

Developmental changes in learning resemble stochastic optimization

Anna P. Giron^{1,*}, Simon Ciranka^{2,3,*}, Eric Schulz⁴, Wouter van den Bos^{5,6}, Azzurra Ruggeri^{7,8,9}, Björn Meder^{2,10}, and Charley M. Wu^{1,2,+†}

¹Human and Machine Cognition Lab, University of Tübingen, Tübingen, Germany

²Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

³Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany

⁴MPRG Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics

⁵Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

⁶Amsterdam Brain and Cognition, University of Amsterdam, Amsterdam, The Netherlands

⁷MPRG iSearch, Max Planck Institute for Human Development, Berlin, Germany

⁸School of Education, Technical University Munich, Munich, Germany

⁹Central European University, Vienna, Austria

¹⁰Health and Medical University, Potsdam, Germany

*These authors contributed equally to this work

+Corresponding author: charley.wu@uni-tuebingen.de

ABSTRACT

Analogies to stochastic optimization are common in developmental psychology, describing a gradual reduction in randomness over the lifespan. Yet for lack of concrete empirical comparison, there is ambiguity in how to interpret this analogy. Using data from $n=281$ participants ages 5 to 55, we show that “cooling off” does not only apply to the single dimension of randomness. Rather, development resembles a stochastic optimization process in the space of learning strategies, which we characterize along key dimensions of reward generalization, uncertainty-directed exploration, and random temperature. What begins as large tweaks in the parameters that define learning during childhood, plateaus and converges in adulthood. The developmental trajectory of human parameters is strikingly similar to several stochastic optimization algorithms, yet we begin to observe a divergence around adolescence. Remarkably, none of the optimization algorithms discovered reliably better regions of the strategy space than adult participants, suggesting an incredible efficiency of human development.

Introduction

Human development has fascinated researchers of both biological and artificial intelligence alike. As the only known process that reliably produces human-level intelligence¹, there is broad interest in understanding the mechanisms of human development^{2–4} and why we observe specific patterns of developmental change⁵.

One influential hypothesis describes development as a “cooling off” process^{4,6,7}, comparable to *simulated annealing*^{8,9} (SA). SA is a stochastic optimization algorithm named in analogy to a piece of metal that becomes harder to manipulate as it cools off. Initialized with high temperature, SA starts off highly flexible and likely to consider worse solutions as it explores the optimization landscape. But as the temperature cools down, the algorithm becomes increasingly greedy and more narrowly favoring only local improvements, eventually converging on an (approximately) optimal solution. Other stochastic optimization algorithms employing analogous cooling mechanisms are abundant in Machine Learning, and have played a pivotal role in the rise of deep learning^{10,11}.

This analogy of stochastic optimization applied to human development is quite alluring: young children start off highly stochastic and flexible in generating hypotheses^{4,12–14} and

selecting actions¹⁵, which gradually tapers off over the lifespan. This allows children to catch information that adults overlook¹⁶, and learn unusual causal relationships adults might never consider^{4,13}. Yet this high variability also results in large deviations from reward-maximizing behavior^{3,17–20}, with gradual improvements during development. Adults, in turn, are well-calibrated to their environment and quickly solve familiar problems, but at the cost of flexibility, since they experience difficulty adapting to novel circumstances^{21–24}.

Yet, for lack of a direct empirical comparisons between human and algorithmic trajectories, the implications and possible boundaries of the metaphor remain ambiguous. Here, we point out that stochastic optimization can have at least two meanings when it is applied to human development.

The most direct interpretation is to apply “cooling off” to the single dimension of random temperature, controlling the amount of decision noise when selecting actions or sampling hypotheses^{6,14,25}. Evidence from experimental studies suggest that young children are harder to predict than adults^{15,26}, implying greater stochasticity, which is amplified in neurodevelopmental disorders such as ADHD²⁷. Yet, this interpretation is only part of the story, since developmental differences in choice variability can be traced to a host of changes to differ-

ent aspects of learning and choice behavior. Aside from only randomness, development is also related to changes in more systematic, uncertainty-directed exploration^{17,25,28}, which is also reduced over the lifespan. Additionally, changes in generalization²⁵ and the integration of new experiences^{29,30} affect how beliefs are formed and different actions are valued, which also influences choice variability. While decision noise certainly diminishes over the lifespan, this is only a single aspect of human development.

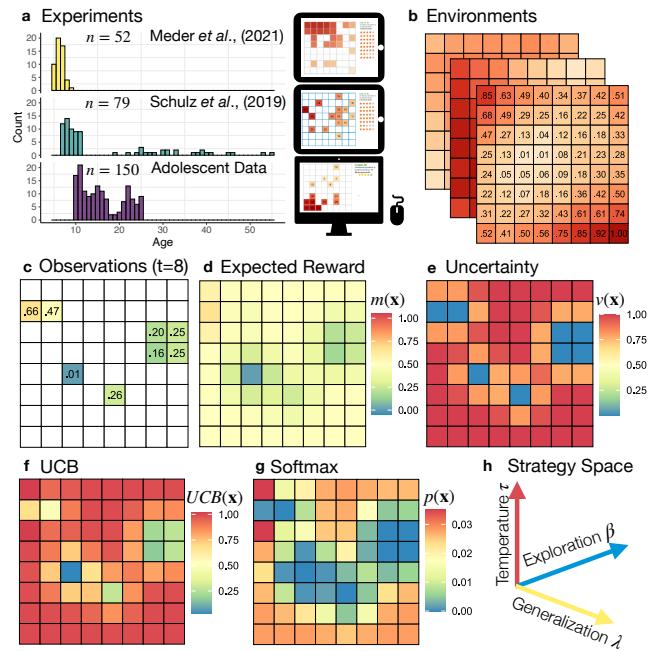


Figure 1. Experiments and model overview. **a)** Description of the three experiments, showing the number and age distribution of participants (after filtering; see Methods). **b)** Fully-revealed examples of the spatially correlated reward environments (showing expected reward), where nearby tiles tended to have similar rewards. **c)** Illustrative example of the task after 8 observations (showing normalized reward). Conditioned on these observations, the GP-UCB model (see Methods) makes Bayesian estimates about **d)** the expected rewards and **e)** uncertainty for each option, where the free parameter λ (Eq. 2) controls the extent that past observations *generalize* to new options. **f)** The expected rewards $m(\mathbf{x})$ and uncertainty estimates $v(\mathbf{x})$ are combined using Upper Confidence Bound (UCB) sampling (Eq. 3) to produce a valuation for each option. The *exploration* bonus β governs the value of exploring uncertain options relative to exploiting high reward expectations. **g)** Lastly, UCB values are entered into a softmax function (Eq. 4) to make probabilistic predictions about where the participant will search next. The *temperature* parameter τ governs the amount of random (undirected) exploration (Eq. 4). **h)** Taken together, generalization (λ), exploration (β), and temperature (τ) define a space of learning strategies, which are distinctly necessary (Figs. 3a-c) and recoverable (Fig. S4).

In contrast, we can also apply the “cooling off” metaphor to an optimization process in the space of learning strategies (Fig. 1h). Development can thus be framed as a search process, whereby individuals tune their “learning engine”, starting off by making large tweaks in childhood, followed by gradually lesser and more refined adjustments. Thus, not only are the outcomes of behavior more stochastic during childhood, but changes to behavior as we learn to learn might also be more stochastic. This interpretation connects the metaphor

of stochastic optimization with Bayesian models of cognitive development, which share a common notion of gradual convergence^{31,32}. Early in development, individuals possess only broad priors and vague theories about the causal structure of the world, which become refined with experience³¹. Bayesian principles dictate that over time, novel experience will have lesser impact on future beliefs or behavior as one’s priors become more narrow through experience^{32,33}.

In this work, we provide much needed clarity to commonly used analogies to stochastic optimization in developmental psychology. While previous work has applied the metaphor of stochastic optimization to the single dimension of random temperature^{3,4,26,34} (i.e., a reduction of decision-noise over the lifespan), here we test whether development can be understood as a stochastic optimization process operating over the space of learning strategies (Fig. 1h). Rather than characterizing the phenotypes of learning at different stages of development^{17,26,28,34}, we characterize the mechanisms of change across a broad range of ages from 5 to 55. Through direct empirical comparisons, our results reveal a striking similarity between how the parameters governing learning develop over the lifespan and how stochastic optimization converges on an optimal solution (Fig. 4).

Results

We analyze experimental data from $n = 281$ participants between the ages of 5 and 55 (Fig. 1a), performing a spatially-correlated multi-armed bandit task³⁵. The task is both intuitive and richly complex, allowing us to simultaneously characterize the distinct (Fig. 3a-c) and recoverable (see Fig. S4) dimensions of reward *generalization* (λ), uncertainty-directed *exploration* (β) and random *temperature* (τ). Participants were given a limited search horizon (25 choices) to maximize rewards by either selecting an unobserved or previously revealed tile on a 8×8 grid. Each choice yielded normally distributed rewards, with reward expectations correlated based on spatial proximity (Fig. 1b), such that tiles close to one another tended to have similar rewards. Since the search horizon was significantly smaller than the number of unique options, generalization and efficient exploration were required to obtain high rewards.

Our dataset combines openly available data from two previously published experiments^{25,26} targeting children and adult participants, along with new unpublished data ($n = 150$) targeting the missing gap of adolescents. Although experimental designs differed in minor details (see Methods), the majority of differences were removed by filtering participants. Reliability tests revealed no differences in performance, model accuracy, and parameter estimates across experiments (all $p > .128$ and $BF < .73$; see Fig. S1 and Table S1). Additionally, robust model and parameter recovery (Figs S3-S4) provide high confidence in our ability to capture the key components of learning across the lifespan.

Behavioral Analyses

We first analyzed participant performance and behavioral patterns in the number of unique options and distance between choices (Fig. 2). We treat age as a continuous variable when possible, but also discretize participants into 7 similarly sized age groups ($n \in [30, 50]$). These behavioral results reveal clear age-related trends in learning and exploration captured by our task.

Performance

Participants monotonically achieved higher rewards as a function of age (Pearson correlation: $r = .51$, $p < .001$, $BF > 100$), with even the youngest age group (5-6 year olds) strongly outperforming chance (one-sample t -test: $t(29) = 5.1$, $p < .001$, $d = 0.9$, $BF > 100$; Fig. 2a). The learning curves in Figure 2b show average reward as a function of trial, revealing a similar trend, with older participants displaying steeper increases in average reward. Notably, the two youngest age groups (5-6 and 7-8 year olds) displayed decaying learning curves with decreasing average reward on later trials, suggesting a tendency to over-explore (supported by subsequent analyses below). We also analyzed the maximum reward (up until a given trial) as a measure of exploration efficacy, where older participants reliably discovered greater maximum rewards (Kendall rank correlation: $r_\tau = .23$, $p < .001$, $BF > 100$) and showed steeper increases on a trial-by-trial basis (Fig. 2c). Thus, lower average reward amongst the youngest age groups (Fig. 2b) did not correspond to any advantages in terms of maximum reward.

Behavioral patterns

Next, we looked at search patterns to better understand the behavioral signatures of age-related changes in exploration. The youngest participants (5-6 year olds) sampled more unique options than chance ($t(29) = 18.0$, $p < .001$, $d = 3.3$, $BF > 100$), but reliably less than the upper-bound on exploration (i.e., unique options on all 25 trials: $t(29) = -4.7$, $p < .001$, $d = 0.9$, $BF > 100$). The number of unique options decreased strongly as a function of age ($r = -.39$, $p < .001$, $BF > 100$; Fig. 2d), consistent with the overall pattern of reduced exploration over the lifespan. Note that all participants were informed and tested about the fact they could repeat choices.

We then classified choices into repeat, near ($\text{distance}=1$) and far ($\text{distance}>1$), and compared this pattern of choices to a random baseline (red dashed line; Fig. 2e). 5-6 year olds started off with very few repeat choices (comparable to the random model: $t(29) = 1.5$, $p = .133$, $d = 0.3$, $BF = .56$) and a strong preference for near choices (more than chance: $t(29) = 14.7$, $p < .001$, $d = 2.7$, $BF > 100$). Over the lifespan, the rate of repeat choices increased, while near decisions decreased, gradually reaching parity for 14-17 year olds (comparing repeat vs. near: $t(41) = -1.3$, $p = .192$, $d = 0.4$, $BF = .38$) and remaining equivalent for all older age-groups (all $p > .378$, $BF < .24$). In contrast, the proportion of far choices remained unchanged over the lifespan ($r_\tau = -.08$,

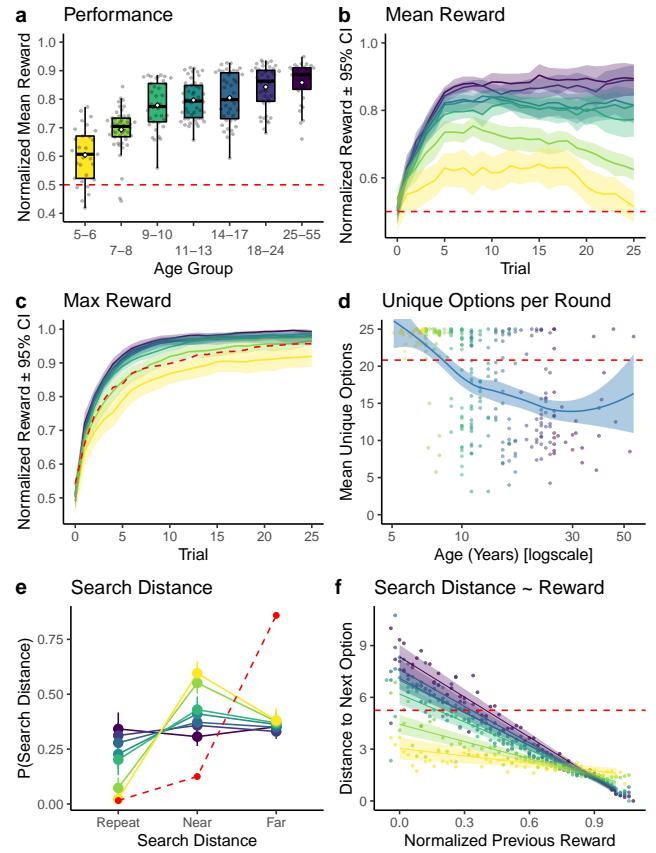


Figure 2. Behavioral results. **a**) Mean reward across age groups. Each dot is one participant, Tukey boxplots showing median and $1.5 \times \text{IQR}$, with the white diamond indicating the group mean. The red dashed line indicates a random baseline (in all plots). The same colors are used to indicate age in all plots. **b**) Learning curves showing mean reward over trials, averaged across all rounds. Lines indicate group means, while the ribbons show the 95%CI. **c**) Learning curves showing maximum reward earned up until a given trial, averaged across all rounds. **d**) The number of unique options sampled per round as a function of age. Each dot is one participant, while the line and ribbon show a locally smoothed regression ($\pm 95\%$ CI). **e**) The portion of repeat, near ($\text{distance}=1$), and far ($\text{distance}>1$) choices as a function of age. Each dot indicates a group mean, while the error bars indicate the 95% CI. **f**) Search distance as a function of the previous reward value. Each line is the fixed effect of a hierarchical Bayesian regression (Table S2) with the ribbons indicating 95% CI. Each dot is the mean of the raw data.

$p = .062$, $BF = .48$). These choice patterns indicate young children do not simply behave like a random model, becoming less random over time. Rather, younger participants exploit past options less than older participants (repeat choices), preferring instead to explore unknown tiles within a local radius (near choices). While the tendency to prefer exploring near rather than far options gradually diminished over the lifespan, this preference for local search distinguished participants of all age groups from the random model.

Lastly, we analyzed how reward outcomes influenced search distance using a Bayesian hierarchical regression (Fig. 2f; Table S2). This model predicted search distance as a function of the previous reward value and age group (including their interaction), with participants treated as random effects. This can be interpreted as a continuous analogue to

past work using a win-stay lose-shift strategy^{15,35}. We found a negative linear relationship in all age-groups, with participants searching locally following high rewards and searching further away after low rewards. This trend becomes stronger over the lifespan with monotonically more negative slopes over the lifespan (see Table S2). While all age groups adapted their search patterns in response to reward, older participants were more adaptive.

Behavioral summary

To summarize, younger children tended to explore unobserved tiles instead of exploiting options known to have good outcomes. This can be characterized as over-exploration, since increase in exploration did not translate into higher maximum rewards (Fig. 2c). Older participants explored less but more effectively, and were able to adapt their search patterns to maximize reward. We now turn to model-based analyses to complement these results with a more precise characterization of how the different mechanisms of learning and exploration change over the lifespan.

Model-Based Analyses

We conducted a series of reinforcement learning³⁶ analyses. We first compared models in their ability to predict out-of-sample choices (Fig. 3a-b) and simulate human-like learning curves across age groups (Fig. 3c). We then analyzed the parameters of the winning model (Fig. 3d-e), which combined Gaussian process (GP) regression with Upper Confidence Bound (UCB) sampling (described below). These parameters allow us to describe how three different dimensions of learning change with age: generalization (λ ; Eq. 2), uncertainty-directed exploration (β ; Eq. 3), and random temperature (τ ; Eq. 4). We then compare the developmental trajectory of these parameters to different stochastic optimization algorithms (Fig. 4).

Modeling learning and exploration

We first describe the GP-UCB model combining all three components of generalization, exploration, and temperature. We then lesion away each component to demonstrate all are necessary for describing behavior. We describe the key concepts below, while Figure 1c-g provides a visual illustration of the model (see Methods for details).

Gaussian process (GP) regression³⁷ provides a reinforcement learning model of value generalization³⁵, where past reward observations can be generalized to novel choices. Here, we describe generalization as a function of spatial location, where closer observations exhibit a larger influence. However, the same model can also be used to generalize based on the similarity of visual features³⁸ or based on graph-structure relationships³⁹.

Given previously observed data $\mathcal{D}_t = \{\mathbf{X}_t, \mathbf{y}_t\}$ of choices $\mathbf{X}_t = [\mathbf{x}_1, \dots, \mathbf{x}_t]$ and rewards $\mathbf{y}_t = [y_1, \dots, y_t]$ at time t (e.g., observations in Fig. 1c), the GP uses Bayesian principles to compute posterior predictions about the expected rewards r_t for

any option \mathbf{x} :

$$p(r_t(\mathbf{x}) | \mathcal{D}_t) \sim \mathcal{N}(m_t(\mathbf{x}), v_t(\mathbf{x})). \quad (1)$$

The posterior in Eq. 1 takes the form of a Gaussian distribution, allowing it to be fully characterized by posterior mean $m_t(\mathbf{x})$ and uncertainty $v_t(\mathbf{x})$ (i.e., variance; see Eqs. 6-7 for details and Figs 1d-e for an illustration).

The posterior mean and uncertainty predictions critically depend on the choice of kernel function $k(\mathbf{x}, \mathbf{x}')$, for which we use a radial basis function (RBF) to describes how observations from one option \mathbf{x} generalize to another option \mathbf{x}' as a function of their distance:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\lambda^2}\right). \quad (2)$$

The model thus assumes nearby options generate similar rewards, with the level of similarity decaying exponentially over increased distances. The lengthscale λ is a free parameter describing the rate at which generalization decays, with larger estimates corresponding to stronger generalization over greater distances.

We then use *Upper Confidence Bound* (UCB) sampling to describe the value of each option $q(\mathbf{x})$ as a weighted sum of expected reward $m(\mathbf{x})$ and uncertainty $v(\mathbf{x})$:

$$q(\mathbf{x}) = m(\mathbf{x}) + \beta \sqrt{v(\mathbf{x})}. \quad (3)$$

β captures uncertainty-directed exploration, determining the extent that uncertainty is valued positively, relative to exploiting options with high expectations of reward.

Lastly, we use a softmax policy to translate value $q(\mathbf{x})$ into choice probabilities:

$$p(\mathbf{x}) \propto \exp(q(\mathbf{x})/\tau) \quad (4)$$

The temperature parameter τ controls the amount of random exploration. Larger values for τ introduce more choice stochasticity, where $\tau \rightarrow \infty$ converges on a random policy.

Lesioned models

To ensure all components of the GP-UCB model play a necessary role in capturing behavior, we created model variants lesioning away each component. The λ lesion model removed the capacity for generalization, by replacing the GP component with a Bayesian reinforcement learning model that assumes independent reward distributions for each option (Eqs. 8-11). The β lesion model removed the capacity for uncertainty-directed exploration by fixing $\beta = 0$, thus valuing options solely based on expectations of reward $m(\mathbf{x})$. Lastly, the τ lesion model swapped the softmax policy for an epsilon-greedy policy, as an alternative form of choice stochasticity: with probability $p(\epsilon)$ a random option is sampled, and with probability $p(1 - \epsilon)$ the option with the highest UCB value is sampled (Eq. 12).

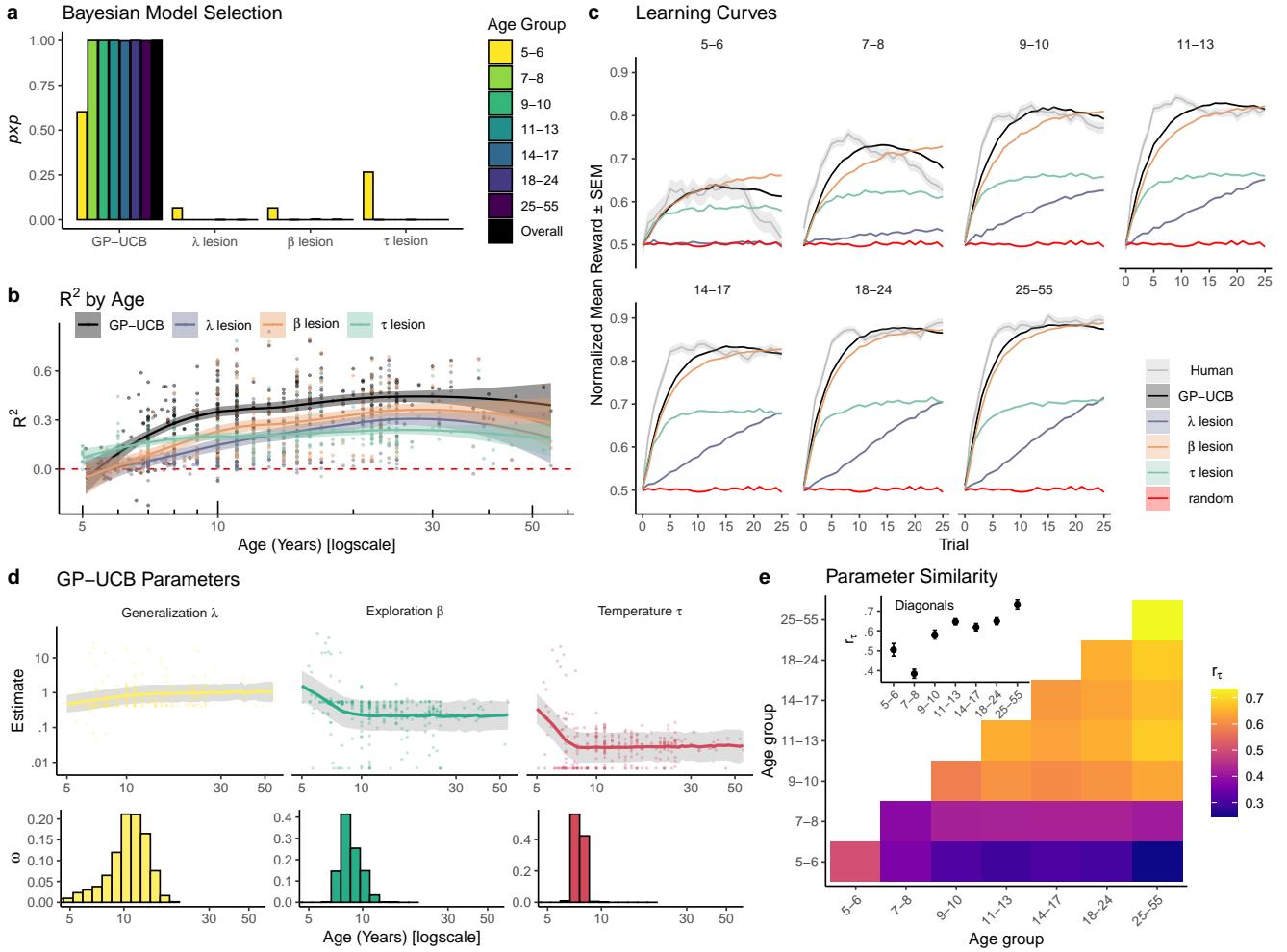


Figure 3. Model results. **a)** Hierarchical Bayesian model selection, where ppx defines the probability of each model being predominant in the population (see Fig. S2 and Table S3 for more details). **b)** Predictive accuracy (R^2) as a function of age. Lines indicate the smoothed conditional mean ($\pm 95\%$ CI) and each dot is a participant. **c)** Simulated learning curves, using participant parameter estimates. Human data illustrates the mean ($\pm 95\%$ CI), while model simulations reports the mean. **d)** Top: Participant parameter estimates as a function of age. Each dot is a single participant, with the line and ribbon showing the posterior predictions from a Bayesian change-point regression model. λ : generalization; β : uncertainty-directed exploration; τ : random exploration. Bottom: Posterior distribution of which age the change-point (ω) is estimated to occur. **e)** Similarity-matrix of parameter estimates. Using Kendall's rank correlation (r_τ), we report similarity of parameter estimates both within and between age-groups. The within age-group similarities (diagonals) are also visualized in the inset plot, where error bars indicate the 95% CI.

Model comparison

We fit all models using leave-one-round-out cross validation (see Methods). We then conducted hierarchical Bayesian model selection⁴⁰ to compute the *protected exceedence probability* (ppx) for describing which model is most likely in the population. We found that GP-UCB was the best model for each individual age group and also aggregated across all data (Fig. 3a). There is still some ambiguity between models in the 5-6 year old group, but this quickly disappears in all subsequent age groups ($ppx_{GP-UCB} > .99$). Figure 3b describes the out-of-sample predictive accuracy of each model as a continuous function of age, where a pseudo- R^2 provides an intuitive comparison to random chance (Eq. 13). Intuitively, $R^2 = 0$ indicates chance-level predictions and $R^2 = 1$ indicates theoretically perfect predictions. While there is again some ambiguity among 5-6 year olds, GP-UCB quickly dominates

and remains the best model across all later ages.

Aside from only predicting choices, we also simulated learning curves for each model (using participant parameter estimates) and compared them against human performance for each age group (Fig. 3c). The full GP-UCB model provides the best description across all age groups, although the β lesion model also produces similar patterns. However, only GP-UCB by virtue of the exploration bonus β is able to recreate the decaying learning curves for 5-6 and 7-8 year olds.

Altogether, these results reveal that all three components of generalization (λ), uncertainty-directed exploration (β), and random temperature (τ) play a vital role in describing behavior. Next, we analyze how each of these parameters change over the lifespan.

Parameters

We use both a uni- and a multivariate approach to interpret age-related changes in GP-UCB parameters. The univariate analysis (Fig. 3d) modeled age-related changes in each parameter separately, using a hierarchical Bayesian change-point regression⁴¹. This approach models the relationship between age and the target parameter as separate linear functions, before and after an estimated change-point ω , at which point the regression slope changes from b_1 to b_2 (Eq. 15). This analysis (Fig. 3d) revealed how all parameters changed rapidly during childhood, but then plateaued such that there were no credible change in parameters after the estimated change point (all b_2 CIs overlapped with 0; see Table S4). More specifically, generalization increased ($b_1(\lambda) = 0.15 [0.03, 0.76]$) until around 10 years of age ($\omega(\lambda) = 10.6 [5.38, 14.8]$), whereas directed exploration ($b_1(\beta) = -0.62 [-1.12, -0.25]$) and random temperature ($b_1(\tau) = -1.04 [-1.56, -0.61]$) decreased sharply, until around 9 ($\omega(\beta) = 8.65 [7.28, 10.80]$) and 8 ($\omega(\tau) = 7.51 [6.93, 8.10]$) years of age, respectively. Using leave-one-out cross-validation, we established that this simple change-point model predicted all GP-UCB parameters better than linear or complex regression models (up to fourth degree polynomials; see Table S5).

In a multivariate similarity analysis, we computed the pairwise similarity of all parameter estimates between participants (Kendalls' τ), which we then averaged over age-groups (Fig. 3e). This shows that older participants were more similar to each other than younger subjects' (Fig. 3e inset), suggesting development produces a convergence towards a more similar set of parameters. Whereas older participants achieved high rewards using similar learning strategies, younger participants tended to over-explore and acquired lower rewards, but each in their own fashion, with more diverse strategies.

These results highlight how development produces changes to all parameters governing learning, not only a uni-dimensional reduction in random sampling. An initially steep but plateauing rate of change across model parameters is consistent with the metaphor of stochastic optimization in the space of learning strategies (Fig. 3d). The increasing similarity in participants' parameters again speaks for a developmental process that gradually converges on a configuration of learning parameters (Fig. 3e), which also generates better performance (Fig. 3c).

Comparison to stochastic optimization

Beyond verbal analogies, we now present a direct comparison between human development and stochastic optimization. We first computed a fitness landscape (Fig. S5) across 1 million combinations of plausible parameter values of the GP-UCB model (see Methods), with each parameter combination yielding a mean reward based on 100 simulated rounds. We then simulated different optimization algorithms on the fitness landscape, using each of the cross-validated parameter estimates of the 5-6 year old age group as initialization points. Specifically, we compared the trajectories of simulated annealing

(SA) and stochastic hill climbing (SHC; a discrete analogue of stochastic gradient descent) in combination with three common cooling schedules (fast cooling, exponential cooling, and linear cooling; see Methods).

While we do not attempt to curve-fit the exact cooling schedule that describes human development, we observe that fast and exponential cooling generally performed better than linear cooling (Fig. 4a). This provides a normative explanation for why we do not observe linear changes in development, but rather rapid initial changes during childhood, followed by a gradual plateau and convergence. Yet remarkably, neither SA nor SHC converged on a better solution than adult 25-55 year old participants (SA-fast cooling: $t(149) = -0.4, p = .675, d = 0.08, BF = .23$; SHC-fast cooling: $t(149) = 0.8, p = .447, d = 0.2, BF = .27$).

Figure 4b visualizes the developmental trajectory of human participants (labeled dots) to the trajectories of both SA and SHC with fast cooling (blue lines; see Fig. S6 for all algorithms and all parameter comparisons). We focus on changes in generalization and exploration parameters, since rewards decrease monotonically with increased random temperature. Particularly for younger age groups, age-related changes in parameters follow a similar trajectory as the optimization algorithms. However, a notable divergence emerges around adolescence (ages 9-13 for SA and ages 14-17 for SHC). Ultimately, both SA and SHC trajectories converge on higher values of λ and β than human participants, and lower values of τ (Fig. S6b). Since these deviations nevertheless fail to translate into reliable differences in performance, this may point towards resource rational constraints on human development^{42,43}: aside from only optimizing for the best performance, the cognitive costs of different strategies may also be considered (see Discussion).

Discussion

From a rich data set of $n = 281$ participants ages 5 to 55 (Fig. 1a), our results reveal human development does not only "cool off" in the search for rewards or hypotheses, but also in the search for the best learning strategy. Thus, stochastic optimization best applies to how we learn to learn, rather than only how we select actions or hypotheses. What begins as large tweaks to the cognitive mechanisms of learning and exploration during childhood, gradually plateaus and converges in adulthood (Fig. 3d-e). This process is remarkably effective, resembling the trajectory of a stochastic optimization algorithm (Fig. 4b) as it optimizes the parameters of a reinforcement learning model (GP-UCB). While human development and the optimization algorithm begin to diverge around adolescence (9-17), none of the optimization algorithms converged on reliably better performance than adult human participants (25-55 year olds; Fig. 4a). This work provides important insights into the nature of developmental changes in learning and offers normative explanations for why we observe these specific patterns of development.

Rather than a uni-dimensional transition from exploration

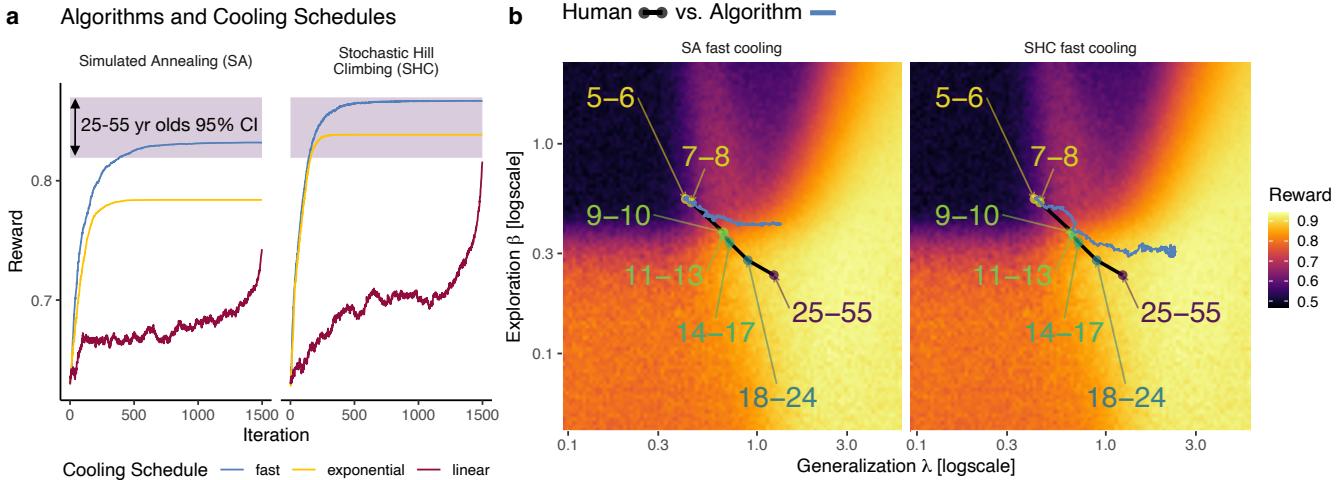


Figure 4. Developmental Trajectory. **a**) Mean reward in each iteration for simulated annealing (SA) and stochastic hill climbing (SHC) algorithms, combined with fast, exponential, and linear cooling schedules. The shaded purple band indicates the 95% CI of human performance in the 25–55 age group. **b**) Comparison of human and algorithm trajectories, focusing on the best performing fast cooling schedule. The blue lines show the trajectory of each optimization algorithm (median of each simulated participant) and the annotated dots show the median parameter estimates of each age group. For this 2D illustration, the fitness landscape uses the median τ estimate across all human data ($\bar{\tau} = .03$; see Fig. S6 for all algorithms and all parameter comparisons).

to exploitation over the lifespan²³, we observe refinements in both the ability to explore (Fig. 2c) and exploit (Fig. 2f), with monotonic improvements in both measures as a function of age. While even 5–6 year olds perform better than chance (Fig. 2a), exploration becomes more effective over the lifespan, as larger reward values are discovered despite sampling fewer unique options (Fig. 2c–d). Meanwhile, exploitation becomes more responsive, with older participants adapting their search distance more strongly based on reward outcomes (Fig. 2e). This resembles a developmental refinement of a continuous win-stay lose-shift strategy^{15,35}. Thus, we reaffirm past work showing a reduction of stochasticity over the lifespan^{4,12,14}, but expand the scope of developmental changes across multiple dimensions of learning.

With a reinforcement learning model (GP-UCB), we are able to characterize age-related changes in learning through the dimensions of *generalization* (λ), uncertainty-directed *exploration* (β), and random *temperature* (τ). All three dimensions play an essential role in predicting choices (Fig. 3a–b) and simulating realistic learning curves across all ages (Fig. 3c), with recoverable models and parameter estimates (Figs S3–S4). Changes in all parameters occur rapidly during childhood (increase in generalization and decrease in both exploration and temperature), but then plateau around adolescence (Fig. 3d). Younger participants tend to be more diverse, whereas adults are more similar to one another, with continued convergence of parameter estimates until adulthood (25–55 year olds; Fig. 3e). Both the reduction of age-related differences and increasing similarity of parameters support the analogy of development as a stochastic optimization process, which gradually converges upon an optimal configuration of learning parameters.

Our direct comparison between the developmental trajectory of human parameters to various stochastic optimization

algorithms (Fig. 4) revealed both striking similarities and intriguing differences. The best performing algorithms (SHC and SA with fast cooling) also most resemble the trajectory of human development (Fig S6). While humans begin to deviate from the algorithm trajectories from adolescence onwards in key respects (lower generalization and exploration, but higher temperature), none of the optimization algorithms achieved significantly better performance than adult participants. These differences might point towards cognitive costs, which are not justified for lack of increased performance benefits. Generalization over a greater extent may require remembering and performing computations over a larger set of past observations^{39,44}, which is why some GP approximations reduce the number of inputs to save computational costs⁴⁵. Deploying uncertainty-directed exploration is also associated with increased cognitive costs⁴⁶, and can be systematically diminished through working memory load⁴⁷ or time pressure⁴⁸ manipulations. Lastly, random temperature has been related to cognitive control⁴⁹ and complexity costs⁵⁰, where adapting to the context-specific demands of an experimental task incurs costs proportional to the extent of the adaptation from some default policy^{51,52}.

Our findings also have important clinical implications for understanding maladaptive development and the origin of psychological disorders that emerge during development. Through comparison to stochastic optimization, we find that childhood is a sensitive period for configuring learning and exploration parameters. Similarly, life history theory⁵³ suggests organisms utilize early life experience to configure strategies for interacting with their environment, which for most species remain stable throughout the life span⁵⁴. Once the configuration of learning strategies has cooled off, there is less flexibility for adapting to novel circumstances in later developmental stages. Adverse childhood experiences can thus

produce a mismatch between training and test environments, to borrow a phrase from machine learning. Such a mismatch would set the developmental trajectory towards regions of the parameter space that are poorly suited for adult environments. This dovetails with findings from computational psychiatry, showing that a variety of mental disorders are characterized by aberrant learning mechanisms^{55–57}. Rather than focusing on adult phenotypes, our results add a developmental perspective and provide a more complete understanding of the computational dynamics underlying psychopathology.

Stochastic optimization also offers the intriguing possibility for "re-heating"⁸ by adding more flexibility in later optimization stages. Re-heating is often used in dynamic environments or when insensitivities of the fitness landscape can cause the algorithm to get stuck. Since deviations from the algorithm trajectories start in adolescence, this may coincide with a second window of developmental plasticity during adolescence^{58,59}. While we observed relatively minor changes in the parameters governing individual learning, plasticity in adolescents is thought to specifically target social learning mechanisms^{7,60}. Thus, different aspects of development may fall under different cooling schedules and similar analyses could also be applied to other learning domains and parameter combinations more broadly.

Scientists often look to statistical and computational tools for explanations and analogies⁶¹. With recent advances in Machine Learning and Artificial Intelligence, these tools are increasingly vivid mirrors into the nature of human cognition. We can understand idiosyncrasies of hypothesis generation through Monte Carlo sampling⁶², individual learning through optimization^{63–65}, and development as programming or "hacking"⁶⁶. An important advantage of computational explanations is that they offer direct empirical demonstrations, instead of remaining as vague, verbal comparisons. Here, we provided such a demonstration, and added much needed clarity to commonly used analogies of stochastic optimization in developmental psychology.

Methods

Experiments

We combined open data from two previously published experiments (Meder et al²⁶ and Schulz et al³⁴) targeting children and adult participants, together with new unpublished data targeting adolescent participants (see below). The experimental designs differed in a few details, the majority of which were removed by filtering participants. The combined and filtered data consisted of 281 participants between the ages of 5 and 55 ($M_{age}=14.46$, $SD=8.61$, 126 female). Informed consent was obtained from all participants or their legal guardians prior to participation.

Generic materials and procedure

All participants performed a spatially-correlated bandit task³⁵ on a 8×8 grid of 64 options (i.e., tiles). A random tile was revealed at the beginning of each round, with participants given a limited search horizon of 25 trials to acquire as many cumulative rewards as possible by either choosing new or previously revealed tiles. After

each round, participants were rewarded a maximum of five stars reflecting their performance relative to always selecting the optimal tile. The number of stars earned in each round stayed visible until the end of the experiment.

When choosing a tile, participants earned rewards corrupted by normally distributed noise $\epsilon \sim \mathcal{N}(0, 1)$. Reward expectations were spatially-correlated across the grid, such that nearby tiles had similar reward expectations (described below). Earned rewards were depicted numerically along with a corresponding color (only colors in Meder et al., ²⁶; see below), with darker colors depicting higher rewards. Figure 1a provides screenshots of each task and Figure 1b depicts the distribution of rewards on a fully revealed environment.

All experiments (after filtering, see below) used the same set of underlying 40 reward environments, which define a bivariate function on the grid, mapping each tile's location on the grid to an expected reward value. The environments were generated by sampling from a multivariate Gaussian distribution $\sim \mathcal{N}(0, \Sigma)$, with covariance matrix Σ defined by a RBF kernel (Eq. 2) with a lengthscale of $\lambda = 4$. In each round, a new environment was chosen without replacement from the list of environments. To prevent participants from knowing when they found the highest reward, a different maximum range was sampled from a uniform distribution $\sim \mathcal{U}(30, 40)$ for each round and all reward values were rescaled accordingly. The rescaled rewards were then shifted by +5 to avoid reward observations below 0. Hence, the effective rewards ranged from 5 to 45, with a different maximum in each round. All experiments included an initial training round designed to interactively explain the nature of the task, and ended with a bonus round in which they were asked to predict the rewards of unseen tiles. All analyses exclude the training and bonus rounds.

Differences across experiments

Participants from the Meder et al²⁶ and Schulz et al³⁴ studies were recruited from museums in Berlin and paid with stickers (Meder et al²⁶) or with money (Schulz et al³⁴) proportional to their performance in the task. The new adolescent data includes 150 participants ($M_{age}=16.1$, $SD=4.97$, 69 female) who completed the task at the Max Planck Institute for Human Development in Berlin along with a battery of 10 other decision-making tasks on a desktop computer. These participants were given a fixed payment of €10 per hour. This study was approved by the Ethics Committee of the Max Planck Institute for Human Development (A 2018/23).

Both Meder et al²⁶ and Schulz et al³⁴ studies used a between-subject manipulation of the strength of rewards correlations (smooth vs. rough environments: $\lambda_{smooth} = 4$, $\lambda_{rough} = 1$). Because only minimal differences in model parameters were found in previous studies^{25,26,35}, the rough condition was omitted in the adolescent sample. Thus, we filtered out all participants assigned to the rough condition such that only participants assigned to the smooth environments were included in the final sample. Lastly, both Schulz et al³⁴ and the adolescent experiment used ten rounds, while Meder et al²⁶ included only six rounds to avoid lapses in attention in the younger age group. In addition, numerical depictions of rewards were removed in the Meder et al²⁶ experiment, and participants were instructed to focus on the colors (deeper red indicating more rewards) to avoid difficulties with reading large numbers.

After filtering, the remaining differences in modality (tablet vs. computer), incentives (stickers vs. variable money vs. fixed money), number of rounds (six vs. ten), and visualization of rewards (numbers+colors vs. only colors) did not result any differences in performance (Fig. S1a), model fits (Fig. S1b), or parameter estimates (Fig. S1c).

Computational models

Gaussian process generalization

Gaussian process (GP) regression³⁷ provides a non-parametric Bayesian framework for function learning, which we use as a method of value generalization³⁵. We use the GP to infer a value functions $f : \mathcal{X} \rightarrow \mathbb{R}^n$ mapping input space \mathcal{X} (all possible options on the grid) to a real-valued scalar outputs r (reward expectations). The GP performs this inference in a Bayesian manner, by first defining a prior distribution over functions $p(r_0)$, which is assumed to be multivariate Gaussian:

$$p(r_0(\mathbf{x})) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \quad (5)$$

with the prior mean $m(\mathbf{x})$ defining the expected output of input \mathbf{x} , and with covariance defined by the kernel function $k(\mathbf{x}, \mathbf{x}')$, for which we use an RBF kernel (Eq. 2). Per convention, we set the prior mean to zero, without loss of generality³⁷.

Conditioned on a set of observations $\mathcal{D}_t = \{\mathbf{X}_t, \mathbf{y}_t\}$, the GP computes a posterior distribution $p(r_t(\mathbf{x}_*) | \mathcal{D}_t)$ (Eq 1) for some new input \mathbf{x}_* , which is also Gaussian, with posterior mean and variance defined as:

$$m(\mathbf{x}_* | \mathcal{D}_t) = \mathbf{k}_{*,t}^\top (\mathbf{K} + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathbf{y}_t \quad (6)$$

$$v(\mathbf{x}_* | \mathcal{D}_t) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_{*,t}^\top (\mathbf{K} + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathbf{k}_{*,t} \quad (7)$$

$\mathbf{k}_{*,t} = k(\mathbf{X}_t, \mathbf{x}_*)$ is the covariance matrix between each observed input and the new input \mathbf{x}_* and $\mathbf{K} = k(\mathbf{X}_t, \mathbf{X}_t)$ is the covariance matrix between each pair of observed inputs. \mathbf{I} is the identity matrix and σ_ϵ^2 is the observation variance, corresponding to assumed i.i.d. Gaussian noise on each reward observation.

Lesioned models

The λ lesion model removes the capacity for generalization, by replacing the GP with a *Bayesian Mean Tracker* (BMT) as a Bayesian reinforcement learning model that learns reward estimates for each option independently. Reward estimates are updated as a function of prediction error, thus the BMT can be interpreted as Bayesian variant of the classic Rescorla-Wagner model^{67,68} and has been used to describe human behavior in a variety of learning and decision-making tasks^{48,69,70}.

The BMT also defines a Gaussian prior distribution of the reward expectations, but does so independently for each option \mathbf{x} :

$$p(r_0(\mathbf{x})) \sim \mathcal{N}(m_0(\mathbf{x}), v_0(\mathbf{x})). \quad (8)$$

The BMT computes an equivalent posterior distribution for the expected reward for each option (Eq. 1), also in the form of a Gaussian, but where the posterior mean $m_t(\mathbf{x})$ and posterior variance $v_t(\mathbf{x})$ are defined independently for each option and computed by the following updates:

$$m_{t+1}(\mathbf{x}) = m_t(\mathbf{x}) + \delta_t(\mathbf{x}) G_t(\mathbf{x})(y_t(\mathbf{x}) - m_t(\mathbf{x})) \quad (9)$$

$$v_{t+1}(\mathbf{x}) = v_t(\mathbf{x})(1 - \delta_t(\mathbf{x}) G_t(\mathbf{x})) \quad (10)$$

Both updates use $\delta_t(\mathbf{x}) = 1$ if option \mathbf{x} was chosen on trial t , and $\delta_t(\mathbf{x}) = 0$ otherwise. Thus, the posterior mean and variance are only updated for the chosen option. The update of the mean is based on the prediction error $y_t(\mathbf{x}) - m_t(\mathbf{x})$ between observed and anticipated reward, while the magnitude of the update is based on the Kalman gain $G_t(\mathbf{x})$:

$$G_t(\mathbf{x}) = \frac{v_t(\mathbf{x})}{v_t(\mathbf{x}) + \theta_\epsilon^2}, \quad (11)$$

analogous to the learning rate of the Rescorla-Wagner model. Here, the Kalman gain is dynamically defined as a ratio of variance terms, where $v_t(\mathbf{x})$ is the posterior variance estimate and θ_ϵ^2 is the error variance, which (analogous to the GP) models the level of noise associated with reward observations. Smaller values of θ_ϵ^2 thus result in larger updates of the mean.

The β lesion simply fixes $\beta = 0$, making the valuation of options solely defined by the expected rewards $q(\mathbf{x}) = m(\mathbf{x})$.

The τ lesion model swaps the softmax policy (characterized by temperature τ) for an epsilon-greedy policy³⁶, since it is not feasible to simply remove the softmax component or fix $\tau = 0$ making it an argmax policy (due to infinite log loss from zero probability predictions). Instead, we use take the opportunity to compare the softmax policy against epsilon-greedy as an alternative mechanism of random exploration²⁷. We still combine epsilon-greedy with GP and UCB components, but rather than choosing options proportional to their UCB value, the τ lesion estimates ε as a parameter controlling the probability of choosing an option at random vs. the highest UCB option:

$$p(\mathbf{x}) = \begin{cases} \arg \max q(\mathbf{x}), & \text{with probability } 1 - \varepsilon \\ 1/64, & \text{with probability } \varepsilon \end{cases} \quad (12)$$

Model cross-validation

Each model was fit using leave-one-round-out cross validation for each individual participant using maximum likelihood estimation (MLE). Model fits are described using negative log likelihoods (nLLs) summed over all out-of-sample predictions, while individual participant parameter estimates are based on averaging over the cross-validated MLEs. Figure 3b reports model fits in terms of a pseudo- R^2 , which compares the out-of-sample nLLs for each model k against a random model:

$$R^2 = 1 - \frac{\log \mathcal{L}(M_k)}{\log \mathcal{L}(M_{\text{random}})} \quad (13)$$

Changepoint regression

We use a change-point model to quantify univariate changes in model parameters as a function of age:

$$\log(\text{estimate}) \sim \mathcal{N}(\mu_{\text{age}}, \sigma_{\text{age}}^2), \quad (14)$$

$$\begin{aligned} \mu_{\text{age}} = b_0 + b_1 * (\text{age} - \omega) * I(\text{age} < \omega) \\ + b_2 * (\text{age} - \omega) * I(\text{age} > \omega). \end{aligned} \quad (15)$$

$I(\cdot)$ is an indicator function, b_0 is the intercept, and ω is the age at which the slope b_1 changes to b_2 .

Fitness landscape

We used Tukey's fence to define credible interval for each GP-UCB parameter $[\lambda, \beta, \tau]$ based on participant estimates and created a grid of 100 equally-sized log-space intervals for each parameter. This defines a space of plausible learning strategies corresponding to 1 million parameter combinations. We then ran 100 simulations of the GP-UCB model for each parameter combination (sampling one of the 40 reward environments with replacement each time) and computed the mean reward across iterations (Fig. S5).

Optimization algorithms

Using this fitness landscape defined across learning strategies, we simulated the trajectories of various optimization algorithms. Specifically, we tested simulated annealing^{4,9} (SA) and stochastic hill climbing^{71,72} (SHC), the latter of which provides a discrete analogue to the better known stochastic gradient descent method commonly used to optimize neural networks^{10,11}. Each optimization algorithm (SA vs. SHC) was combined with one of three common cooling schedules⁷³ defining how the temperature (temp) changes as a function of the iteration number i . Fast cooling uses $\text{temp}_i = 1/(1+i)$, exponential cooling use $\text{temp}_i = \exp(-i^{1/3})$, and linear cooling uses $\text{temp}_i = 1 - (i+1)/\max(i)$. As the temperature decreases over iterations, there is a general decrease in the amount of randomness or stochasticity.

SA is a stochastic sampling algorithm, which is more likely to select solutions with lower fitness when the temperature is high. After initialization, SA iteratively selects a random neighboring solution s_{new} in the fitness landscape (i.e., one step in the grid of 1 million parameter combinations), and either deterministically accepts it if it corresponds to higher fitness than the current solution s_{old} , otherwise, it accepts worse solutions with probability:

$$p(\text{accept}) \propto \exp\left(\frac{s_{\text{new}} - s_{\text{old}}}{\text{temp}_i}\right), \quad (16)$$

where temp_i is the current temperature.

SHC is similar, but considers all neighboring solutions $s' \in S_{\text{neighbors}}$ and selects a new solution proportional its fitness:

$$p(s') \propto \exp(s'/\text{temp}_i) \quad (17)$$

For each combination of optimization algorithm and cooling function, we simulated optimization trajectories over 1500 iterations. Each simulated was initialized on each of the cross-validated parameter estimates of all participants of the youngest age group as starting points. This resulted in 120 (30 participants \times 4 rounds of cross validation) trajectories for each combination of algorithm and cooling schedule.

Data and Code Availability

Code and data are publicly available at https://github.com/AnnaGiron/developmental_trajectory.

References

1. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *science* **331**, 1279–1285 (2011).
2. Moran, R. J., Symmonds, M., Dolan, R. J. & Friston, K. J. The brain ages optimally to model its environment: evidence from sensory learning over the adult lifespan. *PLoS computational biology* **10**, e1003422 (2014). DOI 10.1371/journal.pcbi.1003422.
3. Nussenbaum, K. & Hartley, C. A. Reinforcement learning across development: What insights can we draw from a decade of research? *Dev. cognitive neuroscience* **40**, 100733 (2019).
4. Gopnik, A. *et al.* Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proc. Natl. Acad. Sci.* **114**, 7892–7899 (2017).
5. Walasek, N., Frankenhus, W. E. & Panchanathan, K. Sensitive periods, but not critical periods, evolve in a fluctuating environment: a model of incremental development. *Proc. Royal Soc. B* **289**, 20212623 (2022).
6. Gopnik, A., Griffiths, T. L. & Lucas, C. G. When younger learners can be better (or at least more open-minded) than older ones. *Curr. Dir. Psychol. Sci.* **24**, 87–92 (2015).
7. Ciranka, S. & van den Bos, W. Adolescent risk-taking in the context of exploration and social influence. *Dev. Rev.* **61**, 100979 (2021).
8. Du, K.-L. & Swamy, M. N. S. Simulated annealing. In Du, K.-L. & Swamy, M. N. S. (eds.) *Search and Optimization by Metaheuristics: Techniques and Algorithms Inspired by Nature*, 29–36 (Springer International Publishing, Cham, 2016).
9. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *science* **220**, 671–680 (1983).
10. Robbins, H. & Monro, S. A stochastic approximation method. *The annals mathematical statistics* 400–407 (1951).
11. Bottou, L., Curtis, F. E. & Nocedal, J. Optimization methods for large-scale machine learning. *Siam Rev.* **60**, 223–311 (2018).
12. Buchsbaum, D., Bridgers, S., Skolnick Weisberg, D. & Gopnik, A. The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philos. Transactions Royal Soc. B: Biol. Sci.* **367**, 2202–2212 (2012).
13. Lucas, C. G., Bridgers, S., Griffiths, T. L. & Gopnik, A. When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition* **131**, 284–299 (2014).
14. Denison, S., Bonawitz, E., Gopnik, A. & Griffiths, T. L. Rational variability in children’s causal inferences: The sampling hypothesis. *Cognition* **126**, 285–300 (2013).
15. Bonawitz, E., Denison, S., Gopnik, A. & Griffiths, T. L. Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cogn. psychology* **74**, 35–65 (2014).
16. Sumner, E. *et al.* The exploration advantage: Children’s instinct to explore allows them to find information that adults miss. *PsyArXiv* (2019). DOI 10.31234/osf.io/h437v.
17. Somerville, L. H. *et al.* Charting the expansion of strategic exploratory behavior during adolescence. *J. experimental psychology: general* **146**, 155 (2017).
18. Jepma, M., Schaaf, J. V., Visser, I. & Huizenga, H. M. Uncertainty-driven regulation of learning and exploration in adolescents: A computational account. *PLoS computational biology* **16**, e1008276 (2020).
19. Palminteri, S., Kilford, E. J., Coricelli, G. & Blakemore, S.-J. The computational development of reinforcement learning during adolescence. *PLoS computational biology* **12**, e1004953 (2016).
20. Rosenbaum, G. M., Venkatraman, V., Steinberg, L. & Chein, J. M. The influences of described and experienced information on adolescent risky decision making. *Dev. Rev.* **47**, 23–43 (2018).
21. Baltes, P. B. Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Dev. Psychol.* (1987).
22. Baltes, P. B., Staudinger, U. M. & others. Lifespan psychology: Theory and application to intellectual functioning. *Annu. review* (1999).
23. Gopnik, A. Childhood as a solution to explore-exploit tensions. *Philos. Transactions Royal Soc. B* **375**, 20190502 (2020).
24. Gopnik, A. Scientific thinking in young children: theoretical advances, empirical research, and policy implications. *Science* **337**, 1623–1627 (2012).

- 25.** Schulz, E., Wu, C. M., Ruggeri, A. & Meder, B. Searching for rewards like a child means less generalization and more directed exploration. *Psychol. Sci.* **30**, 1561–1572 (2019). DOI 10.1177/0956797619863663.
- 26.** Meder, B., Wu, C. M., Schulz, E. & Ruggeri, A. Development of directed and random exploration in children. *Dev. Sci.* e13095 (2021). DOI 10.1111/desc.13095.
- 27.** Dubois, M. *et al.* Tabula-rasa exploration decreases during youth and is linked to adhd symptoms. *BioRxiv* (2020).
- 28.** Blanco, N. J. & Sloutsky, V. M. Systematic exploration and uncertainty dominate young children’s choices. *Dev. Sci.* **24**, e13026 (2021).
- 29.** Van den Bos, W., Cohen, M. X., Kahnt, T. & Crone, E. A. Striatum–medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cereb. cortex* **22**, 1247–1255 (2012).
- 30.** Blanco, N. J. *et al.* Exploratory decision-making as a function of lifelong experience, not cognitive decline. *J. Exp. Psychol. Gen.* **145**, 284 (2016).
- 31.** Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
- 32.** Stamps, J. A. & Frankenhuys, W. E. Bayesian models of development. *Trends Ecol. Evol.* **31**, 260–268 (2016).
- 33.** Frankenhuys, W. E. & Panchanathan, K. Balancing sampling and specialization: an adaptationist model of incremental development. *Proc. Biol. Sci.* **278**, 3558–3565 (2011).
- 34.** Schulz, E., Wu, C. M., Huys, Q. J., Krause, A. & Speekenbrink, M. Generalization and search in risky environments. *Cogn. Sci.* **42**, 2592–2620 (2018). DOI 10.1111/cogs.12695.
- 35.** Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D. & Meder, B. Generalization guides human exploration in vast decision spaces. *Nat. Hum. Behav.* **2**, 915—924 (2018). DOI 10.1038/s41562-018-0467-4.
- 36.** Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).
- 37.** Rasmussen, C. E. & Williams, C. Gaussian processes for machine learning the mit press. *Cambridge, MA* (2006).
- 38.** Wu, C. M., Schulz, E., Garvert, M. M., Meder, B. & Schuck, N. W. Similarities and differences in spatial and non-spatial cognitive maps. *PLOS Comput. Biol.* **16**, 1–28 (2020). DOI 10.1371/journal.pcbi.1008149.
- 39.** Wu, C. M., Schulz, E. & Gershman, S. J. Inference and search on graph-structured spaces. *Comput. Brain & Behav.* **4**, 125–147 (2021). DOI 10.1007/s42113-020-00091-x.
- 40.** Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—revisited. *Neuroimage* **84**, 971–985 (2014).
- 41.** Simonsohn, U. Two lines: A valid alternative to the invalid testing of U-Shaped relationships with quadratic regressions. *Adv. Methods Pract. Psychol. Sci.* **1**, 538–555 (2018).
- 42.** Bhui, R., Lai, L. & Gershman, S. J. Resource-rational decision making. *Curr. Opin. Behav. Sci.* **41**, 15–21 (2021).
- 43.** Lieder, F. & Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43** (2020).
- 44.** Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. review psychology* **68**, 101–128 (2017).
- 45.** Liu, H., Ong, Y.-S., Shen, X. & Cai, J. When Gaussian process meets big data: A review of scalable GPs. *IEEE Transactions on Neural Networks Learn. Syst.* (2020).
- 46.** Otto, A. R., Knox, W. B., Markman, A. B. & Love, B. C. Physiological and behavioral signatures of reflective exploratory choice. *Cogn. Affect. & Behav. Neurosci.* **14**, 1167–1183 (2014).
- 47.** Cogliati Dezza, I., Cleeremans, A. & Alexander, W. Should we control? the interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *J. Exp. Psychol. Gen.* **148**, 977 (2019).
- 48.** Wu, C. M., Schulz, E., Pleskac, T. J. & Speekenbrink, M. Time pressure changes how people explore and respond to uncertainty. *PsyArXiv* (2021). DOI 10.31234/osf.io/dsw7q.
- 49.** Schwartenbeck, P. *et al.* Computational mechanisms of curiosity and goal-directed exploration. *Elife* **8**, e41703 (2019).
- 50.** Gershman, S. J. Origin of perseveration in the trade-off between reward and complexity. *Cognition* **204**, 104394 (2020).
- 51.** Zenon, A., Solopchuk, O. & Pezzulo, G. An information-theoretic perspective on the costs of cognition. *Neuropsychologia* **123**, 5–18 (2019).
- 52.** Piray, P. & Daw, N. D. Linear reinforcement learning in planning, grid fields, and cognitive control. *Nat. communications* **12**, 1–20 (2021).
- 53.** Del Giudice, M., Gangestad, S. W. & Kaplan, H. S. Life history theory and evolutionary psychology. (2016).
- 54.** Frankenhuys, W. E., Panchanathan, K. & Nettle, D. Cognition in harsh and unpredictable environments. *Curr. Opin.* (2016).
- 55.** Powers, A. R., Mathys, C. & Corlett, P. R. Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science* **357**, 596–600 (2017).
- 56.** de Berker, A. O. *et al.* Computations of uncertainty mediate acute stress responses in humans. *Nat. Commun.* **7**, 10996 (2016).
- 57.** Schwartenbeck, P. *et al.* Optimal inference with suboptimal models: addiction and active bayesian inference. *Med. hypotheses* **84**, 109–117 (2015).
- 58.** Laube, C., van den Bos, W. & Fandakova, Y. The relationship between pubertal hormones and brain plasticity: Implications for cognitive training in adolescence. *Dev. Cogn. Neurosci.* **42**, 100753 (2020).
- 59.** Dahl, R. E., Allen, N. B., Wilbrecht, L. & Suleiman, A. B. Importance of investing in adolescence from a developmental science perspective. *Nature* **554**, 441–450 (2018).
- 60.** Blakemore, S.-J. & Mills, K. L. Is adolescence a sensitive period for sociocultural processing? *Annu. Rev. Psychol.* **65**, 187–207 (2014).
- 61.** Gigerenzer, G. From tools to theories: A heuristic of discovery in cognitive psychology. *Psychol. Rev.* **98**, 254–267 (1991).
- 62.** Dasgupta, I., Schulz, E. & Gershman, S. J. Where do hypotheses come from? *Cognitive Psychology* **96**, 1–25 (2017).
- 63.** Barry, D. N. & Love, B. C. Human learning follows the dynamics of gradient descent (2021).
- 64.** Ritz, H., Leng, X. & Shenhav, A. Cognitive control as a multivariate optimization problem. *J. Cogn. Neurosci.* 1–23 (2022).

65. Hennig, J. A. *et al.* How learning unfolds in the brain: toward an optimization view. *Neuron* **109**, 3720–3735 (2021).
66. Rule, J. S., Tenenbaum, J. B. & Piantadosi, S. T. The child as hacker. *Trends cognitive sciences* **24**, 900–915 (2020).
67. Rescorla, R. A. & Wagner, A. R. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Class. conditioning II: Curr. research theory* **2**, 64–99 (1972).
68. Gershman, S. J. A unifying probabilistic view of associative learning. *PLoS Comput. Biol* **11**, e1004567 (2015).
69. Gershman, S. J. Deconstructing the human algorithms for exploration. *Cognition* **173**, 34–42 (2018).
70. Dayan, P., Kakade, S. & Montague, P. R. Learning and selective attention. *Nat. neuroscience* **3**, 1218–1223 (2000).
71. Rieskamp, J., Busemeyer, J. R. & Laine, T. How do people learn to allocate resources? comparing two learning theories. *J. Exp. Psychol. Learn. Mem. Cogn.* **29**, 1066 (2003).
72. Moreno-Bote, R., Ramírez-Ruiz, J., Drugowitsch, J. & Hayden, B. Y. Heuristics and optimal solutions to the breadth–depth dilemma. *Proc. Natl. Acad. Sci.* **117**, 19799–19808 (2020).
73. Nourani, Y. & Andresen, B. A comparison of simulated annealing cooling strategies. *J. Phys. A: Math. Gen.* **31**, 8373 (1998).
74. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. & Rev.* **16**, 225–237 (2009).
75. Jeffreys, H. *The Theory of Probability* (Oxford, UK: Oxford University Press, 1961).
76. Ly, A., Verhagen, J. & Wagenmakers, E.-J. Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *J. Math. Psychol.* **72**, 19–32 (2016).
77. van Doorn, J., Ly, A., Marsman, M. & Wagenmakers, E.-J. Bayesian inference for kendall's rank correlation coefficient. *The Am. Stat.* **72**, 303–308 (2018).
78. Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *Elife* **8**, e49547 (2019).
79. Vehtari, A., Gelman, A. & Gabry, J. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *arXiv [stat.CO]* (2015).

Acknowledgments

We thank Kou Murayama, Michiko Sakaki, Philipp Schwartenbeck and Mani Hamidi for helpful feedback and Chantal Wysocki for data collection. CMW is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC2064/1–390727645.

Author Contributions

CMW, SC, and APG conceived the study with feedback from all authors. SC collected the data using materials provided by CMW. CMW, APG, and SC performed the analyses. CMW, SC, and APG wrote the paper, with feedback from all authors.

Competing financial interests

The authors declare no competing financial interests.

Supplementary Information for Developmental changes in learning resemble stochastic optimization

Anna Giron, Simon Ciranka, Eric Schulz, Wouter van den Bos, Azzurra Ruggeri, Björn Meder, & Charley M. Wu

Comparability Across Experiments

To assess whether the data from the different experiments are equivalent, we compared participants' performance (Fig. S1a), the mean predictive accuracy of the GP-UCB model (Fig. S1b), and parameter estimates for all parameters of the GP-UCB model (Fig. S1c). Both, frequentist statistics and Bayes factors (BF) are reported (see Statistics). We compared data from the different experiments for overlapping age ranges: participants between 7-9 from Experiment 1 ($n = 36$) and Experiment 2 ($n = 22$) and participants older than 20 years from Experiment 1 ($n = 24$) and Experiment 3 ($n = 43$). Participants younger than or equal to 20 years were excluded from the comparisons, because of expected developmental changes over the course of adolescence.

Plot S1a shows participants' performance separated by experiments. We found no difference in performance for both age groups (children between 7-9: $t(56) = 1.5, p = .128, d = 0.4, BF = .73$; adults older than 20: $t(65) = -0.7, p = .460, d = 0.2, BF = .33$). Additionally, a comparison of the predictive accuracy showed no differences in model performance between different experiments (children between 7-9: $t(56) = 0.7, p = .467, d = 0.2, BF = .34$; adults older than 20: $t(65) = -0.9, p = .394, d = 0.2, BF = .35$; see Figure S1b). Furthermore, we compared parameter estimates for overlapping age ranges. Since parameter estimates were not normally distributed, we performed rank-based tests (Mann–Whitney U) to look for differences in parameter estimates across different experiments. The results are reported in Table S1 and Figure S1c), where we did not find any significant differences.

From these results, we concluded that data from the different experiments can be integrated and used for joint analyses of behavioral changes over the lifespan.

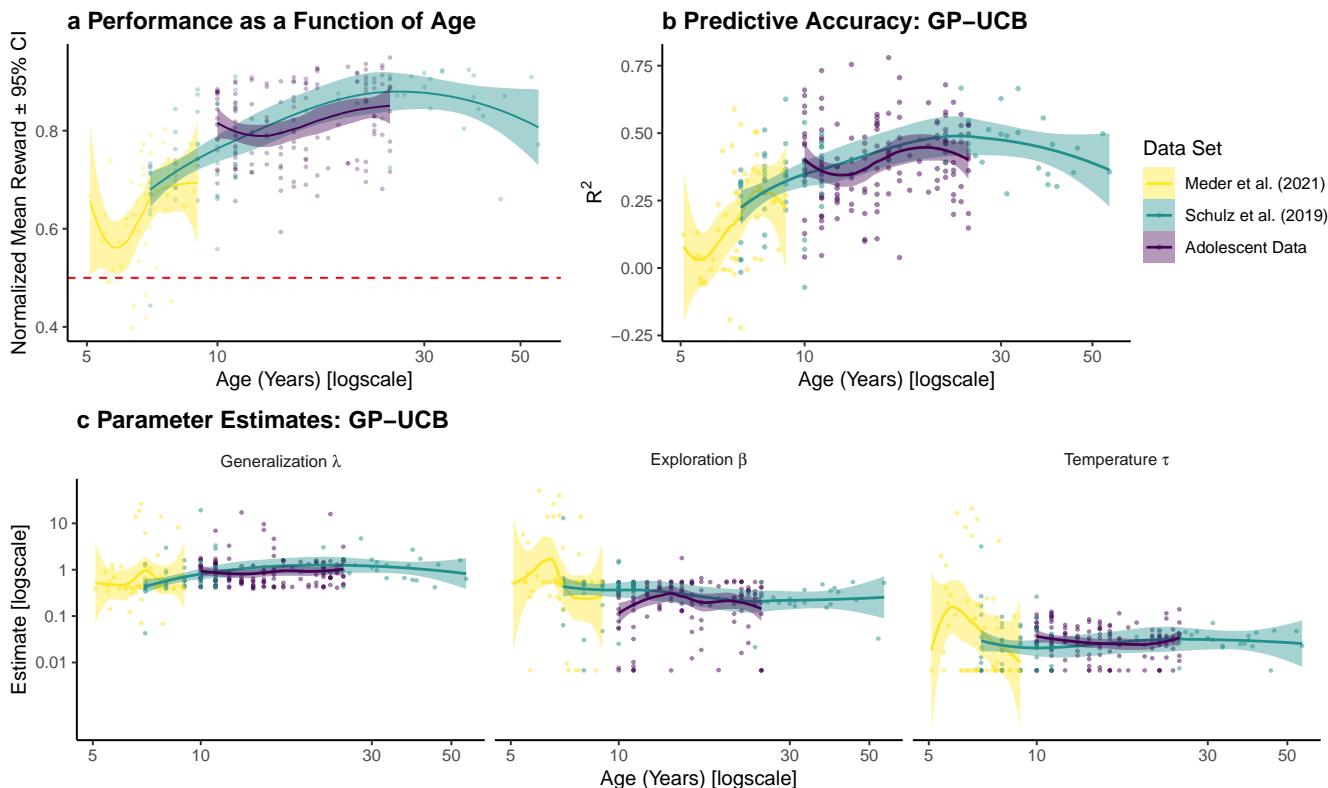


Figure S1. Reliability checks. **a** Average reward as a function of age, lines show the smoothed conditional means and 95% confidence interval for data from each experiment. Dots represent the mean reward for each participant and the red dashed line shows the performance of a random choice model. **b**) Predictive accuracy of the GP-UCB model as a function of age and **c**) Parameter estimates of the GP-UCB model as a function of age, separated by experiment. Lines show the smoothed conditional means and 95% confidence interval and dots represent the predictive accuracy per participant.)

Statistics

Comparisons

Both frequentist and Bayesian statistics are reported throughout this paper. Frequentist tests are reported as Student's t -tests (specified as either paired or independent). Each of these tests are accompanied by a Bayes factors (BF) to quantify the relative evidence the data provide in favor of the alternative hypothesis (H_A) over the null (H_0). This is done using the default two-sided

Table S1. Comparison of GP-UCB parameter estimates across different experiments

Parameter	(Age)	<i>U</i>	<i>p</i>	r_τ	<i>BF</i>
Generalization λ	(7-9)	390	.930	-.01	.29
Generalization λ	(>20)	438	.313	-.10	.38
Exploration β	(7-9)	373	.721	-.04	.29
Exploration β	(>20)	554	.626	.05	.28
Temperature τ	(7-9)	422	.685	.05	.31
Temperature τ	(>20)	496	.800	-.03	.27

Note: We conducted non-parametric two-sample tests to compare the estimates for children between 7-9 from Experiment 1 ($n = 36$) and Experiment 2 ($n = 22$) and the estimates for adult participants older than 20 from Experiment 1 ($n = 24$) and Experiment 3 ($n = 43$).

Bayesian *t*-test for either independent or dependent samples, where both use a Jeffreys-Zellner-Siow prior with its scale set to $\sqrt{2}/2$, as suggested by Ref⁷⁴. All statistical tests are non-directional as defined by a symmetric prior.

Correlations

For testing linear correlations with Pearson's *r*, the Bayesian test is based on Jeffreys⁷⁵ test for linear correlation and assumes a shifted, scaled beta prior distribution $B(\frac{1}{k}, \frac{1}{k})$ for *r*, where the scale parameter is set to $k = \frac{1}{3}$ ⁷⁶. For testing rank correlations with Kendall's tau, the Bayesian test is based on parametric yoking to define a prior over the test statistic⁷⁷, and performing Bayesian inference to arrive at a posterior distribution for r_τ . The Savage-Dickey density ratio test is used to produce an interpretable Bayes Factor. Note that when performing group comparisons of correlations computed at the individual level, we report the mean correlation and the statistics of a single-sample *t*-test comparing the distribution of *z*-transformed correlation coefficients to $\mu = 0$.

Supplementary Behavioral results

Search distance regression

We used a Bayesian hierarchical linear regression model to analyze the relation between the obtained reward and search distance in the next trial. In this model, participants were treated as random effects. Age group, reward obtained in the previous round, and their interaction were treated as fixed effects with random slopes. Table S2 provides the model results. They indicate an effect of age group, older participants showed higher search distances. Additionally, we found an effect of previous reward on search distance, higher rewards lead to sampling of closer tiles. Furthermore, results suggest an interaction between age group and previously obtained reward: older participants adjust their search distance more in line with the previously obtained reward than younger participants do.

Supplementary Model Results

Table S3. Summary of model results over all participants and separated by age group

Model	<i>R</i> ²	<i>n</i>	nLL	<i>p</i> _{pxp}	Gener. λ	Error Var. $\sqrt{\theta_\epsilon^2}$	Explor. β	Temp. τ	Epsi. ϵ
Overall									
GP-UCB	0.34	178	483.94	>.99	0.66		0.38	0.03	
λ lesion	0.18	23	609.91	<.01		4.3	0.3	0.05	
β lesion	0.24	47	563.62	<.01	1.66			0.13	
τ lesion	0.2	33	599.93	<.01	0.32		1.94		0.6
Age 5-6									
GP-UCB	0.11	14	371.15	.60	0.42		0.55	0.07	
λ lesion	0.02	3	405.8	.07		2.42	12.91	1.03	

Table S3. Summary of model results over all participants and separated by age group

Model	R^2	n	nLL	p_{xp}	Gener. λ	Error Var. $\sqrt{\theta_\epsilon^2}$	Explor. β	Temp. τ	Epsi. ϵ
β lesion	0.03	2	401.43	.07	1.43			0.23	
τ lesion	0.11	11	372.18	.27	0.15		5.23		0.75
Age 7-8									
GP-UCB	0.26	32	478.06	>.99	0.45		0.52	0.02	
λ lesion	0.05	0	611.19	<.01		2.79	3.94	0.18	
β lesion	0.1	4	581.7	<.01	1.57			0.18	
τ lesion	0.17	11	530.1	<.01	0.15		7.17		0.61
Age 9-10									
GP-UCB	0.36	24	525.57	>.99	0.67		0.38	0.03	
λ lesion	0.17	1	682.01	<.01		3.26	0.29	0.07	
β lesion	0.25	8	614.91	<.01	1.64			0.14	
τ lesion	0.21	4	650.85	<.01	0.27		2.18		0.6
Age 11-13									
GP-UCB	0.36	32	533.65	>.99	0.67		0.37	0.03	
λ lesion	0.17	5	687.85	<.01		4.9	0.28	0.04	
β lesion	0.26	10	616.47	<.01	1.63			0.12	
τ lesion	0.19	3	671.15	<.01	0.51		0.54		0.62
Age 14-17									
GP-UCB	0.4	25	498.37	>.99	0.72		0.34	0.03	
λ lesion	0.25	5	627.3	<.01		4.25	0.26	0.03	
β lesion	0.29	9	587.46	<.01	1.72			0.1	
τ lesion	0.23	3	640.55	<.01	0.95		0.3		0.58
Age 18-24									
GP-UCB	0.43	32	470.78	>.99	0.89		0.28	0.03	
λ lesion	0.28	6	595.40	<.01		5.45	0.28	0.01	
β lesion	0.34	6	550.87	<.01	1.82			0.07	
τ lesion	0.24	0	634.72	<.01	1.26		0.22		0.53
Age 25-55									
GP-UCB	0.43	19	471.26	>.99	1.23		0.24	0.03	
λ lesion	0.29	3	590.8	<.01		3.96	0.24	0.02	
β lesion	0.36	8	532.53	<.01	2.1			0.07	
τ lesion	0.22	1	646.14	<.01	1.19		0.21		0.57

Note: We report the average predicted accuracy per model (R^2), the amount of participants best described by the respective model (n), the average out-of-sample negative log likelihood (nLL), the protected exceedance probability (p_{xp}), and median parameter estimates of Generalization λ , Error Variance $\sqrt{\theta_\epsilon^2}$, Exploration β , Temperature τ , and the epsilon-greedy parameter ϵ .

Model recovery

A prerequisite for interpreting computational models is to ensure each model is uniquely identifiable⁷⁸. To establish that this is the case for models under consideration, we simulated experiments using the subject-level parameter estimates from each model. These simulations were performed using the same set of environments experienced by participants and generated the same data structure as recorded from participants. We then re-estimated each model to the simulated data sets and evaluated how often each model provided the best account of the generated data (based on the summed nLLs over from leave-one-round-out cross-validation). This provides an estimate of the probability with which each model best fits the data given a known simulating model (Figure S3 top row; in each plot, the columns sum to 1). Based on these probabilities we also computed the inverse

Table S2. Bayesian linear multilevel regression: search distance as a function of previous reward.

	Estimate	95% HDI
Intercept	8.35	[7.64 - 9.04]
Previous reward	-7.76	[-8.51 -- -6.99]
Age group 18-24	-0.75	[-1.65 - 0.17]
Age group 14-17	-1.25	[-2.18 -- -0.33]
Age group 11-13	-1.35	[-2.22 -- -0.44]
Age group 9-10	-2.15	[-3.05 -- -1.23]
Age group 7-8	-3.91	[-4.82 -- -3.01]
Age group 5-6	-5.30	[-6.25 -- -4.36]
Previous reward × Age group 18-24	0.70	[-0.31 - 1.67]
Previous reward × Age group 14-17	1.28	[0.27 - 2.29]
Previous reward × Age group 11-13	1.47	[0.48 - 2.42]
Previous reward × Age group 9-10	2.37	[1.33 - 3.39]
Previous reward × Age group 7-8	4.43	[3.45 - 5.44]
Previous reward × Age group 5-6	6.32	[5.21 - 7.42]
Random Effects		
σ^2	3.17	
τ_{00}	3.52	
τ_{11} previous reward	4.24	
τ_{11} age group 18-24	0.04	
τ_{11} age group 14-17	0.16	
τ_{11} age group 11-13	0.10	
τ_{11} age group 9-10	0.33	
τ_{11} age group 7-8	0.05	
τ_{11} age group 5-6	0.40	
ICC	0.11	
N	281	
Observations	51000	
Bayesian R^2	0.40	

Note: Posterior mean and 95% highest density interval (HDI) are reported for all coefficients. Age group 25-55 is the reference level for the categorical variable age group. σ^2 indicates the individual-level variance and τ the variation between individual intercepts and average intercept. ICC is the intraclass correlation coefficient.

matrix, using Bayes theorem to define the probability that a model in fact generated the data given that observation that it provided the best fit to the data (Figure S3 bottom row; in each plot, the rows sum to 1).

Parameter recovery

We simulated choices using the winning GP-UCB in order to establish whether the models' parameter estimates are reliable. We evaluated the recoverability of the GP-UCB parameters based on two procedures.

First, following ref⁷⁸, we produced a synthetic data-set by simulating choices the base of all empirical subject-level parameter estimates obtained from fitting model to the adolescent data set (Figure S4a; for recovery analyses of the other data-sets see²⁵). In all cases, recovered parameters were highly correlated with the parameters used to generate the data (all $r_\tau > .91$, $p < .001$, $BF > 100$).

Second, we performed an even more rigorous parameter recovery by iteratively varying each parameter of the GP-UCB model over 20 linearly spaced values within a credible interval of parameter estimates in all experiments (using Tukeys' fence). Again, we find a high degree of correlation between recovered and generating parameter value (all $r_\tau > .86$, $p < .001$, $BF > 100$).

While the first parameter recovery provides evidence that empirically observed parameter values are recoverable, the second analysis provides additional evidence that even counterfactual parameters (within a credible interval across all datasets) are also recoverable.

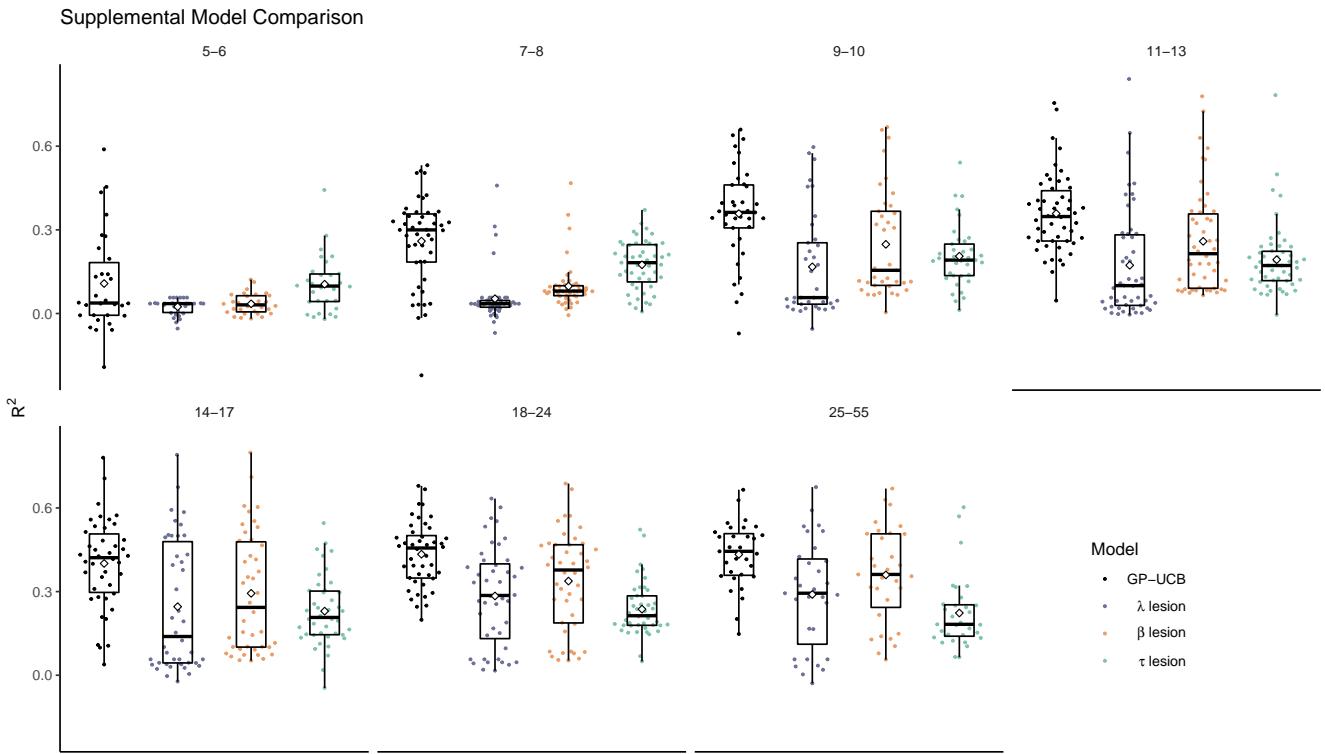


Figure S2. Supplemental model comparisons. Tukey's boxplots showing the predictive accuracy for all models and age groups. Each dot is the predictive accuracy for one participant, white diamonds indicate the group mean.

Parameter Regression Models

Comparison of Parameter Regression Models

For all regression models quantifying the relationship between the GP-UCB parameters and age we computed approximate leave one out cross validation using pareto-importance sampling as described in ⁷⁹. We then used the resulting expected log pointwise predictive densities as model selection criterion.

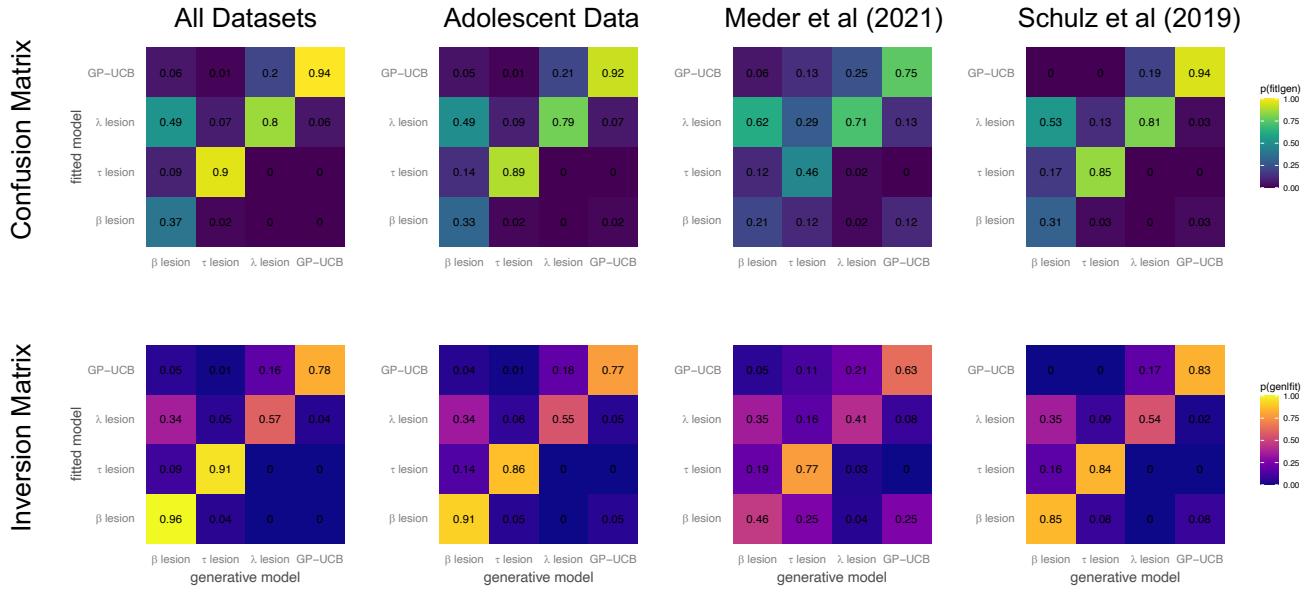


Figure S3. Model Recovery. **Top row:** Confusion matrices showing the probability that a model provides the best out-of-sample predictions (y-axes), given an underlying generative model (x-axes), where behavior was generated using participant parameter estimates. Columns sum to 1 (subject to rounding error). **Bottom row:** Inversion matrices showing the probability that a model did in fact generate the data (x-axes), given that it was found to provide the best out-of-sample predictions (y-axes). Rows sum to 1 (subject to rounding error). The first column refers to all experiments combined, subsequent columns show individual experiments

Table S4. Parameters of the changepoint regression models

Model	Parameter	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
λ (Generalization)								
	Intercept	-0.10000	0.14	-0.440000	0.15000	1	919.82	653.44
	b1	0.15000	0.28	0.030000	0.76000	1	1307.89	716.85
	b2	0.00000	0.01	-0.010000	0.02000	1	1302.60	1892.00
	ω	10.59928	0.64	3.924681	15.58052	1	801.83	363.16
β (Directed Exploration)								
	Intercept	-1.530000	0.16	-1.850000	-1.22000	1	2906.99	3721.15
	b1	-0.610000	0.24	-1.120000	-0.25000	1	2561.90	4001.08
	b2	0.000000	0.01	-0.020000	0.03000	1	3291.90	4158.29
	ω	8.658142	0.23	7.132697	11.19427	1	2250.68	3320.49
τ (Random exploration)								
	Intercept	-3.63000	0.14	-3.91000	-3.35000	1	3836.08	4766.03
	b1	-1.04000	0.25	-1.56000	-0.61000	1	4236.22	4023.70
	b2	0.00000	0.01	-0.02000	0.03000	1	4017.78	4547.85
	ω	7.50387	0.08	6.814792	8.219191	1	3749.10	3796.53

Note: Each model was fit using Hamiltonian Monte-carlo sampling with 4 Markov chains that each drew 4000 samples, 2000 of which were discarded as warm-up. The first column (estimate) refers to the maximum-a-posteriori estimate of the respective parameters. The second column (est-error) denotes the standard deviation of the posterior. The third and fourth column denote the lower and upper credible interval of the posterior. The fifth column denotes the Gelman Rubin Statistic (Rhat) indicating chain convergence. The sixth and seventh columns shows the number of effective samples from the bulk and the tail of the posterior.

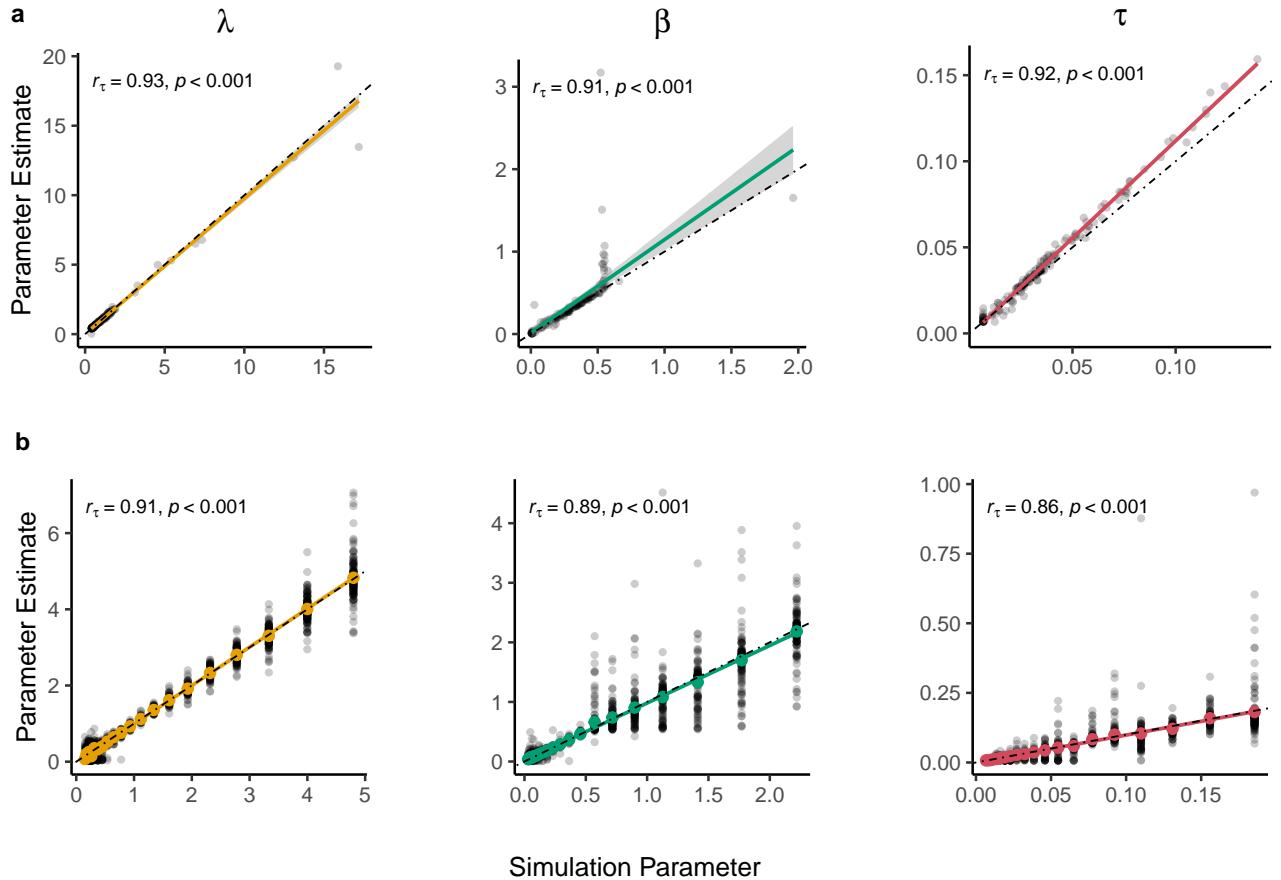


Figure S4. Parameter recovery (GP-UCB). **a)** Parameter recovery using participant parameter estimates to generate data. The black dots denote individual parameter estimates. The colored lines show linear fit, while the dotted line denotes a diagonal indicative of perfect recovery. **b)** Augmented parameter recovery, using participant parameters but with systematic variation across a log-space grid of plausible counterfactual parameter values. This provides additional robustness by simulating data across a wider range of possible parameters. Black dots denote each simulated and generating parameter, while the colored dots show the mean 95%CI. The colored lines show linear fit, while the dotted line denotes a perfect recovery. The insetted statistics refer to the rank correlation between generative and fitted parameters, Kendalls τ .

Table S5. Comparison of Parameter Regression Models

Parameter	Model	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
λ (Generalization)	Changepoint	0.0000000	0.000000	-363.2834	25.26063	8.740013	2.578332	726.5669	50.52127
	4th degree polynomial	-0.8773263	1.160450	-364.1608	24.76685	9.715186	2.385898	728.3215	49.53369
	3rd degree polynomial	-1.2568625	1.412706	-364.5403	24.67081	8.748335	2.283367	729.0806	49.34163
	2nd degree polynomial	-1.3478547	1.928127	-364.6313	24.47876	7.770409	2.065985	729.2626	48.95753
	Linear	-2.0762882	2.231558	-365.3597	24.33395	7.707049	2.046198	730.7195	48.66789
β (Exploration)	Changepoint	0.000000	0.000000	-510.0002	17.49575	8.197045	1.597759	1020.000	34.99151
	Linear	-1.950623	4.249482	-511.9508	17.76379	6.150672	0.938005	1023.902	35.52758
	4th degree polynomial	-3.012897	2.552438	-513.0131	17.30195	8.017924	1.258345	1026.026	34.60390
	2nd degree polynomial	-3.343272	4.245685	-513.3435	17.76542	7.377139	1.161516	1026.687	35.53084
	3rd degree polynomial	-4.147480	3.786323	-514.1477	17.56514	8.279207	1.511922	1028.295	35.13028
τ (Temperature)	Changepoint	0.00000	0.000000	-509.2732	19.38446	8.942122	1.641917	1018.546	38.76892
	4th degree polynomial	-8.16587	4.037310	-517.4390	19.59777	9.192070	1.548149	1034.878	39.19554
	3rd degree polynomial	-9.53390	5.278764	-518.8071	20.12364	8.674504	1.512973	1037.614	40.24728
	Linear	-10.30415	6.541017	-519.5773	20.84232	7.901712	1.515522	1039.155	41.68465
	2nd degree polynomial	-10.55901	6.308428	-519.8322	20.65004	8.156683	1.498387	1039.664	41.30008

Note: Comparison of regression models predicting log-transformed parameter estimates as a function of age. Models are described in descending order of fit (best models first). elpd_diff describes the difference in expected log point-wise predictive density, relative to the model with the best predictive accuracy, while se_diff denotes the standard error of the difference. elpd_loo describes the Bayesian LOO estimate of the expected log pointwise predictive density (Eq 4 in Ref⁷⁹), which is the sum of $n = 281$ pointwise predictive densities, with se_elpd_loo denoting the standard error. p_loo, is the difference between the elpd-loo and the non-cross-validated predictive density., with se_p_loo denoting the standard error. Lastly, looic is the loo information criterion $-2 * \text{elpd}$ and can be interpreted similarly to other information criteria such as BIC or AIC, with se_looic denoting its standard error.

Simulated Reward (Faceted by Temperature τ)

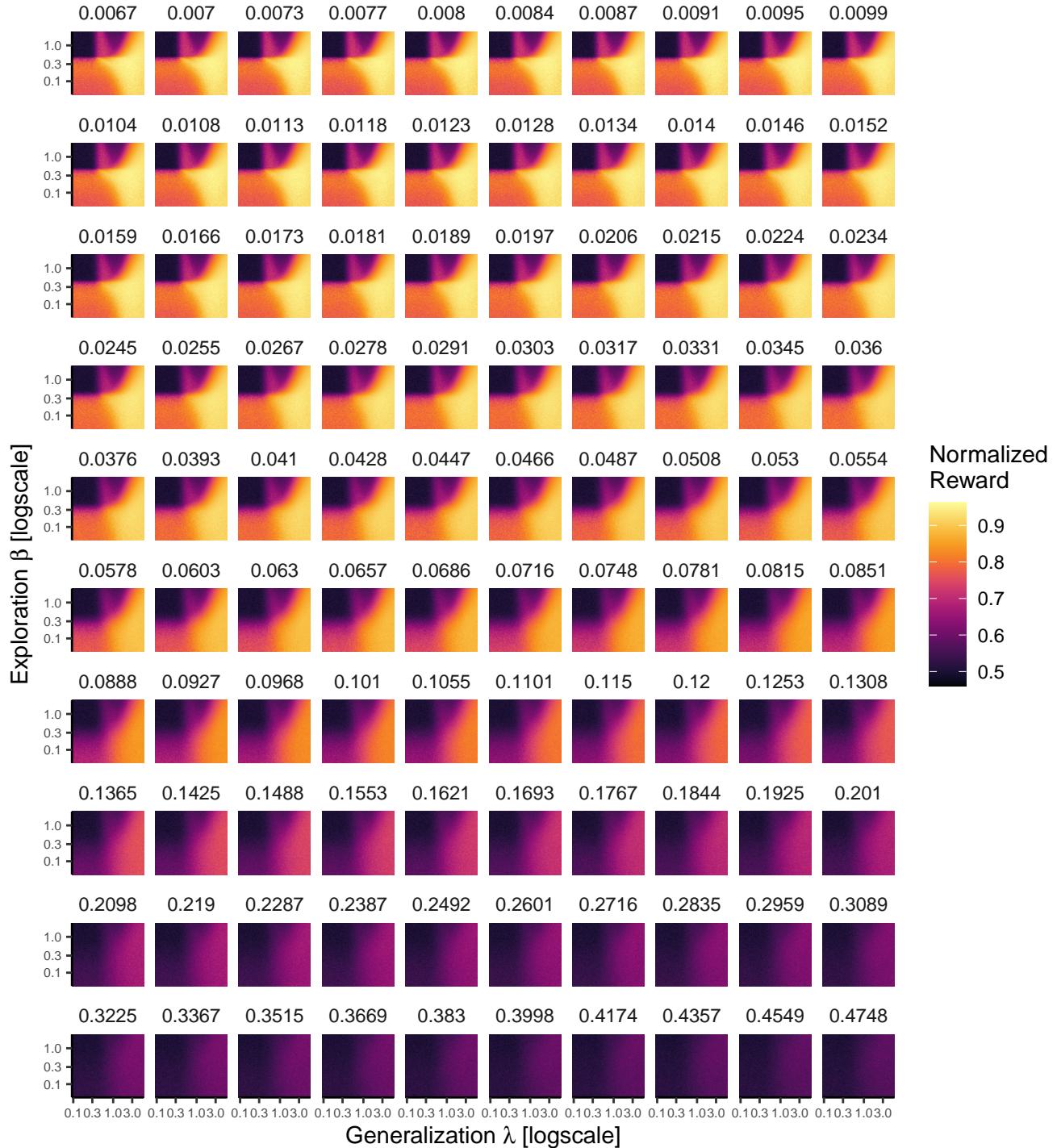
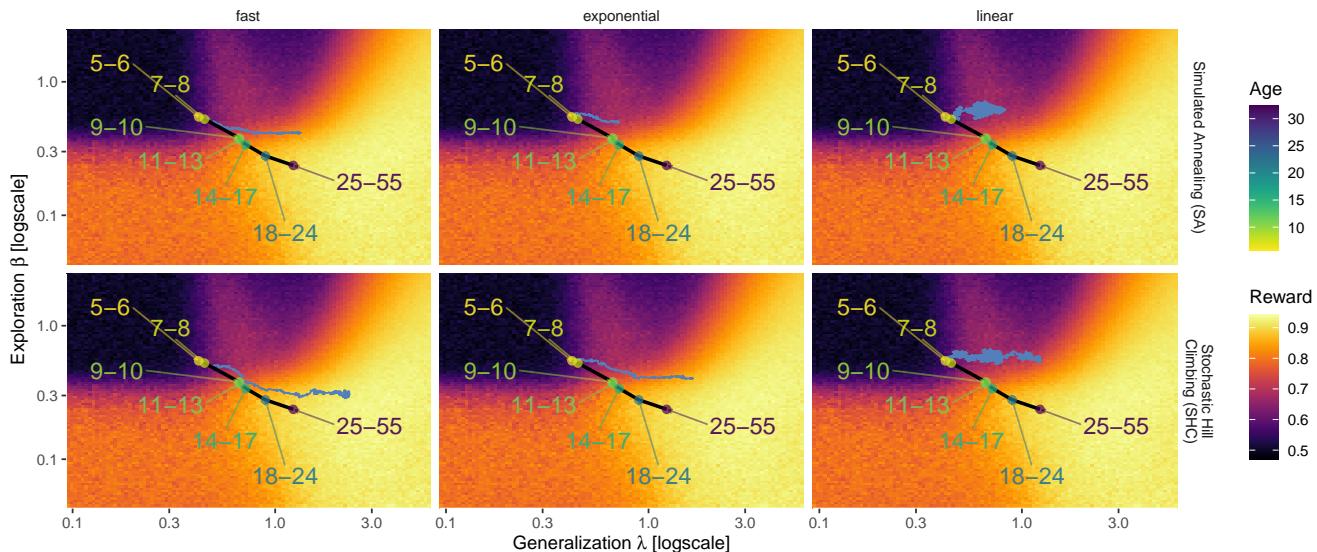


Figure S5. Fitness landscape. The 3-dimensional fitness landscape is computed over 1 million parameter combinations across λ (x-axis), β (y-axis), and τ (facets). For each parameter, we defined 100 equally log-spaced values over a credible range of participant parameter estimates (using Tukey's fence). We then simulated 100 rounds of the task using each combination of parameters (sampling with replacement from the same set of 40 environments given to participants) in order to compute an average reward, which we then normalized to a max of 1.

a Optimization Trajectories



b

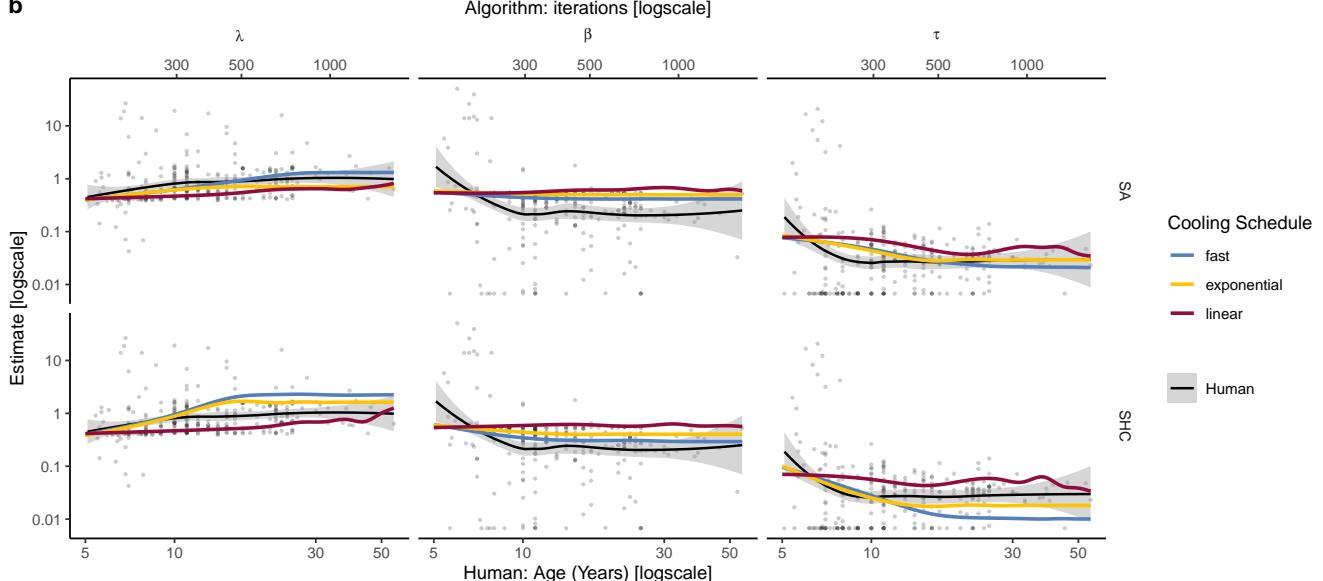


Figure S6. Supplementary results of the optimization algorithms. **a)** Trajectories of λ and β for each combination of optimization algorithm (rows) and cooling function (columns). For the background, we used the τ value with the smallest difference to the mean simulated τ value across all trajectories and iterations ($\bar{\tau} = .03$). Human estimates (labeled dots) are also provided for comparison. **b)** Comparison of the human developmental trajectory (black line indicates smoothed means \pm 95% CI and dots show each individual parameter) to each optimization algorithm (colored lines). Human data is plotted along age in years (bottom axis), while the algorithm results are plotted in terms of iteration number (top axis), both in log scale.