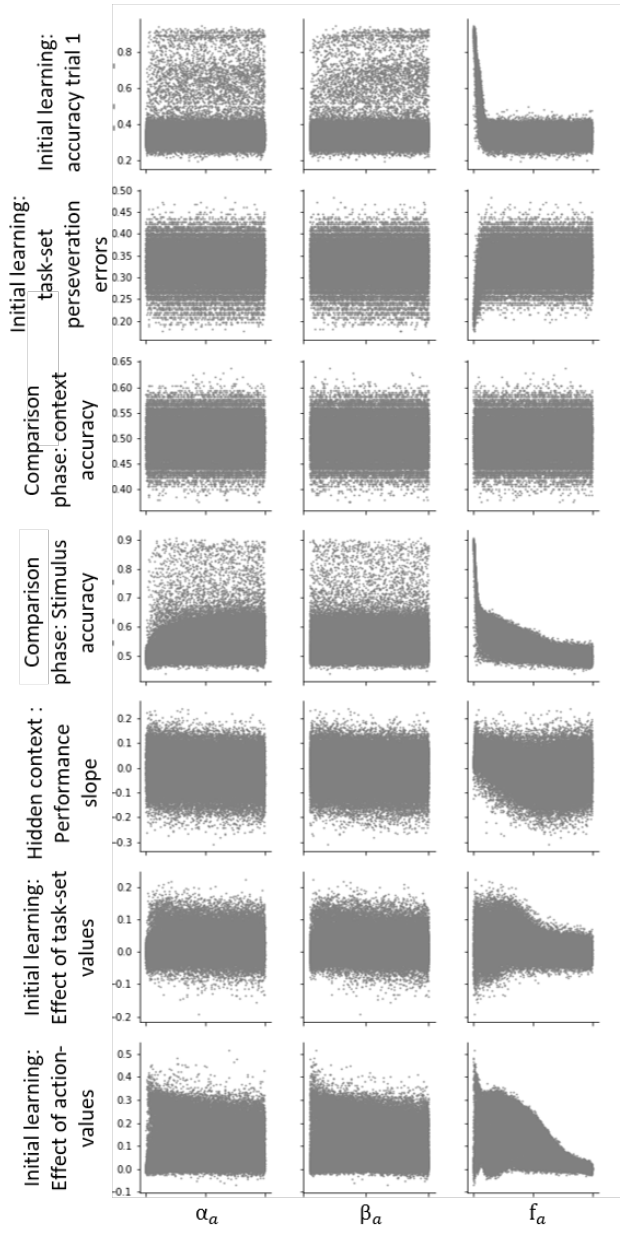


Figure 4.13: Relationship between flat model parameters and behavioral markers across 50,000 simulations.

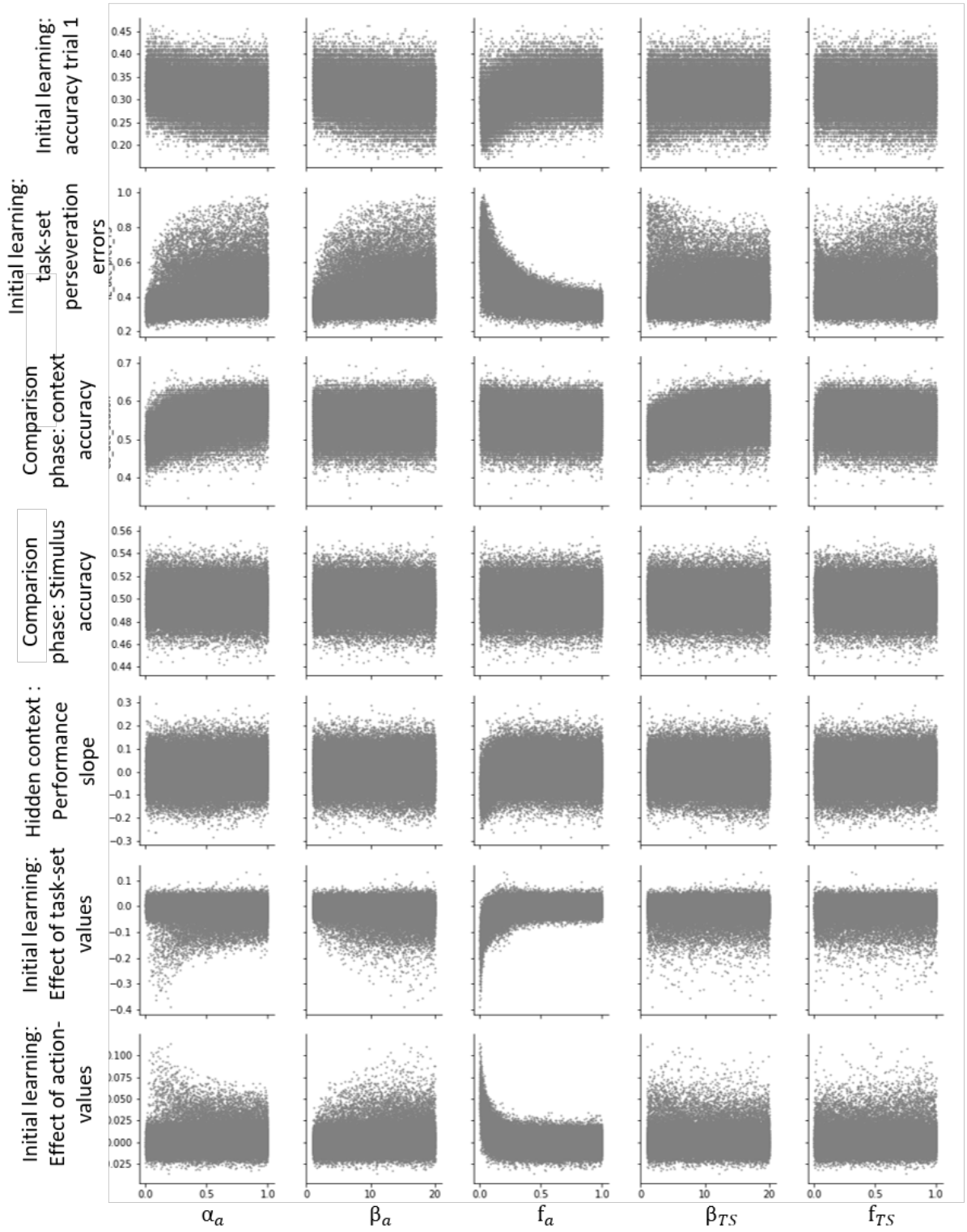


Figure 4.14: Relationship between hierarchical model parameters and behavioral markers across 60,000 simulations.

Chapter 5

Hierarchical Learning of Complex Action Sequences

This chapter delves into the creation of structured representations in humans. We present a task that motivates participants to use hierarchical representations to solve it, and investigate how participants form hierarchical representations. [1]

Abstract

Humans have the astonishing capacity to quickly adapt to varying environmental demands and reach complex goals in the absence of extrinsic rewards. Part of what underlies this capacity is the ability to flexibly reuse and recombine previous experiences, and to plan future courses of action in a psychological space that is shaped by these experiences. Decades of research have suggested that humans use hierarchical representations for efficient planning and flexibility, but the origin of these representations has remained elusive. This study investigates how hierarchical representations can be learned through experience. Seventy-three participants completed a task in which they had to perform complex action sequences to obtain rewards. Crucially, complex action sequences were composed of simpler action sequences, which were not rewarded, but whose completion was signaled to participants. We found that participants learned to perform the simpler action sequences and combined them into complex action sequences. Strikingly, participants actively taught themselves the simpler sequences, pausing after their first completion and actively rehearsing them thereafter, despite the lack of extrinsic rewards. This suggests that intrinsic motivation to understand environmental dynamics can subserve the creation of hierarchical representations. Response times also revealed hierarchical structure, showing segmentation at the borders of simple sequences. Response times within simple sequences dropped markedly after their first discovery, suggesting chunking of individual actions into sequences; similarly, response times between simple sequences dropped once complex sequences were discovered, suggesting hierarchical chunking of simple sequences into complex sequences. After a learning phase, par-

[1] Anne G.E. Collins, Aram Moghadassi, and Amy Zou contributed to the research presented in this chapter.

ticipants completed a transfer phase in which either simple sequences or complex sequences were modified without notice. Relearning progressed slower when simple sequences were changed than complex sequences. This suggests that in hierarchical representations, lower levels might be less malleable than higher levels, in accordance with a role of lower levels to stabilize and structure exploration, and of higher levels to flexibly recombine elements.

5.1 Introduction

Complex problems can only be solved by using hierarchy

The everyday world poses problems of a level of complexity that require sophisticated, abstract representations to solve. Consider the amount of information that arrives at your senses every second, and the number of muscles in your body that you could activate in response at any moment and to any degree. The number of possibilities of just a single choice already seems infinite, but this number multiplies with each additional time step because each action can be followed by any other action, and so on. This combinatorial explosion of possible trajectories makes it seem like planning even a small number of steps into the future is impossible. To make things worse, the everyday world is sparse in rewards and requires sophisticated exploration strategies to achieve even simple goals. Consider how many independent muscle activations were necessary to prepare your last breakfast, let alone write your last research paper, or plan your vacation. The complexity of our goals makes it almost impossible to achieve them by random exploration, and this means that we cannot rely on simple reinforcement to learn how to achieve them. A solution to these problems is the use of abstract representations. Abstract representations reduce the dimensionality of problems by recognizing patterns (e.g., automatizing common movements like reaching and grasping), and thereby reduce the problem of combinatorial explosion. Hierarchical representations also alleviate the problem of sparse rewards. Instead of exploring the vast space of possibilities based on random muscle movements, an abstract representation allows for exploration based on complex action, such as—in the example of planning a vacation—searching online for places with a beach, calling a friend who recently came back from a vacation, or checking for destinations that offer cheap flights.

Example: 4-rooms domain

The potential benefits of hierarchy are famously illustrated in the 4-rooms example (Sutton et al., 1999). An artificial agent is dropped at a random location in a gridworld with four rooms, and needs to navigate to a random goal in a different room. All neighboring rooms are connected by narrow doorways. When the agent only has access to basic actions (one step up, down, left, or right), it often fails to reach the goal because it gets stuck exploring just the initial room. If the agent reaches the goal, learning the correct policy to find it again takes many iterations. On the other hand, when the agent has access to additional hierarchical actions that can reach each door, it explores all rooms more evenly, finds the goal faster and more reliably, and also learns the correct

policy much faster. This example highlights how seemingly complex problems can become very simple with the right form of abstraction—in this case, temporal abstraction, the ability to execute multi-step actions.

Previous Research

Previous research across disciplines has shown that hierarchy is a powerful ingredient for artificial intelligence, and a fundamental part of natural intelligence. Research in machine learning and Artificial Intelligence (AI) has employed hierarchical representations to solve increasingly complex problems (e.g., Bacon et al., 2017; Dayan, n.d.; Dietterich, 2000; Duan et al., 2016; Finn et al., 2017; Parr and Russell, 1998; Sutton et al., 1999; Vezhnevets et al., 2017; Wang et al., 2016). In Psychology, decades of research have investigated the hierarchical structure of the mind, including research on cognitive control, expertise, and sequential action (Broadbent, 1977; Chase and Simon, 1973; Cohen, 2000; Cooper and Shallice, 2006; Lashley, 1951; Newell, 1994; Rosenbaum, 1987; Schank and Abelson, 1977). Recent research has increasingly focused on formalizing models of hierarchical cognition, using hierarchical Bayesian models (Collins and Koechlin, 2012; Gershman and Niv, 2010; Griffiths et al., 2019; Huys et al., 2015; Kemp and Tenenbaum, 2008; Schapiro et al., 2013; Solway et al., 2014; Tomov et al., 2019), and hierarchical Reinforcement Learning (RL) models (Botvinick and Weinstein, 2014; Cushman and Morris, 2015; Dezfouli et al., 2014; Diuk, Schapiro, et al., 2013; Eckstein and Collins, 2018; Frank and Badre, 2012; Momennejad et al., 2017; Ribas Fernandes et al., 2011). Research in neuroscience has provided strong evidence for the hierarchical organization of the brain, both in the sense of "processing hierarchies", in which superordinate levels operate over longer time scales and asymmetrically modulate subordinate processing, and of "representational hierarchies", in which superordinate representations form abstractions over subordinate representations, favoring generality over detail, and allowing information to be inherited asymmetrically from higher to lower levels (for reviews, see Badre, 2008; Balleine et al., 2015; Graybiel and Grafton, 2015; E. K. Miller and Cohen, 2001).

The Problem of Creating Hierarchy

Despite the near-universal conviction that hierarchical representations are necessary to solve problems of real-world complexity, the question remains unanswered how to create appropriate hierarchical representations. In AI, this is called the "option discovery problem" because abstract, multi-step actions are often called "options" (Sutton et al., 1999). Hierarchical representations are only beneficial when they condense the important aspects of a task in the right way, but they can be disadvantageous otherwise and even hurt performance. (For example, in the 4-rooms domain, adding multi-step actions that lead to doorways boosts performance, but multi-step actions that lead to room corners hurts performance.) Humans have been shown to discover the Bayes-optimal task decomposition when solving complex problems (Solway et al., 2014), but it is unclear how they discover these decompositions, lacking access to the full state space, and with limited computational resources. Research in AI has investigated several promising avenues for how to create appropriate hierarchical representations. Some approaches equip agents with intrinsic motivation,

a form of motivation that is independent of extrinsic rewards and often tries to mimic novelty seeking and curiosity observed in humans and animals (Gershman and Niv, 2015; Lieshout et al., 2018). Other approaches analyze the abstract problem structure and try of locate bottlenecks of the state-space or locations of advantageous graph-theoretical measures such as maximum betweenness (e.g., Machado et al., 2017; Pathak et al., 2017; Schmidhuber, 2010; Singh et al., 2005; for review, see Konidaris, 2019). The goal of both approaches is to identify states that would make appropriate targets for multi-step actions, and thereby create a hierarchical representation.

Our Take

In the current study, we propose that humans create hierarchical representations piece-by-piece, continuously learning new, ever more complex actions. We propose that, starting from a set of basic actions (e.g., stretch arm, move fingers), we explore the world around us by randomly executing one action at a time. Some combinations of actions lead to unexpected events in our surroundings (e.g., hitting a rattle makes a sound). Such events trigger curiosity(—defined as an interest in events that are not rewarding, but potentially provide information that are relevant for obtaining reward in the future). Curiosity motivates further exploration of the event, and once the event can be reproduced reliably by using appropriate combinations of basic actions, a new skill has been learned (e.g., grasp the rattle). Adding skills to the action repertoire allows for more targeted exploration, and can speed up the acquisition of more abstract skills by combining less abstract skills (e.g., shake and throw the rattle), following the same curiosity-guided process. Curiosity-based explanations like this one deal elegantly with the problem of sparse rewards because they move the role of teaching signal from rewards to other environmental signals. And it reduces the dimensionality of the space by creating multi-step actions, which reach further into the future and constrain all subsequent actions once the chain has been picked.

The Task

To test this prediction, we created a task in which participants learned to execute complex action sequences, which were composed of simpler action sequences, which themselves were composed of basic actions (Fig. 5.1A-B). The goal of the task was to create a specific star on each trial, using a star-making machine. The machine accepted 4 key presses per trial, and created a star when a correct 4-key sequence was typed in. Crucially, stars' "complex" 4-key sequences were composed of "simple" 2-key sequences, and the execution of a "valid" 2-key sequence was signaled by an item appearing in a window on the machine. Four different stars, learned in successive blocks, required a different combination of two of four possible simple 2-key sequences. This paradigm has a clear hierarchical structure: basic actions (individual key presses) form the lowest level; simple skills (2-key sequences), which are not rewarded and have to be learned through intrinsic motivation, form the intermediate level; and complex skills (4-key sequences), which are composed of simple skills and lead to reward, form the most abstract level. This task was designed to elicit curiosity-driven learning and the creation of hierarchical structure. The goal of the study was to investigate whether participants could leverage the hierarchical structure of the task, and if so, how

they learned to create the necessary hierarchical representations. We predicted that participants would acquire and practice 2-key sequences before 4-key sequences, and that they would explore 4-key sequences with 2-key sequences, rather than key by key.

The task also included a transfer phase in which either simple or complex key sequences changed without notice, to investigate whether actions at different levels of abstraction played different roles. After 300 trials of the initial paradigm, participants either entered a "low" or a "high" transfer phase. In the low transfer phase, some 2-key sequences were modified by replacing individual keys; in the high transfer phase, some 4-key sequences were modified by replacing entire 2-key sequences. Even though both manipulations affected similar numbers of individual keys in the tested stars, they should have different effects when using a hierarchical representation. We predicted that sequences at the low level, being more consolidated, would be difficult to re-learn, whereas sequences at high levels, still flexible and malleable, should be less affected by transfer. Furthermore, we predicted that participants would attempt to reuse learned simple chunks for learning, rather than exploring in the "single key press" space.

Paper Outline

The goal of this study was to investigate the process by which humans *create* hierarchical representations, and how these representations are *used*, using a task with sparse rewards and a relatively large combinatorial space, but clear hierarchical structure. The study addresses both questions in turn: In part 1, we will investigate how participants added layers of abstraction to create hierarchy, that is, how they combined basic-level actions into skills, and how these skills were added to the action repertoire. We investigated whether participants were *intrinsically* motivated to learn skills / 2-key sequences, in the absence of extrinsic reward. In part 2, we will shed light on how participants used the hierarchical structure just created. Did the existence of complex actions help explore the space of possibilities more efficiently? What happened when participants reached a star, did they build 4-key sequences? Did actions at the low (key press) and the high level (2-key sequences) play different roles? We also investigated what it meant to use a hierarchical rather than a flat representation, in terms of how transfer affected choices and learning.

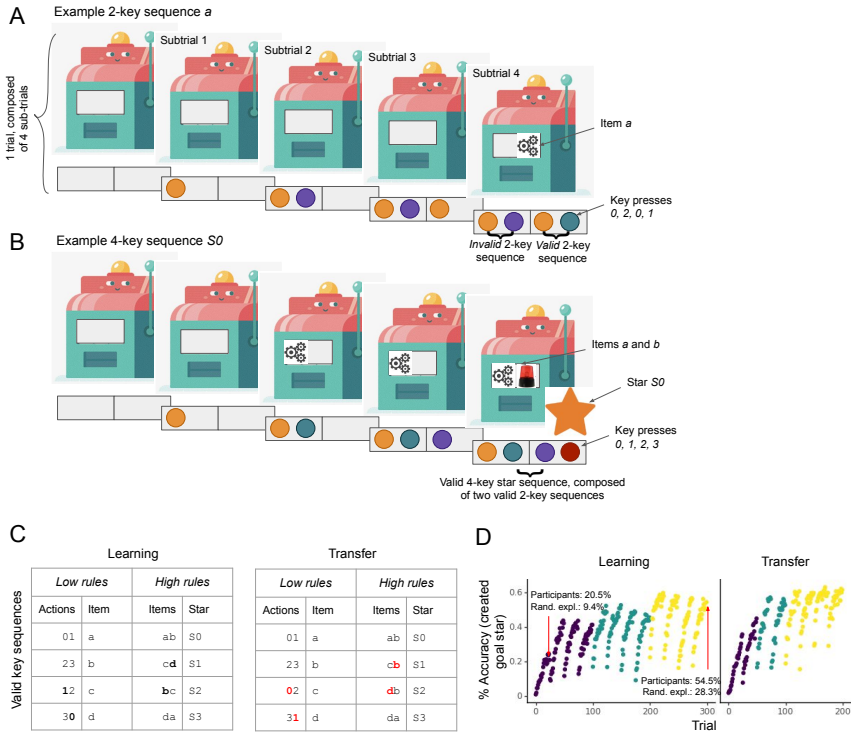


Figure 5.1: Task design. (A) On each trial, participants sequentially entered four key presses in a self-paced manner (maximum response time: 2.5 seconds). Each key press was acknowledged by the appearance of a colored circle in the response board. Each valid 2-key sequence (see rules in part C of this figure) was acknowledged by the appearance of a unique item: in the example shown, keys 0 (subtrials 3) and 1 (subtrial 4) led to item *a*. All other "invalid" 2-key sequences did not lead to items. (B) Four specific combinations of 2-key sequences led to the appearance of a star (see part C for rules). Each trial indicated a specific goal star (not shown here). Achieving this star led to reward. (C) Rules for valid 2-key and 4-key sequences. Left table "Learning": Key sequences for the learning phase. The left column "Low rules" shows valid 2-key sequences. The left sub-column "Actions" shows the identity and order of actions that need to be executed, and the right sub-column "Item" shows the resulting item. Keyboard keys were randomly assigned to actions, and images were randomly assigned to items. The right column "High rules" shows valid 4-key sequences. Each 4-key sequence was composed of two 2-key sequences, shown in the left sub-column "Items", and led to the star shown in the right sub-column "Star". Right table "Transfer": In the transfer phase, either the low rules or the high rules changed. In the low transfer phase, the low rules shown in "Learning" table were replaced by the low rules shown in the "Transfer" table. In the high transfer phase, the high rules of the "Learning" table were replaced by the high rules in the "Transfer" table. The table highlights differences between learning and transfer rules in red. (D) Learning curves. Trials 1-300 were part of the learning phase, 301-501 of the transfer phase. Accuracy was determined by whether the correct star was created.

5.2 Results

Introductory Explanations

Complexity of the Task

Despite the complexity of the task, most participants were able to learn it well. With four keys allowed on each of four subtrials, the task allowed for $4^4 = 256$ different 4-key sequences, and only 4 of them (1.6%) led to a star, only one of which (0.4%) was rewarded in any trial. A strategy of trying a different sequence on each trial has a $\frac{1}{256} + \frac{1}{255} + ... + \frac{1}{231} = 10.3\%$ probability of discovering the correct star within 25 trials (the length of a block).Participants, on the other hand, achieved 20.5% accuracy by the end of the first block (25 trials), and 54.5% by the end of the last block (75 trials per star; trials highlighted by red arrows in Fig. 5.1D), roughly doubling both probabilities, which suggests that participants used more efficient strategies. Individual performance varied widely. Working on a total of 1,000 trials, participants earned anywhere between 106 (10.6% average accuracy) and 702 points (70.2% average accuracy; population average: 426.4 points, sd: 135.1). This spread facilitated the analysis of individual differences and of how different strategies related to performance.

How to Interpret the Learning Curves

Participants showed consistent patterns of learning and forgetting. In the first few trials of the task, participants were unable to find the goal star, but they learned quickly and achieved 20.5% accuracy by the end of the first block (Fig. 5.1D). When the goal star changed for the first time on trial 26, accuracy dropped to 10.9%, suggesting that participants were able to reuse knowledge they had gained when working on the first star for the second star. Final performance in a block improved with every repetition of a star (repetition 1, average of all four star: 36.8% accuracy on last trial of block; repetition 2: 48.9%; repetition 3: 56.5%; effect of repetition on performance in mixed-effects regression: $\beta = 0.098$, $z = 6.1$, $p < 0.001$). At the beginning of the transfer phase, when rules changed without notice, performance dropped back to 0% accuracy, but participants recovered fast, which suggests successful re-learning of rules (Fig. 5.1D).

Part 1: Creating the Hierarchy by Learning Action Sequences

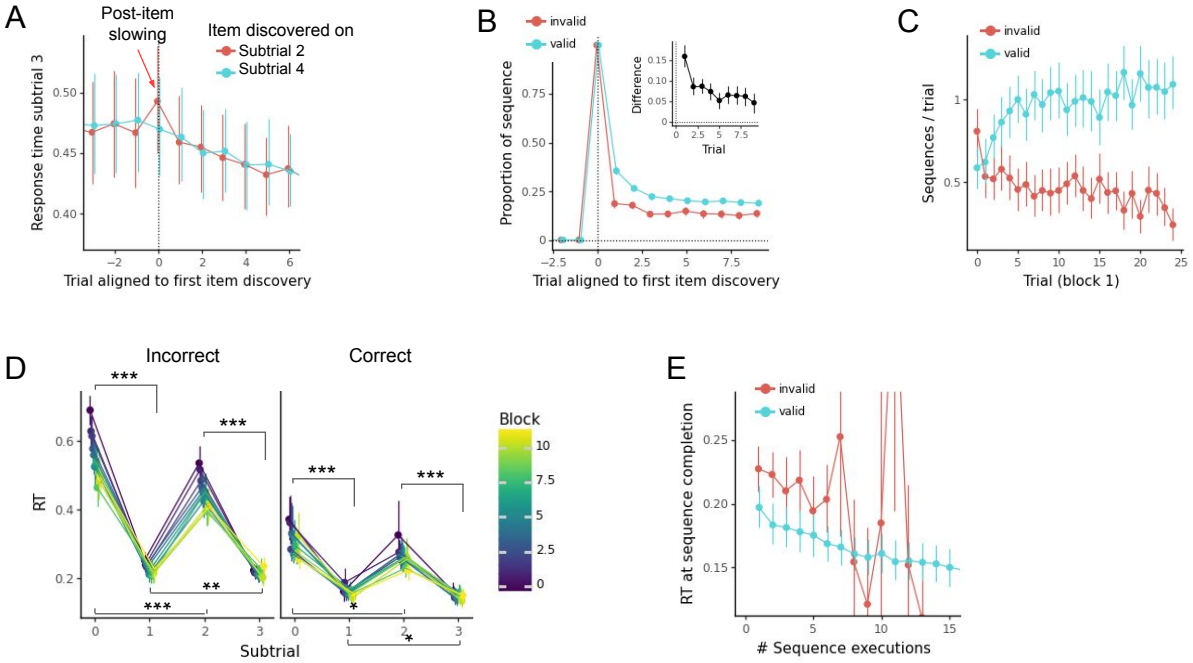


Figure 5.2: (A) Post-item slowing. Response times on subtrial 3 were uniquely elevated on the trial in which a new 2-key sequence was discovered on subtrials 1 and 2 (red). Slowing on subtrial 3 was specific to sequence discovery on subtrials 1 and 2 as discovery on subtrials 3 and 4 (blue) did not elicit slowing. Dots represent means and error bars show between-participant 95% confidence intervals, as in the remaining figure. (B) Repetition of valid and invalid sequences after first discovery. Trial 0 shows the first execution of any particular 2-key sequence in a block. Subsequent trials show the proportion of trials in which the same sequence was repeated, separately for valid (blue) and invalid (red) 2-key sequences. The inset shows the within-participant difference between the proportions of valid and invalid sequences. (C) The number of 2-key sequences executed per trial, over time in block 1. The blue line shows the average over the four valid 2-key sequences (signaled by item appearance), and the red line shows the average over four matched invalid 2-key sequences (not signaled by items). The maximum number of 2-key sequences per trial is two because each trial allows for four key presses. The red and blue lines do not add up to two because many 2-key sequences were neither categorized as valid nor invalid for this analysis. (D) Execution time for each key press within a trial. Statistical comparisons refer to repeated-measures t-tests, as described in the main text. (E) The change of execution time for the final key of 2-key sequences over sequence repetitions, for valid (blue) and invalid (red) sequences. The larger error bars for invalid sequences on later repetitions stem from the fact that participants repeated invalid sequences less than valid ones.

In part I, we investigated how participants learned new action sequences. We tested whether participants slowed down after discovering an item for the first time, as a sign of processing the item as feedback. We also inquired how often participants reused valid sequences compared to invalid ones, as a measure of intrinsic motivation to execute valid action sequences. We assessed how many valid compared to invalid sequences participants executed per trial, hypothesizing that valid sequences would increase and invalid sequences would decrease as participants started exploring the task based on 2-key sequences rather than individual keys. Finally, we analyzed response time patterns within trials as an indicator of whether individual key presses were chunked into action sequences, and examined the development of these patterns over time. All analyses were restricted to the learning phase of the experiment.

Participants slowed down after seeing a new item for the first time, suggesting that items elicited feedback processing. We assessed response times on the subtrial after an item was discovered for the first time in a block during the learning phase, and compared them between the trial in which the item was discovered, and the preceding and subsequent trial (Fig. 5.2A, red line). Repeated-measures t-tests, controlling for multiple comparison using the Bonferroni correction, revealed that participants showed significantly longer post-item response times on the trial of item discovery, compared to the preceding ($t(54) = 4.1$, $p = 0.0003$) and subsequent trial ($t(54) = 6.9$, $p < 0.001$). This slowing on subtrial 3 was specific to item discovery on trial 2, as it was absent on trials in which an item was discovered on trial 4 (Fig. 5.2A, blue line). This result shows that participants slowed down in the middle of a trial if they had just discovered a new item for the first time. Similar slowing commonly arises after participants make errors (Danielmeier and Ullsperger, 2011), receive rewards (Raio et al., 2020), or observe a surprising event [cite], and this slowing is commonly interpreted as an orienting response and potentially related to the processing of prediction errors and to learning [cite]. In our task, items implicitly indicated that participants had executed valid 2-key sequences, and additional processing is expected if participants use this information to learn 2-key sequences.

We next compared what effect observing a valid 2-key sequence (i.e. a sequence that lead to an item, independently of whether it was a useful sequence for the current goal star) had on subsequent 2-key sequence selections. After discovering them for the first time, participants were more likely to repeat valid compared to invalid 2-key sequences, suggesting a role of intrinsic motivation in the execution of valid 2-key sequences (Fig. 5.2B). To directly compare valid to invalid sequences, we randomly selected 4 invalid 2-key sequences that had the same characteristics as valid sequences (e.g., the two keys of the sequence differ), and subjected them to the same analyses as the four valid sequences. From participants' perspective, nothing discriminated the invalid sequences we selected for this analysis from other invalid sequences, and using a different set of invalid sequences (with the same restrictions) led to similar results. For the following reasons, we limited this analysis to incorrect trials only. On correct trials, participants necessarily executed two valid 2-key sequences to achieve the goal star, and because accuracy increased within each block (Fig. 5.1D), the inclusion of correct trials would necessarily lead to an increase in the proportion of valid 2-key sequences per trial. To avoid this issue, we only included incorrect trials. To compare the repetition of valid and invalid sequences, we assessed the proportion of trials in which participants repeated a 2-key sequence, that they had executed for the first time in a block on trial 0, in the

subsequent trials, for both valid and invalid sequences (Fig. 5.2B). To compare valid and invalid sequences statistically, we calculated the differences between their proportions for each participant and each trial (Fig. 5.2B, inset). We then used mixed-effects regression to predict the proportion differences from the trial since sequence discovery. The regression revealed that the difference was significantly different from zero (Intercept $\beta = 0.13$, $z = 14.7$, $p < 0.001$), with a negative effect of trial ($\beta = -0.01$, $z = -6.6$, $p < 0.001$). This confirms that the proportions of valid and invalid sequences differed significantly, controlling for trial post-discovery. In other words, participants were more likely to repeat valid 2-key sequences after they first discovered them, compared to invalid sequences. This suggests that the appearance of items motivated participants to repeat key sequences, potentially in a similar way as explicit rewards would, a typical example of intrinsic motivation.

Over the course of the first block, participants increased the number of valid 2-key sequences per trial while decreasing the number of invalid sequences. This is in accordance with a transition from exploration based on individual keys to exploration based on 2-key sequences. For the reasons explained above, we again limited our analysis to incorrect trials only. In addition, this analysis only included data of the first block of the experiment. We assessed how many valid and invalid sequences participants executed on each trial in the first block. For statistical comparison between valid and invalid sequences, we conducted a mixed-effects regression predicting the number of sequences from sequence validity (valid vs invalid) and trial (1-25), as well as their interaction. The regression showed a significant interaction between sequence validity and trial ($\beta = 0.036$, $z = 7.1$, $p < 0.001$), confirming that the trajectories of valid and invalid sequences differed (Fig. 5.2C). Valid sequences showed a positive slope ($\beta = 0.011$, $p = 0.04$), indicating that participants increased the number of valid sequences per trial, while negative sequences showed a negative slope ($\beta = -0.025$, $p < 0.001$), indicating that participants decreased the number of invalid sequences. This pattern suggests a shift in exploration strategies: In the beginning, participants necessarily explored the task based on individual keys, but once they discovered valid 2-key sequences, they started replacing individual keys by sequences. Executing such temporally-extended actions is a signature of hierarchy.

Participants showed a pronounced slow-fast-slow-fast response pattern (Fig. 5.2D), which suggests that they executed two distinct 2-key sequences, rather than four distinct key presses. For statistical comparison, we calculated the differences in response times between pairs of subtrials within each trial, separately for each participant. We then averaged these response time differences within participants and conducted one-sample t-tests to determine whether the differences were significant. All eight tested differences were significant, even after stringent Bonferroni correction (all $t(55)s > 2.9$, all $ps < 0.04$; Fig. 5.2D). That participants executed the second (fourth) key press faster than the first (third) confirms that they chunked the pair of key presses into a single unit, in accordance with the consolidation of 2-key sequences into distinctive, temporally-extended actions. That participants also responded faster on the third (fourth) compared to the first (second) subtrial might suggest frontloading of processing, such that in part, the second 2-key sequence was already prepared before or during the first sequence.

Participants became faster at executing valid 2-key sequences with each execution, but the same was not true for invalid sequences. We examined how the response time of the second key

of a 2-key sequence developed with each execution of that sequence (Fig. 5.2E). If sequences were increasingly automatized, the execution time should decrease over time. Indeed, mixed-effects regression models, restricted to ten sequence executions to maximize the amount of data for invalid sequences, revealed a significant effect of the number of repetitions on execution time ($\beta = -4.1$, $z = -2.4$ $p = 0.014$), revealing a decrease in execution times with repetition. Furthermore, execution times were significantly slower in invalid compared to valid sequences ($\beta = -30.9$, $p = 0.001$), confirming that invalid sequences were less automatized.

Part 2: Using the Hierarchy for Exploration and Planning

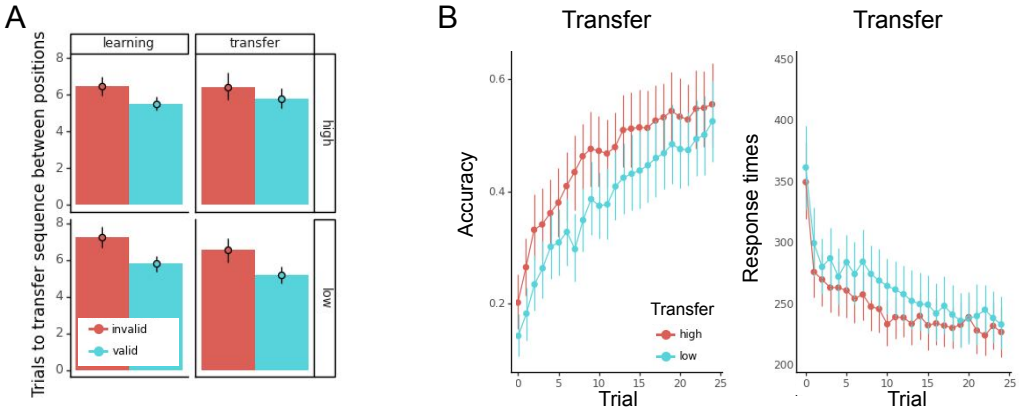


Figure 5.3: (A) Number of trials to use a 2-key sequence that was first discovered in one position in the opposite position of a trial. (B) Performance in the transfer phase. Accuracy (left) and response times (right) over trials, averaged over blocks, for both high (red) and low (blue) transfer phases.

In part II, we investigated how participants used the hierarchical representations whose creation we investigated in part I, assessing whether it benefited exploration and planning, as hypothesized. We tested whether participants reused sequences learnt in one position in the other position, a sign of adaptive exploration based on 2-key sequences. We also inquired how the transfer phase affected performance, hypothesizing that changing the "low" rules for 2-key sequences would affect performance more than changing the "high" rules of 4-key sequences.

Participants quickly transferred valid action sequences from the position in which they were originally discovered to the opposite position, suggesting flexible reuse and exploration. We counted how many trials occurred between the first execution of a 2-key sequence within a block, and the first execution of the same sequence in the opposite position, e.g., a repetition on subtrials 3 and 4 when the sequence was initial discovered in subtrials 1 and 2. On average, participants took 5.6 trials to transfer valid sequences (Fig. 5.3A), and mixed-effects regression revealed that significantly fewer trials were needed to transfer valid 2-key sequences compared to invalid 2-key sequences (Intercept $\beta = -0.87$, $se = 0.33$, $z = -2.63$, $p = 0.008$), with no effect of block ($\beta = -0.004$, $p = 0.94$). This is in accordance with the notion that when participants acquired valid 2-key sequences, these were consolidated and added the action repertoire, such that they could be used for future exploration.

Performance suffered more in the low transfer phase than in the high transfer phase (Fig. 5.3B), in accordance with a role of 2-key sequences as "building blocks" for planning and complex action. We probed differences between accuracy in the high and low transfer phase using mixed-effects regression, predicting accuracy from transfer type (high vs low), trial (1-25), and their interaction. The model revealed main effects of transfer type ($\beta = 0.10$, $z = 6.4$, $p < 0.001$) and trial ($\beta = 0.1$, $z = 15.7$, $p = 0.10$), but no interaction ($\beta = 0.002$, $z = 1.6$, $p = 0.11$). We used a similar model to test differences in response time, and found main effects of transfer type ($\beta = 28.2$, $z = 6.8$, $p < 0.001$) and trial ($\beta = -2.8$, $z = -13.5$, $p < 0.001$), as well as an interaction ($\beta = 0.60$, $z = 2.0$, $p = 0.044$). This confirms that participants performed better in the high compared to the low transfer phase, and shows that it was more difficult to relearn "low-level" 2-key sequences compared to "high-level" sequences of 2-key sequences. This is in accordance with the view that lower level of hierarchical representations are more consolidated and less accessible. Bluntly, once a key sequence becomes an action, it becomes difficult to change it.

5.3 Discussion

Abstraction has long been argued to facilitate complex problem solving, including planning over long time horizons, responding to new situations by leveraging old experiences, and learning in the absence of direct rewards. Abstraction provides a handle on these problems because it directs exploration by constricting the space of possibilities, and combating the curse of dimensionality. It also facilitates learning with sparse rewards because it aims to identify alternative targets of learning. Given that the right abstractions can solve some of the hardest problems, the crucial –yet unsolved– issue is how to create such abstractions. The possibility we investigate here is to use environmental signals to bootstrap learning, i.e., learning complex action sequences that elicit environmental signals. Such acquired action sequences then form the basis for future exploration, facilitating the discovery of rewards.

Summary of Results

In part I, we investigated how participants learned new action sequences, creating temporal hierarchy. Participants slowed down after discovering an item for the first time, suggesting they processed the appearance of the item as feedback, as would be expected for explicit rewards. Thereafter, participants reused valid action sequences –i.e., those that result in items appearing– more often than invalid ones, revealing intrinsic motivation to execute valid action sequences. The frequency of valid sequences increased, while invalid sequences decreased, suggesting that participants progressively explored the task based on 2-key sequences rather than individual keys. Response time patterns within trials confirmed that participants clustered individual key presses into action sequences.

In part II, we investigated how participants used the hierarchical representation whose creation we investigated in part I, asking whether it benefited exploration and planning, as hypothesized. Participants successfully reused sequences they learnt in one position in the other position, a sign of active exploration. In the transfer phase, changing the "low" rules –i.e., the recipe for 2-key sequences– affected performance more than changing the "high" rules –i.e., the recipe for 4-key sequences, based on 2-key sequences. This confirms that adapting 2-key sequences was more difficult than adapting 4-key sequences, in accordance with participants' integration of 2-key sequences into their action repertoire.

Future Directions

We focused on one way of creating hierarchical representations in this study, using environmental signals as the targets of complex action sequences. Future research is needed to assess whether this is a general framework in human learning. Two factors should be manipulated to investigate this question: the reactivity of the environment and its modularity. With "reactivity", we refer to whether an environment signals partial solutions, i.e., whether valid action sequences lead to items appearing whereas invalid ones do not. Our task design provided maximum reactivity because all valid (and no invalid) sequences were signaled. Would participants still use environmental signals to create action sequences when only a subset of valid sequences were signaled, or when some invalid led to signals as well? "Modularity" refers to the degree to which rewarded action sequences are composed of the same set of sub-sequences. Our task provided maximum modularity because all 4-key sequences were composed of the same set of four 2-key sequences. How would learning be impacted under lower modularity, i.e., if there were more 2-key sequences and each was used less in the 4-key sequences? At the extreme, would participants still acquire hierarchical representations if there was no modularity at all, i.e., if each 4-key sequence was unique, without repeated sub-sequences? Previous research has suggested that humans create hierarchical representations even in the absence of hierarchical structure in a feature-based task (Collins, 2017), and research is needed to show whether this tendency extends to sequential paradigms.

Future research should also employ computational modeling to formally test the process of hierarchy creation we described qualitatively in this study. We have already presented our intended computational model elsewhere, including behavioral predictions across a variety of task variations

(Eckstein and Collins, 2017). Future steps entail formal model comparison to identify the model components that are required to capture human behavior, and estimating the model parameters of the best-fitting model to investigate individual differences, and their relation to task behavior.

5.4 Methods

Participant Sample

Seventy-three undergraduate participants provided informed consent and completed the task online for course credit (58 females, 13 males, 2 declined to answer). Two participants were excluded because they reported present or past psychological illness; 2 more were excluded because they had experienced head trauma or loss of consciousness. Six participants were excluded because they missed more than 50 trials, an elbow point in the number of missed trials (mean missed trials after excluding: 11.6, sd: 9.2, min: 1, max: 35). Two participants were excluded because they took more than 60 minutes to respond (mean duration after excluding: 36 minutes, min: 26, max: 46, sd: 5.3). Eleven participants were excluded because they used pen and paper or other external devices to help with the task. Because of the COVID pandemic, the study was conducted online and no experimenters were present to monitor subjects. We therefore asked participants in a post-experiment questionnaire whether they had used pen and paper, and excluded participants who indicated doing so because the use of pen and paper most likely obscures the memory processes we aimed to investigate. In total, 17 participants were excluded (some fulfilled more than one of the above criteria), leading to a final sample of 56 participants (45 females, 10 males, 1 declined to answer; mean age: 20.6, min: 18.1, max: 31.8, sd: 1.96).

Task procedure

During their experimental session, participants first provided online informed consent, in accordance with the Institutional Review Board of the University of California, Berkeley. Then, participants completed a standard demographics form that included questions about sex, age, race, and medical exclusion criteria. They then worked on the task, which consisted of a tutorial, a learning phase, and an unsignaled transfer phase. After the task, participants completed a strategy questionnaire that asked specific and open-ended questions about their strategies employed during the task.

On each trial, participants pressed four keys with the goal of finding the current trial's goal star. The goal star was shown at the top of the screen and participants were told that they would receive a point each time they achieved a goal star. A point counter at the top of the screen kept track of participants' collected points throughout the task. Participants were allowed a maximum of 2.5 seconds for each trial. Available keys were Q, W, E, and R when participants were using their left hand, and U, I, O, and P when participants were using their right hand. When the four key presses took more than 2.5 seconds, participants were reminded to respond faster next time and the trial was counted as missed. Each trial was followed by a 0.5-second inter-trial interval,

after which the next trial started. Each key press was visualized as a colored circle that appeared in a response box underneath the star machine without delay, with a one-to-one match between key and color. When participants executed a valid 2-key sequence within the first (last) two slots, an item appeared on the left (right) side of the machine's window without delay. Each of the four valid 2-key sequences was represented by a unique item. When participants executed a valid 4-key sequence, a star appeared immediately after the last key press. When the star coincided with the goal star, participants received a point, which was signaled by an increment of the point counter. When a trial did not form a valid 4-key sequence, no star appeared. No other message signaled an incorrect trial.

Valid 2-key and 4-key sequences were constructed to maximize similarity between high-level and low-level transfer, and the same abstract rules were used for all participants. Abstract rules assigned a number to each action, and we avoided systematic biases by randomizing the assignment of actions to keys, 2-key sequences to items, and 4-key sequences to star color. For example, the action sequence "0, 1, 2, 3" could map onto "E, R, W, Q", "W, R, Q, E", or any other permutations of the allowed keys. Valid 2-key sequences are shown and explained in Fig. 5.1C.

Participants completed 12 blocks of 25 trials during the training phase, and 8 blocks of 25 trials during the transfer phase. Within each block, all trials had the same goal star. The order of blocks was pseudo-randomized to avoid the presentation of the same goal star in two subsequent blocks. In addition, pseudo-randomization entailed that each goal star was presented once in each mega-block of 4 blocks, for a total of 3 blocks per star in the learning phase, and 4 blocks per star in the transfer phase. The transition between learning and transfer block was not signaled.

After completing their first machine (including learning and transfer phase), participants took a 1-minute break. After the break, participants were presented with a new machine that differed in color, and were instructed to use the opposite hand from the one they used for the first machine, on a different set of keys. Order of hands was randomized between participants (number of participants who used their right hand for the low-transfer machine: $n = 32$; right hand for high-transfer machine: $n = 26$). Participants who had received the low transfer phase in the first machine, received the high transfer phase in the second one ($n = 30$ after exclusion), and vice versa ($n = 26$). The new machine followed the same abstract rules as the old machine, but keys were randomly re-assigned to the new set of keys to avoid biases and minimize transfer effects. A novel set of items indicated valid 2-key sequences, and a novel set of stars indicated valid 4-key sequences.

The task was written in jsPsych (de Leeuw, 2015), a JavaScript library that facilitates online data collection.

Data analysis

We used Python for data analysis and visualization. Regression models were conducted using the statsmodels package (Seabold and Perktold, 2010). Unless otherwise specified, we used mixed-effects models and defined each participant as a group.

Chapter 6

Conclusion

This chapter links the four experiments presented in this thesis, raises open questions and discusses them, suggests future directions of research, and adds broader points of discussion.

6.1 Developmental Changes in Learning

The study in chapter 2 revealed that some aspects of learning change non-monotonically during development, and that these aspects can be illuminated using computational modeling. Both Reinforcement Learning and Bayesian Inference models captured human behavior, and each provided a sound and coherent explanation for the observed development. This finding raises deeper questions about computational modeling: the goal when fitting computational models is often stated as investigating cognitive processes. If two –mechanistically very different– models fit the same dataset equally well, what can we conclude about the cognitive process? Does neither model capture the cognitive process, does only one capture it but the other does not, or do both capture it but in different ways? In chapter 2, we tried to make progress on this question by assessing whether both models captured different aspects of behavior –using model recovery approaches–, and how similar parameters were between models. Interestingly, the models captured distinguishable aspects of behavior, but many model parameters showed large overlaps, suggesting that the models were not identical, but similar.

Even after these analyses, the question about cognitive processes remains unresolved. Approaching the issue more rigorously, when the same behavior can be described equally well using different cognitive models, a variety of conclusions could be drawn: (a) There is a direct contradiction, and therefore both models must be wrong and neither can be said to capture the underlying cognitive process. (b) There is no direct contradiction because the models describe the cognitive process at a different level of analysis (Marr, 1982). (c) There is no contradiction because the concept of "cognitive processes" does not have a factual counterpart. Humans perceive there to be cognitive processes, but this does not prove that they exist in reality. (d) There is no contradiction because computational models only describe behavior, and never cognitive processes. (e) There is no contradiction because both models capture different aspects of the cognitive process. (f) There

is no contradiction because each model proposes a different theory about the cognitive process, and the task is not designed to discriminate between both theories. (g) There is no contradiction because each model captures different aspects of the cognitive process, which are then mixed in behavior, possibly to different degrees in different individuals. (h) There is no contradiction because "all models are wrong" (Box, 1976).

Options (a), (c), and (d) would be detrimental for cognitive modeling. For (a), if the existence of an equally-good model led to the invalidation of both models, all existing models would most likely be invalid if we assume that we can always construct an equally-good competitor model. For (c), if cognitive processes do not actually exist, the whole endeavor of cognitive research might have been misguided (Churchland and Haldane, 1988). And for (d), if computational models are unable to capture cognitive processes, computational modeling has never been the right tool for cognitive science. Options (b), (e), (f), (g), and (h) rescue computational modeling, but (b) might not apply to our case because both models were process models at the algorithmic level, and (e) is unlikely in our case because both models captured very similar aspects of behavior and parameters showed large overlaps between models. (f) shares spirit with the theory of paradigm shifts (Kuhn, 1996). Different theories –which sometimes are practiced at the same time in science– explain the same phenomena in different ways, but all explanations are, in the end, just views based in different scientific paradigms. Similarly, (h) draws a clear distinction between models and the phenomena they describe, demanding caution when drawing conclusions from models to phenomena.

In conclusion, our results raise questions that seem to go beyond the realm of empirical psychology. Nevertheless, this shows how important it is to understand the strengths and limitations of the method of computational modeling, and the next chapter investigated this issue in more depth.

6.2 What Can We Learn from Computational Modeling?

The research in chapter 3 revealed that the same parameters did not capture the same cognitive processes across tasks, and that the same participants did not show the same parameters across tasks. This contradicts important implicit assumptions of the computational modeling community, including that computational models can be compared between tasks, and that *generic* model parameters are related to psychological traits, real-world behavior, and brain function.

Our research instead shows that model parameters in reinforcement learning play a very different role than is commonly assumed. A model parameter is not an inherent trait that can be measured by fitting a computational model, like someone's inherent intelligence can be measured using an intelligence test (even though there is also considerable controversy on that), or someone's inherent iron deficiency can be measured using a blood test. Instead, each model parameter, in each task, likely reflects a different compilation of behaviors, which might include temporary strategies (e.g., favoring speed over accuracy) or persistent personality traits (e.g., intelligence), participants' psychological qualities (e.g., attentiveness, memory, reasoning style) or task demands (e.g., speed, number of stimuli, volatility), or different aspects altogether.

Bluntly put, different tasks lead to different results in computational modeling because the same parameters simply measure different things in each task. It is important to note that this does

not question the entire endeavor of computational modeling (like our discussion of chapter 2), and it does not negate that model parameters capture meaningful aspects of cognition or behavior. What our result show is just that each task's parameters might need to be interpreted differently, just like each task's accuracy needs to be interpreted differently, in accordance with the task's characteristics. Back to our earlier example on blood tests, we would not expect an iron deficiency test to measure liver health or vice versa.

One limitation of this study was the inability to identify which exact cognitive processes were captured by each model parameter. If it were possible to pinpoint each parameter's cognitive processes, it would be possible to determine which aspects were constant across tasks, and which differed. Without this possibility, we were only able to assess the amount of variance that was shared between parameters, and had to guess which cognitive processes the shared and non-shared aspects corresponded to. Future research needs to devise a way of associating model parameters with cognitive processes, both to assess similarities in parameters between tasks, and to answer the more fundamental question of what parameters measure in each task. To do this, computational models will likely need to be refined to better match complex cognitive processes, and to be carefully validated against underlying mechanisms and individual differences in multiple tasks.

6.3 Hierarchically-Structured Reinforcement Learning in Humans

In chapter 4, we showed that human behavior in a context-based learning task is reproduced well by a hierarchical reinforcement learning model, but not by a flat reinforcement learning model or a hierarchical Bayesian model. Many open questions remain for future research: (1) Are task-sets discrete entities, as assumed in our model, or should they be considered continuous? In real life, no two contexts are ever exactly the same, and we constantly have to generalize beyond previous experience (Schulz, 2017; Shepard, 1987). Can a theory of discrete task-sets explain our flexibility in responding to a continuous array of situations? I think that two answers can be given. One possibility is that even though there is an infinite number of contexts in real life, our mind combines contexts into a discrete number of clusters, and applies the same task-set to each. In this way, a discrete number of task-sets is sufficient for a potentially infinite number of contexts. Another possibility is that both discrete contexts and discrete task-sets (like in our task) are a simplification, and in reality, both exist in a continuous space. In this view, future models will need to incorporate task-set generalization, potentially using function learning (Schulz, 2017).

(2) How separate are task-sets from each other? Incidental observations suggest that different task-sets might share more similarities with each other than would be assumed of they were fully independent. For example, meta-rules acquired in one task-set readily translate to other task-sets (e.g., which responses are allowed, the meaning of feedback) and individuals' characteristics are evident across task-sets (e.g., risk-proneness vs caution). This suggests that task-sets might have access to the same information, or be fed by the same underlying processes, and represent the final, refining step of information processing. Future research should investigate which kinds of

information are shared between task-sets, and at which processing step this sharing occurs.

(3) How many levels of hierarchy are there? In our model, we assumed a lower level of task-sets, and a higher level of master strategy over task-sets. But task-sets themselves are composed of button presses, and button presses are composed of muscle activations. Which structure do these component processes of task-sets have, and is it similar to task-sets? Moving in the opposite direction, might there be several master strategies, each for a different master context, such that we acquire master master strategies to select between them? Future research needs to determine if there is a limit on the number of levels humans are able to represent, and how processes at each level differ from each other.

6.4 Hierarchical Learning of Complex Action Sequences

In chapter 5, we took a detailed look at how participants create hierarchical representations when learning action sequences. We showed that not only explicit rewards can be powerful learning signals, but also other forms of environmental feedback. Participants learned hierarchical representations of the task by decomposing complex action sequences into simpler ones, and thereby scaffolding learning and exploration. We are currently developing a computational model of this task to assess this process in more detail (Eckstein and Collins, 2017). Even though we show behavioral evidence for each step taken by the proposed algorithm, constructing the model is necessary to prove that it will capture all aspects of behavior as a whole.

The task in this study was both compositional –complex elements were composed of a small number of simpler elements– and responsive –simpler elements elicited environmental responses that could be observed by participants. Future research needs to investigate each of these two aspects in more detail. In the case of compositionality, how would participants respond to a non-compositional task, in which complex elements do not share common elements? We designed the task to be compositional under the assumption that the world around us is compositional, but it might be the case that this compositionality itself is an interpretation of our mind. Using a non-compositional task might shed light on whether humans impose compositionality even when it is not present in the environment, in an effort to structure the task and break it up into manageable sub-tasks. In the case of responsiveness, how would participants respond to a more or less responsive task? Arguably, the real world does not signal each meaningful sub-response, and many non-meaningful responses elicit responses. It would be interesting to investigate how participants' strategies will change in a non-responsive task condition. In this case, participants would need to infer which 2-key sequences are valid retroactively, based on the composition of valid 4-key sequences. Partly-responsive and overly-responsive conditions might also shed light on processing. In the former, only some valid 2-key sequences elicit responses, whereas in the latter, not only valid, but also some invalid sequences elicit responses.

My prediction is that the lack of both compositionality and responsiveness would hurt participant performance. I also hypothesize that participants would perceive compositional structure in a task in which there is none. There is evidence that our minds choose to represent tasks hierarchically that are not intrinsically hierarchical in structure (Collins, 2017), and humans seem prone

to perceive pattern when there are none (e.g., superstition). In the case of varying responsiveness, I hypothesize that participants would deal elegantly with both intermediate versions (partly-responsive and overly-responsive), but not the non-responsive version. The reason is that as long as environmental responses are used as teaching signals for the creation of 2-key sequences, both intermediate versions will speed up the learning of at least some or all valid 2-key sequences. Even if not all (or too many) have been pre-learned, their availability should aid in learning compositional 4-key sequences.

6.5 Summary

We conducted four studies with the goal of understanding complex human cognition. Learning is a crucial component of every intelligent system, and we investigated how this skill changes during human development. Computational modeling provided insights into individual strategies, shedding light on the potential mechanisms behind age changes; but it also raised questions about computational modeling more fundamentally, like which questions the method is able to answer. The second study focused on the method of computational modeling, specifically trying to understand the role of model parameters. Instead of capturing the same individual differences each time, many model parameters seem to reflect different processes depending on the task.

Hierarchical representations have the potential to solve a range of otherwise unsolvable problems, and might explain aspects of human intelligence. Indeed, a hierarchical reinforcement learning model explained human behavior in our context-based task better than flat reinforcement learning or hierarchical Bayesian inference. Hierarchical representations themselves might be learned through scaffolding and creating more-and-more complex actions.

Bibliography

Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, *31*(2), 790–795. https://doi.org/10.1016/j.neuroimage.2006.01.001

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2016). Computational Psychiatry: Towards a mathematically informed understanding of mental illness [Publisher: BMJ Publishing Group Ltd Section: Neuropsychiatry]. *Journal of Neurology, Neurosurgery & Psychiatry*, *87*(1), 53–63. https://doi.org/10.1136/jnnp-2015-310737

Albert, D., Chein, J., & Steinberg, L. (2013). The Teenage Brain: Peer Influences on Adolescent Decision Making. *Current Directions in Psychological Science*, *22*(2), 114–120. https://doi.org/10.1177/0963721412471347

Alexander, G., DeLong, M., & Strick, P. (1986). Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex. *Annual Review of Neuroscience*, *9*(1), 357–381. https://doi.org/10.1146/annurev.ne.09.030186.002041

Alexander, W. H., & Brown, J. W. (2015). Hierarchical Error Representation: A Computational Model of Anterior Cingulate and Dorsolateral Prefrontal Cortex. *Neural Computation*, *27*(11), 2354–2410. https://doi.org/10.1162/NECO_a_00779

Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. *Psychology of learning and motivation* (pp. 89–199). Retrieved November 21, 2020, from https://cogs.sitehost.iu.edu/FestschriftForRichShiffrin/pubs/1968%20Human%20Memory.%20Atkinson,%20Shiffrin.pdf

Bacon, P.-L., Harb, J., & Precup, D. (2017). The Option-Critic Architecture. *AAAI*, 1726–1734.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200. https://doi.org/10.1016/j.tics.2008.02.004

Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, *10*(9), 659–669. https://doi.org/10.1038/nrn2667

Badre, D., & Frank, M. J. (2012). Mechanisms of Hierarchical Reinforcement Learning in Cortico–Striatal Circuits 2: Evidence from fMRI. *Cerebral Cortex*, *22*(3), 527–536. https://doi.org/10.1093/cercor/bhr117

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron*, *66*(2), 315–326. https://doi.org/10.1016/j.neuron.2010.03.025

Balleine, B. W., Dezfouli, A., Ito, M., & Doya, K. (2015). Hierarchical control of goal-directed action in the cortical–basal ganglia network. *Current Opinion in Behavioral Sciences*, *5*, 1–7. https://doi.org/10.1016/j.cobeha.2015.06.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron*, *47*(1), 129–141. https://doi.org/10.1016/j.neuron.2005.05.020

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Bernardo, J. M., & Smith, A. F. M. (2009). *Bayesian Theory* [Google-Books-ID: 11nSgIcd7xQC]. John Wiley & Sons.

Bolenz, F., Reiter, A. M. F., & Eppinger, B. (2017). Developmental Changes in Learning: Computational Mechanisms and Social Influences [Publisher: Frontiers]. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02048

Boorman, E. D., Behrens, T. E., & Rushworth, M. F. (2011). Counterfactual Choice and Learning in a Neural Network Centered on Human Lateral Frontopolar Cortex (M. L. Platt, Ed.). *PLoS Biology*, *9*(6), e1001093. https://doi.org/10.1371/journal.pbio.1001093

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*(7), 997–1003. https://doi.org/10.1038/nn.4573

Botvinick, M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*(6), 956–962. https://doi.org/10.1016/j.conb.2012.05.008

Botvinick, M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280. https://doi.org/10.1016/j.cognition.2008.08.011

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Phil. Trans. R. Soc. B*, *369*(1655), 20130480. https://doi.org/10.1098/rstb.2013.0480

Botvinick, M., Weinstein, A., Solway, A., & Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences*, *5*, 71–77. https://doi.org/10.1016/j.cobeha.2015.08.009

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Broadbent, D. E. (1977). Levels, hierarchies, and the locus of control [Place: United Kingdom Publisher: Taylor & Francis]. *The Quarterly Journal of Experimental Psychology*, *29*(2), 181–201. https://doi.org/10.1080/14640747708400596

Brown, V. M., Chen, J., Gillan, C. M., & Price, R. B. (2020). Improving the Reliability of Computational Analyses: Model-Based Planning and Its Relationship With Compulsivity. *Bio-*

logical Psychiatry: Cognitive Neuroscience and Neuroimaging, *5*(6), 601–609. https://doi.org/10.1016/j.bpsc.2019.12.019

Cazé, R. D., & van der Meer, M. A. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, *107*(6), 711–719. https://doi.org/10.1007/s00422-013-0571-5

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., & Rubia, K. (2013). Neural and psychological maturation of decision-making in adolescence and young adulthood. *Journal of Cognitive Neuroscience*, *25*(11), 1807–1823. https://doi.org/10.1162/jocn_a_00447

Churchland, P., & Haldane, J. (1988). Folk Psychology and the Explanation of Human Behaviour [Publisher: [Aristotelian Society, Wiley]]. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *62*, 209–254. Retrieved November 22, 2020, from https://www.jstor.org/stable/4106765

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality? *European Journal of Cognitive Psychology*, *12*(1), 1–36. https://doi.org/10.1080/095414400382181

Collins, A. G. E. (2017). The cost of structure learning. *Journal of cognitive neuroscience*, *29*(10), 1646–1655.

Collins, A. G. E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432. https://doi.org/10.1162/jocn_a_01238

Collins, A. G. E. (2019). Reinforcement learning: Bringing together computation and cognition. *Current Opinion in Behavioral Sciences*, *29*, 63–68. https://doi.org/10.1016/j.cobeha.2019.04.011

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. *Biological Psychiatry*, *82*(6), 431–439. https://doi.org/10.1016/j.biopsych.2017.05.017

Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *34*(41), 13747–13756. https://doi.org/10.1523/JNEUROSCI.0989-14.2014

Collins, A. G. E., Cavanagh, J. F., & Frank, M. J. (2014). Human EEG Uncovers Latent Generalizable Rule Structure during Learning. *The Journal of Neuroscience*, *34*(13), 4677–4685. https://doi.org/10.1523/JNEUROSCI.3900-13.2014

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working Memory Load Strengthens Reward Prediction Errors. *The Journal of Neuroscience*, *37*(16), 4332–4342. https://doi.org/10.1523/JNEUROSCI.2700-16.2017

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis:

Working memory in reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. https://doi.org/10.1037/a0030852

Collins, A. G. E., & Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, *121*(3), 337–366. https://doi.org/10.1037/a0037015

Collins, A. G. E., & Frank, M. J. (2017). Within and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *bioRxiv*, 184812.

Collins, A. G. E., & Koechlin, E. (2012). Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLOS Biology*, *10*(3), e1001293. https://doi.org/10.1371/journal.pbio.1001293

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Cools, R., Clark, L., Owen, A. M., & Robbins, T. W. (2002). Defining the Neural Mechanisms of Probabilistic Reversal Learning Using Event-Related Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *22*(11), 4563–4567. https://doi.org/10.1523/JNEUROSCI.22-11-04563.2002

Cools, R., Frank, M. J., Gibbs, S. E., Miyakawa, A., Jagust, W., & D'Esposito, M. (2009). Striatal Dopamine Predicts Outcome-Specific Reversal Learning and Its Sensitivity to Dopaminergic Drug Administration. *Journal of Neuroscience*, *29*(5), 1538–1543. https://doi.org/10.1523/JNEUROSCI.4467-08.2009

Cooper, R. P., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*(4), 887–916, discussion 917–931. https://doi.org/10.1037/0033-295X.113.4.887

Cox, J., & Witten, I. B. (2019). Striatal circuits for reward learning and decision-making. *Nature Reviews. Neuroscience*, *20*(8), 482–494. https://doi.org/10.1038/s41583-019-0189-2

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, 201506367. https://doi.org/10.1073/pnas.1506367112

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675. https://doi.org/10.1038/s41586-019-1924-6

Dahl, R. E., Allen, N. B., Wilbrecht, L., & Suleiman, A. B. (2018). Importance of investing in adolescence from a developmental science perspective. *Nature*, *554*(7693), 441–450. https://doi.org/10.1038/nature25770

Danielmeier, C., & Ullsperger, M. (2011). Post-Error Adjustments [Publisher: Frontiers]. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00233

Davidow, J. Y., Foerde, K., Galvan, A., & Shohamy, D. (2016). An Upside to Reward Sensitivity: The Hippocampus Supports Enhanced Reinforcement Learning in Adolescence. *Neuron*, *92*(1), 93–99. https://doi.org/10.1016/j.neuron.2016.08.031

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and Performance XXIII*. https://doi.org/10.1093/acprof:oso/9780199600434.003.0001

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans [Number: 7095 Publisher: Nature Publishing Group]. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Dayan, H. (n.d.). Feudal Reinforcement Learning, 8.

Decker, J. H., Lourenco, F. S., Doll, B. B., & Hartley, C. A. (2015). Experiential reward learning outweighs instruction prior to adulthood. *Cognitive, Affective & Behavioral Neuroscience*, *15*(2), 310–320. https://doi.org/10.3758/s13415-014-0332-5

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Delevich, K., Piekarski, D., & Wilbrecht, L. (2019). Neuroscience: Sex Hormones at Work in the Neocortex. *Current Biology*, *29*(4), R122–R125. https://doi.org/10.1016/j.cub.2019.01.013

Delevich, K., Thomas, A. W., & Wilbrecht, L. (2018). Adolescence and "late Blooming"synapses of the Prefrontal Cortex. *Cold Spring Harbor Symposia on Quantitative Biology*, *83*, 37–43. https://doi.org/10.1101/sqb.2018.83.037507

DePasque, S., & Galván, A. (2017). Frontostriatal development and probabilistic reinforcement learning during adolescence. *Neurobiology of Learning and Memory*, *143*, 1–7. https://doi.org/10.1016/j.nlm.2017.04.009

Dezfouli, A., Lingawi, N. W., & Balleine, B. W. (2014). Habits as action sequences: Hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130482–20130482. https://doi.org/10.1098/rstb.2013.0482

Dickstein, D. P., Finger, E. C., Brotman, M. A., Rich, B. A., Pine, D. S., Blair, J. R., & Leibenluft, E. (2010). Impaired probabilistic reversal learning in youths with mood and anxiety disorders. *Psychological Medicine*, *40*(7), 1089–1100. https://doi.org/10.1017/S0033291709991462

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, *13*, 227–303.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., & Niv, Y. (2013). Hierarchical Learning Induces Two Simultaneous, But Separable, Prediction Errors in Human Basal Ganglia. *Journal of Neuroscience*, *33*(13), 5797–5805. https://doi.org/10.1523/JNEUROSCI.5445-12.2013

Diuk, C., Schapiro, A., Córdova, N., Ribas-Fernandes, J., Niv, Y., & Botvinick, M. (2013). Divide and Conquer: Hierarchical Reinforcement Learning and Task Decomposition in Humans.

Computational and Robotic Models of the Hierarchical Organization of Behavior (pp. 271–291). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39875-9_12

Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486. https://doi.org/10.1126/science.1252254

Drzewiecki, C. M., Willing, J., & Juraska, J. M. (2016). Synaptic number changes in the medial prefrontal cortex across adolescence in male and female rats: A role for pubertal onset. *Synapse (New York, N.Y.)*, *70*(9), 361–368. https://doi.org/10.1002/syn.21909

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL2: Fast Reinforcement Learning via Slow Reinforcement Learning.

Eckstein, M. K., & Collins, A. G. E. (2017). CHRL: Combining intrinsic motivation and hierarchical reinforcement learning. *Advances in Neural Information Processing Systems, workshop*.

Eckstein, M. K., & Collins, A. G. E. (2018). Evidence for hierarchically-structured reinforcement learning in humans. *Proceedings of the Annual Meeting of the Cognitive Science Meeting*, 6.

Eckstein, M. K., Master, S. L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2020). Understanding the Unique Advantage of Adolescents in Stochastic, Volatile Environments: Combining Reinforcement Learning and Bayesian Inference [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 2020.07.04.187971. https://doi.org/10.1101/2020.07.04.187971

Farashahi, S., Rowe, K., Aslami, Z., Lee, D., & Soltani, A. (2017). Feature-based learning improves adaptability without compromising precision. *Nature Communications*, *8*(1), 1768. https://doi.org/10.1038/s41467-017-01874-w

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135.

Fitts, P. M. (Ed.). (1951). *Human engineering for an effective air-navigation and traffic-control system* [Pages: xxii, 84]. National Research Council, Div. of.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322. https://doi.org/10.1073/pnas.1800755115

Frank, M. J., & Badre, D. (2012). Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 1: Computational Analysis. *Cerebral Cortex*, *22*(3), 509–526. https://doi.org/10.1093/cercor/bhr114

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*(2), 300–326. https://doi.org/10.1037/0033-295X.113.2.300

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Frankenhuis, W. E., & Walasek, N. (2020). Modeling the evolution of sensitive periods. *Developmental Cognitive Neuroscience*, *41*, 100715. https://doi.org/10.1016/j.dcn.2019.100715

Geddes, C. E., Li, H., & Jin, X. (2018). Optogenetic Editing Reveals the Hierarchical Organization of Learned Action Sequences. *Cell*, *174*(1), 32–43.e15. https://doi.org/10.1016/j.cell.2018.06.012

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3 edition). Chapman; Hall/CRC.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. https://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J. (2017a). Dopamine, Inference, and Uncertainty. *Neural Computation*, *29*(12), 3311–3326. https://doi.org/10.1162/neco_a_01023

Gershman, S. J. (2017b). On the Blessing of Abstraction. *Quarterly Journal of Experimental Psychology*, *70*(3), 361–365. https://doi.org/10.1080/17470218.2016.1159706

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256. https://doi.org/10.1016/j.conb.2010.02.008

Gershman, S. J., & Niv, Y. (2015). Novelty and Inductive Generalization in Human Reinforcement Learning. *Topics in Cognitive Science*, *7*(3), 391–415. https://doi.org/10.1111/tops.12138

Gershman, S. J., & Uchida, N. (2019). Believing in dopamine [Number: 11 Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *20*(11), 703–714. https://doi.org/10.1038/s41583-019-0220-7

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., Paus, T., Evans, A. C., & Rapoport, J. L. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience*, *2*(10), 861–863. https://doi.org/10.1038/13158

Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex (New York, N.Y.: 1991)*, *19*(2), 483–495. https://doi.org/10.1093/cercor/bhn098

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(3), 15647–15654.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., Nugent, T. F., Herman, D. H., Clasen, L. S., Toga, A. W., Rapoport, J. L., & Thompson, P. M. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences*, *101*(21), 8174–8179. https://doi.org/10.1073/pnas.0402680101

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–119. https://doi.org/10.1037/a0021336

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899. https://doi.org/10.1073/pnas.1700811114

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Science*, *10*(3), 281–287. https://doi.org/10.1111/j.1467-7687.2007.00584.x

Graybiel, A. M., & Grafton, S. T. (2015). The Striatum: Where Skills and Habits Meet. *Cold Spring Harbor Perspectives in Biology*, *7*(8), a021691. https://doi.org/10.1101/cshperspect.a021691

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30. https://doi.org/10.1016/j.cobeha.2019.01.005

Gullone, E., Moore, S., Moss, S., & Boyd, C. (2000). The Adolescent Risk-Taking Questionnaire: Development and Psychometric Evaluation [Publisher: SAGE Publications Inc]. *Journal of Adolescent Research*, *15*(2), 231–250. https://doi.org/10.1177/0743558400152003

Guskjolen, A., Josselyn, S. A., & Frankland, P. W. (2017). Age-dependent changes in spatial memory retention and flexibility in mice. *Neurobiology of Learning and Memory*, *143*, 59–66. https://doi.org/10.1016/j.nlm.2016.12.006

Harden, K. P., & Tucker-Drob, E. M. (2011). Individual differences in the development of sensation seeking and impulsivity during adolescence: Further evidence for a dual systems model. *Developmental Psychology*, *47*(3), 739–746. https://doi.org/10.1037/a0023279

Haruno, M., & Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, *19*(8), 1242–1254.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hauser, T. U., Iannaccone, R., Walitza, S., Brandeis, D., & Brem, S. (2015). Cognitive flexibility in adolescence: Neural and behavioral mechanisms of reward prediction error processing in adaptive decision making during development. *NeuroImage*, *104*, 347–354. https://doi.org/10.1016/j.neuroimage.2014.09.018

Hauser, T. U., Will, G.-J., Dubois, M., & Dolan, R. J. (2019). Annual Research Review: Developmental computational psychiatry. *Journal of Child Psychology and Psychiatry*, *60*(4), 412–426. https://doi.org/https://doi.org/10.1111/jcpp.12964

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(10), 3098–3103. https://doi.org/10.1073/pnas.1414219112

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

Izquierdo, A., Brigman, J. L., Radke, A. K., Rudebeck, P. H., & Holmes, A. (2017). The neural basis of reversal learning: An updated perspective. *Neuroscience*, *345*, 12–26. https://doi.org/10.1016/j.neuroscience.2016.03.021

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art [Place: US Publisher: American Psychological Association]. *Journal of Experimental*

Psychology: Human Perception and Performance, *20*(6), 1311–1334. https://doi.org/10.1037/0096-1523.20.6.1311

Javadi, A. H., Schmidt, D. H. K., & Smolka, M. N. (2014). Adolescents adapt more slowly than adults to varying reward contingencies. *Journal of Cognitive Neuroscience*, *26*(12), 2670–2681. https://doi.org/10.1162/jocn_a_00677

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-Mediated Reinforcement Learning Signals in the Striatum and Ventromedial Prefrontal Cortex Underlie Value-Based Choices. *Journal of Neuroscience*, *31*(5), 1606–1613. https://doi.org/10.1523/JNEUROSCI.3904-10.2011

Johnson, C., & Wilbrecht, L. (2011). Juvenile mice show greater flexibility in multiple choice reversal learning than adults. *Developmental Cognitive Neuroscience*, *1*(4), 540–551. https://doi.org/10.1016/j.dcn.2011.05.008

Juraska, J. M., & Willing, J. (2017). Pubertal onset as a critical transition for neural development and cognition. *Brain Research*, *1654*(Pt B), 87–94. https://doi.org/10.1016/j.brainres.2016.04.012

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, *73*, 37–58. https://doi.org/10.1016/j.jmp.2016.03.007

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321. https://doi.org/10.1111/j.1467-7687.2007.00585.x

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kleibeuker, S. W., Dreu, C. K. W. D., & Crone, E. A. (2013). The development of creative cognition across adolescence: Distinct trajectories for insight and divergent thinking. *Developmental Science*, *16*(1), 2–12. https://doi.org/10.1111/j.1467-7687.2012.01176.x

Koechlin, E. (2016). Prefrontal executive function and adaptive behavior in complex environments. *Current Opinion in Neurobiology*, *37*, 1–6. https://doi.org/10.1016/j.conb.2015.11.004

Konidaris, G. (2019). On the necessity of abstraction. *Current Opinion in Behavioral Sciences*, *29*, 1–7. https://doi.org/10.1016/j.cobeha.2018.11.005

Kraemer, H. C., Yesavage, J. A., Taylor, J. L., & Kupfer, D. (2000). How can we learn about developmental processes from cross-sectional studies, or can we? *The American Journal of Psychiatry*, *157*(2), 163–171. https://doi.org/10.1176/appi.ajp.157.2.163

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd edition). University of Chicago Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/S0140525X16001837

Lashley, K. (1951). The Problem of Serial Order in Behavior. In L. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–131). Wiley.

Laube, C., Lorenz, R., & van den Bos, W. (2020). Pubertal testosterone correlates with adolescent impatience and dorsal striatal activity. *Developmental Cognitive Neuroscience*, *42*, 100749. https://doi.org/10.1016/j.dcn.2019.100749

Lee, D., Seo, H., & Jung, M. W. (2012). Neural Basis of Reinforcement Learning and Decision Making. *Annual review of neuroscience*, *35*, 287–308. https://doi.org/10.1146/annurev-neuro-062111-150512

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448. https://doi.org/10.1364/JOSAA.20.001434

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, *93*(2), 451–463. https://doi.org/10.1016/j.neuron.2016.12.040

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 1–85. https://doi.org/10.1017/S0140525X1900061X

Lieshout, L. L. F. v., Vandenbroucke, A. R. E., Müller, N. C. J., Cools, R., & Lange, F. P. d. (2018). Induction and relief of curiosity elicit parietal and frontal activity. *Journal of Neuroscience*, 2816–17. https://doi.org/10.1523/JNEUROSCI.2816-17.2018

Lin, W. C., Delevich, K., & Wilbrecht, L. (2020). A role for adaptive developmental plasticity in learning and decision making. *Current Opinion in Behavioral Sciences*, *36*, 48–54. https://doi.org/10.1016/j.cobeha.2020.07.010

Lourenco, F., & Casey, B. (2013). Adjusting behavior to changing environmental demands with development. *Neuroscience & Biobehavioral Reviews*, *37*(9), 2233–2242. https://doi.org/10.1016/j.neubiorev.2013.03.003

Machado, M. C., Bellemare, M. G., & Bowling, M. (2017). A Laplacian Framework for Option Discovery in Reinforcement Learning [arXiv: 1703.00956]. *arXiv:1703.00956 [cs]*. Retrieved November 14, 2020, from http://arxiv.org/abs/1703.00956

MacKay, D. J. C. (1992). Bayesian Interpolation. *Neural Computation*, *4*(3), 415–447. https://doi.org/10.1162/neco.1992.4.3.415

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt; Co., Inc.

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. E. (2020). Disentangling the systems contributing to changes in learning during adolescence. *Developmental Cognitive Neuroscience*, *41*, 100732. https://doi.org/10.1016/j.dcn.2019.100732

McDougle, S., & Collins, A. G. E. (2019). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. https://doi.org/10.31234/osf.io/gcwxn

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information [Place: US Publisher: American Psychological Association]. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. https://doi.org/10.1038/s41562-017-0180-8

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Natterson-Horowitz, D. B., & Bowers, K. (2019). *Wildhood: The Astounding Connections between Human and Animal Adolescents*. Scribner.

Navarro, D. J. (2019). Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection. *Computational Brain & Behavior*, *2*(1), 28–34. https://doi.org/10.1007/s42113-018-0019-z

Newell, A., Shaw, J., & Simon, H. (1959). Report on a general problem-solving program. *Proceedings of the International Conference on Information Processing.*, 256–264. Retrieved November 21, 2020, from http://bitsavers.informatik.uni-stuttgart.de/pdf/rand/ipl/P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf

Newell, A. (1994). *Unified Theories of Cognition* [Google-Books-ID: 1lbY14DmV2cC]. Harvard University Press.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Nussenbaum, K., & Hartley, C. A. (2019). Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental Cognitive Neuroscience*, *40*, 100733. https://doi.org/10.1016/j.dcn.2019.100733

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning [Publisher: American Association for the Advancement of Science Section: Report]. *Science*, *304*(5669), 452–454. https://doi.org/10.1126/science.1094285

O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, *1*, 94–100. https://doi.org/10.1016/j.cobeha.2014.10.004

Palminteri, S., Kilford, E. J., Coricelli, G., & Blakemore, S.-J. (2016). The Computational Development of Reinforcement Learning during Adolescence. *PLoS Computational Biology*, *12*(6). https://doi.org/10.1371/journal.pcbi.1004953

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Parr, R., & Russell, S. J. (1998). Reinforcement learning with hierarchies of machines. *Advances in neural information processing systems*, 1043–1049.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *arXiv preprint arXiv:1705.05363*.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321. https://doi.org/10.1016/j.cognition.2010.11.015

Petersen, A. C., Crockett, L., Richards, M., & Boxer, A. (1988). A self-report measure of pubertal status: Reliability, validity, and initial norms. *Journal of Youth and Adolescence*, *17*(2), 117–133. https://doi.org/10.1007/BF01537962

Peterson, D. A., Elliott, C., Song, D. D., Makeig, S., Sejnowski, T. J., & Poizner, H. (2009). Probabilistic reversal learning is impaired in Parkinson's disease. *Neuroscience*, *163*(4), 1092–1101. https://doi.org/10.1016/j.neuroscience.2009.07.033

Piekarski, D. J., Boivin, J. R., & Wilbrecht, L. (2017). Ovarian Hormones Organize the Maturation of Inhibitory Neurotransmission in the Frontal Cortex at Puberty Onset in Female Mice. *Current biology: CB*, *27*(12), 1735–1745.e3. https://doi.org/10.1016/j.cub.2017.05.027

Piekarski, D. J., Johnson, C. M., Boivin, J. R., Thomas, A. W., Lin, W. C., Delevich, K., M Galarce, E., & Wilbrecht, L. (2017). Does puberty mark a transition in sensitive periods for plasticity in the associative neocortex? *Brain Research*, *1654*(Pt B), 123–144. https://doi.org/10.1016/j.brainres.2016.08.042

Premack, D. (2007). Human and animal cognition: Continuity and discontinuity [Publisher: National Academy of Sciences Section: Review]. *Proceedings of the National Academy of Sciences*, *104*(35), 13861–13867. https://doi.org/10.1073/pnas.0706147104

Raio, C. M., Konova, A. B., & Otto, A. R. (2020). Trait impulsivity and acute stress interact to influence choice and decision speed during multi-stage decision-making [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *10*(1), 7754. https://doi.org/10.1038/s41598-020-64540-0

RCoreTeam. (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Ribas Fernandes, J., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. (2011). A Neural Signature of Hierarchical Reinforcement Learning. *Neuron*, *71*(2), 370–379. https://doi.org/10.1016/j.neuron.2011.05.042

Romer, D., & Hennessy, M. (2007). A Biosocial-Affect Model of Adolescent Sensation Seeking: The Role of Affect Evaluation and Peer-Group Influence in Adolescent Drug Use. *Prevention Science*, *8*(2), 89. https://doi.org/10.1007/s11121-007-0064-7

Rosenbaum, D. A. (1987). Hierarchical organization of motor programs. *Higher brain functions: Recent explorations of the brain's emergent properties* (pp. 45–66). John Wiley & Sons.

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3rd edition). Pearson.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55. https://doi.org/10.7717/peerj-cs.55

Sarkka, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press. https://doi.org/10.1017/CBO9781139344203

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures* (1st edition). Psychology Press.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. https://doi.org/10.1038/nn.3331

Schmidhuber, J. (2010). Formal Theory of Creativity, Fun, and Intrinsic Motivation. *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230–247. https://doi.org/10.1109/TAMD.2010.2056368

Schultz, W. (2013). Updating dopamine reward signals. *Current Opinion in Neurobiology*, *23*(2), 229–238. https://doi.org/10.1016/j.conb.2012.11.012

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schulz, E. (2017). *Towards a unifying theory of generalization* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/rzj2m

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python, 5.

Sercombe, H. (2014). Risk, adaptation and the functional teenage brain. *Brain and Cognition*, *89*, 61–69. https://doi.org/10.1016/j.bandc.2014.01.001

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Simon, N. W., Gregory, T. A., Wood, J., & Moghaddam, B. (2013). Differences in response initiation and behavioral flexibility between adolescent and adult rats. *Behavioral Neuroscience*, *127*(1), 23–32. https://doi.org/10.1037/a0031328

Singh, S., Barto, A. G., & Chentanez, N. (2005). *Intrinsically Motivated Reinforcement Learning:* (tech. rep.). Defense Technical Information Center. Fort Belvoir, VA. https://doi.org/10.21236/ADA440280

Skinner, B. F. (1977). Why I am not a cognitive psychologist. *Behaviorism*, 1–10.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. (2014). Optimal behavioral hierarchy. *PLoS computational biology*, *10*(8), e1003779.

Sowell, E. R., Peterson, B. S., Thompson, P. M., Welcome, S. E., Henkenius, A. L., & Toga, A. W. (2003). Mapping cortical change across the human life span. *Nature Neuroscience*, *6*(3), 309–315. https://doi.org/10.1038/nn1008

Starkweather, C. K., Gershman, S. J., & Uchida, N. (2018). The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. *Neuron*, *98*(3), 616–629.e6. https://doi.org/10.1016/j.neuron.2018.03.036

Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences*, *9*(2), 69–74. https://doi.org/10.1016/j.tics.2004.12.005

Steinberg, L. (2013). The influence of neuroscience on US Supreme Court decisions about adolescents' criminal culpability. *Nature Reviews Neuroscience*, *14*(7), 513–518. https://doi.org/10.1038/nrn3509

Steinberg, L., Graham, S., O'Brien, L., Woolard, J., Cauffman, E., & Banich, M. (2009). Age Differences in Future Orientation and Delay Discounting. *Child Development*, *80*(1), 28–44. https://doi.org/10.1111/j.1467-8624.2008.01244.x

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2016). Bayes factors for reinforcement-learning models of the Iowa gambling task. *Decision*, *3*(2), 115–131. https://doi.org/10.1037/dec0000040

Sunnaker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian Computation. *PLOS Computational Biology*, *9*(1), e1002803. https://doi.org/10.1371/journal.pcbi.1002803

Sutton, R. S., & Barto, A. G. (2017). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1), 181–211. https://doi.org/10.1016/S0004-3702(99)00052-1

Swainson, R., Rogers, R. D., Sahakian, B. J., Summers, B. A., Polkey, C. E., & Robbins, T. W. (2000). Probabilistic learning and reversal deficits in patients with Parkinson's disease or frontal or temporal lobe lesions: Possible adverse effects of dopaminergic medication. *Neuropsychologia*, *38*(5), 596–612. https://doi.org/10.1016/S0028-3932(99)00103-7

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*(3), 439–471. https://doi.org/10.1037/a0033138

Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A., & Wilbrecht, L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neuroscience*, *15*(9), 1281–1289. https://doi.org/10.1038/nn.3188

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Toga, A. W., Thompson, P. M., & Sowell, E. R. (2006). Mapping brain maturation. *Trends in neurosciences*, *29*(3), 148–159. https://doi.org/10.1016/j.tins.2006.01.007

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2019). Discovery of Hierarchical Representations for Efficient Planning. *bioRxiv*, 499418. https://doi.org/10.1101/499418

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, New Series*, *59*(236), 433–460. http://www.jstor.org/stable/2251299

Turner, B. M., & Sederberg, P. B. (2014). A Generalized, Likelihood-Free Method for Posterior Estimation. *Psychonomic bulletin & review*, *21*(2), 227–250. https://doi.org/10.3758/s13423-013-0530-0

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69–85. https://doi.org/10.1016/j.jmp.2012.02.005

van den Bos, W., Bruckner, R., Nassar, M. R., Mata, R., & Eppinger, B. (2017). Computational neuroscience across the lifespan: Promises and pitfalls. *Developmental Cognitive Neuroscience*. https://doi.org/10.1016/j.dcn.2017.09.008

van den Bos, W., Cohen, M. X., Kahnt, T., & Crone, E. A. (2012). Striatum–Medial Prefrontal Cortex Connectivity Predicts Developmental Changes in Reinforcement Learning. *Cerebral Cortex*, *22*(6), 1247–1255. https://doi.org/10.1093/cercor/bhr198

van den Bos, W., & Hertwig, R. (2017). Adolescents display distinctive tolerance to ambiguity and to uncertainty during risky decision making [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *7*(1), 40962. https://doi.org/10.1038/srep40962

van der Schaaf, M. E., Warmerdam, E., Crone, E. A., & Cools, R. (2011). Distinct linear and non-linear trajectories of reward and punishment reversal learning during development: Relevance for dopamine's role in adolescent decision making. *Developmental Cognitive Neuroscience*, *1*(4), 578–590. https://doi.org/10.1016/j.dcn.2011.06.007

Verharen, J. P. H., Adan, R. A. H., & Vanderschuren, L. J. M. J. (2019). Differential contributions of striatal dopamine D1 and D2 receptors to component processes of value-based decision making [Number: 13 Publisher: Nature Publishing Group]. *Neuropsychopharmacology*, *44*(13), 2195–2204. https://doi.org/10.1038/s41386-019-0454-0

Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., & Kavukcuoglu, K. (2017). FeUdal Networks for Hierarchical Reinforcement Learning. *arXiv:1703.01161*.

Vezhnevets, A. S., Wu, Y. T., Eckstein, M., Leblond, R., & Leibo, J. Z. (2020). OPtions as REsponses: Grounding Behavioural Hierarchies in Multi-Agent Reinforcement Learning. *Proceedings of ICML*, 10.

Waltz, J. A., & Gold, J. M. (2007). Probabilistic reversal learning impairments in schizophrenia: Further evidence of orbitofrontal dysfunction. *Schizophrenia Research*, *93*(1), 296–303. https://doi.org/10.1016/j.schres.2007.03.010

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Watabe-Uchida, M., Eshel, N., & Uchida, N. (2017). Neural circuitry of reward prediction error. *Annual review of neuroscience*, *40*, 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109

Watanabe, S. (2013). A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, *14*(Mar), 867–897. Retrieved October 30, 2019, from http://www.jmlr.org/papers/v14/watanabe13a.html

Watson, J. B. (1913). Psychology as the Behaviorist Views it. *Psychological review*, *20*(2), 158.

Wechsler, D., & Matarazzo, J. D. (1972). *Wechsler's Measurement and Appraisal of Adult Intelligence* (5th and enlarged). Williams & Wilkins.

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2015). 8-Month-Old Infants Spontaneously Learn and Generalize Hierarchical Rules. *Psychological science*, *26*(6), 805–815. https://doi.org/10.1177/0956797615571442

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2016). Role of Prefrontal Cortex in Learning and Generalizing Hierarchical Rules in 8-Month-Old Infants. *The Journal of Neuroscience*, *36*(40), 10314–10322. https://doi.org/10.1523/JNEUROSCI.1351-16.2016

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *arxiv*. https://doi.org/10.31234/osf.io/46mbn

Wilson, R. C., & Niv, Y. (2012). Inferring Relevance in a Changing World. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00189

Wimmer, G. E., Braun, E. K., Daw, N. D., & Shohamy, D. (2014). Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(45), 14901–14912. https://doi.org/10.1523/JNEUROSCI.0204-14.2014

Xia, L., Master, S., Eckstein, M., Wilbrecht, L., & Collins, A. G. E. (2020). Learning under uncertainty changes during adolescence. *Proceedings of the Cognitive Science Society*.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science (New York, N.Y.)*, *345*(6204), 1616–1620. https://doi.org/10.1126/science.1255514