Searching for rewards like a child means less generalization and more directed exploration

Eric Schulz[a,1], **Charley M. Wu**[b,1], **Azzurra Ruggeri**[c], **and Björn Meder**[c]

[a]Department of Psychology, Harvard University, Cambridge, Massachusetts, USA; [b]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany; [c]MPRG iSearch | Information Search, Ecological and Active learning Research with Children, Max Planck Institute for Human Development, Berlin, Germany

This manuscript was compiled on May 21, 2018

How do children and adults differ in their search for rewards? We consider three different hypotheses that attribute developmental differences to either children's increased random sampling, more directed exploration towards uncertain options, or narrower generalization. Using a search task in which noisy rewards are spatially correlated on a grid, we compare 55 younger children (age 7-8), 55 older children (age 9-11), and 50 adults (age 19-55) in their ability to successfully generalize about unobserved outcomes and balance the exploration-exploitation dilemma. Our results show that children explore more eagerly than adults, but obtain lower rewards. Building a predictive model of search to disentangle the unique contributions of the three hypotheses of developmental differences, we find robust and recoverable parameter estimates indicating that children generalize less and rely on directed exploration more than adults. We do not, however, find reliable differences in terms of random sampling.

Exploration-Exploitation | Generalization | Developmental Psychology

Alan Turing famously believed that in order to build a General Artificial Intelligence, one must create a machine that can learn like a child (1). Indeed, recent advances in machine learning often contain references to child-like learning and exploration (2, 3). Yet little is known about how children actually explore and search for rewards in their environments, and in what way their behavior differs from adults (4).

In the course of learning through interactions with the environment, all organisms, biological or machine, are confronted with the *exploration-exploitation dilemma* (5). This dilemma highlights two (frequently) opposing goals. The first is to explore unfamiliar options that might lead to low rewards, yet provide useful information for future decisions. The second is to exploit options known to have high expectations of reward, but potentially forgo learning about unexplored options.

Exploration is not the only ingredient required for adaptive search behavior. Another crucial component is a mechanism that can *generalize* beyond observed outcomes, thereby guiding search by forming inductive beliefs about novel options. For example, from a purely combinatorial perspective, it only takes a few features and a small range of values to generate a pool of options vastly exceeding what could ever be explored in a lifetime. Nonetheless, humans of all ages manage to generalize from limited experiences in order to choose from amongst a set of potentially unlimited possibilities (6). Thus, a model of human search also needs to incorporate our ability to generalize.

Previous research has found extensive variability and developmental differences in children's and adults' search behavior that not only result from a progressive refinement of basic cognitive functions (e.g., memory or attention), but also derive from systematic changes in the computational principles driving behavior (7). In particular, developmental differences in learning and decision making have been explained by appealing to three hypothesized mechanisms: children sample more randomly, explore more eagerly, or generalize more narrowly than adults.

In this paper we use a search task in which noisy and continuous rewards are spatially correlated on a grid such that similar rewards are clustered together. We investigate whether and how three potential sources of developmental change (generalization, random and directed exploration) explain the differences in the exploratory behavior of younger children (age 7-8), older children (age 9-11), and adults (age 19-55). Moreover, we provide a precise characterization of these competing ideas in a unified formal model. This enables us to distinguish the relative contributions of the suggested mechanisms, and map the developmental trajectories of generalization and exploration using both behavioral markers and parameter estimates from computational models. Both behavioral and modeling results converge on the finding that children generalize less, but explore in a more directed manner than adults. By contrast, we do not find reliable evidence of increased random sampling.

A tale of three mechanisms

Before introducing our models and experiment, we briefly review the three hypotheses of developmental differences.

Development as cooling off. Because solutions to the explore-exploit dilemma are generally intractable (8), heuristic alter-

Significance Statement

Theories of developmental differences in exploration-exploitation behavior broadly coalesce around three postulated mechanisms, describing children as either sampling more randomly, as exploring uncertain options more eagerly, or generalizing more narrowly than adults. We test these hypotheses using a paradigm that combines generalization and search together with a model whose parameters uniquely correspond to each of the postulated mechanisms. We find that children generalize less and engage in more directed exploration than adults. We do not, however, find reliable developmental differences in random exploration. These results enrich our understanding of maturation in reinforcement learning, demonstrating that children explore using uncertainty-guided mechanisms rather than simply behave in a more random fashion.

E.S., C.M.W, A.R., and B.M. designed the experiment, analyzed the data, and wrote the paper.

No conflict of interest declared.

[1]To whom correspondence should be adressed: ericschulz@fas.harvard.edu or cwu@mpib-berlin.mpg.de

natives are frequently employed. In particular, learning under the demands of the exploration-exploitation trade-off has been described using at least two distinct strategies (9).

One such strategy is increased *random exploration*, which uses noisy, random sampling to learn about new options. A key finding in the psychological literature is that children tend to try out more options than adults (10, 11). The same tendency to explore more options during young age has also been observed in other animals such as wasps (12) and rats (13). Recent theoretical proposals assume that children can be described by higher temperature parameters resulting in noisier sampling behavior, where the learner initially samples very randomly across a large set of possibilities, before eventually focusing on a smaller subset (14). This temperature parameter is expected to "cool off" with age, leading to lower levels of random exploration in late childhood and adulthood. The "cooling off" of random search over the lifespan has been loosely compared to algorithms of simulated annealing from computer science (15), where random exploration is encouraged during earlier trials and then gradually tapers off.

Development as reduction of directed exploration. Simply behaving more randomly is not the only way to tackle the exploration-exploitation dilemma. A second strategy uses *directed exploration* by preferentially sampling highly uncertain options in order to gain more information and reduce uncertainty about the environment. Directed exploration has been formalized by introducing an "uncertainty bonus" that values the exploration of lesser known options (16), with behavioral markers found in a number of studies (cf., 17, 18).

Directed exploration treats information as intrinsically valuable by inflating rewards by their estimated uncertainty (16). This leads to a more sophisticated and uncertainty-guided sampling strategy that could also explain developmental differences. Indeed, the literature on self-directed learning (19) shows that children are clearly capable of exploring their environment in a systematic, directed fashion. Already infants tend to value the exploration of uncertain options (20), and children can balance theory and evidence in simple exploration tasks (21) and are able to efficiently adapt their search behavior to different environmental structures (22, 23). Moreover, children can sometimes even outperform adults in self-directed learning of unusual relationships (24). Both directed and random exploration do not have to be mutually exclusive mechanisms, with recent research finding signatures of both types of exploration in adolescent and adult participants (25).

Development as refined generalization. Rather than explaining development as a change in how we explore given some beliefs about the world, generalization-based accounts attribute developmental differences to the way we form our beliefs in the first place. Thus, differences of exploratory behavior may emerge through the development of more complex cognitive processes, leading to broader generalizations (26). Because many studies have shown that human learners use structured knowledge about the environment to guide exploration (27–29), the quality of these representations and the way that people utilize them to generalize across experiences may have a crucial impact on search behavior.

The notion of generalization as a mechanism explaining developmental differences has a long standing history in psychology. For instance, Piaget's model of cognitive development (30) assumes that children learn and adapt to different situational demands by the processes of assimilation (applying a previous concept to a new task) and accommodation (changing a previous concept in the face of new information). Expanding on Piaget's idea, (31) proposed generalization as a crucial developmental process, in particular the mechanism of regularity detection, which supports generalization and improves over the course of development. More generally, the implementation of various forms of decision making (32) could be constrained by the capacity for complex cognitive processes, which become more refined over the life span. For example, although younger children attend more frequently to irrelevant information than older children (33), they can be prompted to attend to the relevant information by marking the most relevant cues, whereupon they eventually select the best alternative (34). Thus, children may indeed be able to apply uncertainty-driven exploratory strategies, but lack the appropriate task representation to successfully implement them.

A task to study generalization and exploration

We study children's and adults' behavior in a *spatially-correlated multi-armed bandits task* (Fig. 1A; 18), in which rewards are distributed on a grid characterized by spatial correlation (i.e., high rewards cluster together; see Fig 1G). Efficient search and accumulation of rewards in such an environment requires two critical components. First, participants need to make use of the underlying spatial correlation in order to generalize from observed rewards to unseen options. This is crucial because there are considerably more options than can be explored given a limited search horizon. Second, participants need a sampling strategy that achieves a balance between exploring new options and exploiting known options with high rewards. Our task is thus designed to assess the contributions of both generalization and exploration, in terms of how they may explain behavioral differences.

A combined model of generalization and exploration

Understanding the origins of developmental differences requires a model that can assess the individual contributions of random exploration, directed exploration, and generalization. We introduce and assess such a model (cf. 18), which combines a mechanism for generalization with a sampling strategy that accounts for both directed and random exploration. The model of generalization is based on a Bayesian approach to function learning called *Gaussian Process* (GP; 35) regression. GP regression is theoretically capable of learning any stationary function through Bayesian inference, and has been found to effectively describe human behavior in explicit function learning tasks (36).

The GP prior is completely determined by the choice of a kernel function $k(\mathbf{x}, \mathbf{x}')$, which encodes assumptions about how points in the input space are related to each other. A common choice of this function is the *radial basis function*:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{\lambda}\right), \qquad [1]$$

where the length-scale parameter λ encodes the extent of spatial generalization between options in the grid. The assumptions of the kernel function are similar to the gradient of

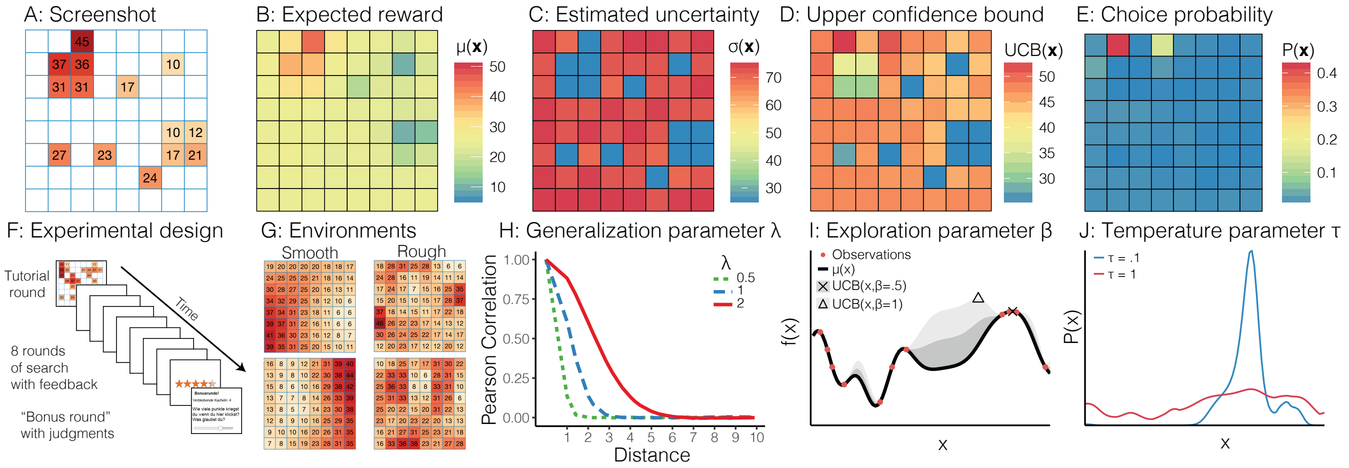

Fig. 1. Overview of task and model. **A:** Screenshot of experiment with partially revealed grid. Expected reward (**B**) and estimated uncertainty (**C**) based on observations in A using Gaussian Process regression. **D:** Upper confidence bounds of each option based on a weighted sum of panels B and C. **E:** Choice probabilities of softmax function. Panels **B-E** use median participant parameter estimates. Overview of the experimental design (**F**) and types of environments (**G**). **H:** Correlations between different options decay exponentially as a function of their distance, where higher values of λ (generalization parameter) lead to slower decays and broader generalizations. **I:** An illustration of UCB sampling using a univariate example, where the expected reward (black line) and estimated uncertainty (gray ribbon; for different values of β) are summed up, with higher values of β valuing the exploration of uncertain options more strongly (compare the argmax of the two beta values, indicated by the cross and the triangle). **J:** Overview of softmax function, where higher values of the temperature parameter τ lead to increased random exploration.

generalization historically described by Shepard (37), which also models the correlations between inputs $\mathbf{x}$ and $\mathbf{x}'$ as an exponentially decaying function of their distance (see Fig. 1H), and is found across a wide range of stimuli and organisms.

As an example, a generalization to the extent of $\lambda = 1$ corresponds to the assumption that the rewards of two tiles next to each other are correlated by $r = 0.6$, and that this correlation decays to zero if options are further than three tiles apart from each other. We treat λ as a free parameter in our model comparison in order to assess age-related differences in the capacity for generalization.

Given different possible options $\mathbf{x}$ to sample from (i.e., tiles in the grid), GP regression can be used to generate normally distributed beliefs about rewards with expectation $\mu(\mathbf{x})$ and the estimated uncertainty $\sigma(\mathbf{x})$. A sampling strategy is then used to map the beliefs of the GP onto a valuation for sampling each option at a given time. Crucially, such a sampling strategy must address the exploration-exploitation dilemma. One frequently applied heuristic for solving this dilemma is *Upper Confidence Bound* (UCB) sampling (38), which evaluates each option based on a weighted sum of expected reward and estimated uncertainty:

$$UCB(\mathbf{x}) = \mu(\mathbf{x}) + \beta\sigma(\mathbf{x}) \qquad [2]$$

where β models the extent to which uncertainty is valued positively and therefore directly sought out. This strategy corresponds to directed exploration because it encourages the sampling of options with higher uncertainty according to the underlying generalization model (see Fig. 1I). As an example, an exploration bonus of $\beta = 0.5$ means participants would prefer an option x_1 expected to have reward $\mu(x_1) = 30$ and uncertainty $\sigma(x_1) = 10$, over option x_2 expected to have reward $\mu(x_2) = 34$ and uncertainty $\sigma(x_2) = 1$. This is because sampling x_1 is expected to reduce a larger amount of uncertainty, even though x_2 has a higher expected reward ($\text{UCB}(x_1) = 35$ vs. $\text{UCB}(x_2) = 34.5$).

Although seemingly naïve, UCB sampling can lead to competitive performance guarantees when paired with Gaussian Process regression (38). We treat the exploration parameter β as a free parameter to assess how much participants value the reduction of uncertainty (i.e., engage in directed exploration).

Finally, our proposed Gaussian Process-Upper Confidence Bound sampling model only produces valuations of different options (sometimes called utilities or propensities), which need to be mapped onto choice probabilities. A common choice rule is the softmax function,

$$p(\mathbf{x}) = \frac{\exp(UCB(\mathbf{x})/\tau)}{\sum_{j=1}^{N}\exp(UCB(\mathbf{x}_j)/\tau)}, \qquad [3]$$

which transforms values (e.g., the upper confidence bound value $UCB(\mathbf{x})$) into a probability distribution over options, where τ is the temperature parameter governing the amount of randomness or noise in sampling behavior. Importantly, τ encodes the tendency towards random exploration (Fig. 1J). If τ is high (higher temperatures), then participants are assumed to sample more randomly, whereas if τ is low (cooler temperatures), the choice probabilities are concentrated on the highest valued options. We treat τ as a free parameter to assess the extent of random exploration (see SI for alternative implementations).

In summary, GP-UCB contains three different parameters: the length-scale λ capturing the extent of generalization, the exploration bonus β describing the extent of directed exploration, and the temperature parameter τ modulating random exploration. These three parameters directly correspond to the three postulated mechanisms of developmental differences in complex decision making task and can also be robustly recovered (see SI).

Experiment and Results

Participants sampled tiles on a two-dimensional grid to gain rewards (Fig. 1A). On each grid, rewards were spatially correlated (18), with the level of correlation manipulated between subjects (smooth or rough, corresponding to higher or lower spatial correlation; see Fig. 1G). There were 25 trials (i.e., tile choices) per round and participants completed ten rounds (i.e., grids) in total, instructed to "gain as many points as

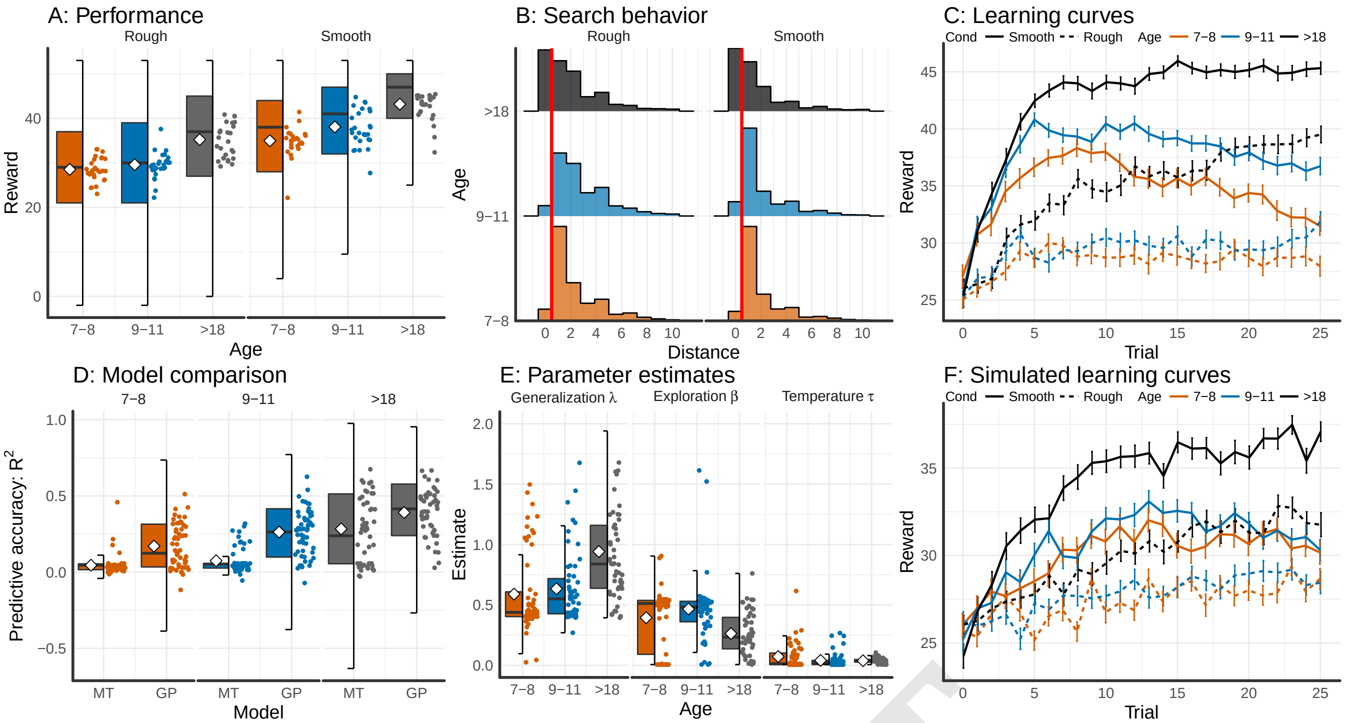

Fig. 2. Main results. **A:** Tukey box plots of rewards, showing the distribution of all choices for all participants. Each dot is the participant-wise mean and diamonds indicate group means. **B:** Histograms of distances between consecutive choices by age group and condition, with a distance of zero corresponding to a repeat click. The vertical red line marks the difference between a repeat click and sampling a different option. **C:** Mean reward over trials by condition (solid lines for smooth and dashed lines for rough) and age group (color). Error bars indicate the standard error of the mean. **D:** Tukey box plots showing the results of the model comparison between Gaussian Process (GP) and Mean Tracker (MT) models by age group. Each point is a single subject and group means are shown as a diamond. **E:** Tukey box plot of cross-validated parameters retrieved from the GP-UCB model by age group, where each point is a single median estimate per subject and diamonds indicate the mean. Outlier values higher than 5 have been removed (see SI). **F:** Learning curves simulated by GP-UCB model using mean participant parameter estimates. Error bars indicate the standard error of the mean.

possible”. The first round was a tutorial round, where participants interacted with a grid alongside the task instructions and had to complete three comprehension questions. The last round paused after 15 trials and asked participants to enter reward predictions for five previously unobserved tiles (randomly chosen), including how certain they were about their predictions on a scale from 0 to 10. Afterwards, participants chose one of the five selected tiles, received the reward of that tile, and then continued the round as before. All behavioral and modeling results exclude the tutorial and bonus round, with the exception of the analysis of bonus round judgments.

Behavioral results. Participants gained higher rewards in smooth than in rough environments (Fig. 2A; $t(158) = 10.51$, $p < .001$, $d = 1.66$, $BF > 100$), suggesting they made use of the spatial correlations in the environment and performed better when correlations were stronger. Adults performed better than older children (Fig. 2A; $t(103) = 4.91$, $p < .001$, $d = 0.96$, $BF > 100$), who in turn performed better than younger children ($t(108) = 2.42$, $p = .02$, $d = 0.46$, $BF = 2.68$). Analyzing the distance between consecutive choices (Fig. 2B) revealed that participants sampled more locally (smaller distances) in smooth compared to rough environments ($t(158) = -3.83$, $p < .001$, $d = 0.61$, $BF > 100$). Adults sampled more locally than older children ($t(103) = -3.9$, $p < .001$, $d = 0.76$, $BF > 100$), but there was no difference between younger and older children ($t(108) = 1.76$, $p = .08$, $d = 0.34$, $BF = 0.80$). Importantly, adults sampled less unique options than older children (14.5 vs. 21.7; $t(103) = 6.77$, $d = 1.32$, $p < .001$, $BF > 100$), whereas the two children groups did not differ in how many options they sampled (21.7 vs. 22.7; $t(108) = 1.27$, $d = 0.24$ $p = .21$, $BF = 0.4$).

Looking at the learning curves (i.e., averaged rewards over trials), we found a positive rank-correlation between mean rewards and trial number (Spearman's $\rho = .12$, $t(159) = 6.12$, $p < .001$, $BF > 100$). Although this correlation did not differ between the rough and smooth condition ($t(158) = -0.43$, $p = .67$, $d = 0.07$, $BF = 0.19$), it was significantly higher for adults than for older children ($t(103) = 5.90$, $p < .001$, $d = 1.15$, $BF > 100$). The correlation between trials and rewards did not differ between younger and older children ($t(108) = -1.87$, $p = .06$, $d = 0.36$, $BF = 0.96$). Therefore, adults learned faster, while children explored more extensively (see SI for further behavioral analyses).

Model comparison. We compared the GP-UCB model with an alternative model that does not generalize across options but is a powerful Bayesian model for reinforcement learning across independent reward distributions (*Mean Tracker*; MT). Model comparisons are based on leave-one-round-out cross-validation error, where we fit each model combined with the Upper Confidence Bound sampling strategy to each participant using a training set omitting one round, and then assessing predictive performance on the hold out round. Repeating this procedure for every participant and all rounds, we calculated the standardized prediction error by evaluating how much each model's out-of-sample likelihood performed better than chance, with 0 indicating random performance and 1 perfect predictions (see SI for full model comparison including other sampling strategies). The results of this comparison are shown in Fig. 2D. The GP-UCB model predicted participants' behavior better overall ($t(159) = 13.28$, $p < .001$, $d = 1.05$, $BF > 100$), and also for

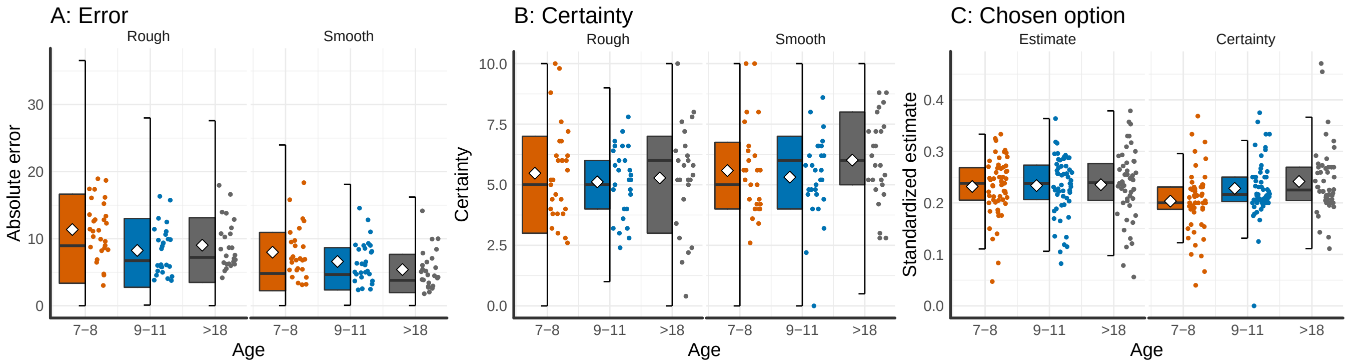

Fig. 3. Bonus round results. **A:** Absolute error of participant predictions about the rewards of unobserved tiles. **B:** Certainty judgments, where 0 is least certain and 10 is most certain. **C:** Standardized predictions and certainty estimates, where we divided judgments about the chosen option (both reward and certainty) by the sum of judgments for all five options. Thus, the standardized value indicates how much the estimated reward and certainty influenced choice (relative to judgments about non-chosen options). All figures show Tukey box plots including participant-wise data points and overall means indicated as diamonds.

adults ($t(49) = 5.98$, $p < .001$, $d = 0.85$, $BF > 100$), older ($t(54) = 10.92$, $p < .001$, $d = 1.48$, $BF > 100$) and younger children ($t(54) = 6.77$, $p < .001$, $d = 0.91$, $BF > 100$). The GP-UCB model predicted adults' behavior better than that of older children ($t(103) = 4.33$, $p < .001$, $d = 0.85$, $BF > 100$), which in turn was better predicted than behavior of younger children ($t(108) = 3.32$, $p = .001$, $d = 0.63$, $BF = 24.8$).

Testing computational differences. We analyzed the mean participant parameter estimates of the GP-UCB model (Fig. 2E) to assess the contributions of the three mechanisms (generalization, directed exploration, and random exploration) towards developmental differences. We found that adults generalized more than older children, as indicated by larger λ-estimates (Mann-Whitney-$U = 2001$, $p < .001$, $r_\tau = 0.32$, $BF > 100$), whereas the two groups of children did not differ significantly in their extent of generalization ($U = 1829$, $p = .06$, $r_\tau = 0.15$, $BF = 1.7$). Furthermore, older children valued uncertainty higher (i.e., had higher β-values leading to more directed exploration) than adults ($U = 629$, $p < .001$, $r_\tau = 0.39$, $BF > 100$), whereas there was no difference between younger and older children ($U = 1403$, $p = .51$, $r_\tau = 0.05$, $BF = 0.2$). Critically, whereas there were strong differences between age groups for the parameters capturing generalization and directed exploration, there was no reliable difference in the softmax temperature parameter τ, with no difference between older children and adults ($W = 1718$, $p = .03$, $r_\tau = 0.17$, $BF = 0.7$) and only anecdotal differences between the two groups of children ($W = 1211$, $p = .07$, $r_\tau = 0.14$, $BF = 1.4$).

This suggests that the amount of random exploration did not reliably differ by age group. Thus, our modeling results converge on the same conclusion as the behavioral results, namely that children explore more than adults. However, children's exploration behavior seems to be directed toward options with high uncertainty instead of merely more random. Finally, we used mean participant parameter estimates to simulate learning curves by letting the GP-UCB model produce outputs in the exact same setting as participants had encountered (see Fig. 2F). This showed that the GP-UCB can generate human-like behavior by reproducing the differences between the age groups as well as between smooth and rough conditions.

Bonus round. In the bonus round, each participant estimated expected rewards and the associated uncertainty for five unrevealed tiles after having made 15 choices on the grid. We first calculated the mean absolute error between predictions and the actual expected value of rewards (Fig. 3A). Prediction error was higher for rough compared to smooth environments ($t(158) = 4.93$, $p < .001$, $d = 0.78$, $BF > 100$), reflecting the lower degree of spatial correlation that could be utilized to evaluate unseen options. Surprisingly, older children were as accurate as adults ($t(103) = 0.28$, $p = .77$, $d = 0.05$, $BF = 0.2$), but younger children performed worse than older children ($t(108) = 3.14$, $p = .002$, $d = 0.60$, $BF = 15$). Certainty judgments did neither differ between the smooth and rough environments ($t(158) = 1.13$, $p = .26$, $d = 0.18$, $BF = 0.2$) nor between the different age groups (max-$BF = 0.1$).

Of particular interest is how judgments about the expectation of rewards and perceived uncertainty related to the eventual choice from amongst the five options. We standardized the estimated reward and confidence judgment of each participant's chosen tile by dividing by the sum of the corresponding estimates for all five options (i.e., dividing each predicted reward by the sum of all predicted rewards and dividing each certainty judgment by the sum of all certainty judgments; Fig. 3C). There was no difference between age groups in terms of the predicted rewards they chose (max-$BF = 0.1$).

By contrast, there was a difference between the age groups in terms of the certainty of the chosen option: younger children preferred options with higher uncertainty marginally more than older children ($t(108) = 2.22$, $p = .03$, $d = 0.42$, $BF = 1.8$), and substantially more than adults ($t(103) = 2.82$, $p = .006$, $d = 0.55$, $BF = 6.7$). This corroborates our previous analyses, showing that children's sampling behavior is directed more toward highly uncertain options than adults'.

Discussion

We examined three potential sources of developmental differences in a complex learning and decision-making task: random exploration, directed exploration and generalization. Using a paradigm that combines both generalization and search, we found that adults gained higher rewards and exploited more strongly, whereas children sampled more unique options, thereby gaining lower rewards but exploring the environment more extensively. Using a computational model with parameters directly corresponding to the three hypothesized mechanisms of developmental differences, we found that children generalized less and were guided by directed exploration more strongly than adults. They did not, however, explore substantially more randomly than adults. Our results paint a rich picture of developmental trajectories in generalization and exploration, casting children decision makers not as merely prone to noisy sampling behavior, but as directed explorers who are hungry for information in their environment. Our results suggest that to fulfill Alan Turing's dream of creating a child-like AI, we need to incorporate generalization and curiosity-driven exploration mechanisms (39).

Materials and Methods

Participants. We recruited 55 younger children (range: 7 to 9, 26 female, M_{age}=7.53; SD=0.50), 55 older children (range: 9 to 11, 24 female, M_{age}=9.95; SD=0.80), and 50 adults (25 female, M_{age}=33.76; SD= 8.53) from museums in Berlin, Germany. Participants were paid up to €3.50 for taking part in the experiment, contingent on performance (range: €2.00 to €3.50, M_{reward}=€2.67; SD=0.50). Informed consent was obtained from all participants.

Design. The experiment used a 2-groups between-subjects design, where participants were randomly assigned to one of two different classes of environments sampled from a bivariate Gaussian Process parameterized by a RBF kernel (*smooth* with $\lambda = 4$ vs. *rough* with $\lambda = 1$). Each grid world represented a bivariate function, with each observation including normally distributed noise, $\epsilon \sim \mathcal{N}(0, 1)$. The task was presented over 10 rounds on different grid worlds drawn from the same class of environments. The first round was a tutorial round in which the task was introduced and the last round was a bonus round in which participants sampled for 15 trials and then had to generate predictions for 5 randomly chosen tiles on the grid. Participants had a search horizon of 25 trials per grid, including repeat clicks.

Materials and procedure. Participants were introduced to the task through a tutorial round and were required to correctly complete three comprehension questions prior to continuing the task. At the beginning of each round, one random tile was revealed and participants could click on any of the tiles (including re-clicks) in the grid until the search horizon was exhausted. Clicking an unrevealed tile displayed the numerical value of the reward along with a corresponding color aid, where darker colors indicated higher rewards. Per round, observations were scaled to a randomly drawn maximum value in the range of 35 to 45, so that the value of the global optima could not be easily guessed. Re-clicked tiles could show some variations in the observed value due to noise. For repeat clicks, the most recent observation was displayed numerically, while the color of the tile corresponded to the mean of all previous observations. In the bonus round, participants sampled for 15 trials as before and were then asked to generate predictions for 5 randomly selected and previously unobserved tiles. Additionally, participants had to indicate how certain they were about their prediction on a scale from 0 to 10. Afterwards, they had to select one of the 5 tiles before then continuing with the round.

ACKNOWLEDGMENTS. We thank all of the families who participated in this research, the Berlin Museum of Natural History where we conducted the study, and Andreas Sommer for collecting the data. This work was supported by the Max Planck Society and DFG-grant ME 3717/2-2 to BM. ES is supported by the Harvard Data Science Initiative. Code and data are available at https://git.io/vppsK.

1. Turing A (1950) Computing intelligence and machinery. *Mind* 59(2236):433–460.
2. Riedmiller M, et al. (2018) Learning by playing-solving sparse reward tasks from scratch. *arXiv preprint arXiv:1802.10567.*
3. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. *Behavioral and Brain Sciences* 40:e253.
4. Gopnik A (2017) Making AI more human. *Scientific American* 316(6):60–65.
5. Mehlhorn K, et al. (2015) Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision* 2(3):191–215.
6. Sutton RS (1996) Generalization in reinforcement learning: Successful examples using sparse coarse coding in *Advances in neural information processing systems*. pp. 1038–1044.
7. Palminteri S, Kilford EJ, Coricelli G, Blakemore SJ (2016) The computational development of reinforcement learning during adolescence. *PLoS Computational Biology* 12(6):e1004953.
8. Gittins JC, Jones DM (1979) A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika* 66(3):561–565.
9. Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD (2014) Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General* 143(6):2074–2081.
10. Cauffman E, et al. (2010) Age differences in affective decision making as indexed by performance on the iowa gambling task. *Developmental Psychology* 46(1):193–207.
11. Mata R, Wilke A, Czienskowski U (2013) Foraging across the life span: is there a reduction in exploration with aging? *Frontiers in Neuroscience* 7:53.
12. Thiel A, Driessen G, Hoffmeister TS (2006) Different habitats, different habits? response to foraging information in the parasitic wasp venturia canescens. *Behavioral Ecology and Sociobiology* 59(5):614–623.
13. Lalonde R (2002) The neurobiological basis of spontaneous alternation. *Neuroscience & Biobehavioral Reviews* 26(1):91–104.
14. Gopnik A, Griffiths TL, Lucas CG (2015) When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science* 24(2):87–92.
15. Gopnik A, et al. (2017) Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences* 114(30):7892–7899.
16. Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3(Nov):397–422.
17. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature neuroscience* 12(8):1062–1068.
18. Wu CM, Schulz E, Speekenbrink M, Nelson JD, Meder B (2018) Exploration and generalization in vast spaces. *bioRxiv* p. 171371.
19. Gureckis TM, Markant DB (2012) Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science* 7(5):464–481.
20. Schulz LE (2015) Infants explore the unexpected. *Science* 348(6230):42–43.
21. Bonawitz EB, van Schijndel TJ, Friel D, Schulz L (2012) Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology* 64(4):215–234.
22. Ruggeri A, Lombrozo T (2015) Children adapt their questions to achieve efficient search. *Cognition* 143(1):203–216.
23. Nelson JD, Divjak B, Gudmundsdottir G, Martignon LF, Meder B (2014) Children's sequential information search is sensitive to environmental probabilities. *Cognition* 130(1):74–80.
24. Lucas CG, Bridgers S, Griffiths TL, Gopnik A (2014) When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition* 131(2):284–299.
25. Somerville LH, et al. (2017) Charting the expansion of strategic exploratory behavior during adolescence. *Journal of Experimental Psychology: General* 146(2):155–164.
26. Blanco NJ, et al. (2016) Exploratory decision-making as a function of lifelong experience, not cognitive decline. *Journal of Experimental Psychology: General* 145(3):284–297.
27. Acuna D, Schrater PR (2009) Structure learning in human sequential decision-making in *Advances in Neural Information Processing Systems.* pp. 1–8.
28. Schulz E, Konstantinidis E, Speekenbrink M (2017) Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology, Learning, Memory, and Cognition. Advance online publication.*
29. Wu CM, Schulz E, Speekenbrink M, Nelson JD, Meder B (2017) Mapping the unknown: The spatially correlated multi-armed bandit in *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.* pp. 1357–1362.
30. Piaget J (1964) Part i: Cognitive development in children: Piaget development and learning. *Journal of Research in Science Teaching* 2(3):176–186.
31. Klahr D (1982) Nonmonotone assessment of monotone development: An information processing analysis in *U-shaped behavioral growth*, eds. Strauss S, Stavy R. (Academic Press, New York), pp. 63–86.
32. Hartley CA, Somerville LH (2015) The neuroscience of adolescent decision-making. *Current Opinion in Behavioral Sciences* 5(9):108–115.
33. Hagen JW, Hale GA (1973) The development of attention in children. *ETS Research Report Series* 1973(1).
34. Davidson D (1996) The effects of decision characteristics on children's selective search of predecisional information. *Acta Psychologica* 92(3):263–281.
35. Rasmussen C, Williams C (2006) *Gaussian Processes for Machine Learning.* (MIT Press, Cambridge, MA, USA).
36. Lucas CG, Griffiths TL, Williams JJ, Kalish ML (2015) A rational model of function learning. *Psychonomic Bulletin & Review* 22(5):1193–1215.
37. Shepard RN (1987) Toward a universal law of generalization for psychological science. *Science* 237(4820):1317–1323.
38. Srinivas N, Krause A, Kakade SM, Seeger M (2009) Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995.*
39. Stadie BC, Levine S, Abbeel P (2015) Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814.*
40. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* 16(2):225–237.
41. van Doorn J, Ly A, Marsman M, Wagenmakers EJ (2017) Bayesian latent-normal inference for the rank sum test, the signed rank test, and Spearman's ρ. *arXiv preprint arXiv:1712.06941.*
42. Mullen K, Ardia D, Gil D, Windover D, Cline J (2011) DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software* 40(6):1–26.
43. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. *Neuroimage* 46(4):1004–1017.
44. Patil A, Huard D, Fonnesbeck CJ (2010) PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software* 35(4):1.