Supplemental Information

Reinforcement-Learning Modeling

The modeling approach used in this paper has been described in detail previously by Otto et al. (Otto, Raio, Chiang, Phelps, & Daw, 2013). The model is a variation of the hybrid reinforcement-learning model used by Daw et al. (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). The reinforcement learning model consists of a weighted combination of first, a model-free SARSA(λ) temporal difference algorithm that incrementally updates a fixed value for the first-stage choice based on reward history, and second, a model-based "tree-search" reinforcement learning algorithm (explicit computation of Bellman's equation), which represents all possible choice options and associated outcomes (Sutton & Barto, 1998).

The hybrid model consists of model-free and model-based algorithms, both of which estimate state-action value functions Q(s,a) that map each state-action pair to its expected future value. The task consists of three states (first stage s_A; second stage s_B and s_C), each with two possible actions (a_A and a_B), and a reward (r). For every trial (t), the first and second stages, actions, and rewards are denoted as $s_{1,t}$, $s_{2,t}$, $a_{1,t}$, $a_{2,t}$, $r_{1,t}$ (always zero), and $r_{2,t}$.

Model-free component

The model-free temporal difference algorithm updates the state action values according to the following formula:

$$Q_{MF}(s_{i,t}, a_{i,t}) = Q_{MF}(s_{i,t-1}, a_{i,t-1}) + \alpha\delta_{i,t}$$

with

$$\delta_{i,t} = r_{i,t} + Q_{MF}(s_{i+1,t}, a_{i+1,t}) - Q_{MF}(s_{i,t-1}, a_{i,t-1})$$

As such, δ is the reward-prediction error (RPE), and α is the learning rate parameter. At stage one, reward $r = 0$ and the RPE is driven by the estimate of the second stage action value. At the second stage, $r = 0$ or 1. The eligibility trace λ, which is only carried over across stages for one trial, is used to further update the first-stage action by the second-stage RPE according to:

$$Q_{MF}(s_{1,t}, a_{1,t}) = Q_{MF}(s_{1,t}, a_{1,t}) + \alpha\lambda\delta_{2,t}$$

Model-based component

For the model-based algorithm, the first-stage learning function differed from the model-free algorithm in that it took into account the 70/30-transition probability structure and computed cumulative state-action values from all possible outcomes. The second-stage action value estimate is the same across the model-free and model-based algorithms. As such, the model-based algorithm updates the first stage action values according to the following formula:

$$Q_{MB}(s_{1,t}, a_{j,t}) = P(s_B|s_A, a_j) \max_{a\in\{a_A,a_B\}} Q_{TD}(s_B, a) + P(s_C|s_A, a_j) \max_{a\in\{a_A,a_B\}} Q_{TD}(s_c, a)$$

Choice rule

Finally, to connect the values to choices, we use a softmax choice rule, which assigns a probability to each action according to the combination of both Q_{MB} and Q_{MF}, each weighted with a separate inverse temperature parameter, β_{MB} and β_{MF}, to calculate the stay probability at each stage:

$$P(a_{i,t} = a|s_{i,t}) = \frac{\exp[\beta_{MF} \cdot Q_{MF}(s_{i,t}, a) + \beta_{MB} \cdot Q_{MB}(s_{i,t}, a) + p \cdot \text{rep}(a)]}{\sum_{a'} \exp[\beta_{MF} \cdot Q_{MF}(s_{i,t}, a') + \beta_{MB} \cdot Q_{MB}(s_{i,t}, a') + p \cdot \text{rep}(a')]}$$

where rep*(a)* is defined as 1 for a first-stage action that repeated the action of the previous trial, which when combined by the "stickiness" parameter (p), captures first-order perseveration (p *>0*) or switching (p *<0*). At the second stage, as there is no model-based learning possible, the equation simplifies to only contain the model-free value function Q_{MF} with its own inverse temperature term (β_2).

Group level modeling

The above single-subject modeling was embedded within a multi-level random effects model. All of the free parameters of the model (α, λ, β_{MB}, β_{MF}, β_2, p) were taken as random effects, instantiated separately for each subject s from a common group level distribution. Parameters with infinite support (β_2 and p), the group level distributions were Gaussian with free mean and standard deviation:

$$\beta_{2_s} \sim N\left(\mu_{\beta_2}, \sigma_{\beta_2}\right)$$

To test the dependence of the model-based and model-free effects on age, age was entered into a regression at the group level :

$$\beta_{\text{MB}_s} \sim N\left(\mu_{\text{MB}} + \beta_{MBage} \cdot age(s), \sigma_{\beta_{MB}}\right)$$

And similarly for β_{MF}.

The parameters with support in [0,1] (α and λ) were assumed to be drawn form the group level beta distribution:

$$\alpha_s \sim Beta(A_\alpha, B_\alpha)$$

Finally, we used uninformative priors to estimate the parameters of the group level distributions: for all means, the broad Gaussian *N(0,100)*, for all standard-deviations, the heavy-tailed *Cauchy(0,2.5)*. The priors for the A and B parameters of the beta distributions were given using a change of variables that characterizes the distribution's mean M=A/(A+B) and spread S=1/sqrt(a+b), the latter approximating its standard deviation. This allowed us to take as uninformative hyperpriors the uniform distributions M~*U(0,1)* and *S~U(0,∞)*.

Estimation

We estimated the joint distribution of the parameters of the model, conditional on all subjects' observed choices and rewards. We used Markov Chain Monte Carlo (MCMC) techniques (specifically No-U-Turn variant of Hamiltonian Monte Carlo) as implemented in the Stan modeling language (Stan Development Team, 2015). Given a probabilistic generative model (the above equations) and a subset of observed variables, MCMC techniques provide samples from the conditional joint distribution over the remaining random variables. We ran four chains of 2,000 samples each, discarding the first 1,000 samples of each chain for burn-in. We examined the chains visually for convergences and also computed Gelman and Rubin's (Gelman & Rubin, 1992) potential scale reduction factors. For this, large values indicate convergence problems, whereas values near 1 are consistent with convergence. We ensured that these diagnostics were less than 1.1 for all variables.

Results

Table S1 reports the free parameters of the model by their group-level means and variances over individual subjects. We also report the regression slopes estimating how individuals' parameter settings covaried with age. This uncertainty is reported via 25th, median, and 75th percentiles of the distribution. Of note, the group-level mean α was centered on 0.35, indicative of a more gradual learning process than is ascribed by the regression analysis in the main text, which assumes a learning rate of 1 (that is, only the most recent trial influences choice), supporting the conclusion that our reported effects reflect longer-term incremental learning, and are not limited to patterns of win-stay-lose-shift adjustments.

Table S1. Group level estimates for the free parameters of the reinforcement-learning model and estimated slopes for the covariates.

		Group-level means				
Percentile	β_{MB}	β_{MF}	p	β_2	λ	α
25	0.273	0.268	0.832	0.940	0.915	0.334
50	0.314	0.294	0.894	0.989	0.942	0.355
75	0.356	0.321	0.958	1.036	0.965	0.379
		Group-level variances				
	β_{MB}	β_{MF}	p	β_2	λ	α
25	0.366	0.231	0.678	0.473	0.378	0.485
50	0.401	0.252	0.725	0.508	0.542	0.524
75	0.439	0.278	0.779	0.548	0.720	0.565
	Covariate Slopes					
	β_{MB}-age	β_{MF}-age				
25	0.164	0.033				
50	0.205	0.058				
75	0.246	0.085				

Regression Analysis

In the main text, we used categorical age groupings (i.e., children, adolescents, adults) to visually depict age-related changes in behavioral patterns reflecting reinforcement learning strategies (1st stage stay choices). However, in our statistical analysis of age effects, we chose to treat age continuously, as a categorical division of age groups is somewhat arbitrary and a continuous analysis would better capture gradual developmental changes in reinforcement learning processes.

Here, we additionally repeated the mixed-effects analysis treating age categorically and using linear contrasts (doBy package in R) to test for categorical age group (child, adolescent, and adult) differences in model-free and model-based behavior. Children showed no difference in model-free behavior from adolescents (p=0.49, uncorrected) or adults (p=0.13, uncorrected). However, the model-free effect as significantly higher in adults than adolescents (p=0.033, uncorrected). This difference does not remain significant after correcting for multiple post-hoc comparisons. Moreover, this categorical grouping clearly reveals the absence of a linear increase from childhood to adulthood in model-free choice. As for the model-based comparison, children show significantly less model-based behavior than adolescents (p=0.012, uncorrected), and adults (p=0.00012, uncorrected), and there is no difference between adolescents and adults (p=0.15, uncorrected).

References:

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.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–511. doi:10.1214/ss/1177011136

Otto, a R., Raio, C. M., Chiang, A., Phelps, E. a, & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(52), 20941–6. doi:10.1073/pnas.1312011110

Stan Development Team. (2015). Stan: A C++ Library for Probability and Sampling, Version 2.7.0.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press. doi:10.1109/TNN.1998.712192