Modeling the development of decision making in volatile environments using strategies, reinforcement learning, and Bayesian inference

Maria K. Eckstein[1] **(maria.eckstein@berkeley.edu)**
Sarah L. Master[1] **(sarah.master@berkeley.edu)**
Ronald Dahl[2] **(rondahl@berkeley.edu)**
Linda Wilbrecht[13] **(wilbrecht@berkeley.edu)**
Anne G.E. Collins[1] **(annecollins@berkeley.edu)**

[1]Department of Psychology, 2121 Berkeley Way West
[2]School of Public Health, 233 University Hall
[3]Helen Wills Neuroscience Institute, 175 Li Ka Shing Center
Berkeley, California 94720 USA

Abstract

Continuously adjusting behavior in changing environments is a crucial skill for intelligent creatures, but we know little about how this ability develops in humans. Here, we investigate this question in a large sample using behavioral analyses and computational modeling. We assessed over 200 participants (ages 8-30) on a probabilistic, volatile reinforcement learning task, and measured pubertal development status and salivary testosterone. We used three classes of models to analyze behavior on the task: fixed strategies, incremental reinforcement learning, and Bayesian inference. All model classes provided converging evidence for a decrease in decision noise or exploration with age. Individual models also provided insight into unique aspects of decision making, such as changes in estimated reward probabilities, and sed-specific changes in the sensitivity to positive versus negative outcomes. Our results show that the combination of models can provide detailed insight into the development of decision making, and into complex cognition more generally.

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

Introduction

Adjusting behavior to new circumstances is a crucial skill for intelligent creatures. The current study assesses how this ability develops in human children and adolescents. The maturation of higher-level cortical regions, which occurs late in development, has been associated with increases in IQ (Sowell et al., 2003). We hypothesized that the maturation of subcortical regions, which extends throughout development (Dennison et al., 2013), might similarly be linked to the development of reinforcement learning and decision making (Niv, 2009). To test this, we assessed participants aged 8-18 and 25-30 on a probabilistic choice task in a volatile environment, in which they learned through trial and error which of two stimuli was rewarding at any given time. Once participants selected the rewarding stimulus consistently, it changed, forcing participants to switch behavior. Probabilistic feedback precluded absolute certainty as to which stimulus was the correct one.

We explored a large number of computational models to describe the cognitive processes involved in this task, and to assess developmental changes. A first class of models implemented specific, fixed strategies; a second class employed reinforcement learning (RL); and a third Bayesian inference. We used each class of models to shed light on different aspects of the decision making process, employing their conjunction to provide a more complete picture of human adaptive behavior in volatile environments.

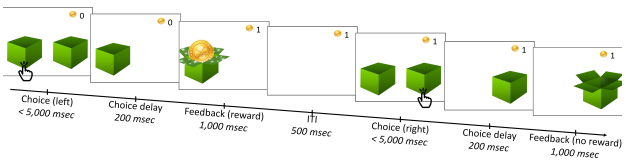

Figure 1: Task design. Participants were given 5 seconds to choose a box, which either revealed a golden coin (reward) or was empty (no reward).

Methods

Participants We recruited 233 participants, 93 children and adolescents (ages 8-18) and 54 adults (ages 25-30), from the community, using protocols approved by the institutional review board of UC Berkeley. Participants were free of present or past psychological and neurological disorders. Compensation consisted in 25$ for the in-lab part and 25$ for completing optional take-home saliva samples. We created equal-sized age groups within non-adults based on quantile splits; adults formed a separate group. Four participants were excluded because they ended the task early.

Experimental design During the two-hour lab visit, participants completed 4 computerized tasks, three questionnaires, and a saliva sample. In the probabilistic switching task, participants were asked to select one of two boxes on each of 150 trials, with the goal of collecting golden coins (reward). One box was correct (75% reward probability), and the other was

incorrect at any time (0% reward probability) (Fig. 1). Contingencies switched without notice after participants reached a criterion of 7-15 rewards: the previously incorrect box became correct. Switches occurred after rewarded trials only and the first correct choice after a switch was also always rewarded. Participants underwent an average of 7.3 switches in the task (range 2-9, sd=1).

Computational models

Strategy models We first implemented a noisy win-stay lose-shift (WSLS) strategy. When participants received a reward for box a, denoted $(a,1)$, the value of choosing a again ("staying") was set to 1, $Q(a|a,1)=1$, and the value of choosing the other box ("shifting") was set to 0, $Q(a_{ns}|a,1)=0$. If participants did not receive a reward $(a,0)$, the value of a was set to 0 and the value of the non-chosen box, a_{ns}, was set to 1. Choice on the subsequent trial was determined by a softmax function, $p(a)=\frac{1}{1+exp(\beta(Q(a_{ns})-Q(a)))}$, where β was fit to individual participants.

We also implemented 2-trial WSLS, an extension of WSLS that switched when two trials were unrewarded. The value of switching was 1 when two consecutive trials for the same box failed to produce reward, $Q(a_{ns}|a,0,a,0)=1$. Otherwise, the value of staying was 1 (and the value of switching was 0). When different boxes were chosen, the rewarded one was repeated, $Q(a|a,1,a_{ns},0)=1$ and $Q(a|a_{ns},0,a,1)=1$. When none was rewarded, the more recent one was repeated, $Q(a|n_{ns},0,a,0)=1$.

RL models In reinforcement learning, learned Q-values guide choices (Sutton and Barto, 2017). We tested distinct RL models with different state spaces and different parameters. In the basic α-β model, values were updated according to the observed outcome $o\in(0,1)$: $Q(a)=Q(a)+\alpha(o-Q(a))$. Values were initialized at 0.5.

In the 1-back α-β model, separate values were learned for different 1-trial histories of actions and outcomes, $Q(a_t|a_{t-1},o_{t-1})=Q(a_t|a_{t-1},o_{t-1})+\alpha(o_t-Q(a_t|a_{t-1},o_{t-1}))$. The 2-back α-β model was an extension with 2-trial history, with values of the form $Q(a_t|a_{t-2},o_{t-2},a_{t-1},o_{t-1})$.

The multi-parameter RL model was based on basic α-β RL, with additional parameters cα, nα, and d. cα allowed for updating of non-selected actions a_{ns}, based on counterfactual outcomes $1-o$: $Q(a_{ns})=Q(a_{ns})+c\alpha(1-o-Q_{ns})$, where $0\leq c\leq 1$. nα introduced a separate learning rate for negative outcomes: $Q(a)=Q(a)+\alpha(1-Q(a))$ and $Q(a)=Q(a)+n\alpha(0-Q(a))$, $0\leq n\alpha\leq 1$. d shifted the softmax indecision point such that at equal values of $Q(a)$ and $Q(a_{ns})$, staying was more likely when $d<0$, and switching was more likely when $d>0$: $p(a=a_{t-1})=\frac{1}{1+exp(\beta(0.5-d-Q(a_{t-1}))}$.

Bayesian models Using Bayes rule, we estimated the probability that a chosen box was correct given the observed outcome o: $p(a=cor|o)=\frac{p(o|a=cor)p(a=cor)}{p(o|a=cor)+p(o|a=inc)}$. The likelihood was $p(o=1|a=cor)=p_{reward}$ and $p(o=0|a=cor)=1-p_{reward}$, where p_{reward} was truthfully set to 75%. Choice in the subsequent trial took contingency switches into account: $p(a)=(1-p_{switch})p(a=cor)+p_{switch}(1-p(a=cor))$, where p_{switch} was set to the empirical switch probability 0.05.

In the Bayesian multi-parameter model, p_{switch} and p_{reward} were free parameters, and probabilities were softmax-transformed for action selection, with parameters d and β.

Model fitting and comparison Parameters were fitted using maximum likelihood estimation. Model fits were calculated using the Akaike information criterion, $AIC=-2LL+2log(|\theta|)$, with number of parameters $|\theta|$. We interpreted fitted parameters in the best-fit model of each class, assessing age-related changes through the correlation between parameters and age. These tests were based on non-adult participants only, although results including adults were similar.

Results

Human behavior We first assessed participants' responses to switch trials. Younger participants switched faster than older participants, as revealed by the effect of age on staying in the "reward, no reward" condition (Fig. 2B), in a mixed-effects regression model (both females and males: $p<0.001$). At the same time, younger participants reached lower asymptotic accuracy, as evident in the significant effect of age on accuracy in a logistic regression in trials 3-7 post-switch (both females and males: all $\beta's>0.1$, all $p's<0.02$; Fig. 2A). This suggests that younger participants responded more strongly to negative outcomes, which led them to switch quickly after negative outcomes at the expense of asymptotic accuracy. Indeed, younger children showed greater sensitivity to negative reward in logistic regression models predicting action on trial $t+i$ from action and outcome on trial t (Fig. 2), revealed by a positive effect of age on regression coefficients (females: $p=0.003$, males: $p=0.02$).

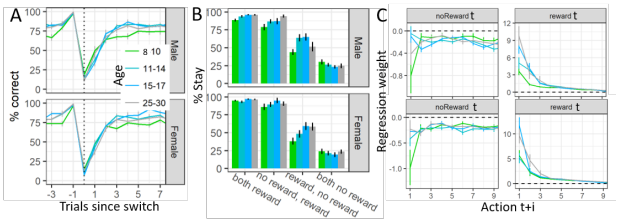

Figure 2: Human Behavior. A) Percent correct choices aligned to switch trials. Colors denote age group. B) Percent trials in which actions were repeated ("stay"), for different reward histories. C) Effect of outcome (no reward, reward) on future choices, based on the regression model in the main text.

Strategy models We first assessed whether models of fixed strategies captured human behavior. Fixed strategies are an

intuitive way to describe decision making, e.g., "I selected the same box when it produced a coin, otherwise I switched to the other box" (WSLS strategy, see methods). Nevertheless, simulations based on the WSLS strategy, with parameter β fitted to individuals (Fig. 3C, see methods), led to poor task performance and failed to capture characteristic human behavior (Fig. 3A-B). Model fit was poor (AIC 34,442), but better than chance behavior (AIC 41,345).

The more complex 2-back WSLS strategy provided a better model fit (AIC 31,744). Two-state WSLS can be summarized as "I usually repeated my previous choice; I only switched when the same box failed to produce a coin twice in a row" (see methods). Simulations performed better and mimicked some human age-based differences, such as faster switching and worse long-term performance in the youngest age group (Fig. 3D-F). The model captured age differences as a significant increase in β with age (females: $r = 0.3$, $p < 0.001$, males: $r = 0.4$, $p < 0.001$), suggesting that decision noise decreased with age.

In summary, the 2-back WSLS strategy captured some aspects of human behavior, but both strategy models failed to capture the shape of learning curves (Fig. 3A, D) and performance differences based on reward history (Fig. 3B, E).

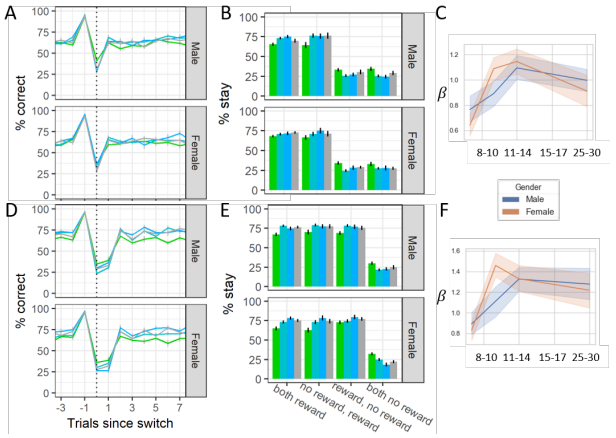

Figure 3: Simulations for WSLS (A-C) and 2-back WSLS (D-F). A, B, D, E as in Fig. 2. C, F) β fitted to participants of different age groups.

History-based RL models As opposed to fixed strategy models, RL can capture changes in behavior based on continuous, feedback-based learning. The 1-back α-β model learned the values that were fixed in WSLS through reinforcement (see methods). The model fit the data better than WSLS (AIC: 30,608), but simulated behavior was indistinguishable from WSLS (data not shown). The same was true for 2-back α-β, the RL version of 2-back WSLS (better AIC: 30,180; indistinguishable behavior).

We also assessed the ability of classic, stateless RL to capture human behavior (see methods). Surprisingly, a simple α-β model, which updated the values of each box based on feedback in the previous trial, fit human data better than all previous models (AIC: 25,051).

In summary, history-based RL models fit human data better than fixed strategies, especially when the history encompassed more trials. Nevertheless, stateless RL fit the data even better, suggesting that participants did not explicitly differentiate values between histories. A different cognitive process might explain human behavior better.

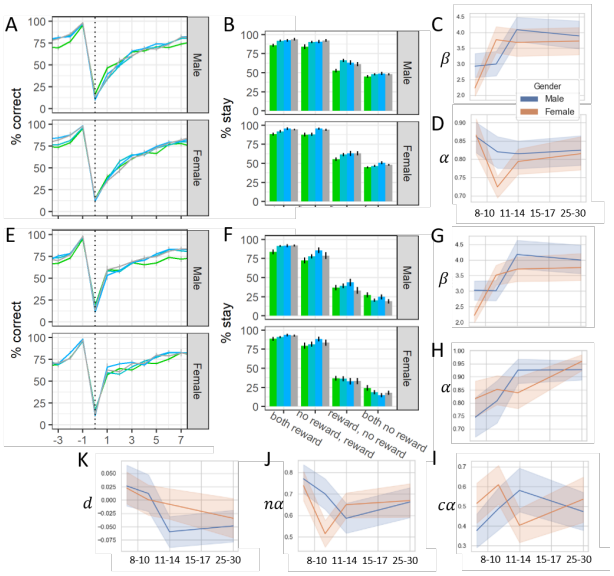

Figure 4: Simulations for RL models. A-D) α-β model. E-K) Multi-parameter model. A, B, E, F as in Fig. 2. C-K) Fitted model parameters for each age group.

Bayesian inference models One candidate process is inferential reasoning. Our parameter-free Bayesian model estimated the probability that each box was correct given the previous outcome (see methods). This model fit human data better than all previous models except simple RL (AIC: 27,258), but failed to produce the learning curves characteristic of human participants (Fig. 5A), and their sensitivity to reward history (Fig. 5B). Lacking free parameters, the model was also unable to capture age differences.

This basic Bayesian model was based on the true probabilities of switch trials and of obtaining reward, but those numbers were unknown to human participants. In the Bayesian multi-parameter model, we treated these probabilities as free parameters, in addition to softmax parameters d and β. Simulated behavior under this model was closer to human learning curves (Fig. 5C) and history-dependent stay behavior (Fig. 5D), and also replicated major age effects. Model fit surpassed the basic model (AIC: 23,335). Several age-based parameter changes gave rise to this behavior. β increased significantly with age (females: $r = 0.3$, $p < 0.001$,

males: $r = 0.3$, $p < 0.001$); p_{reward} decreased from around 80% to around 60% (females: $r = -0.2$, $p = 0.01$, males: $r = -0.2$, $p = 0.02$), suggesting that the youngest children over-estimated reward probabilities, whereas adults underestimated them; d decreased in males only (females: $r = -0.09$, $p = 0.2$, males: $r = -0.3$, $p < 0.001$), suggesting an increased tolerance for staying in the face of decision uncertainty; there were no changes in p_{switch} (females: $r = 0.008$, $p = 0.9$, males: $r = 0.06$, $p = 0.5$).

In summary, Bayesian inference captured the central characteristics of human behavior, and provided insight into age-related changes in terms of estimated reward probabilities, decision noise, and undecision point.

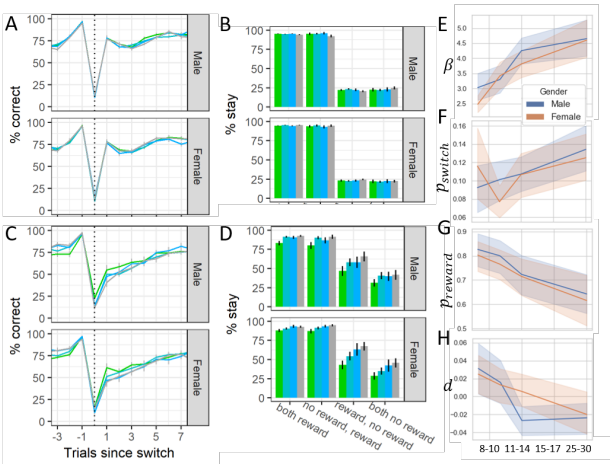

Figure 5: Simulations for Bayesian models. A-B) Basic model without free parameters. C-H) Multi-parameter Bayesian model. A-D as in Fig. 2. E-H) Fitted parameters.

Multi-parameter RL models Benefiting from their flexibility, RL models can mimic inferential reasoning by factoring in the structure of the task and updating the value of unchosen options based on counter-factual outcomes (see methods). We tested RL models with parameter cα for counter-factual updating, in addition to the ability to differentiate between positive and negative feedback (α vs nα), and other parameters, to identify the overall best model, a state-less model with five free parameters, α, β, cα, nα, and d (AIC: 23,271).

Model simulations reproduced human-like learning curves (Fig. 4E) and sensitivity to reward history (Fig. 4F), and replicated some age-related changes, such as low long-term performance in the youngest age group. Age-related differences were related to increasing β (females: $r = 0.4$, $p < 0.001$, males: $r = 0.4$, $p < 0.001$), decreasing learning rate nα (females: $r = -0.1$, $p = 0.06$, males: $r = -0.3$, $p < 0.001$), decreasing d in males only (females: $r = -0.08$, $p = 0.3$, males: $r = -0.3$, $p < 0.001$), increasing learning rate α in males (females: $r = 0.06$, $p = 0.4$, males: $r = 0.3$, $p < 0.001$), and decreasing c, the counter-factual learning parameter, in females (females: $r = -0.2$, $p = 0.04$, males: $r = -0.05$, $p = 0.5$).

Taken together, the multi-parameter RL model achieved the best model fit, and provided insight into developmental changes related to decreasing sensitivity to negative feedback, and gender-specific changes in sensitivity to positive feedback and counter-factual learning.

Discussion

Decision making can be described in various ways, including specific strategies, incremental feedback-based learning, and inferential reasoning. Different models can replicate different behavioral patterns observed in humans, and instead of selecting just one model based on a measure of fit, we integrated the information from several models to obtain a more complete picture of cognitive development in this task. We found that participants' behavior was better described by a 2-back strategy than a simple win-stay lose-shift strategy, and that incremental history-based learning provided an even better fit. Although pure Bayesian inference provided the optimal strategy for the task, the model described human behavior poorly. Parameters that captured uncertainty about the task design were necessary to capture human behavior in a multi-parameter Bayes model. This model revealed that the youngest participants systematically overestimated reward probabilities, whereas adults underestimated them. A multi-parameter RL model, with the best overall model fit, revealed decreasing sensitivity to negative outcomes in both genders, and to positive outcomes in males only, whereas counter-factual updating changed only in females. Both Bayesian and RL models also revealed changes in decision threshold in males, and all models revealed pronounced decreases in decision noise or exploration.

Acknowledgments

We thank L. Kriegsfeld, C. Ford, J. Pfeifer, M.M. Johnson, L. Xia, R. Arsenault, J. Christon, S. Edelman, L. Eletel, H. Keglovits, J. Liu, J. Morillo, N. Rajakumar, N. Spence, T. Smith, B. Tang, T. Welte, L.B. Whitmore, and A. Zou, as well as our participants and their families. The work was funded by NSF SL-CN grant 1640885 to RD, AGEC, and LW.

References

Dennison, M., Whittle, S., Yucel, M., Vijayakumar, N., Kline, A., Simmons, J., & Allen, N. B. (2013). Mapping subcortical brain maturation during adolescence: Evidence of hemisphere- and sex-specific longitudinal changes. *Developmental Science*, *16*(5), 772–791.

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

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.

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