Development of directed and random exploration in children

Björn Meder
University of Erfurt and Max Planck Institute for Human Development

Charley M. Wu
Harvard University and Max Planck Institute for Human Development

Eric Schulz
Max Planck Institute for Biological Cybernetics

Azzurra Ruggeri
Max Planck Institute for Human Development and Technical University Munich

Are young children just random explorers who learn serendipitously? Or are even young children guided by uncertainty-directed sampling, seeking to explore in a systematic fashion? We study how children between the ages of 4 and 9 search in an explore-exploit task with spatially-correlated rewards, where exhaustive exploration is infeasible and not all options can be experienced. By combining behavioral data with a computational model that decomposes search into similarity-based generalization, uncertainty-directed exploration, and random exploration, we map out developmental trajectories of generalization and exploration. The behavioral data show strong developmental differences in children's capability to exploit environmental structure, with performance and adaptiveness of sampling decisions increasing with age. Through model-based analyses, we disentangle different forms of exploration, finding signatures of both uncertainty-directed and random exploration. The amount of random exploration strongly decreases as children get older, supporting the notion of a developmental "cooling off" process that modulates the randomness in sampling. However, even at the youngest age range, children do not solely rely on random exploration. Even as random exploration begins to taper off, children are actively seeking out options with high uncertainty in a goal-directed fashion, and using inductive inferences to generalize their experience to novel options. Our findings provide critical insights into the behavioral and computational principles underlying the developmental trajectory of learning and exploration.

Keywords: exploration-exploitation dilemma, directed exploration, random exploration, generalization, search, multi-armed bandit task

Children are natural born explorers. While exploration and active learning are quintessential features of development and maturation, they also pose fundamental challenges to children and adults alike. In particular, efficiently searching for information and rewards requires balancing the dual goals of exploring unknown options to learn something new, and exploiting familiar options to obtain known rewards. At a restaurant, should you go with your usual favorite or should you try the chef's latest creation? As a child, should you play your favorite game again or try out something new? Exploring novel options can potentially reveal new and even better rewards, but could also lead to disappointment. Known as the *explore-exploit dilemma*, this fundamental problem contrasts the goals of gaining knowledge to reduce uncertainty with immediately acquiring rewards.

Optimal solutions to explore-exploit dilemmas are unattainable in all but limiting cases (Bellman, 1952; Gittins & Jones, 1979), making heuristic strategies an active area of research in many fields, including cognitive and developmental psychology. Whereas many studies have investigated how adults balance exploration and exploitation (for reviews, see Cohen, McClure, & Angela, 2007; Hills et al., 2015; Mehlhorn et al., 2015), less is known about the developmental processes that shape learning and exploration during childhood. Studying how children, who have fewer cognitive resources and less experience, approach such problems can provide critical insights into the computational and behavioral principles that drive learning and develop-

All data and code for reproducing the analyses is available at `https://osf.io/eq2bk/?view_only=fed4735fa15a4f3d8dd56db385b845b1`. We thank all families who participated in this research, Calvin Paulus and Jeanette Blümel for collecting the data, and Federico Meini for help with programming the experiment. Correspondence should be addressed to Björn Meder, Department of Psychology, University of Erfurt, Nordhäuser Str. 63, 99089 Erfurt, Germany. Email: `bjoern.meder@uni-erfurt.de` or `meder@mpib-berlin.mpg.de`

ment more generally. Here, we investigate developmental trajectories in learning and exploration between the ages of 4 and 9, an age range where substantial changes in children's exploration behavior have been observed across different tasks (Betsch, Lehmann, Lindow, Lang, & Schoemann, 2016; Ronfard, Zambrana, Hermansen, & Kelemen, 2018; Ruggeri, Markant, Gureckis, Bretzke, & Xu, 2019; Ruggeri, Xu, & Lombrozo, 2019). To map out developmental trajectories we combine behavioral data from a spatial search task with predictions from a computational model that disentangles different forms of exploration. Consistent with previous theories (Gopnik et al., 2017), our results show that the exploration patterns of young children are characterized by high levels of random sampling, which decreases with age. However, even at the youngest age range, children do not rely solely on random exploration, but they actively seek out options with high uncertainty (directed exploration) and use inductive inferences to predict unobserved rewards (generalization).

How to explore: Random exploration, directed exploration, and generalization

Research on explore-exploit problems typically contrasts two distinct classes of exploration strategies (Gershman, 2018; Wilson, Geana, White, Ludvig, & Cohen, 2014). *Random exploration* models exploration by adding noise to the decision process (Luce, 1959; Thompson, 1933). Instead of only making reward-maximizing decisions, this added randomness can lead to the incidental exploration of new options and (better or worse) rewards. Related to this strategy, it has been recently suggested that children's exploration behavior is characterized by "higher temperature" (i.e., noisier) sampling, which "cools off" with age (Gopnik et al., 2017). The metaphor of temperature appeals to methods such as simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983), which is an optimisation algorithm that uses a time-dependent reduction of randomness to avoid getting stuck in a local optimum. On this view, young children exhibit high amounts of random sampling, which results in exploration of a larger set of possibilities compared to adults (Cauffman et al., 2010; Mata, Wilke, & Czienskowski, 2013). As children grow older, temperature decreases, yielding a stronger focus on reward maximization, leading to less diverse sampling behavior (Bonawitz, Denison, Griffiths, & Gopnik, 2014).

Directed exploration (E. Schulz & Gershman, 2019; Wilson et al., 2014) is an alternative strategy, which relies on representing one's uncertainty about the world and then assigning an intrinsic value towards actively reducing this uncertainty (Gottlieb & Oudeyer, 2018). Instead of adding more variability through random (noisy) sampling, directed exploration actively seeks out uncertainty. According to this view, obtaining information is rewarding in and of itself, and the value of an option is inflated through an "uncertainty bonus" (Auer, 2002). By valuing uncertainty positively, directed exploration encourages sampling options with promising but uncertain rewards, rather than focusing merely on exploiting known high-reward options. Computationally, directed exploration is more demanding, since it requires a richer representational structure that encodes both expected rewards and the underlying uncertainty. However, already infants have been shown to value the exploration of uncertain options positively (L. E. Schulz, 2015), 6- and 7-year-olds can integrate prior beliefs and obtained evidence in simple learning and exploration tasks (Bonawitz, van Schijndel, Friel, & Schulz, 2012), and children age 7 to 11 have been shown to rely more on directed exploration than adults when searching for rewards (E. Schulz, Wu, Ruggeri, & Meder, 2019).

In addition to random and directed exploration, the ability to *generalize* (Shepard, 1987) is another important cognitive capacity for navigating the exploration-exploitation dilemma. In particular, generalization provides traction for exploring large problem spaces by making predictions about novel options. For instance, when Italian immigrants came to the US around 1900, they brought with them knowledge and love of the classic Neapolitan pizza. In their search for creating similarly rewarding dishes, they explored a variety of novel, but similar options – giving the world Chicago-, New York-, and California-style pizza, as well as several other new variations. A child encountering a new toy can predict whether or not it will be fun by comparing it to other toys it has encountered. If it appears similar to other fun toys, there is a good chance this new toy is also fun. Thus, generalization provides critical guidance for *which* options to explore – namely those which are similar to known high-reward options. On this view, developmental differences in exploration are tightly connected to the ability to make inductive inferences about unexplored options based on prior experience. As cognitive functions and memory develop, they enable more complex cognitive processes and representations (Blanco et al., 2016), thereby supporting more effective generalization for guiding exploration. For instance, changes in search behavior over the life span may be due to the accumulation of knowledge, with adults having stronger inductive biases than children, who seem to weigh new evidence more strongly (Gopnik, Griffiths, & Lucas, 2015).

Goals and scope

While random and directed exploration are conceptually different, they are not mutually exclusive. Research shows that both types of exploration strategies contribute to search and decision making in adolescent and adult participants (Gershman, 2018; Somerville et al., 2017; Wilson et al., 2014), with dissociable neural signatures underlying the two forms of exploration (Zajkowski, Kossut, & Wilson, 2017). In addition, both children and adults rely on generalization to learn about the environment and make inferences from ex-

perienced to not-yet-explored options (E. Schulz, Wu, Huys, Krause, & Speekenbrink, 2018; E. Schulz et al., 2019; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018).

The goal of the present paper is to investigate how young children, age 4 to 9 years, balance random and directed exploration, using a spatial search task with correlated rewards. In particular, we trace age-related differences in learning and exploration using a computational model that combines similarity-based generalization with both directed and random exploration (Wu et al., 2018). Our data enable a direct test of the "cooling off" hypothesis and offers empirical evidence for the trajectory with which random sampling decreases over the course of childhood development.

Previous studies have shown reliable signatures of generalization and directed exploration in adults, with relatively little random exploration (Wu, Schulz, & Gershman, 2020; Wu et al., 2018). In a comparison of children age 7 to 11 and adults, Schulz and colleagues (2019) found no age-related differences in random exploration. Rather, children differed from adults by having higher levels of directed exploration and narrower generalization. While the lack of differences in random exploration does not support the idea of a "cooling off" process over the lifespan, it could also be the case that children age 7 to 11 had already transitioned to a lower temperature and had already developed the capacity for directed exploration. Therefore, our goal is to investigate a younger age range to search for the developmental stage where random exploration diminishes and directed exploration emerges.

Experiment

We used a simplified version of the spatially-correlated multi-armed bandit paradigm (Wu et al., 2018) to investigate how children learn and search for rewards on a grid world by clicking on different tiles (Fig. 1). Each tile had a different reward distribution, where the goal was to accumulate as many rewards as possible within a limited search horizon (i.e., a fixed number of clicks). Rather than displaying rewards numerically, as in previous experiments (E. Schulz et al., 2019), here the value of rewards was indicated using different shades of red to be interpretable by children as young as 4 (Fig. 1). In this task, rewards were spatially correlated, such that nearby options had a similar mean reward. Thus, participants could use generalization from a sparse number of observations to guide their exploration towards promising regions of the search space. Importantly, the number of available clicks (25) was much smaller than the number of available options (64), requiring searchers to balance clicking novel tiles to discover new rewarding options (exploration) with re-clicking tiles already known to provide high rewards (exploitation).

Methods

Participants. We recruited 54 children in the age range 4 to 7 (M = 72.6 months SD = 7.6, range 51 − 82 months, 24 female), henceforth referred to as 6-year-olds, and 48 children in the age range 8 and 9 (M = 93.1 months, SD = 6.5, range 84 − 108 months, 23 female), henceforth referred to as 8-year-olds. In addition to comparing these age groups, we also conducted analyses that treat age as a continuous variable. Fourteen additional children were excluded from analysis because they failed the instruction check (n = 9), did not want to play anymore (n = 1), were not native speakers (n = 2), or because their parents intervened during the experiment (n = 2). Informed consent was obtained from children's legal guardians prior to participation; average duration was about 12 minutes.

Materials, design, and procedure. Children played six rounds of a spatial search game on a tablet, in which they were presented with an 8×8 grid world with spatially correlated rewards (Fig. 1). The expected reward across all environments was identical (i.e., average reward over all tiles of a grid); what differed between environments was the spatial correlation among rewards. The strength of the spatial correlations was manipulated between subjects, with *smooth environments* having stronger spatial correlations than *rough environments*. On each round, a new environment was sampled without replacement from a set of 40 environments generated for each class from a radial basis function kernel (see below), with $\lambda_{smooth} = 4$ and $\lambda_{rough} = 1$. The sampled environments defined a bivariate reward function on the grid, with each reward including additional normally distributed noise, such that there were slight variations in reward when repeatedly clicking a tile.

At the beginning of each round, one random tile was revealed and children could sequentially sample 25 tiles. On each trial, they could either click a new tile or re-click a tile they had already selected before (clicking was done by touching the desired tile on the tablet). Clicking a tile for the first time revealed its color, with darker colors indicating higher rewards along a continuous, linearly scaled color range (Fig. 1). In each round, the underlying rewards were scaled to a randomly drawn maximum value in the range of 70% to 90% of the darkest reward value. Re-clicked tiles could show small variations in the observed color (i.e., underlying reward) due to normally distributed noise, $\epsilon \sim \mathcal{N}(0, 1)$.

Children were awarded up to five stars at the end of each round (e.g., 4.6 out of 5; see Fig. 1b), based on the ratio of their average reward to the global maximum of the given grid. At the beginning of a round, the stars were empty, then they continuously filled up in accordance with each obtained reward. The instructed goal was to collect as many stars as possible in each round; at the end of the game, children received a number of stickers proportional to the average number of stars earned in each round.

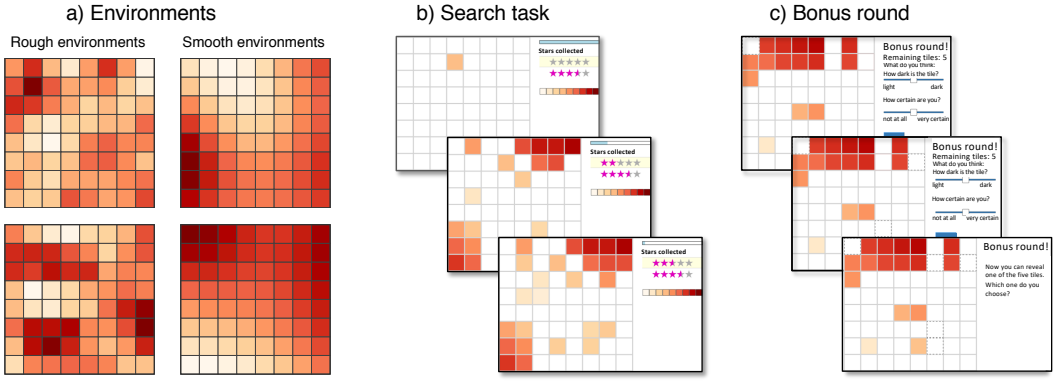


Figure 1. Example environments and screenshots from experiment. *a)* Two rough environments with low spatial correlation and two smooth environments with high spatial correlation. Darker shades of red correspond to higher rewards. *b)* Spatial search game, in which children had 25 clicks in each round to obtain as many stars as possible by finding darker (i.e., more rewarding) tiles. *c)* Bonus round judgments, in which children predicted the rewards for five previously unobserved tiles (tile with dashed border) and made a confidence judgment about their prediction.

The first round was a tutorial round, in which children were familiarized with the goal of the game, the spatial correlation of rewards, the maximum number of clicks allowed per round, and the possibility of re-clicking tiles (Appendix C). After the tutorial round, children were required to answer three comprehension questions. If they failed to answer any of the questions correctly, the relevant part of the instructions was repeated and the questions were asked again. If they failed again, they continued with the experiment, but were later excluded from the analyses.

The sixth and last round was a bonus round, in which children sampled for 15 trials and then made reward predictions for five randomly chosen and previously unobserved tiles (Figure 1c). This was explained to them before the bonus round started. Judgments were made using a continuous slider, asking children to indicate the darkness of the target tile, with the end points labeled as "light" and "dark". When moving the slider, the target tile changed its color accordingly. The underlying reward scale was continuous, ranging from 0 to 50. To assess the level of confidence associated with the reward predictions, children were asked how certain they were about the predicted darkness, using a slider from 0 to 10 in steps of 1, with the endpoints labeled as "not certain at all" to "very certain". After judging five tiles children were asked to select one of them. They received the corresponding reward and then continued the round until the search horizon was exhausted.

Behavioral results

We first analyze the behavioral data in terms of performance and exploration behavior. These analyses exclude the tutorial and bonus rounds, leaving a total of 100 search decisions (4 rounds $\times$ 25 trials) for each of the 102 participants. We then report the results of the bonus round, where we analyze children's reward predictions and confidence judgments. The behavioral data are complemented by model-based analyses, where we disentangle generalization, directed exploration, and random exploration. We report both frequentist statistics and Bayes factors (BF) to quantify the relative evidence of the data in favor of the alternative hypothesis (H_A) over the null hypothesis (H_0) (see Appendix A for details).

Performance: Obtained rewards

Whereas both smooth and rough environments had the same expected rewards, the stronger spatial correlations in the smooth environment facilitated better performance for both age groups (6-year-olds: $t(52) = 3.3$, $p = .002$, $d = 0.9$, $BF = 22$; 8-year-olds: $t(46) = 6.4$, $p < .001$, $d = 1.8$, $BF > 100$; Fig. 2a). Thus, regardless of age, children were able to leverage the spatial correlation of rewards in the environment, and performed better in more correlated environments.

Eight-year-old children obtained higher rewards than 6-year-olds in both rough ($t(48) = 2.6$, $p = .012$, $d = 0.7$, $BF = 4.1$) and smooth environments ($t(50) = 3.3$, $p = .002$, $d = 0.9$, $BF = 19$). Age-related performance differences were also found when treating age as continuous variable (Fig. 2b), with performance increasing with age in both rough (Pearson $r = .36$, 95% CI = [.09, .58], $p = .011$, $BF = 6.0$) and smooth environments ($r = .39$, 95% CI = [.14, .60], $p = .004$, $BF = 14$).

Figure 2c shows the learning curves (average reward over trials; first aggregated within and then across participants). Consistent with the overall performance, learning curves increased more strongly in smooth compared to rough environments. In rough environments, 8-year-olds performed

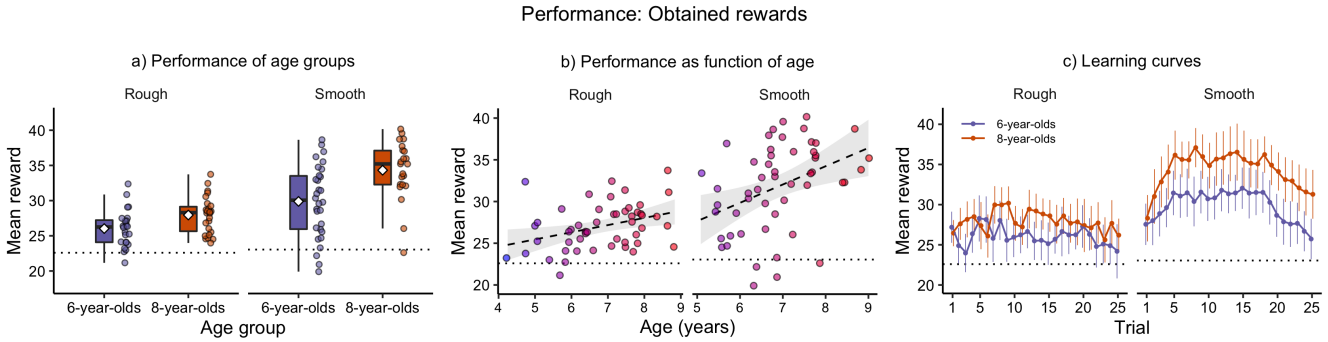


Figure 2. Obtained rewards. *a)* Tukey box plots of the distribution of obtained mean rewards, separately for each age group and environment. Each dot is a participant-wise mean, the horizontal line in the box shows the group median and the diamonds indicate group means. Dotted line is random performance. *b)* Average obtained rewards as a function of age in smooth and rough environments. Each dot represents one participant, the dashed line shows a linear regression (± 95% CI); dotted line is random performance. *c)* Learning curves showing the average rewards over trials, first averaged within participants and then aggregated across participants; error bars are 95% CIs.

slightly better than 6-year-olds, but generally there was only little improvement over trials. In smooth environments, older children learned more quickly than younger children and consistently outperformed them. A notable finding is that in smooth environments, towards the end of the search, the average obtained rewards tended to decrease again, in both age groups, suggesting a tendency to continue exploration even at the cost of foregone rewards.

Search trajectories

Rather than only comparing performance, we also looked for behavioral patterns in how children searched for rewards, by analyzing the distance between consecutive choices and how this was affected by the magnitude of rewards and the subsequent search decisions. Figure 3a shows the distribution of Manhattan distances between consecutive choices. For 8-year-olds, the mean distance was smaller in smooth than in rough environments ($t(46) = -3.1$, $p = .003$, $d = 0.9$, $BF = 13$), indicating they searched more locally in the presence of strong spatial correlations. For 6-year-olds, there was no difference between environments ($t(52) = 1.0$, $p = .31$, $d = 0.3$, $BF = .42$), suggesting a more limited capability to adapt to environmental structure.

We also analyzed search decisions (Fig. 3b) by computing the proportions of *repeat choices*, corresponding to re-clicking the previously revealed tile, *near choices*, corresponding to searching a neighboring tile (i.e., distance of 1), and *far choices*, corresponding to clicking tiles with a distance larger than 1. Older children tended to search more locally in smooth compared to rough environment, while conversely making more far choices in rough compared to smooth environments. This pattern was not observed for 6-year-olds, indicating that younger children did not adapt their search patterns to the correlation structure of rewards in the environment. Notably, the number of repeat clicks is overall rather low, regardless of age group and environment (see General Discussion). This may also explain the learning curves (Figure 2c), which tended to decrease towards the end of each round in smooth environments. This demonstrates that children generally show higher levels of exploration when searching for rewards, and thus less exploitation of high-value options that have already been observed.

Finally, we analysed the relation between the value of a reward obtained at time t and the search distance on the subsequent trial $t + 1$. If a large reward was obtained, searchers should search more locally, while conversely, if a low reward was obtained, searchers should be more likely to search farther away. Using hierarchical Bayesian regression analyses, we predicted search distance using the reward obtained on the previous step, age group, and their interactions as population-level ("fixed") effects, while treating participants as random intercepts. Figure 3c shows how the reward obtained from the previous choice related to subsequent search distance (see Table A1 in Appendix B for detailed results). Both 6- and 8-year-olds tended to search more locally when high rewards were obtained and searched further away when low rewards were obtained. The two age groups were differentially influenced by obtained rewards, such that the search distance of 8-year-olds markedly decreased with the magnitude of reward, in both smooth and rough environments. In comparison, 6-year-olds also tended to decrease search distance with higher rewards, but with a flatter slope. Taken together, these findings indicate that the magnitude of rewards influenced search distance, but 8-year-olds were more responsive in adapting their search behavior than 6-year-olds.

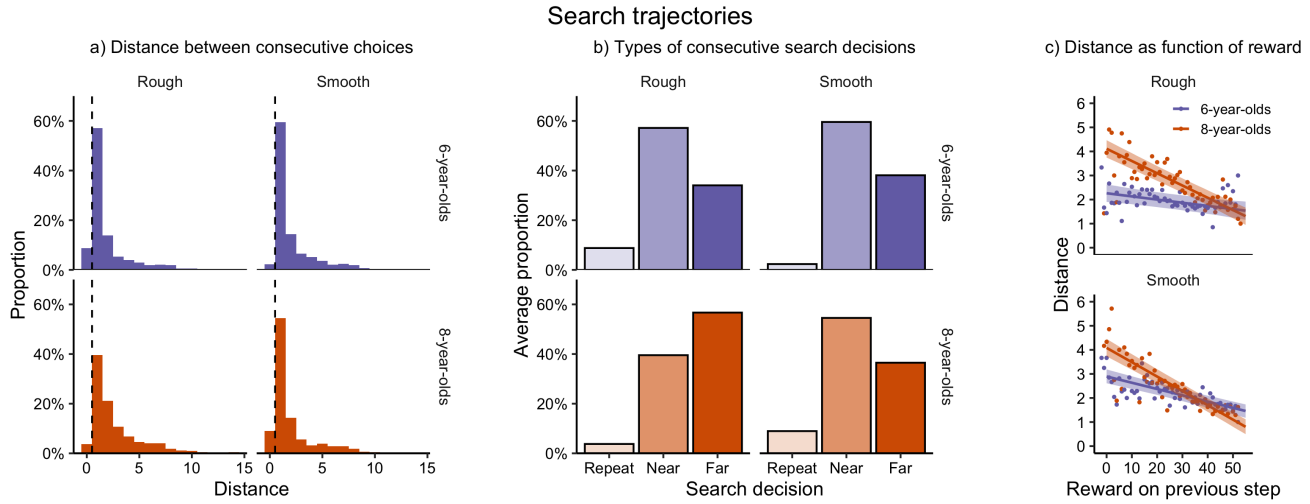


Figure 3. Search trajectories. *a)* Histogram of distances between consecutive search choices. A distance of zero indicates a repeat click; a distance of 1 corresponds to clicks on neighboring tiles; distances > 1 correspond to other clicks on the grid. The vertical dashed line marks the difference between a repeat click and selecting any other tile. *b)* Average proportion of search decisions by age group and environment. Repeat clicks correspond to re-clicking a previously revealed tile, near clicks correspond to directly neighboring tiles, and far clicks are sampling decisions with a distance > 1. *c)* Search distance as function of reward obtained on the previous trial. The lines visualize the relation between search distance and previous reward for each age group and environment, obtained from a Bayesian regression (± 95% CI). The dots show the observed mean distances given previous rewards, aggregated across all decisions and children. One outlier has been removed from the lower plot, but is included in all statistical analyses.

Bonus round judgments

The last round was a bonus round in which children made 15 search decisions and then predicted the expected rewards for five random, unrevealed tiles. Additionally, they were also asked how confident they were about the predicted reward (i.e., darkness of tile).

Figure 4a shows the mean absolute error between children's estimates and the true underlying expected reward. Overall, 8-year-olds had lower prediction error than 6-year-olds ($t(100) = 3.9$, $p < .001$, $d = 0.8$, $BF > 100$). The difference between age groups was found in both environments, albeit less pronounced in the rough ($t(48) = 2.4$, $p = .019$, $d = 0.7$, $BF = 2.9$) compared to the smooth environment ($t(50) = 3.0$, $p = .004$, $d = 0.8$, $BF = 9.1$). Aggregating both age groups, we found no effect of environment on prediction error ($t(100) = -1.0$, $p = .32$, $d = 0.2$, $BF = .32$). Compared to a random baseline, 6-year-olds performed worse than chance level ($t(53) = 2.7$, $p = .009$, $d = 0.4$, $BF = 4.2$) whereas 8-year-olds were better than chance ($t(47) = -3.1$, $p = .003$, $d = 0.4$, $BF = 9.6$). Looking at prediction error as a function of age in months (Fig. 4), we found that in both rough and smooth environments children's prediction error declined with age (rough: $r = -.40$, $p = .004$, $BF = 14$, smooth: $r = -.46$, $p < .001$, $BF = 57$).

Across all judgments and children, we found no systematic relation between confidence and prediction error (Kendall rank correlation: $r_\tau = .07$, $p = .04$, $BF = .67$). A Bayesian regression with confidence, age group, and their interaction as predictors and subject-wise random intercept also showed no reliable relationship (see Table A2 in Appendix B).

In summary, 8-year-olds obtained higher rewards than 6-year-olds, with both groups performing better in smooth compared to rough environments, facilitated by stronger spatial correlations. Participants adapted their search patterns in response to the magnitude of obtained rewards, searching locally upon finding rich rewards, and searching farther away upon finding poor rewards. The responsiveness of this adaptive search pattern was mediated by age, where 8-year-olds exhibited a stronger relationship between reward value and search distance than 6-year-olds. Lastly, prediction accuracy increased reliably with age, but there was no relation between children's subjective confidence in their reward judgments and their prediction error.

A computational analysis of directed and random exploration in children

The behavioral data presented above show strong and systematic differences between the exploration behavior of 6- and 8-year-old children. We next present a computational model that captures key aspects of generalization and sampling strategies in order to map the developmental trajectory of learning and exploration. In particular, the model provides a clear computational framework for estimating to what ex-

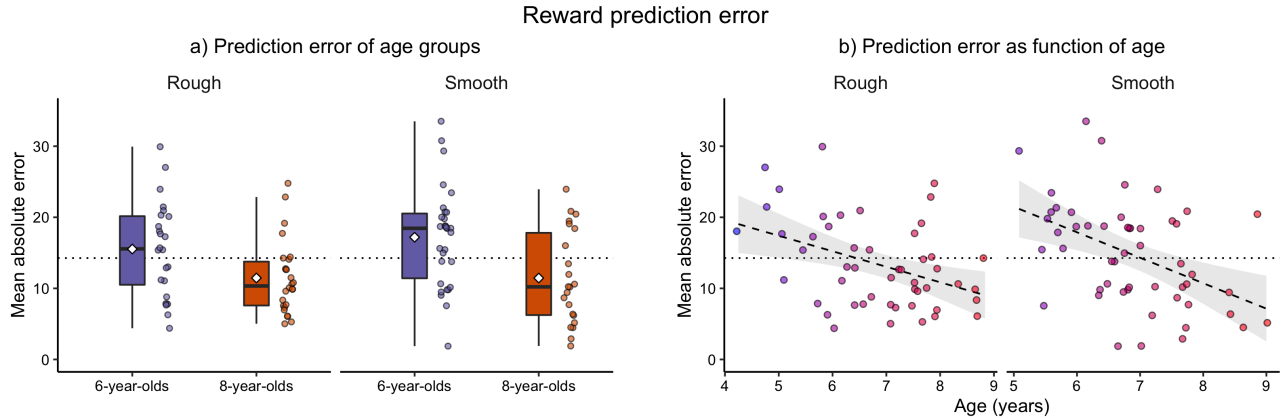


Figure 4. Bonus round judgments. *a)* Mean absolute prediction error for 6- and 8-year-olds. *b)* Mean absolute prediction error as function of age. Each dot is one participant, the dashed line shows a linear regression (± 95% CI). Dotted line is random performance.

tent children generalize about the spatial correlation of rewards, and how their sampling behavior can be decomposed into directed and undirected exploration.

The Gaussian Process Upper Confidence Bound (GP-UCB) model

Our model consists of three building blocks: a *learning model* that makes predictions about the distribution of rewards in the environment, a *sampling strategy*, which maps these predictions onto valuation of options, and a *choice rule*, which converts value into choice probabilities. We now briefly describe these components, with further details provided in Supplement A.

Learning model. To model learning about rewards in the environment we use *Gaussian Process* (GP) regression as a form of Bayesian function learning (Rasmussen & Williams, 2006). The GP uses the principles of Bayesian inference to adaptively learn a value function, mapping the location of each option onto rewards. Generalization about novel options is thus accomplished through interpolation or extrapolation from previous observations (rewards and their locations). This approach has been shown to account for how adults explicitly learn functions (Lucas, Griffiths, Williams, & Kalish, 2015), and has been successfully applied to model the behavior of children and adults in a wide range of learning and search tasks (E. Schulz, Konstantinidis, & Speekenbrink, 2017; E. Schulz et al., 2019; Wu, Schulz, Garvert, Meder, & Schuck, 2020; Wu, Schulz, & Gershman, 2020; Wu et al., 2018).

Formally, a GP defines a distribution over functions $f \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, where each function can be interpreted as a candidate hypothesis about the relationship between spatial location and expected rewards. The GP prior is determined by a mean function $m(\mathbf{x})$ and a kernel function $k(\mathbf{x}, \mathbf{x}')$. We follow the convention of setting the mean function to zero, while using the kernel function to encode the covariance structure. Put simply, the kernel provides an inductive bias about how points in the input space are related to each other as a function of distance (i.e., spatial similarity). A common choice for the kernel is the *radial basis function* (RBF):

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

where $\mathbf{x}$ and $\mathbf{x}'$ denote two inputs (e.g., coordinates of tiles on the grid) and λ is the *length-scale* parameter governing the extent of generalization. Put simply, the RBF kernel models generalization as an exponentially decaying function of the distance between inputs $\mathbf{x}$ and $\mathbf{x}'$. This kernel is closely related to Shepard's (1987) universal law of generalization, which models generalization as an exponentially decaying function of similarity, where similarity is the inverse of distance.

In the present task, GP regression generates normally distributed beliefs about the rewards for any tile $\mathbf{x}$, summarized as expectation $\mu(\mathbf{x})$ and uncertainty $\sigma(\mathbf{x})$. These predictions are modulated by the length-scale parameter λ, which defines the extent to which rewards are assumed to be correlated as a function of distance. For instance, $\lambda = 1$ corresponds to the assumption that the rewards of two neighboring tiles are correlated by $r = 0.6$, and that due to the exponential decay this correlation effectively decreases to zero for options further than three tiles apart. We treat λ as a free parameter, which we estimate for each individual participant. This enables us to assess each child's tendency to generalize.

Sampling strategies. Given a learner's belief about expected reward $\mu(\mathbf{x})$ and estimated uncertainty $\sigma(\mathbf{x})$, we use a sampling strategy to map these beliefs onto a valuation for each option. Specifically, we use *Upper Confidence Bound* (UCB) sampling (Auer, 2002) to model directed exploration as a simple weighted sum:

$$\text{UCB}(\mathbf{x}) = \mu(\mathbf{x}) + \beta\sigma(\mathbf{x}) \quad (2)$$

where μ is the mean expected reward and β represents the extent to which uncertainty σ (measured in terms of the standard deviation of $\mathbf{x}$) is valued positively. The parameter β is an "uncertainty bonus", since it optimistically inflates expected rewards by their degree of uncertainty. UCB provides an effective sampling strategy for balancing the exploration-exploitation dilemma, by mediating between exploring novel options to reduce uncertainty while also prioritizing the exploitation of high-value options.

To illustrate this sampling strategy, consider two options (tiles) $\mathbf{x}_1$ and $\mathbf{x}_2$. Option $\mathbf{x}_1$ has expected reward of $\mu(\mathbf{x}_1) = 50$ and uncertainty $\sigma(\mathbf{x}_1) = 5$. Option $\mathbf{x}_2$ has expected reward of $\mu(\mathbf{x}_2) = 45$ and uncertainty $\sigma(\mathbf{x}_2) = 15$. Thus, option $\mathbf{x}_1$ has higher expected reward than $\mathbf{x}_2$, but $\mathbf{x}_2$ is more uncertain. UCB sampling takes into account both reward and uncertainty to balance the explore-exploit trade-off. For instance, if $\beta = 1$, $UCB(\mathbf{x}_1|\beta = 1) = 50 + 5 = 55$ and $UCB(\mathbf{x}_2|\beta = 1) = 45 + 15 = 60$, meaning that option $\mathbf{x}_2$ is more attractive than option $\mathbf{x}_1$. By contrast, if $\beta = 0.2$, then $UCB(\mathbf{x}_1|\beta = 0.2) = 50 + 1 = 51$ and $UCB(\mathbf{x}_2|\beta = 0.2) = 45 + 3 = 48$. In this case, option $\mathbf{x}_1$ is valued higher than $\mathbf{x}_2$, making it more likely to click this tile. Thus, the higher β, the stronger a searcher values uncertainty positively, nudging them towards sampling uncertain options. Conversely, when $\beta \to 0$ the value of an option is dominated by its expected reward, regardless of the attached uncertainty. In our model, we estimate β for each learner based on their individual search behavior, to assess their level of uncertainty-directed exploration.

Choice rule. The final component of the model is the choice rule, which translates UCB values into choice probabilities with a softmax function:

$$p(\mathbf{x}) = \frac{\exp(\text{UCB}(\mathbf{x})/\tau)}{\sum_{j=1}^{N} \exp(\text{UCB}(\mathbf{x}_j)/\tau)}. \quad (3)$$

Importantly, the softmax choice contains a temperature parameter τ that governs the amount of randomness in the choice probabilities. This enables us to quantify the amount of undirected (random) sampling for each learner. Higher temperature sampling corresponds to noisier predictions, where as $\tau \to \infty$, all options have an equal probability of being chosen. Conversely, lower temperatures produce choice probabilities that are more concentrated on high-value options, where as $\tau \to 0$, it becomes an argmax choice rule (i.e., always choosing the option with the highest value). In our model, τ is estimated from the data, to assess the amount of random sampling for each child.

Model summary. In sum, the GP-UCB model combines i) a learning component that generalizes from limited observations to unobserved options, ii) a UCB sampling strategy that inflates expectations of reward by the associated uncertainties to perform directed exploration, and iii) a softmax choice rule that converts UCB values into choice probabilities and adds decision noise as a form of random exploration. Each model component has a single free parameter that we estimate through cross-validation from children's search decisions: the length-scale parameter λ indicates the extent of generalization, the uncertainty bonus β defines the level of directed exploration, and the temperature parameter τ captures the amount of random exploration Careful analyses of these parameters provides a window into the computational principles of learning and exploration, enabling us to identify age-related changes.

Model comparison

We contrast the predictive accuracy of the GP-UCB model with a Bayesian reinforcement learning model (*Mean Tracker*; MT). This model uses the same UCB and softmax components, but differs in that it does not generalize. Instead, it learns independent reward distributions about each option using the principles of associative learning (see Supplement A and B for details and extended model results including additional sampling strategies).

We used cross validation to assess how well the models predict each searcher's sampling decisions, where—as before—we omit the tutorial round and bonus round. Specifically, we iteratively split each child's data into a training set consisting of three of the four rounds, and holding out the remaining round as a test set. We computed the maximum-likelihood estimates for each model's parameters (range $[exp(-5), exp(4)]$) using differential evolution (Mullen, Ardia, Gil, Windover, & Cline, 2011) and then evaluated each model's predictive accuracy on the held-out test set. This procedure was repeated for each participant for all rounds.

We can describe the objective performance of our models using *predictive accuracy* as a pseudo-R^2, comparing the summed out-of-sample log loss for each model k against a random model (i.e., choosing all options with with equal probability):

$$R^2 = 1 - \frac{\log \mathcal{L}(M_k)}{\log \mathcal{L}(M_{rand})}, \quad (4)$$

where $\log \mathcal{L}$ represents log loss. Intuitively, $R^2 = 0$ indicates chance-level predictions and $R^2 = 1$ indicates theoretically perfect predictions.

Figure 5a shows the predictive accuracy of the two models for both age groups. The GP-UCB model had higher predictive accuracy than the MT-UCB model overall ($t(101) = 6.6$, $p < .001$, $d = 0.7$, $BF > 100$), and also for each age group

(6-year-olds: $t(53) = 3.4$, $p = .001$, $d = 0.5$, $BF = 22$; 8-year-olds: $t(47) = 6.1$, $p < .001$, $d = 1.0$, $BF > 100$). In total, 73 out of 102 participants were best described by the GP-UCB model: 34 out of 54 6-year-olds (63%) and 39 out of 48 8-year-olds (81%). These results demonstrate the importance of generalization, since this component was not present in the MT learning model.

Developmental differences in parameter estimates

To map the developmental trajectories of learning and search, we analyzed the parameter estimates of the GP-UCB model (Fig. 5b). There was no difference in the level of generalization (λ parameter) between 6- and 8-year-olds (Mann-Whitney-U test: $U = 1093$, $p = .18$, $r_\tau = -.11$, $BF = .42$). However, younger children had higher estimates for both the exploration bonus β ($U = 1602$, $p = .041$, $r_\tau = .17$, $BF = 1.6$) and temperature τ ($U = 1688$, $p = .009$, $r_\tau = .21$, $BF = 2.2$), with a stronger age-related decrease for the latter. These results indicate that 6-year-olds exhibited a stronger tendency towards both directed and random exploration than 8-year-olds.

Figures 5c to f provide a more detailed analysis of these findings by treating age as a continuous variable. First, Figure 5a shows that the predictive accuracy of the GP-UCB model increased with age (Kendall's $r_\tau = .27$, $p < .001$, $BF > 100$). Second, consistent with the group-based analyses, there were little changes in the generalization parameter λ as a function of age ($r_\tau = .10$, $p = .14$, $BF = .39$). In contrast, both the uncertainty bonus parameter β and in particular the temperature parameter τ of the softmax function decreased with age. Younger children tended to have higher values of β ($r_\tau = -.14$, $p = .043$, $BF = 1.0$), indicating a somewhat larger value placed on reducing uncertainty, and thus more directed exploration. Whereas the age-related change in directed exploration were rather weak, there was a marked decrease in the temperature parameter τ ($r_\tau = -.23$, $p < .001$, $BF = 46$). Thus, the amount of random sampling strongly decreased with age. These same changes in parameters as a function of age also hold when controlling for the predictive accuracy of the GP-UCB model (see Fig. A2 and Table A3 in Appendix B), although these analyses find a slightly stronger increase in λ as a function of age, indicating broader generalizations as children grow older.

Taken together, these analyses provide a window into the developmental trajectories of exploration behavior, showing how both directed and, in particular, random exploration decrease as children get older.

General Discussion

We investigated how 6- and 8-year-old children search for rewards in a spatial version of the explore-exploit dilemma, focusing on disentangling how generalization, random exploration, and directed exploration contribute to age-related changes. Although general performance increased with age, we found that even younger children could successfully generalize the observed spatial correlations and use this knowledge to guide their search for rewards. Children adapted their exploration behavior depending on the rewards they obtained, with 8-year-olds showing a stronger relationship between obtained rewards and search distance. Finally, while prediction accuracy in the bonus round increased with age, there was no relation between children's confidence and their prediction error.

The model-based analyses showed that the GP-UCB model provided a better account of children's behavior than the MT-UCB model, highlighting the importance of similarity-based generalization. A key finding is a strong age-related decrease of random exploration, represented by the τ parameter of the softmax choice rule, consistent with the hypothesis that children's temperature "cools off" as they get older (Gopnik et al., 2017). However, children's exploration behavior was not solely driven by random exploration, but also by a high amount of uncertainty-directed sampling, as indicated by high levels of the uncertainty-bonus parameter β. The valuation of uncertainty also tended to decrease with age, but this trend was much weaker compared to the tapering off of random exploration.

Our findings extend the developmental investigation of children's exploration behavior, complementing previous research with older children (E. Schulz et al., 2019), as well as adolescent and adult participants, who also show signatures of both types of exploration strategies (Wilson et al., 2014; Wu et al., 2018). Table 1 provides an overview of children and adults' model parameters across different studies using similar versions of the multi-armed spatially-correlated bandit paradigm. The comparison shows that children up to around age 11 show higher levels of directed exploration than adult subjects, whereas adults tend to generalize more strongly. High levels of random exploration were only observed in 6-year-olds, indicating that this form of exploration diminishes earlier in development than uncertainty-guided exploration. Future studies should systematically investigate an even broader age range (e.g., from childhood through adolescence to adulthood, ideally in a longitudinal design) to identify changes in exploration and generalization over the lifespan.

Children are keen explorers – but are they good exploiters? One peculiar finding we obtained was the low number of exploitation decisions (i.e., repeat clicks; Figure 3b). Across all children and rounds (excluding tutorial and bonus round), the proportion of repeat clicks was about 7% (6-year-olds: 6.8%, 8-year-olds: 7.5%). While this proportion was comparable to participants in a similar age range as reported in other studies (e.g., E. Schulz et al., 2019, reported 5.6% repeat clicks for 7-8-year-olds and 6.4% for 9-11-year-olds), this contrasts with the behavior of adults, who typically

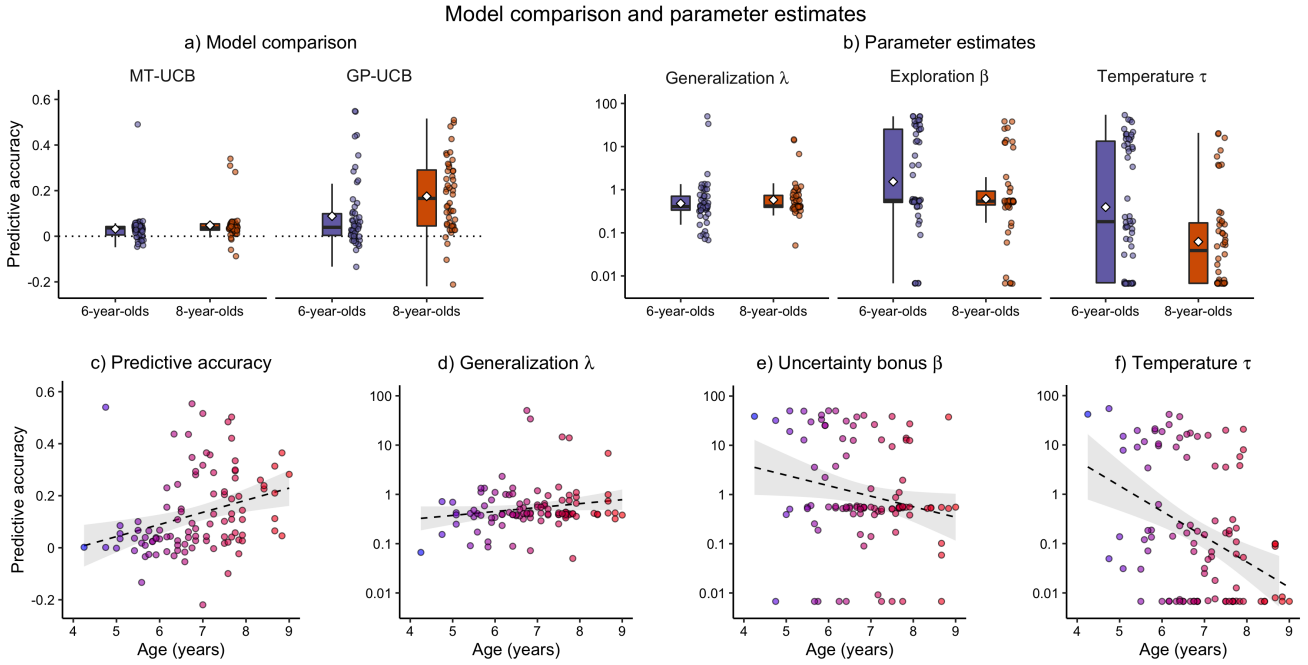


Figure 5. Model comparison and parameter estimates of the GP-UCB model. *a)* Predictive accuracy (pseudo-R^2) of mean tracker (MT) and Gaussian process (GP) learning model combined with upper-confidence-bound (UCB) sampling. Each dot represents one participant with the mean out-of-sample accuracy across rounds (excluding practice and bonus round). Box shows IQR, the line is the median and the diamond is the mean. *b)* Individual parameter estimates of the GP-UCB model by age group. *c)* Predictive accuracy of the GP-UCB model as function of age. *d-f)* Parameter estimates of the GP-UCB model as function of age. Each dot represents one child with their cross-validated median parameter estimates. Dashed line indicates a linear regression (±95% CI).

show a higher proportions of repeat clicks; 12% in Wu et al. (2018, averaged across three experiments) and 32.1% in the study by Schulz and colleagues (2019). Lower exploitation rates for children have also been observed in simpler bandit tasks with fewer options and independent reward distributions (Blanco & Sloutsky, 2019).

The tendency to over-explore might be responsible for the decrease of children's average rewards towards the end of the search horizon (Figure 2c). Indeed, given a fixed search horizon, it is typically better at some point to start exploiting the found high-reward options, rather than keeping on searching for even better options. It is likely that this behavior was driven by the high amount of both random and directed exploration, as captured by a high temperature parameter τ, leading to increased random sampling, and a high exploration bonus β, leading children to optimistically inflate expected rewards of unobserved tiles. While this tendency to over-explore impaired performance in our task, it may nevertheless be adaptive in some settings (Sumner et al., 2019), by allowing children to discover changes that are not obvious and are overlooked by adults (Gopnik et al., 2015; Lucas, Bridgers, Griffiths, & Gopnik, 2014). It could be especially adaptive in dynamic environments where reward structures change over time (Behrens, Woolrich, Walton, & Rushworth, 2007; Speekenbrink & Konstantinidis, 2015). In such non-stationary environments, previously rewarding options may no longer be valuable at a later point in time, thereby benefiting continuous exploration.

An important question for future research concerns the representation of uncertainty in learning and exploration. In our task, the spatial correlation of rewards favors a more complex representation of uncertainty structured around generalization, but in other tasks simpler representations of uncertainty may provide a better account. For instance, count-based exploration strategies operate on simpler representations of uncertainty solely based on the number of experiences with a certain stimulus (e.g., the number of times a tile has been visited; Bellemare et al., 2016; Cogliati Dezza, Cleeremans, & Alexander, 2019). This representation of uncertainty can be used to implement a variant of the GP-UCB model, where the posterior uncertainty $\sigma(\mathbf{x})$ is replaced with a count-based representation of uncertainty (Supplement A). Exploratory analyses with a GP count-based model with our data suggest promising results (Supplement B), yet also present a crucial limitation. Specifically, the uncertainty estimates of the count-based model are decoupled from the generalization component, producing identical uncertainty estimates for all unobserved options. This holds for both near

Table 1
Comparison of Predictive Accuracy and GP-UCB Parameter Estimates Across Different Studies with Children and Adults, Using the Spatially-Correlated Multi-Armed Bandit Paradigm.

Age group	Accuracy R^2	Generalization λ	Uncertainty bonus β	Randomness τ
Current study				
6-year-olds (N=54)	0.09	0.41	0.57	0.18
8-year-olds (N=48)	0.18	0.42	0.54	0.04
Schulz et al. (2019)				
7-8 years (N=55)	0.17	0.44	0.51	0.01
9-11 years (N=55)	0.26	0.53	0.50	0.02
Adults (N=50)	0.39	0.83	0.24	0.03
Wu et al. (2018)				
Adults (N=241)	0.26	0.74	0.40	0.03

Note: R^2 is the mean predictive accuracy of the GP-UCB model. Model parameters λ, β, and τ are the median values of the cross-validated estimates. We report the mean across three experiments from Wu et al. (2018), which used both 1D (Exp. 1) and 2D spatially correlated bandits (Exp. 2-3), with similar smooth and rough environments (Exp. 1-2) or natural environments defined by agricultural data (Exp. 3).

and distant options, disregarding the level of spatial proximity to previous observations. This is also the case for time-based representations, where uncertainty is assumed to increase the longer an option has not been chosen (Blanco & Sloutsky, 2019). In this sense, the count-based account is similar to the MT model, where both the estimates of reward and uncertainty are updated only when a tile is observed. When using a count-based representation of uncertainty, reward estimates are influenced by generalization, but not the uncertainty of rewards which is solely a function of previous visits. By contrast, the GP-UCB model generalizes both reward expectations and attached uncertainty by exploiting the correlation structure of rewards in the environment. In fact, research with adults has shown that confidence judgments are systematically related to the uncertainty estimates predicted by the GP (Wu, Schulz, Garvert, et al., 2020; Wu, Schulz, & Gershman, 2020), as opposed to being uniform across all unobserved options. (We observed a similar relation for 8-year-olds in our study, but the data were rather noisy, so a cautious interpretation is warranted; see Appendix B). Future research should contrast different representations of uncertainty in their ability to predict children's and adults' confidence judgments about expected rewards of novel options, to gain a better understanding of possible developmental trends in the representation of uncertainty across the lifespan.

Conclusions

To conclude, our study provides important new insights into the developmental origins and trajectory of learning and exploration, revealing some of its underlying computational principles. Being able to disentangle the role of generalization, and directed versus random exploration enriches our understanding of how children learn about the world they live in (Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Gopnik, Sobel, Schulz, & Glymour, 2001) and the people they interact with (Bridgers, Jara-Ettinger, & Gweon, 2019; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). It is also important to extend this computational approach to investigate the exploration behavior of even younger preschoolers, toddlers and infants, to identify a more comprehensive developmental trajectory and potentially account for individual differences. Finally, connecting this line of work with the growing body of research and theories on curiosity (Berlyne, 1966; Gottlieb, Oudeyer, Lopes, & Baranes, 2013; Kidd & Hayden, 2015) promises to bring us one step closer to identifying the key to children's impressively successful early learning.

References

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, *3*, 397–422.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*, 1214–1221.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. In *Advances in neural information processing systems* (pp. 1471–1479).

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, *38*, 716–719.

Berlyne, D. E. (1966). Curiosity and exploration. *Science*, *153*, 25–33.

Betsch, T., Lehmann, A., Lindow, S., Lang, A., & Schoemann, M. (2016). Lost in search: (Mal-)adaptation to probabilistic decision environments in children and adults. *Developmental Psychology*, *52*, 311-325.

Blanco, N. J., Love, B. C., Ramscar, M., Otto, A. R., Smayda, K., & Maddox, W. T. (2016). Exploratory decision-making as a function of lifelong experience, not cognitive decline. *Journal of Experimental Psychology: General*, *145*, 284–297.

Blanco, N. J., & Sloutsky, V. M. (2019). Systematic exploration and uncertainty dominate young children's choices. *PsyArXiv*. doi: 10.31234/osf.io/72sfx

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in Cognitive Sciences*, *18*, 497–500.

Bonawitz, E., van Schijndel, T. J., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*, 215–234.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2019). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, 144–152.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*, 331–340.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32.

Cauffman, E., Shulman, E. P., Steinberg, L., Claus, E., Banich, M. T., Graham, S., & Woolard, J. (2010). Age differences in affective decision making as indexed by performance on the Iowa gambling task. *Developmental Psychology*, *46*, 193–207.

Cogliati Dezza, I., Cleeremans, A., & Alexander, W. (2019). Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *Journal of Experimental Psychology: General*, *148*, 977–993.

Cohen, J. D., McClure, S. M., & Angela, J. Y. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*, 933–942.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*, e1004567.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gittins, J. C., & Jones, D. M. (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, *66*, 561–565.

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*, 87–92.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*, 7892–7899.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*, 620–629.

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*, 758–770.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences*, *17*, 585–593.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & the Cognitive Search Research Group. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*, 46–54.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*, 589–604.

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*, 449–460.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & 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*, 284–299.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*, 1193–1215.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

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.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*, 191–215.

Mullen, K., Ardia, D., Gil, D. L., Windover, D., & Cline, J. (2011). DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software*, *40*, 1–26.

Rasmussen, C. E., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.

Ronfard, S., Zambrana, I. M., Hermansen, T. K., & Kelemen, D.

(2018). Question-asking in childhood: A review of the literature and a framework for understanding its development. *Developmental Review*, 101–120.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Ruggeri, A., Markant, D. B., Gureckis, T. M., Bretzke, M., & Xu, F. (2019). Memory enhancements from active control of learning emerge across development. *Cognition*, *186*, 82–94.

Ruggeri, A., Xu, F., & Lombrozo, T. (2019). Effects of explanation on children's question asking. *Cognition*, *191*, 103966.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14.

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*, *44*, 927–943.

Schulz, E., Wu, C. M., Huys, Q. J., Krause, A., & Speekenbrink, M. (2018). Generalization and search in risky environments. *Cognitive Science*, *42*, 2592–2620.

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *Psychological Science*, *30*, 1561–1572.

Schulz, L. E. (2015). Infants explore the unexpected. *Science*, *348*, 42–43.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., & Wilson, R. C. (2017). Charting the expansion of strategic exploratory behavior during adolescence. *Journal of Experimental Psychology: General*, *146*, 155–164.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*, 351–367.

Sumner, E., Li, A. X., Perfors, A., Hayes, B., Navarro, D., & Sarnecka, B. W. (2019). The exploration advantage: Children's instinct to explore allows them to find information that adults miss. *PsyArXiv*. doi: 10.31234/osf.io/h437v

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*, 285–294.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*, *72*, 303–308.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and spearman's ρ. *Journal of Applied Statistics*.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*, 2074–2018.

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2020). Similarities and differences in spatial and non-spatial cognitive maps. *bioRxiv*. doi: 10.1101/2020.01.21.914556

Wu, C. M., Schulz, E., & Gershman, S. J. (2020). Inference and search on graph-structured spaces. *bioRxiv*. doi: 10.1101/2020.01.21.981399

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*, 915–924.

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife*, *6*, e27430.

Appendix A
Statistical analyses

We report both frequentist statistics and Bayes factors (BF) to quantify the relative evidence of the data in favor of the alternative hypothesis (H_A) over the null hypothesis (H_0). All model specifications and R-code are available online at `https://osf.io/eq2bk/?view_only=fed4735fa15a4f3d8dd56db385b845b1`.

Group comparisons

Frequentist tests are reported as t-tests for parametric comparisons, and Mann-Whitney-U or Wilcoxon signed-rank test for non-parametric comparisons. Bayes factors are based on the default two-sided Bayesian t-test for either independent or dependent samples, using a Jeffreys-Zellner-Siow prior with its scale set to $\sqrt{2}/2$ (Rouder, Speckman, Sun, Morey, & Iverson, 2009). All statistical tests are non-directional as defined by a symmetric prior. Bayes factors for the Mann-Whitney-U test are based on performing posterior inference over the test statistic (Kendall's r_τ), assigning a prior using parametric yoking (van Doorn, Ly, Marsman, & Wagenmakers, 2020).Bayes factors for non-parametric comparisons are based on performing posterior inference over the test statistics (Kendall's r_τ for the Mann-Whitney-U test and standardized effect size $r = \frac{Z}{\sqrt{N}}$ for the Wilcoxon signed-rank test), assigning a prior using parametric yoking (van Doorn et al., 2020). The posterior distribution for Kendall's r_τ or the standardized effect size r yields a Bayes factor via the Savage-Dickey density ratio test, where the null hypothesis posits that parameters do not differ between groups and the alternative hypothesis posits an effect and assigns an effect size using a Cauchy distribution with the scale parameter set to $1/\sqrt{2}$.

Correlations

Linear correlations are tested with Pearson's r, the corresponding Bayesian test is based on Jeffrey's test for linear correlation assuming a shifted, scaled beta prior distribution $B(\frac{1}{k}, \frac{1}{k})$ for r, where the scale parameter is set to $k = \frac{1}{3}$ (Ly, Verhagen, & Wagenmakers, 2016). For testing rank correlations with Kendall's tau, the Bayesian test is based on parametric yoking to define a prior over the test statistic (van Doorn, Ly, Marsman, & Wagenmakers, 2018). Bayesian inference is performed to compute a posterior distribution for r_τ, and the Savage-Dickey density ratio test is used to produce an interpretable Bayes Factor.

Bayesian multilevel regressions

Regression analyses were performed in a Bayesian framework with `Stan` (Carpenter et al., 2017), accessed via R-package `brms` (Bürkner, 2017). In all models, participants were treated as a random intercept, the remaining predictors were implemented as population-level ("fixed") effects. For population-level effects, we used a normal prior with a mean of 0 and standard deviation of 10; for group-level ("random") effects, we used a half student-t prior with 3 degrees of freedom, a mean of 0, and a scale parameter of 10; for the intercept a student-t prior with 3 degrees of freedom, a mean of 1, and a scale parameter of 10. All models were estimated over four chains of 4000 iterations, with a burn-in period of 1000 samples.

Appendix B
Bayesian regression analyses

Search distance as function of reward on previous step

We ran separate regression analyses for each environment to assess the influence of reward obtained at trial t on search distance at $t+1$, with population-level ("fixed") effects for previous reward, age group, and their interaction, and by-participant random intercepts. Figure 3c illustrates the population-level effects; Table A1 provides a summary of the results. For both environments, these analyses showed an effect of previously obtained reward on search distance (i.e., lower rewards lead to higher subsequent search distances), an effect of age group (i.e., 8-year-olds showed higher search distances overall), and an interaction (i.e., the search distance of 8-year-olds was stronger influenced by obtained rewards than that of 6-year-olds).

Table A1
Bayesian Regression Results: Search Distance as Function of Reward on Previous Step.

	Rough environment		**Smooth environment**	
Predictor	**Estimate**	**95% HDI**	**Estimate**	**[95% HDI]**
Intercept	2.26	[1.90 – 2.63]	2.89	[2.6. – 3.19]
Previous reward	-0.01	[-0.02 – -0.01]	-0.03	[-0.03 – -0.02]
Age group	1.85	[1.31 – 2.34]	1.19	[0.73 – 1.64]
Previous reward × age group	-0.04	[-0.05 – -0.03]	-0.03	[-0.04. – -0.02]
Random Effects				
σ^2	0.48		0.29	
τ_{00}	4.84		4.14	
N	50		52	
Observations	5000		5200	
Bayesian R^2	0.16		0.13	

Note: Both models were implemented in `brms` (Bürkner, 2017). We report the posterior mean estimates for the coefficients, followed by an 95% uncertainty interval in brackets ("highest density interval", *HDI*). σ^2 indicates the individual-level variance and τ_{00} indicates the variation between individual intercepts and the average intercept. For categorical variable age group, 6-year-olds are the reference level.

Bonus round judgments

In the bonus round, children made reward predictions for five previously unseen tiles and rated their confidence in their predictions. To assess the relation between prediction error (mean absolute deviation between judged and true reward value) and confidence we ran a Bayesian linear regression with prediction error as dependent variable, and confidence, age group and their interaction as population-level ("fixed") effects, and a random intercept for participants. Children's confidence judgments were elicited using an 11-point (0–10) slider with the endpoints labeled as "not at all" and "very sure".

Table A2 provides a summary of the results; Figure A1 show the population-level (fixed) effects of the model, excluding the group-level effects (random intercepts over participants). These data show no systematic relation between children's subjective confidence in their predictions, and the magnitude of their prediction error.

Table A2
Bayesian Regression Results: Prediction Error and Confidence

Predictor	**Estimate**	**95% HDI**
Intercept	13.72	[10.04 – 17.51]
Confidence	0.35	[-0.09 – 0.77]
Age group	-2.12	[-7.57 – 3.29]
Confidence × age group	-0.38	[-1.07 – 0.30]
Random effects		
σ^2	25.09	
τ_{00}	81.36	
N	102	
Observations	510	
Bayesian R^2	0.3	

Note. The model was implemented in `brms` (Bürkner, 2017). We report the posterior mean estimates for the coefficients, followed by an 95% uncertainty interval in brackets ("highest density interval", *HDI*). σ^2 indicates the individual-level variance and τ_{00} indicates the variation between individual intercepts and the average intercept. For variable age group, 6-year-olds are the reference level.

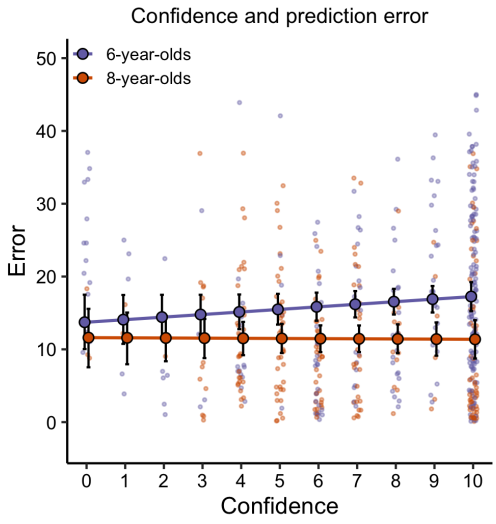


Figure A1. Confidence and prediction error in the bonus round. The lines visualize the expected values of the posterior predictive distribution of a Bayesian regression (± 95% CI); the dots show the raw data.

Regression analyses for age-related trends in parameter estimates

To control for the effect of predictive accuracy R^2 on the age-related changes in the GP-UCB parameter estimates, we ran regression analyses for each parameter with age (in months), individual R^2, and their interaction as predictors for the individual median parameter estimates. Since λ, β, and τ are defined as non-negative, we log-transformed them for the regressions; for plotting the influence of age on parameters we converted the regression models' predictions back to the original scale by exponentiating them, such that all parameters are non-negative. Table A3 shows the results of the regression analyses; Figure A2 visualizes the effects of age on the GP-UCB parameter estimates while taking into account R^2.

Table A3
Bayesian Regression Results: Parameter Estimates with Age and R^2 as Predictors.

	Generalization λ (log)		**Uncertainty bonus β (log)**		**Temperature τ (log)**	
Predictor	**Estimate**	**95% HDI**	**Estimate**	**95% HDI**	**Estimate**	**95% HDI**
Intercept	-2.83	[-4.55 – -1.12]	2.98	[-0.77 – 6.70]	3.76	[0.21 – 7.56]
Age (in months)	0.03	[0.01 – 0.05]	-0.03	[-0.08 – 0.02]	-0.05	[-0.10 – -0.01]
R^2	5.49	[-3.05 – 13.95]	-7.75	[-22.98 – 7.55]	-5.94	[-20.91 – 8.96]
R^2 × Age (in months)	-0.08	[-0.18 – 0.03]	0.04	[-0.14 – 0.22]	-0.06	[-0.24 – 0.13]
Observations	102		102		102	
Bayesian R^2	0.08		0.13		0.69	

Note: All models were implemented in `brms` (Bürkner, 2017). We report the posterior mean estimates for the coefficients, followed by an 95% uncertainty interval in brackets ("highest density interval", HDI).

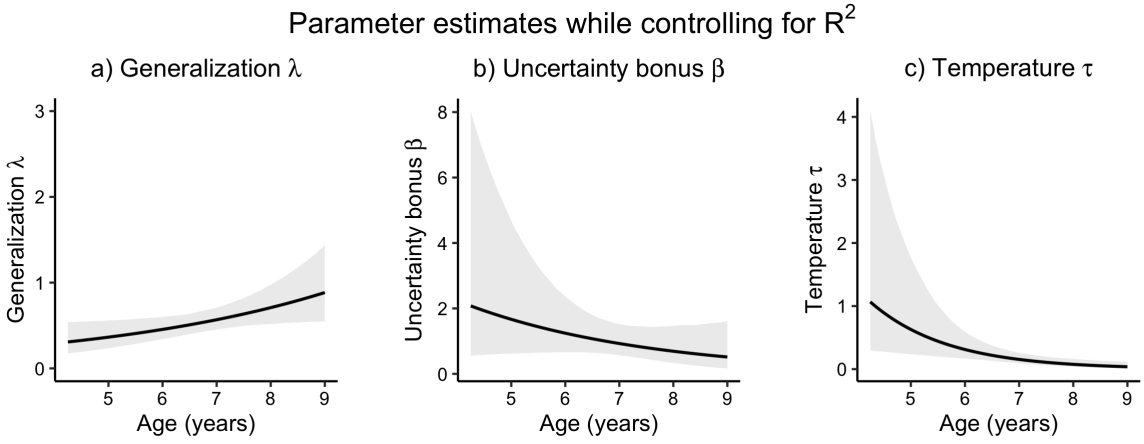


Figure A2. Effect of age on GP-UCB parameters, derived from a Bayesian regression with age (in months), individual model R^2, and their interactions, as predictor for the (log-transformed) median parameter estimates. For plotting we converted the regression models' predictions back to the original scale by exponentiating the parameter estimates, such that all parameters are non-negative.

GP model predictions and bonus round judgments of reward and confidence

We assessed the relation between GP model predictions and participant judgments about expected reward and confidence in the bonus round. In the bonus round, participants selected 15 tiles and then made reward predictions for five unseen tiles and judged their confidence in their predictions. The MT model, which learns independent reward distributions, makes identical predictions for all unseen tiles, as it does not generalize. By contrast, the GP model makes specific predictions for novel options, taking into account the data obtained so far and the spatial correlation of the search ecology.

For each participant, we used parameters estimated estimated from rounds 2 to 5 in order to generate individual GP model predictions (estimated mean reward and variance) for the five randomly selected tiles in the bonus round. These predictions were conditioned on the 15 individual choices and observations made by each child and were generated using each individuals' median λ estimates. This represents a type of out-of-task prediction, where we used parameters estimated from search decisions to prediction out-of-sample judgments. We use the mean reward predictions of the GP model (posterior $\mu(\mathbf{x})$ of tile) as a prediction for each each child's judgment about expected reward and the GP's uncertainty estimates (posterior σ) as a prediction of each child's confidence judgments, where we treat uncertainty as the inverse of confidence.

GP predictions were somewhat correlated with participant predictions (r_τ = .08, p = .013, BF = 1.5), although this disappeared when separating participants into age groups (6-year-olds: r_τ = .06, p = .182, BF = .22; 8-year-olds: r_τ = .08, p = .054, BF = .57). GP uncertainty estimates were negatively correlated with confidence for 8-year-olds (r = −.18, p = .005, BF = 7.5), but not for 6-year-olds (r = .06, p = .330, BF = .23). This suggests that the confidence judgments of 8-year-olds were somewhat accounted for by the GP model, but not those of 6-year-olds.

To analyze these findings in more detail, we conducted Bayesian regression analyses to predict children's reward and confidence judgments based on the outputs of the GP model. Specifically, we used GP model predictions, age group, and their interaction as population-level ("fixed") effects, and by-participant random intercept (Table A4). In the first model (*Reward judgments*), participant reward judgments in the range [0,50] for novel options $\mathbf{x}$ (tiles) were predicted from the GP posterior means of rewards, $\mu(\mathbf{x})$. The second model (*Confidence judgments*) used the GP posterior uncertainty, $\sigma(\mathbf{x})$ to predict children's confidence judgments in the range [0,10]. All GP predictions were computed based on individual participant λ-values and the 15 search decisions they made prior to providing their judgments for five random novel options.

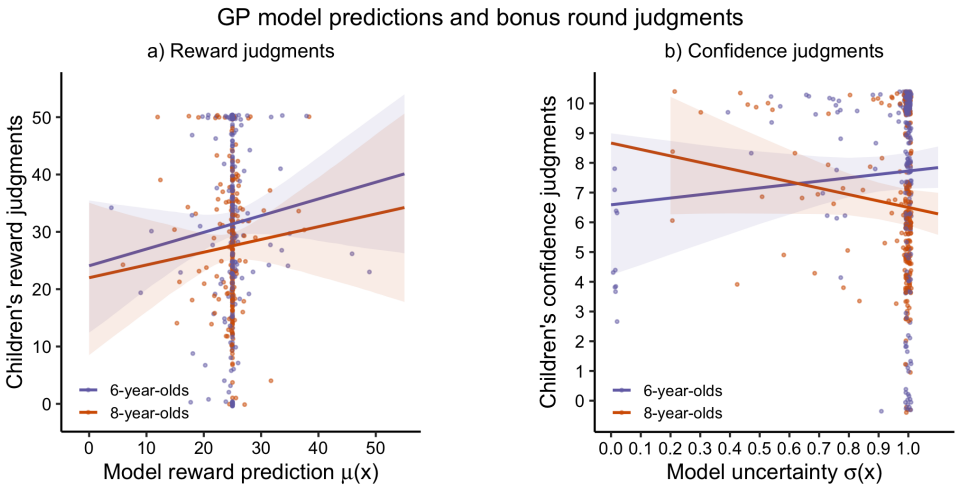


Figure A3. GP model predictions for bonus round judgments. The lines visualize the means of the posterior predictive distribution of the Bayesian regression (± 95%CI); the dots show the raw data points. *a)* Relation between GP model predictions of reward and children's reward judgments. *b)* Relation between GP model uncertainty about expected rewards and children's confidence about their reward judgments.

Table A2 provides a summary of the results; Figure A3 visualizes the population-level (fixed) effects of the model, excluding the group-level effects (random intercepts over participants). The results show a positive but rather weak relation between the GP model's reward predictions and children's reward judgments about unobserved tiles (Figure A3a). The trends for the relation between model uncertainty and children's confidence judgments mirror the overall correlations. For 6-year-olds, there's a weak relation in the wrong direction (i.e., they tend to be more confident when the GP model is more uncertain). By contrast, for 8-year-olds there is a fairly strong trend in that children's confidence declined with increasing

model uncertainty. However, the raw data are very noisy and unevenly distributed, so a cautious interpretation of these results is warranted.

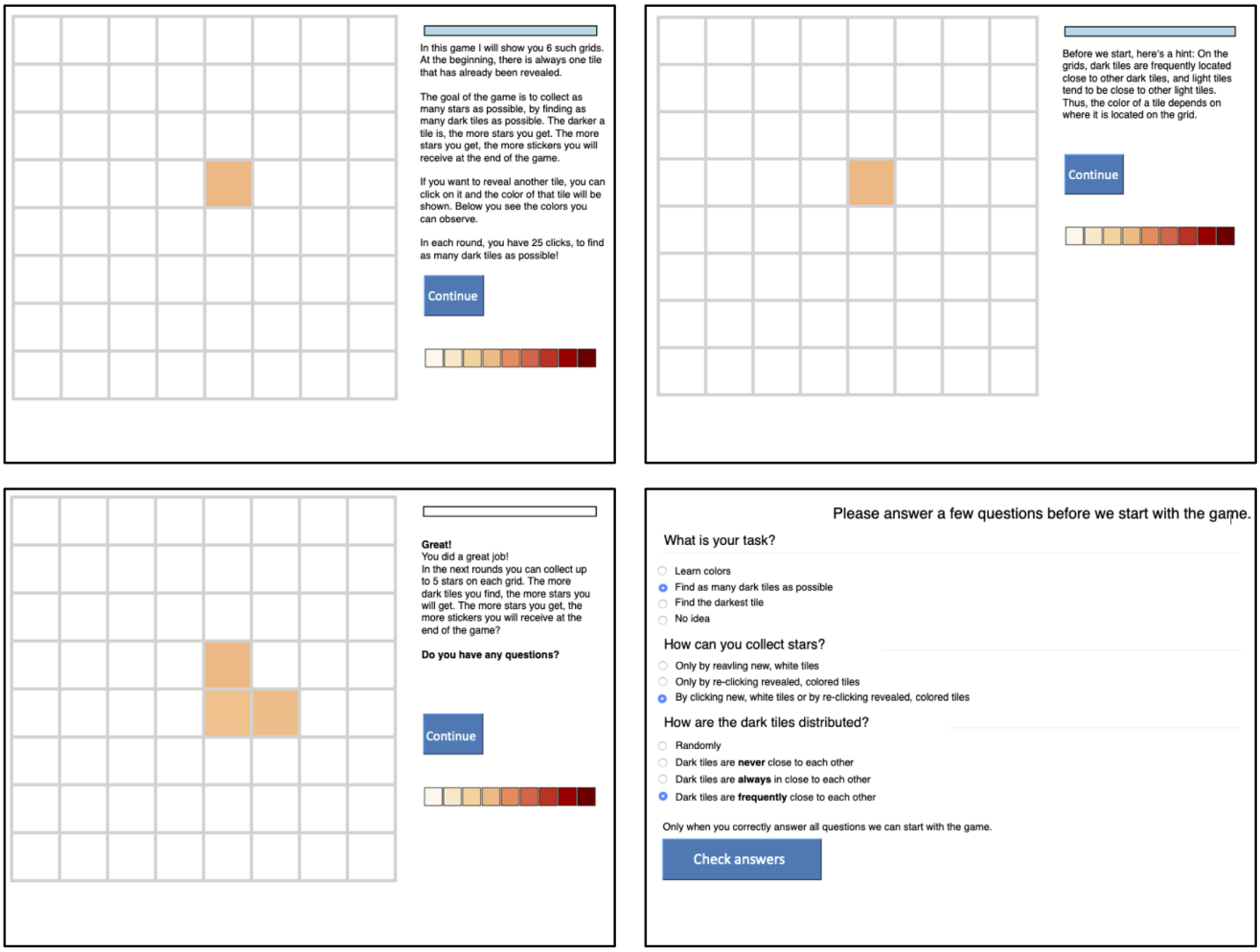
Table A4
Bayesian Regression Results: GP Model Predictions and Bonus Round Judgments.

	Reward judgments		**Confidence judgments**	
Predictor	**Estimate**	**95% HDI**	**Estimate**	**95% HDI**
Intercept	24.08	[12.43 – 35.46]	6.59	[4.23 – 8.99]
GP predictions	0.29	[-0.15 – 0.75]	1.14	[-1.32 – 3.55]
Age group	-2.09	[-16.13 – 12.33]	2.08	[-1.19 – 5.42]
GP predictions × age group	-0.07	[-0.65 – 0.51]	-3.3	[-6.73 – 0.03]
Random effects				
σ^2	32.69		3.49	
τ_{00}	168.25		4.53	
N	102		102	
Observations	510		510	
Bayesian R^2	.19		.49	

Note: Both models were implemented in `brms` (Bürkner, 2017). We report the posterior mean estimates for the coefficients, followed by an 95% uncertainty interval in brackets ("highest density interval", HDI). σ^2 indicates the individual-level variance and τ_{00} indicates the variation between individual intercepts and the average intercept. For categorical variable age group, 6-year-olds are the reference level.

Appendix C
Instructions

The experiment was implemented on a tablet, where children could touch the screen to reveal new tiles. Below are screenshots from the instructions; further screenshots are shown in Fig. 1b) and c).



Supplement A
Computational models

We here provide a formal description of the learning models (Gaussian process regression and Bayesian mean tracker) and sampling strategies we used to model children's exploration behavior. In addition to the UCB sampling strategy we explored three additional sampling functions, namely *Count-based sampling*, *Mean Greedy Exploitation*, and *Variance Greedy Exploration*. Supplement B provides a full comparison of all all 2(learning model)×4(sampling strategies)=8 models.

Gaussian process regression

Gaussian Process (GP) regression is a Bayesian approach to generalization, which we use as a model of learning in our spatially-correlated multi-armed bandit task. Let $f : \mathcal{X} \rightarrow R^n$ denote a function that maps values from the input space $\mathcal{X}$ (e.g., x- and y-coordinates of a tile) to real-valued scalar outputs (e.g., reward values). In our current task, we are mapping locations on the grid to reward values. A GP defines a distribution over functions, where each function f can be understood as a candidate hypothesis about the structure of rewards environment, and is modeled as a random draw from a GP:

$$f \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right), \tag{S1}$$

The GP prior is defined by a mean function specifying the expected output given input $\mathbf{x}$:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \tag{S2}$$

and kernel function specifying the covariance between any two inputs $\mathbf{x}$ and $\mathbf{x}'$:

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right] \tag{S3}$$

In the present study, reward values were visualized using different shades of red, corresponding to the underlying numerical reward values. We scaled rewards to the range [0, 50] and set the prior mean to the median value of unscaled payoffs $m(\mathbf{x}) = 25$. The kernel function $k(\mathbf{x}, \mathbf{x}')$ is used to encode an inductive bias about the expected spatial correlations between rewards (see Radial Basis Function kernel below).

To make predictions about expected rewards, we condition on the observed data $\mathcal{D}_t = \{\mathbf{x}_j, y_j\}_{j=1}^t$, where we assume observations of reward $y_j \sim \mathcal{N}(f(\mathbf{x}_j), \sigma_\epsilon^2)$ have Gaussian noise $\sigma_\epsilon^2 = 1$. The posterior predictive distribution for any new input $\mathbf{x}_*$ is also a Gaussian distribution, with mean and variance given by:

$$\mathbb{E}[f(\mathbf{x}_*)|\mathcal{D}_t] = \mathbf{k}_*^\top (\mathbf{K} + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathbf{y}_t \tag{S4}$$

$$\mathbb{V}[f(\mathbf{x}_*)|\mathcal{D}_t] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_\star^\top (\mathbf{K} + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathbf{k}_*, \tag{S5}$$

where $\mathbf{k}_* = [k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_t, \mathbf{x}_*)]$ is the covariance between each observed input and the new input $\mathbf{x}_*$, $\mathbf{K}$ is the $t \times t$ covariance matrix evaluated at each pair of observed inputs, and $\mathbf{y} = [y_1, \ldots, y_t]^\top$. For simplicity, the main text uses the notation $\mu(\mathbf{x}_*) = \mathbb{E}[f(\mathbf{x}_*)|\mathcal{D}_t]$ and $\sigma(\mathbf{x}_*) = \sqrt{\mathbb{V}[f(\mathbf{x}_*)|\mathcal{D}_t]}$, based on the standard convention for describing the mean and standard deviation of a normal distribution.

Radial Basis Function kernel. The Radial Basis Function (RBF) kernel specifies the correlation between inputs $\mathbf{x}$ and $\mathbf{x}'$ as

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\lambda^2}\right). \tag{S6}$$

where $\mathbf{x}$ and $\mathbf{x}'$ denote two inputs (e.g., coordinates of tiles on the grid) and λ is the *length-scale* parameter governing the rate of correlation decay. The RBF kernel models generalization as an exponentially decaying function of the distance between inputs, such that larger λ-values correspond to slower decays, stronger spatial correlations, and smoother functions. As $\lambda \rightarrow \infty$, the RBF kernel assumes functions approaching linearity; as $\lambda \rightarrow 0$, there ceases to be any spatial correlation, meaning that learning of options' rewards happens independently (similar to the assumption of the Mean Tracker model; see below for details). In the model comparisons, we treat λ as a free parameter estimated through cross-validation to make inferences about the extent to which children generalize.

Bayesian mean tracker

The key difference between the GP model and the Mean Tracker (MT) is that the MT model learns independent reward distributions for the options. We implement the MT model as a Bayesian updating model, which learns the rewards of each option by computing an independent posterior distribution for the mean μ_j of each option j. As in the GP account, the MT model assumes that rewards are normally distributed (as is the case in our experiment), with a known variance but unknown mean. The prior distribution of the mean is a normal distribution, implying that the posterior distribution for each option's mean is also a normal distribution:

$$p(\mu_{j,t}|\mathcal{D}_{t-1}) = \mathcal{N}(m_{j,t}, v_{j,t}) \tag{S7}$$

where $m_{j,t}$ and $v_{j,t}$ denotes the posterior mean and variance, respectively. The mean and variance of an option j are only updated when it has been selected at trial t:

$$m_{j,t} = m_{j,t-1} + \delta_{j,t} G_{j,t} \left[y_t - m_{j,t-1} \right] \tag{S8}$$

$$v_{j,t} = \left[1 - \delta_{j,t} G_{j,t} \right] v_{j,t-1} \tag{S9}$$

where $\delta_{j,t} = 1$ if option j is chosen on trial t, and 0 otherwise. In addition, y_t is the observed reward at trial t, and $G_{j,t}$ is defined as:

$$G_{j,t} = \frac{v_{j,t-1}}{v_{j,t-1} + \theta^2_\epsilon} \tag{S10}$$

where θ^2_ϵ is the error variance, which we estimate as a free parameter.

Intuitively, the estimated mean of the chosen option $m_{j,t}$ is updated based on the difference between the observed value y_t and the prior expected mean $m_{j,t-1}$ (i.e., prediction error), scaled by the Kalman gain $k_{j,t}$. Thus, the Kalman gain acts as a learning rate that is dynamically defined based on the ratio of the estimated uncertainty ($v_{j,t-1}$) and the assumed uncertainty (θ^2_ϵ) in the environment. This form of prediction error learning is shared with a broad range of models from associated learning, where specifically, the MT can be understood as a Bayesian variant of the traditional Rescorla-Wagner (1972) model (Gershman, 2015). As with the GP, we set the prior mean of the MT to the median value of unscaled payoffs $m_{j,0} = 25$, while also setting the prior variance to $\sqrt{v_{j,0}} = 250$.

Sampling strategies

For each option, the GP and MT learning models generate normally distributed posteriors of the expected rewards and associated uncertainty. The posterior predictions of the MT in the form of mean $m_{j,t}$ and standard deviation $\sqrt{v_{j,t}}$ have the same structure as the GP posterior, which is defined by mean $\mu(\mathbf{x})$ and standard deviation $\sigma(\mathbf{x})$. However, the MT uses index j to denote each option, while the GP uses a vector notation $\mathbf{x} = \{x_1, x_2\}$ to denote the coordinates of each option. For simplicity, we will refer to these predictions using mean $\mu(\mathbf{x})$ and standard deviation $\sigma(\mathbf{x})$. We then use various sampling strategies to map these estimates onto valuations for each option, which combined with a softmax choice rule (Eq. 3) provide probabilistic predictions about where each participant would search next. We considered two sampling strategies that take into account both estimated rewards and their uncertainty to balance the exploration-exploitation trade-off, *Upper Confidence Bound* sampling and *count-based sampling*. We additionally tested the performance of two strategies that consider either only rewards or uncertainty, i.e., constitute a pure exploitation strategy (*mean greedy exploitation*) or a pure exploration strategy (*variance greedy exploration*).

Upper Confidence Bound sampling. Given the posterior predictive mean $\mu(\mathbf{x})$ and its standard deviation $\sigma(\mathbf{x})$, the upper confidence bound is given by a weighted sum

$$\text{UCB}(\mathbf{x}) = \mu(\mathbf{x}) + \beta\sigma(\mathbf{x}), \tag{S11}$$

where the "exploration bonus" β determines how much a searcher values the reduction of uncertainty, relative to exploiting known high-value options. We estimate β as a free parameter representing children's tendency towards directed exploration.

Count-based sampling. Similarly to the UCB sampling strategy, count-based sampling also considers both rewards and uncertainty. However, uncertainty is represented in a computationally simpler way, namely solely by the number of experiences with an option (Bellemare et al., 2016):

$$\mathrm{CB}(\mathbf{x}) = \mu(\mathbf{x}) + \beta \frac{1}{F(\mathbf{x}) + 1}, \tag{S12}$$

where $F(\mathbf{x})$ are the number of previous visits to a specific option. This count-based exploration model is similar to UCB, but depends on a simpler representation of uncertainty. Whereas the GP representation of uncertainty may vary across different unobserved options as a function of distance to observed options, the count-based model treats all unobserved (i.e., not visited) options as having the same uncertainty. Thus, this can be understood as a heuristic implementation of the full GP-UCB model, but where representations of uncertainty are not influenced by the same similarity-based generalization mechanism as used to make predictions about rewards.

Mean Greedy Exploitation. Whereas UCB and count-based sampling integrate both estimates of reward and uncertainty, mean greedy exploitation values options solely based on expected rewards:

$$\mathrm{M}(\mathbf{x}) = \mu(\mathbf{x}), \tag{S13}$$

This sampling strategy disregards any uncertainty and only samples options with high expected rewards, i.e. greedily exploits the environment. This strategy is the special case of UCB and count-based sampling with $\beta = 0$.

Variance Greedy Exploration. Another special case of UCB sampling (with $\beta \to \infty$) is to greedily explore options solely according to their uncertainty (i.e., their predictive standard deviation):

$$\mathrm{V}(\mathbf{x}) = \sigma(\mathbf{x}). \tag{S14}$$

This sampling strategy only cares about reducing uncertainty, without taken into account the expected rewards of options.

Supplement B
Full model comparison results

We evaluated all 2(learning model)×4(samplings strategies)=8 models in terms of their predictive accuracy R^2 (Eq. 4). Table S1 shows the full model comparison results in terms of models' predictive accuracy (mean R^2), and participant median parameter values, estimated through leave-one-out cross-validation.

Table S1
Full model comparison results.

Age group	Model	R^2	λ	β	τ	θ^2_ϵ
6-year-olds	GP-UCB	0.0883	0.406	0.567	0.181	—
(N=54)	GP-Counts	0.0847	1.44	3.49	0.170	—
	GP-GM	0.0375	1.37	—	0.241.	—
	GP-GV	0.0274	0.166	—	0.510	—
	MT-UCB	0.0321	—	15.2	0.868	4.44
	MT-Counts	0.0324	—	22.4	7.02	27.5
	MT-GM	0.0084	—	—	54.6	54.6
	MT-GV	0.0267	—	—	0.0565	13.4
8-year-olds	GP-UCB	0.175	0.419	0.540	0.0396	—
(N=48)	GP-Counts	0.188	0.801	5.10	0.0819	—
	GP-GM	0.0656	1.46	—	0.213	—
	GP-GV	0.0306	0.156	—	0.420	—
	MT-UCB	0.0481	—	16.9	0.342	4.66
	MT-Counts	0.0517	—	15.8	4.69	12.2
	MT-GM	0.0205	—	—	54.6	54.6
	MT-GV	0.0305	—	—	0.0364	14.9

Note: Learning models: GP = Gaussian Process regression, MT = Bayesian Mean Tracker. Sampling strategies: UCB = Upper Confidence Bound, GM = greedy mean, GV = greedy variance, Counts = count-based. Parameters: R^2 = predictive accuracy, λ = length-scale of RBF kernel (generalization parameter), β = uncertainty bonus, τ = random exploration, θ^2_ϵ = error variance.