Understanding the Unique Advantage of Adolescents in Stochastic, Volatile Environments: Combining Reinforcement Learning and Bayesian Inference

Maria K. Eckstein[1], Sarah L. Master[1], Ronald E. Dahl[2], Linda Wilbrecht[1,3], and Anne G.E. Collins[1]

[1]Department of Psychology, 2121 Berkeley Way West
[2]Institute of Human Development, 2121 Berkeley Way West
[3]Helen Wills Neuroscience Institute, 175 Li Ka Shing Center
Berkeley, California 94720 USA

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.

Keywords: reinforcement learning, Bayesian inference, computational modeling, development, volatility

2. 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 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. 6), on a task with volatile structure and stochastic outcomes (Fig. 1A, B). The goal of the task was to collect rewards, which were hidden in one of two locations (Fig. 1A). Which location was rewarding changed unpredictably several times ("task switch"), and the rewarded location provided rewards only 75 percent of the time (Fig. 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. 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.

Computational modeling studies often select one type of cognitive 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. This approach 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 qualitatively and quantitatively, and a different model type is unlikely to lead to even better results (Palminteri et al., 2017). 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.

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. 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., 2019; 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., 2019) 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.

3. Results

3.1. 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. 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. 1B). Participants' goal was to collect as many gold coins as possible. More task details are provided in the Methods.

3.2. 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. 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. 1D; for age differences, see suppl. Fig. 15).

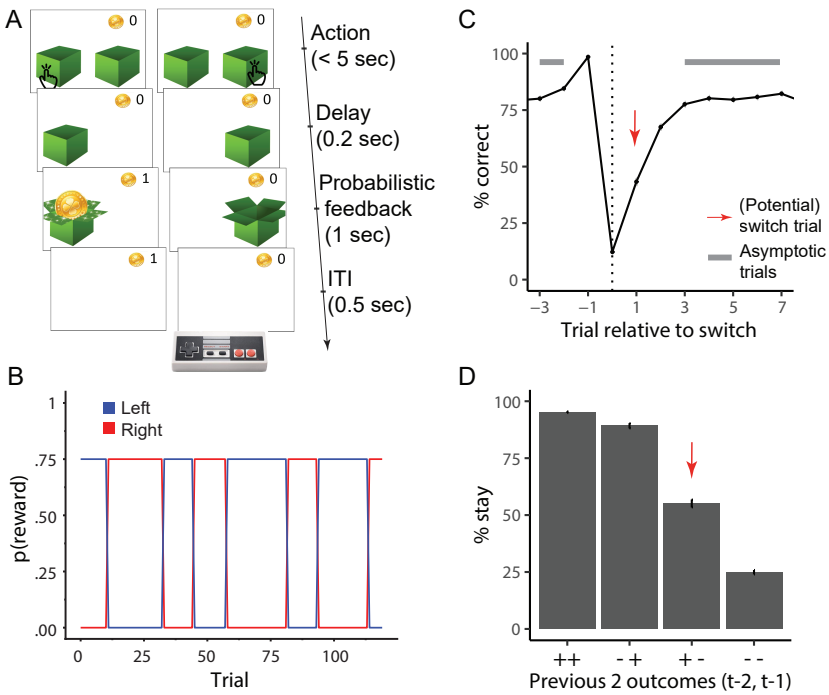

Figure 1: (A) Task design. On each trial, participants chose one of two boxes, using the two red buttons of the shown game controller. The chosen box either revealed a gold coin (left) or was empty (right). The probability of coin reward was 75% on the rewarded side, and 0% on the non-rewarded side. (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, suppl. Fig. 13, Fig. 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 1).

Table 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 (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 (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 (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 (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. 13D-F; Fig. 3C-F). Overall task performance peaked in 13-15 year-olds (mid-adolescence), and declined steeply for both younger and older participants (Fig. 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. 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. 3F).

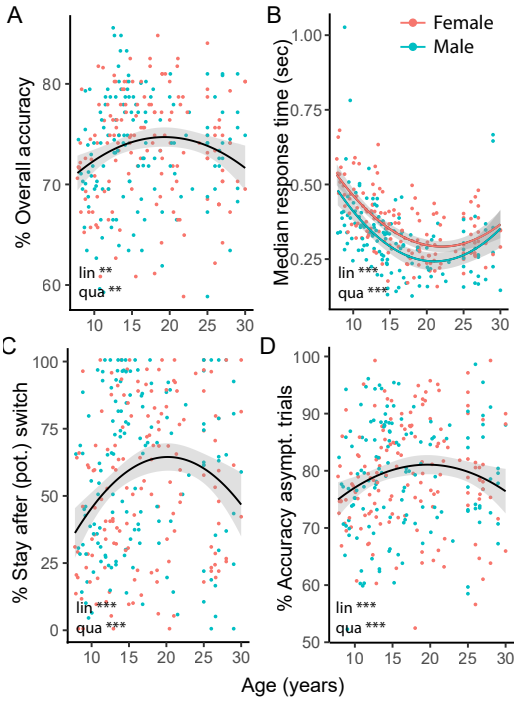

Figure 2: Task performance across age. Each dot shows one participant, color denotes sex. Curves show the fits of linear regression models, with shaded standard errors. "Lin.": significant effect of age on outcome; "qua.": significant effect of squared age on outcome. Stars denote p-values like before. (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. 1C). (D) Accuracy on asymptotic trials (grey bars in Fig. 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 8; Fig. 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. 13B, C, E, and F; suppl. Table

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 8), such that the effect of positive outcomes increased with age and then slowly plateaued (suppl. Fig. 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 8). This shows that the effect of negative outcomes flipped, being weakest in 13-15 year olds for trial $t-1$ (Fig. 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.

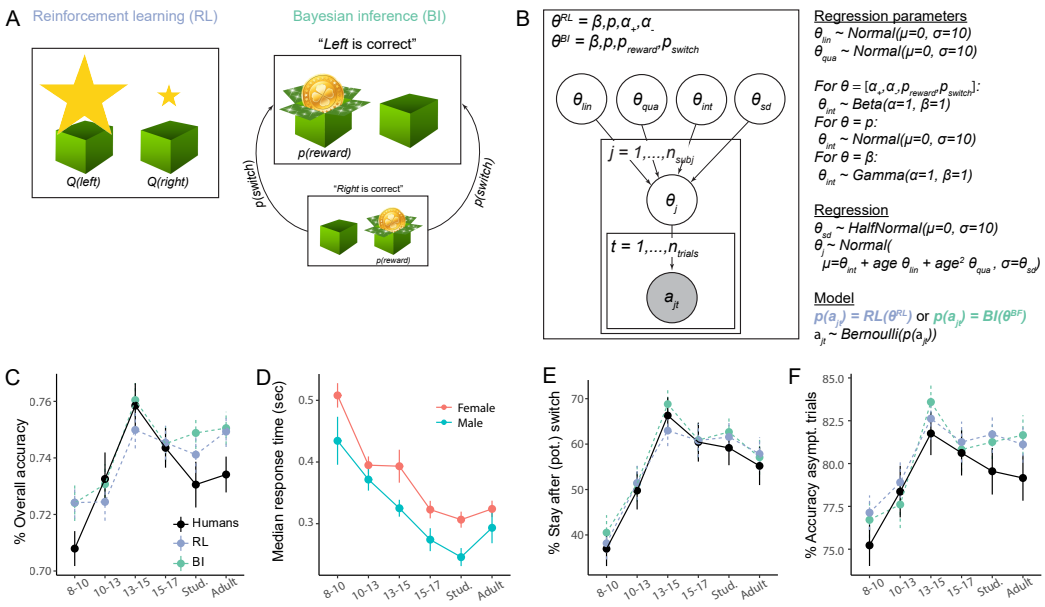

Figure 3: (A) Conceptual depiction of the RL and BI models. In RL (left), actions are selected based on learned values, illustrated by the size of stars ($Q(left)$, $Q(right)$). In BI (right), actions are selected based on a mental model of latent state of the task, which contains task stochasticity ($p(reward)$) 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, binned in age groups. (C), (E), and (F) also show simulated model behavior, verifying that models closely reproduced human behavior and age differences.

3.3. Cognitive Modeling

3.3.1. 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 out-

come, 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., 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. 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. 3B; Methods). This approach recovered individual parameters reliably (suppl. Fig. 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). 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. 3C), the largest proportion of staying after (potential) switch trials (Fig. 3E), best asymptotic performance on non-switch trials (Fig. 3F), and the most efficient use of previous outcomes to adjust future actions (suppl. Fig. 13 D-F). Other tested models (Table 2) did not capture all qualitative patters (suppl. Fig. 16. 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: 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 ± 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

3.3.2. Age Differences in Model Parameters

All model parameters showed age effects (Fig. 4; suppl. Tables 11 and 12). We tested these effects statistically by modeling age explicitly in a hierarchical Bayesian model (Fig. 3B, suppl. Table 11), and also by assessing age-group differences in the posteriors of an age-less hierarchical Bayesian model (suppl. Table 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. 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. 4A, B, E, F). This was reflected in linear and negative quadratic effects of age (suppl. Table 11): Persistence p increased near-linearly from age 8 until 17, and then plateaued around age 18-30 (Fig. 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. 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 12; also see Methods). For p_{switch}, age differences were captured in a significant quadratic effect of age (suppl. Table 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. 4D), captured in a linear effect of age (suppl. Table 11).

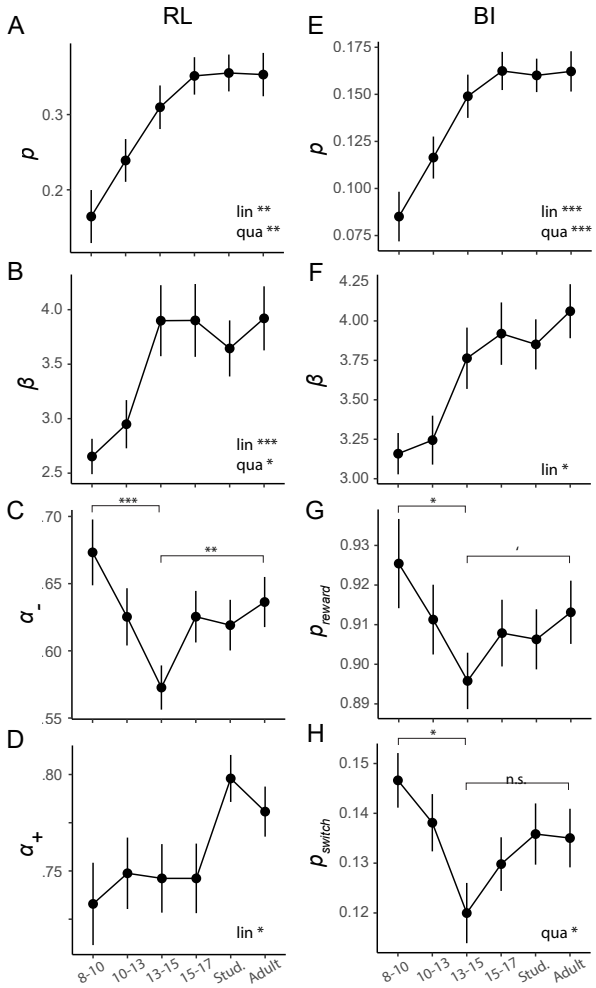

Figure 4: Age-related differences in model parameters for the winning RL (left column) and BI model (right). Stars indicate significant linear ("lin") and quadratic ("qua") effects of age on model parameters, obtained from the hierarchical Bayesian model, and differences between age groups, obtained from the age-less hierarchical Bayesian model (Methods; suppl. Tables 11 and 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.

3.4. 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. 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. 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. 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) and simulated data (Fig. 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.

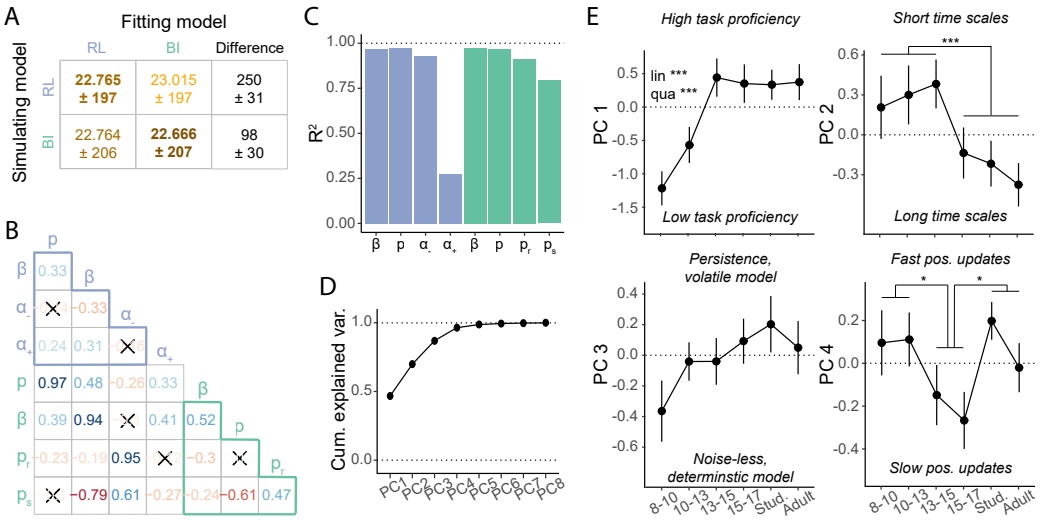

Figure 5: Relating RL and BI models. (A) Model recovery. WAIC scores were worse (larger; lighter colors) when recovering behavior that was simulated from one model (row) using the other model (column), than when using the same model (diagonal), revealing that the models were discriminable. The difference 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. * $p < .05$; ** $p < .01$, *** $p < .001$.

3.5. 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. 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. 18A; suppl. Table 13; 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. 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. 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. 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; suppl. Fig. 13; Fig. 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. 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. 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. 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. 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. 18C). PC3 showed no significant age-related differences (Fig. 5E, bottom-left). PC4 captured the unique variance of α_+ (RL), with a tension between slow and fast updates from positive outcomes (suppl. Fig. 18D). PC4 was lower in 13-17 year olds than both 8-13 year olds and adults (18-30; Fig. 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.

4. 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 vari-

ous 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 all 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).

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., 2019; 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 studies, 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 α_{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 understand-

ing 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. 7, 8, 9; suppl. Tables 4, 5; for a discussion of differences, see suppl. Text). Nevertheless, pubertal measures were so highly correlated with age (suppl. Fig. 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. 10, 11, 12). Thus, we were unable to identify a biological mechanism underlying age besides accumulated experience over time.

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 neu-

ral 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.

5. Methods

5.1. 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 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.

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. 6). For some analyses, we split participants into quantiles based on age. Quantiles were calculated separately within each sex.

5.2. 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.

5.3. 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., 2019; 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. 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).

5.4. Behavioral Analyses

We assessed the effects of age on behavioral outcomes (Fig. 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. 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 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. 13B, C, E, F.

5.5. Computational Models

5.5.1. 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). 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.

5.5.2. 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. 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. 17A) and quantitative model fit (Table 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.

5.5.3. Model Fitting and Comparison

We fitted parameters using hierarchical Bayesian methods (M. D. Lee, 2011; Fig. 3B), and found that the obtained results clearly superseded those of classical maximum-likelihood fitting in terms of parameter recovery (suppl. Fig. 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. 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 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 15, 17, and 16, and to calculate group means in Fig. 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.

5.5.4. Integrating RL and BI Models

Model Recovery between RL and BI (Fig. 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. 5B). We used Spearman correlation, the nonparametric 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. 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. 5C shows the explained variance (R^2) of each model.

5.5.5. 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. 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. 5E. For PC2 and PC4, we also conducted post-hoc t-tests, correcting for multiple comparison using the Bonferroni method (Table 3).

Table 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	*

6. Acknowledgments

Numerous people contributed to this research: Lance Kriegsfeld, Celia Ford, Jennifer Pfeifer, Megan Johnson, Liyu Xia, Rachel Arsenault, Josephine Christon, Shoshana Edelman, Lucy Eletel, Neta Gotlieb, Haley Keglovits, Julie Liu, Justin Morillo, Nithya Rajakumar, Nick Spence, Tanya Smith, Benjamin Tang, Talia Welte, Lucy Whitmore, and Amy Zou. We are also grateful to our participants and their families. The work was funded by National Science Foundation SL-CN grant 1640885 to RD, AGEC, and LW.

References

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

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

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

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

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

Collins, A. G., & 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

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

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., 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. E. (2019). Distentangling the systems contributing to changes in learning during adolescence. *bioRxiv*, 622860. https://doi.org/10.1101/622860

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

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? *Develop-*

mental Cognitive Neuroscience, *40*, 100733. https://doi.org/10.1016/j.dcn.2019.100733

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

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

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

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

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

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, Cambridge University Press. https://doi.org/10.1017/CBO9781139344203

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

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

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

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

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

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

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

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 Rein-

forcement Learning. *Cerebral Cortex*, *22*(6), 1247–1255. https://doi.org/10.1093/cercor/bhr198

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

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

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

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. (2020). Learning under uncertainty changes during adolescence, In *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

7. Supplemental Material

7.1. 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., 2019. PDS scores and testosterone levels were highly correlated with age for both males and females (suppl. Fig. 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. 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. 7, 8, 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. 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. 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. 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. 9).

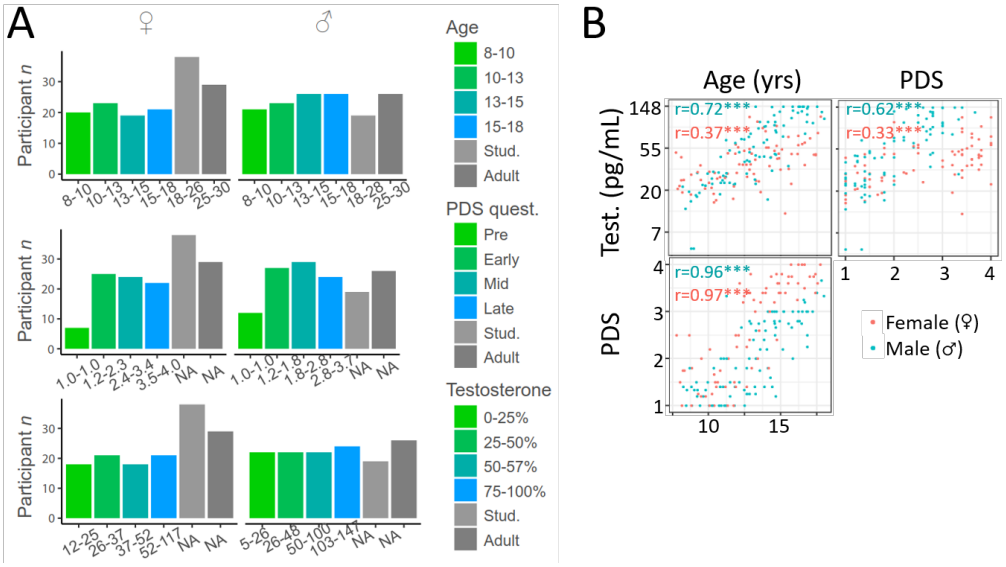

Figure 6: Participant sample and pubertal development. A) Number of participants in each bin, separately for each sex. Top: Age quantiles, which are the basis of Figures 2, 3, 4, and 5, and suppl. Figures 13, 15, 16, and 17. Numbers on the x-axis indicate the age ranges that went into each quantile bin, 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.

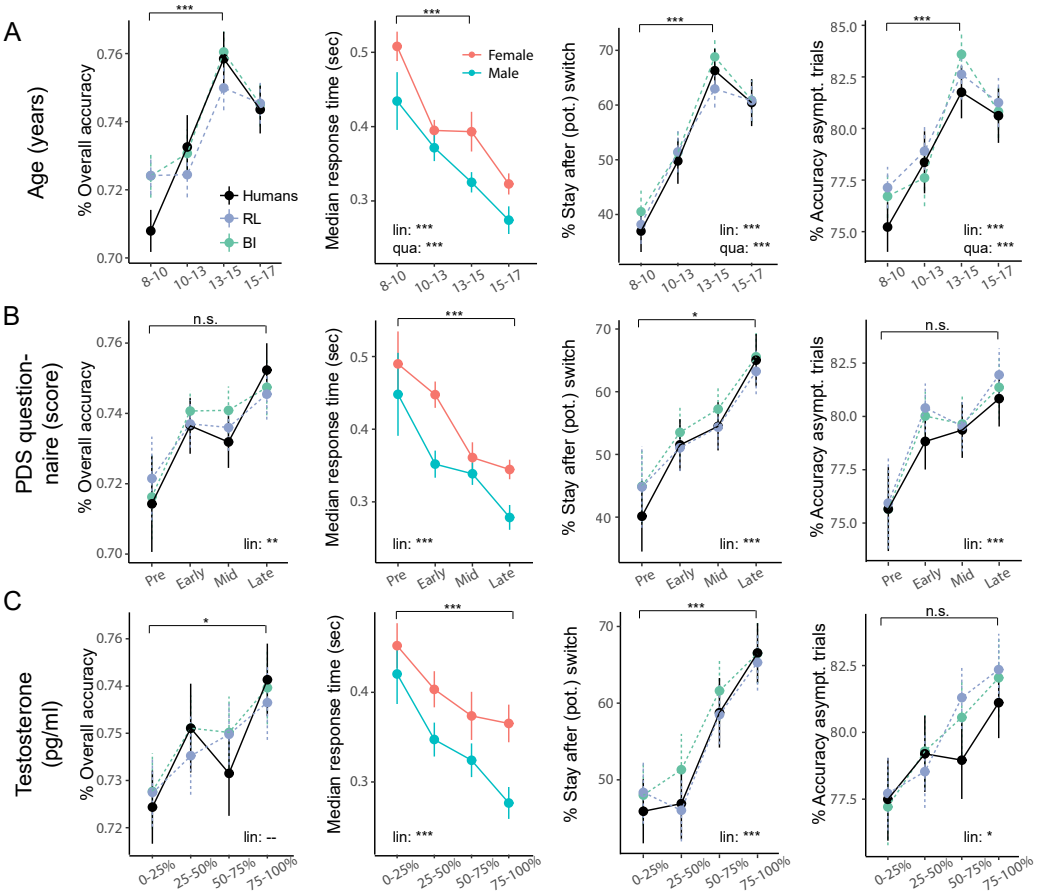

Figure 7: Behavior broken up by age / PDS / Testosterone bins. Significance bars and stars show the results of planned t-tests. A) Same data as in Fig. 3. Planned t-tests compared 8-10 year olds to 13-15 year olds. B) Same measures, but broken up by PDS bins. T-tests compared pre-pubertal to late-pubertal 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 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 (7B, left)	PDS	0.069	2.9	0.0038	**
	Sex	0.017	0.37	0.71	
Response times (7B, 2nd-to-left)	PDS	-0.13	-4.9	< 0.001	***
	Sex	0.25	4.8	< 0.001	***
Stay after (pot.) switch (7B, 2nd-to-right)	PDS	0.48	3.5	< 0.001	***
	Sex	0.76	2.9	0.0036	**
Asymptotic performance (7B, right)	PDS	0.25	4.2	< 0.001	***
	Sex	0.098	0.9	0.39	
Overall accuracy (7C, left)	Test.	< 0.0001	1.2	0.24	
	Sex	0.032	0.69	0.49	
Response times (7C, 2nd-to-left)	Test.	-0.0034	-5.1	< 0.001	***
	Sex	0.010	1.9	0.049	*
Stay after (pot.) switch (7C, 2nd-to-right)	Test.	0.012	3.5	< 0.001	***
	Sex	0.27	1.0	0.29	
Asymptotic performance (7C, right)	Test.	0.0034	2.2	0.029	*
	Sex	0.12	1.0	0.34	

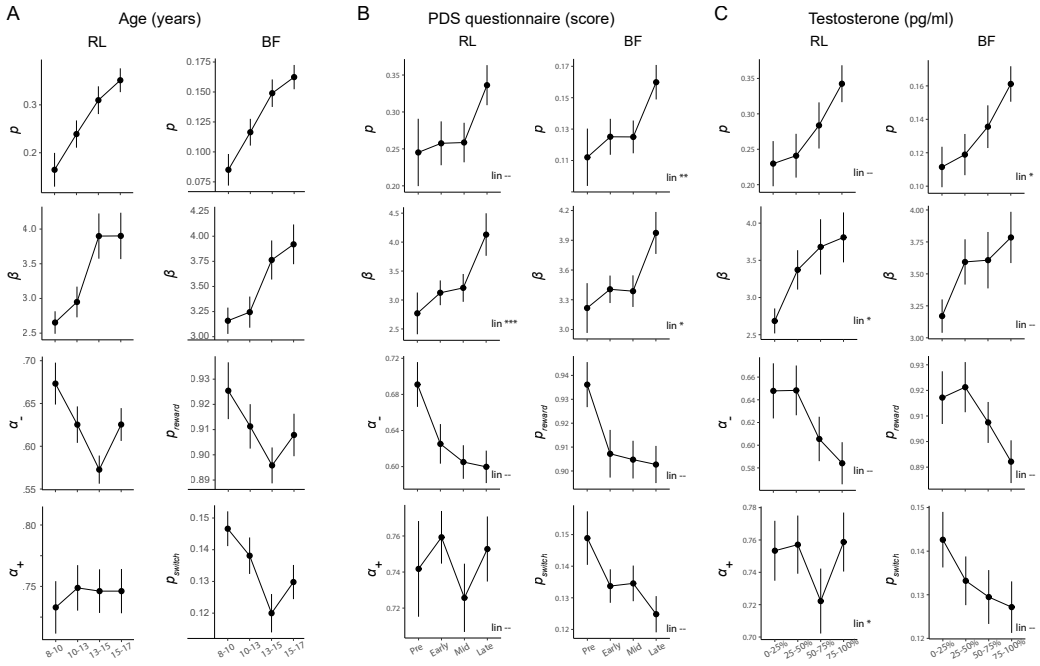

Figure 8: Model parameters broken up by age / PDS / Testosterone bins. A) Participants younger than 18 years of age, reproduced from Fig. 4. B)-C) Same data, broken up by PDS / testosterone bins. ”lin.” indicates whether a linear effect of the measure of interest (PDS / testosterone) reached significance in a linear regression model.

Table 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	*

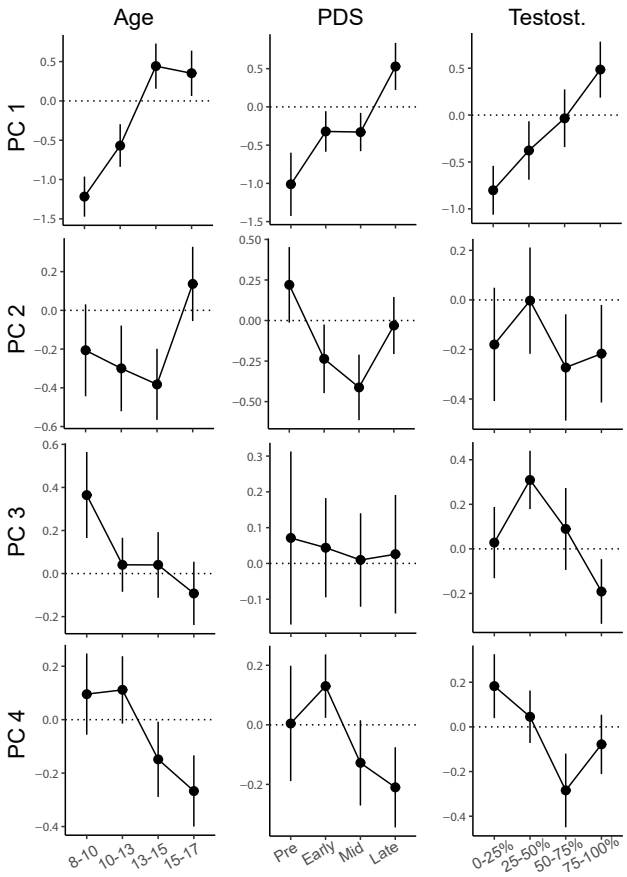

Figure 9: Model parameter PCs broken up by age / PDS / Testosterone bins. Left row: Participants younger than 18 years of age, reproduced from Fig. 5. Middle (right) row: same data, but broken up by PDS (testosterone) bins.

7.2. 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., 2019). 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. 10A, right-most column). Nevertheless, this pattern is difficult to interpret because pubertal status was heavily confounded 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. 10A) or salivary testosterone levels (suppl. Fig. 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 6; suppl. Fig. 11, 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. 11) and testosterone (suppl. Fig. 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 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 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	

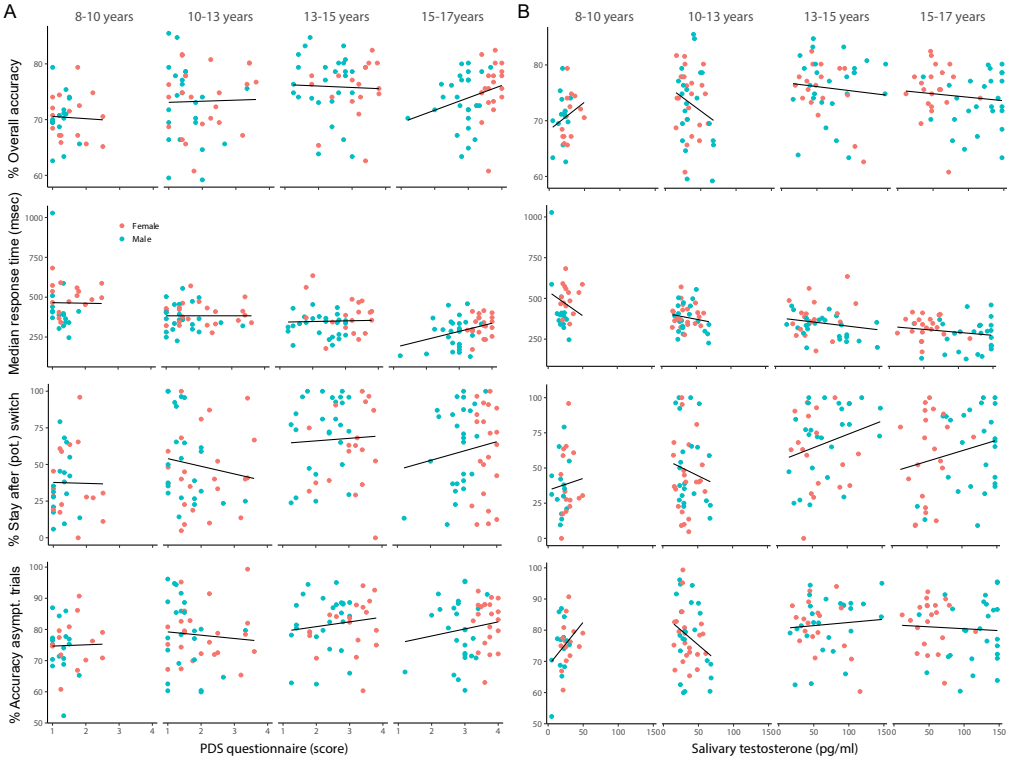

Figure 10: Effects of pubertal status on performance, controlling for age. Each column shows one age group, colors denote sex. Pubertal status was determined by (A) PDS questionnaire, or (B) salivary testosterone.

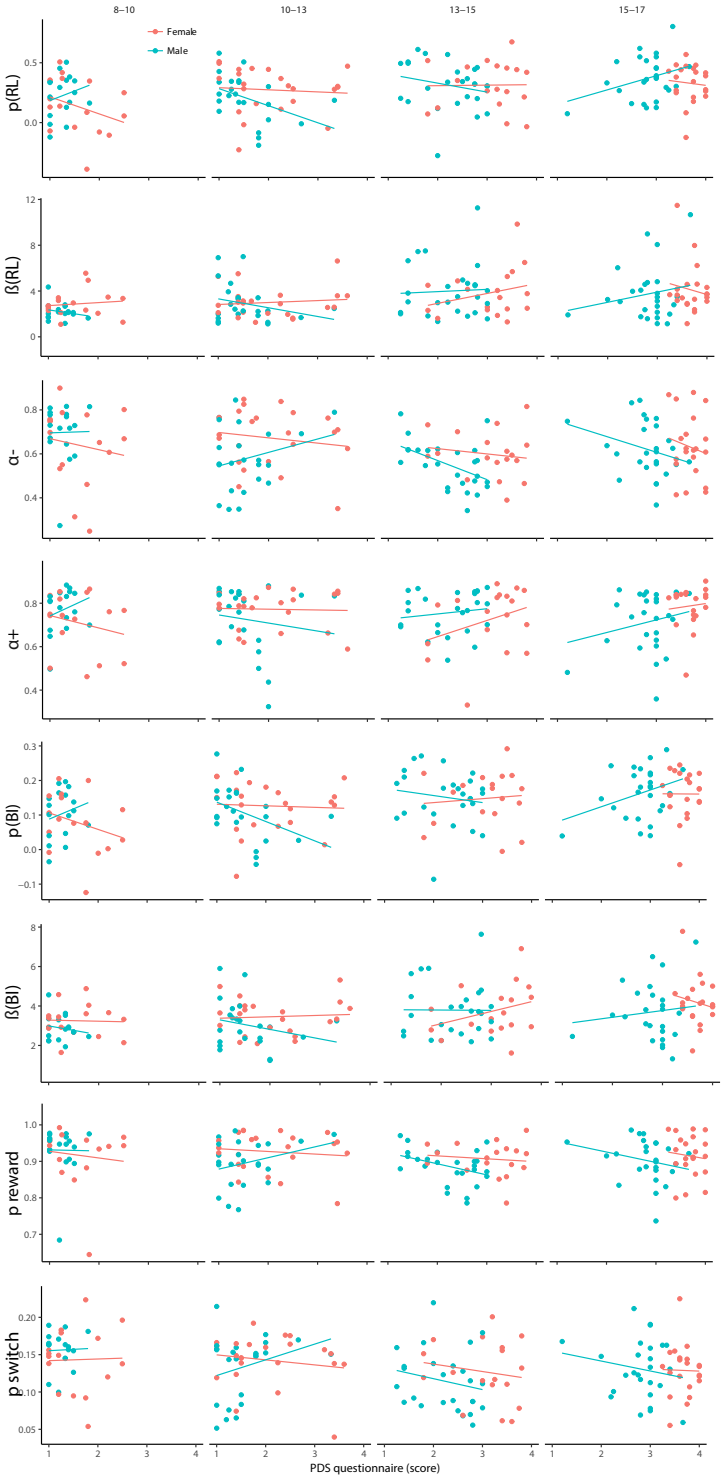

Figure 11: Effects of pubertal status (PDS questionnaire) on model parameters, controlling for age. Each column shows one age group, and colors denote sex.

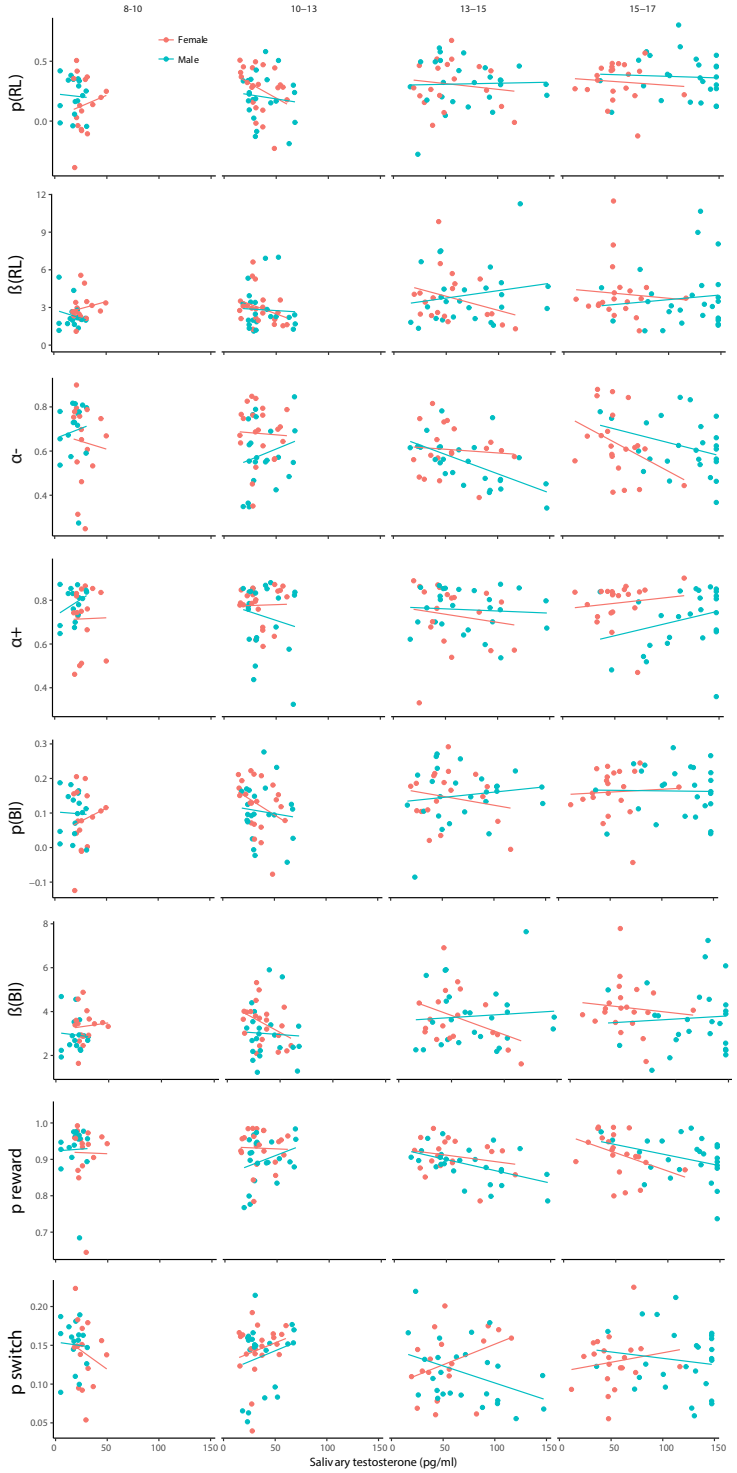

Figure 12: Effects of pubertal status (salivary testosterone levels) on parameters, controlling for age. Each columns shows one parameter. each row one age group, and colors denote sex.

7.3. Additional Behavioral Analyses

The youngest children showed the lowest overall and asymptotic accuracy (Fig. 3C, F) and were the most likely to switch after a single negative outcome (Fig. 3E, suppl. Fig. 15B, middle). This explains why they were also fastest at switching (suppl. Fig. 13A, D; suppl. Table 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. 2B, 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.

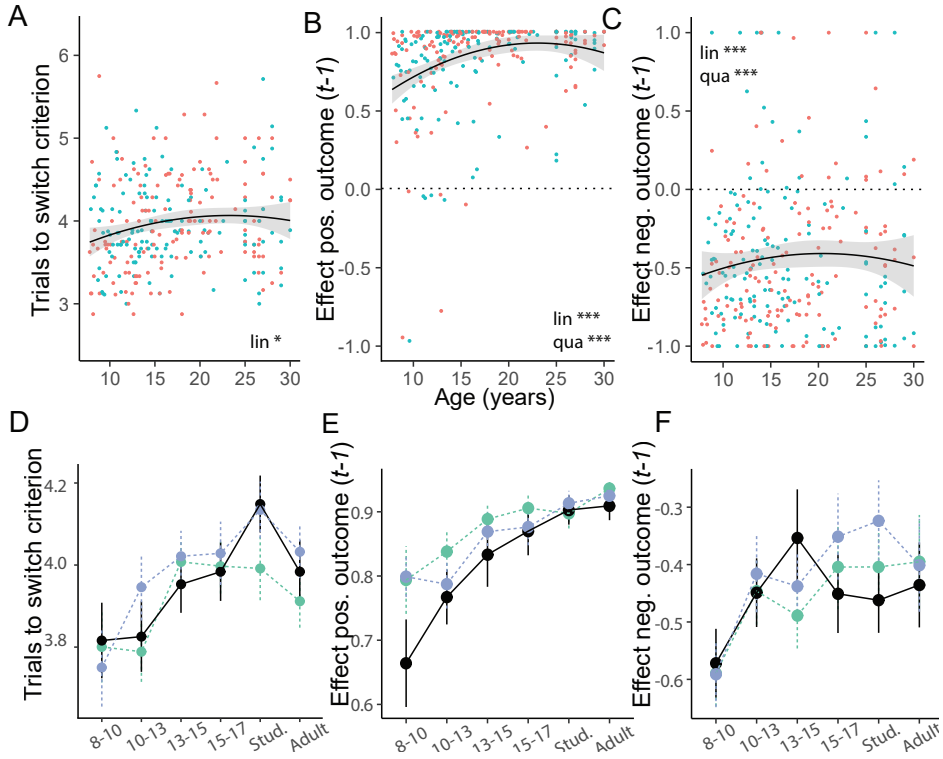

Figure 13: Human behavior (A-C) and model validation (D-F) for additional behavioral measures. (A, D): Number of trials after task switch until participants reached performance criterion (2 correct responses). (B-F): Effect of previous negative (B, E) and positive (C, F) outcomes on choices. "$t-1$": Outcome occurred 1 trial before choice, i.e., delay $i = 1$. Regression weights were tanh transformed for visualization.

Table 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 (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	

7.4. 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 8.

7.5. Meta-Priors for Hierarchical Bayesian Models

When specifying the hierarchical Bayesian models (Fig. 3B and age-less version) for parameter fitting, we chose uninformative priors, as detailed in Table 9.

7.6. Statistics for Hierarchical Bayesian Models

We verified convergence of the Hierarchical Bayesian Model (Fig. 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 10; Salvatier et al., 2016).

Table 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	$[< 0.001; 0.01]$	$[281; 1,576]$	$[1.000; 1.004]$

Table 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 9: Priors for hierarchical Bayesian model (top: age-based; bottom: age-less). For the age-based version, θ is a stand-in for any parameter, i.e., there are eight different distributions θ_{int}, one for each parameter. In the case of the regression intercept θ_{int}, these distributions differ between parameters; in the case of θ_{lin}, θ_{qua}, and θ_{sd}, the same distributions apply for all eight parameters.

	Model parameter	**Distribution (parameters)**	**Notes**
	Age-based model (Fig. 3B)		
θ_{int}	β	Gamma($\alpha = 1$, $\beta = 1$)	Gamma distribution over $[0, \infty[$
	p	Normal($\mu = 0$, $= 10$)	Wide Normal
	α_+, α_-, p_{reward}, p_{switch}	Beta($\alpha = 1$, $\beta = 1$)	Uniform over [0, 1]
θ_{lin}, θ_{qua}	(all parameters)	Normal($\mu = 0$, $= 10$)	Wide Normal over $]-\infty, \infty[$
θ_{sd}	(all parameters)	HalfNormal($\mu = 0$, $= 10$)	Truncated Normal over $[0, \infty[$
	Age-free model (Methods)		
	β	Gamma($\alpha = 1$, $\beta = 1$)	Gamma distribution over $[0, \infty[$
	p	Normal($\mu = 0$, $= 10$)	Wide Normal
	α_+, α_-, p_{reward}, p_{switch}	Beta($\alpha = 1$, $\beta = 1$)	Uniform over [0, 1]

7.7. 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 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. 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. 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.

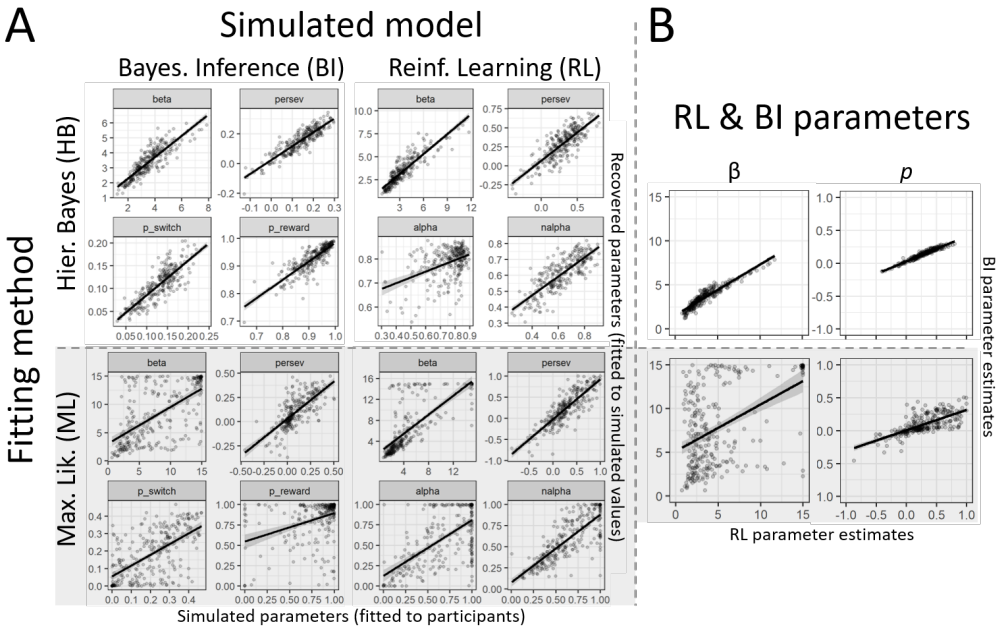

Figure 14: Model validation using hierarchical Bayesian model fitting (top, unshaded), as well as classical Maximum likelihood fitting (bottom, shaded). The results of hierarchical Bayesian fitting are presented in the main text. A) Simulate-and-recover procedure. The x-axes of all graphs show the parameter 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).

7.8. 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.

15). Simpler, non-winning models, on the other hand, failed to capture human characteristics (suppl. Fig. 17, 16).

Raw fitted parameters, obtained from the age-free model (Methods; suppl. Fig. 17, 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. 3B). Significant effects (suppl. Table 11) are shown as lines in suppl. Figures 17 and 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 12.

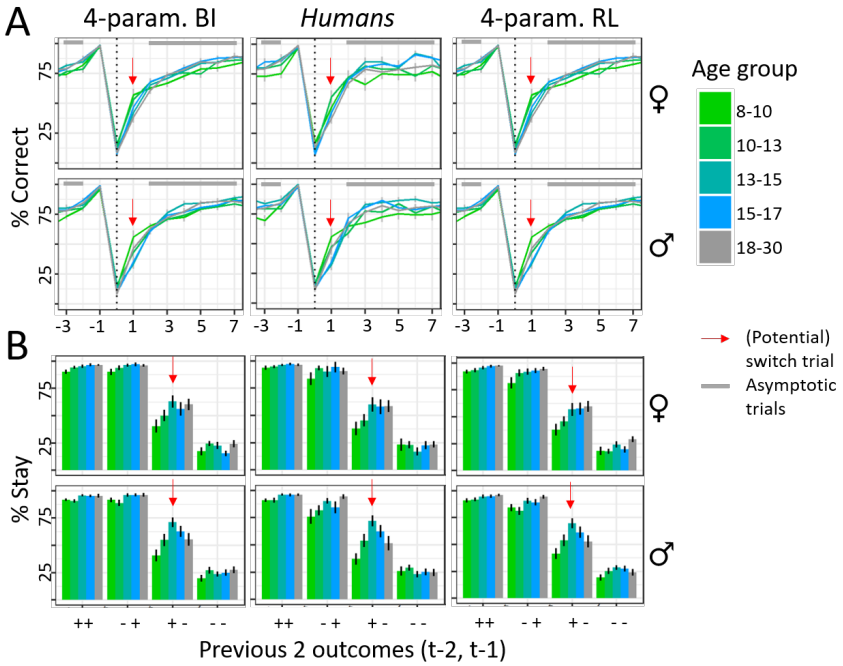

Figure 15: Human and model behavior, showing that models closely reproduced human patterns, A) Behavior in response to switch trials. Colors refer to age groups, red arrows show switch trials, grey bars trials of asymptotic performance. Both models captured quicker switching on switch trials in younger (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.

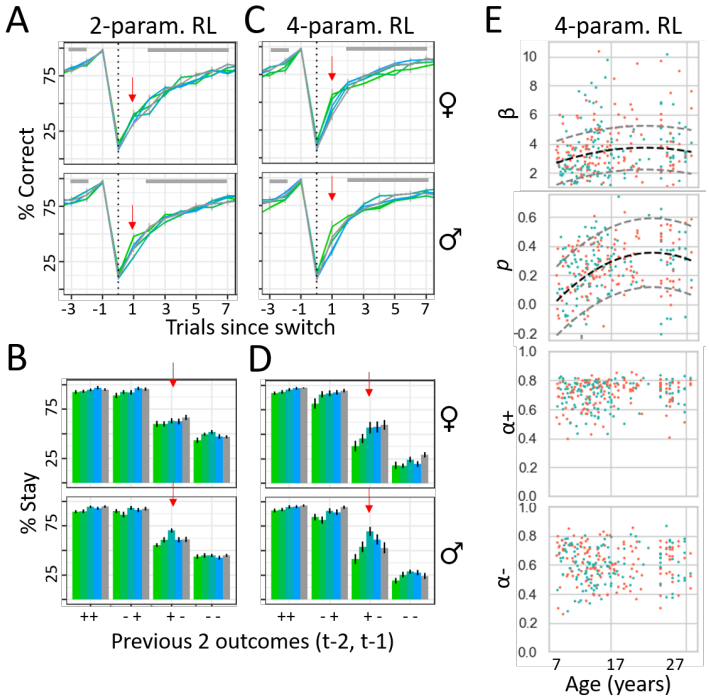

Figure 16: Qualitative fit of different versions of the RL model. Model behavior is shown in the same way as human behavior in suppl. Fig. 15. A-B) Behavior of simulations from the basic, 2-parameter version, with free parameters α and β. Lacking counter-factual updating and the ability to differentiate 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 10), based on the model with access to participants' age (Fig. 3B). This is the same data as summarized in Fig. 4A-D. Colors denote age groups, red arrow (potential) switch trials, and grey bars asymptotic trials, as in suppl. Fig. 15.

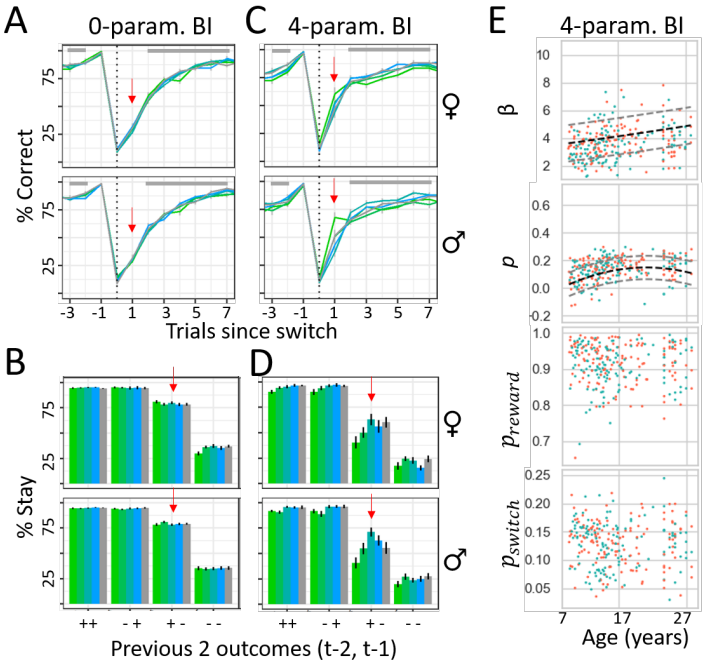

Figure 17: Qualitative fit of different versions of the BI model. Model behavior is shown in the same way as human behavior in suppl. Fig. 15. A-B) Behavior of simulations from the basic, 0-parameter version, in which truthfully $p_{reward} = 0.75$ and $p_{switch} = 0.05$. Lacking free parameters, the 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 10), based on the model with access to participants' age (Fig. 3B). This is the same data as summarized in Fig. 4E-H.

Table 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 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	

7.9. 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. 18, center) from the population mean of each parameter, low levels were determined by adding it (suppl. Table 13). (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. 18).

Table 13: Parameters used to simulate data to visualize the role of each PC.

	p (RL)	β (RL)	α_-	α_+	p (BI)	β (BI)	p_{reward}	p_{switch}
PC1 (plus)	0.57	6.95	0.45	0.87	0.26	5.67	0.84	0.07
PC1 (minus)	0.04	0.1	0.8	0.65	0.02	1.72	0.98	0.2
PC2 (plus)	0.06	2.65	0.31	0.64	0.1	2.98	0.84	0.12
PC2 (minus)	0.54	4.41	0.94	0.89	0.18	4.41	0.98	0.15
PC3 (plus)	0.76	0.49	0.57	0.74	0.29	1.87	0.85	0.19
PC3 (minus)	-0.16	6.56	0.68	0.78	-0.01	5.52	0.97	0.08
PC4 (plus)	0.15	1.68	0.58	1.19	0.1	3.06	0.88	0.14
PC4 (minus)	0.45	5.38	0.67	0.33	0.18	4.33	0.94	0.13
Parameter mean	0.3	3.53	0.62	0.76	0.14	3.69	0.91	0.13

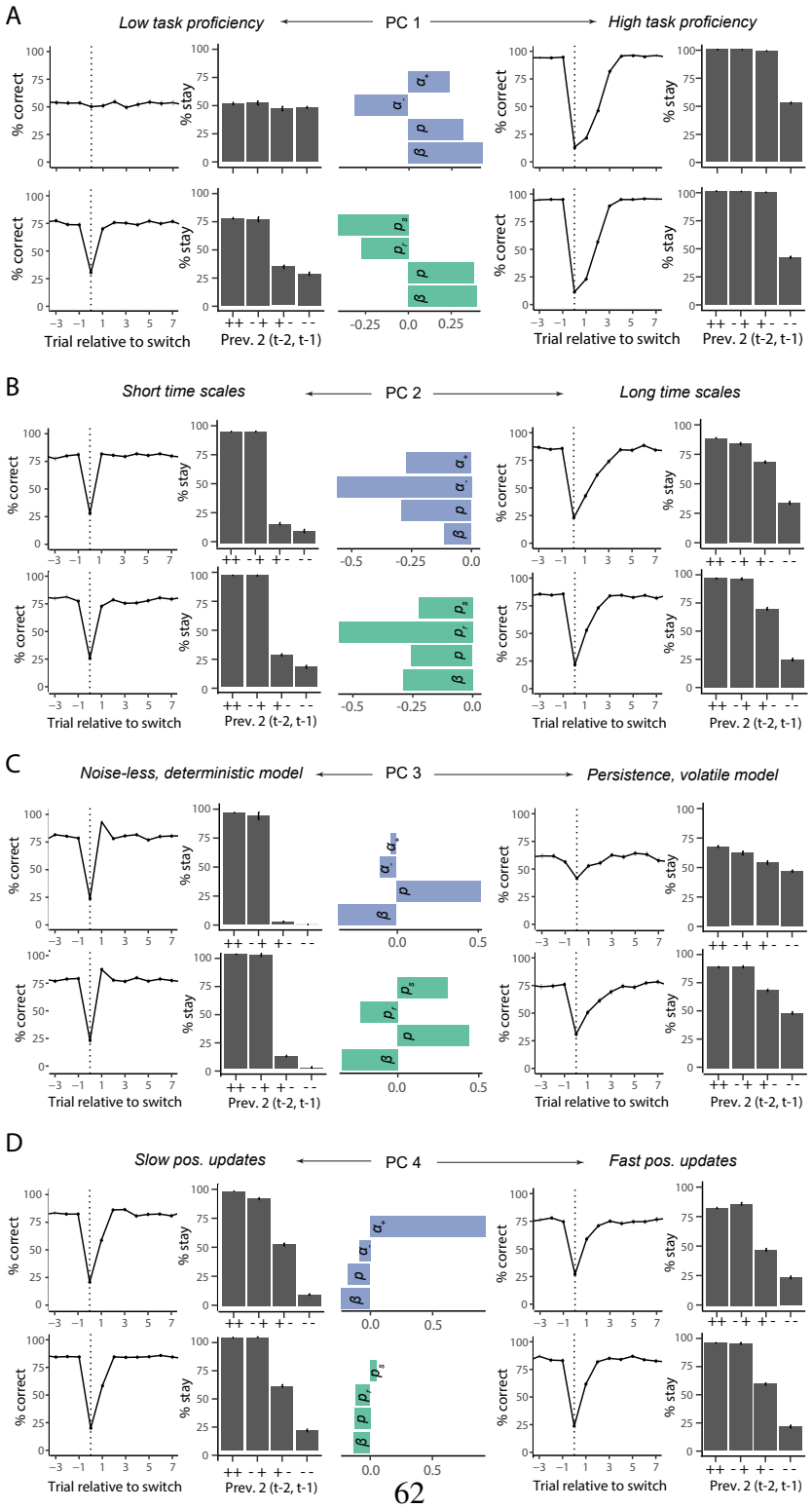

Figure 18: Role of each PC, assessed by simulating behavior at the extremes of each PC. A-D) PC1-4.