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

Eric Schulz
Harvard University

Charley M. Wu
Max-Planck-Institute for Human Development

Azzurra Ruggeri
Max-Planck-Institute for Human Development

Björn Meder
Max-Planck-Institute for Human Development

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.

Keywords: Exploration-Exploitation | Generalization | Developmental Psychology

Introduction

Alan Turing famously believed that in order to build a General Artificial Intelligence, one must create a machine that can learn like a child Turing (1950). Indeed, recent advances in machine learning often contain references to child-like learning and exploration (Lake, Ullman, Tenenbaum, & Gershman, 2017; Riedmiller et al., 2018). 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 (Gopnik, 2017).

In the course of learning through interactions with the environment, all organisms, biological or machine, are confronted with the *exploration-exploitation dilemma* (Mehlhorn et al., 2015). 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 (Sutton, 1996). 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 (Palminteri, Kilford, Coricelli, & Blakemore, 2016). 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

Correspondence should be addressed to Eric Schulz, Department of Psychology, Harvard University, 52 Oxford Street, Cambridge, MA 02138, United States. Email: ericschulz@fas.harvard.edu

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 (Gittins & Jones, 1979), heuristic alternatives are frequently employed. In particular, learning under the demands of the exploration-exploitation trade-off has been described using at least two distinct strategies (Wilson, Geana, White, Ludvig, & Cohen, 2014).

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 (Cauffman et al., 2010; Mata, Wilke, & Czienskowski, 2013). The same tendency to explore more options during young age has also been observed in other animals such as wasps (Thiel, Driessen, & Hoffmeister, 2006) and rats (Lalonde, 2002). 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 (Gopnik, Griffiths, & Lucas, 2015). 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 (Gopnik et al., 2017), 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 (Auer, 2002), with behavioral markers found in a number of studies (cf., Frank, Doll, Oas-Terpstra, & Moreno, 2009; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018).

Directed exploration treats information as intrinsically valuable by inflating rewards by their estimated uncertainty (Auer, 2002). This leads to a more sophisticated and uncertainty-guided sampling strategy that could also explain developmental differences. Indeed, the literature on self-directed learning (Gureckis & Markant, 2012) 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 (L. E. Schulz, 2015), and children can balance theory and evidence in simple exploration tasks (Bonawitz, van Schijndel, Friel, & Schulz, 2012) and are able to efficiently adapt their search behavior to different environmental structures (Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri & Lombrozo, 2015). Moreover, children can sometimes even outperform adults in self-directed learning of unusual relationships Lucas, Bridgers, Griffiths, and Gopnik (2014). 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 Somerville et al. (2017).

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 (Blanco et al., 2016). Because many studies have shown that human learners use structured knowledge about the environment to guide exploration (Acuna & Schrater, 2009; E. Schulz, Konstantinidis, & Speekenbrink, 2017; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2017), 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 Piaget (1964) 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, Klahr (1982) 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 (Hartley & Somerville, 2015) 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 (Hagen & Hale, 1973), they can be prompted to attend to the relevant information by marking the most relevant cues, whereupon they eventually select the best alternative Davidson (1996). 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; Wu et al., 2018), 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. Wu et al., 2018), 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; Rasmussen & Williams, 2006) 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 (Lucas, Griffiths, Williams, & Kalish, 2015).

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*:

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

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 generalization historically described by Shepard Shepard (1987), 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 (Srinivas, Krause, Kakade, & Seeger, 2009), which evaluates each option based on a weighted sum of expected reward and estimated uncertainty:

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

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 (UCB(x_1) = 35 vs. UCB(x_2) = 34.5).

Although seemingly naïve, UCB sampling can lead to competitive performance guarantees when paired with Gaussian Process regression (Srinivas et al., 2009). 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,

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

which transforms values (e.g., the upper confidence bound value $UCB(\mathbf{x})$) into a probability distribution over options,

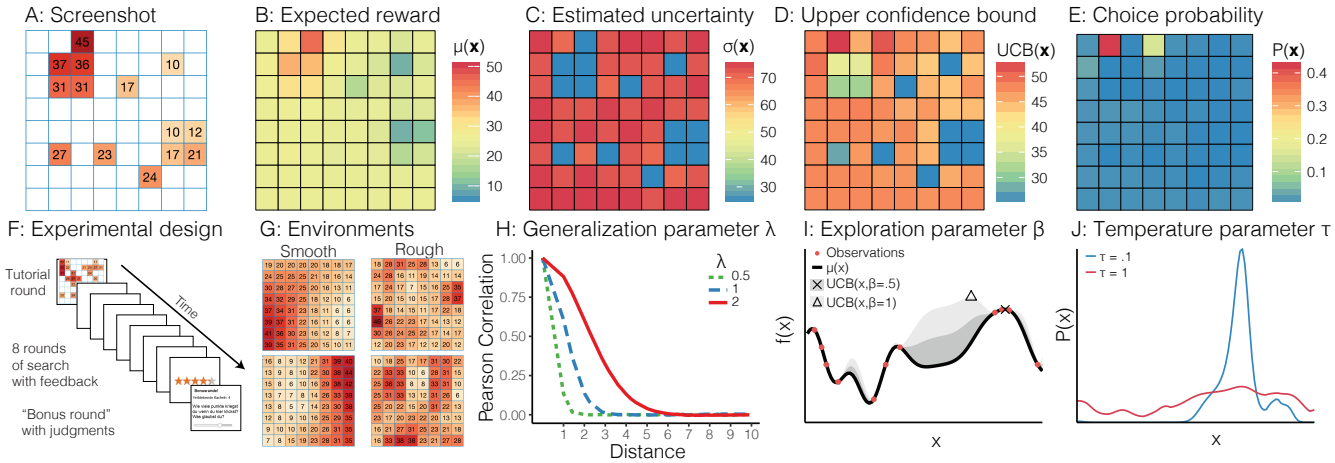

Figure 1. Overview of task and model. **A:** Screenshot of experiment with partially revealed grid. Expected reward (**B**) and estimated uncertainty (**C**) based on observations in A using Gaussian Process regression. **D:** Upper confidence bounds of each option based on a weighted sum of panels B 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.

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 (Wu et al., 2018), 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 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 dif-

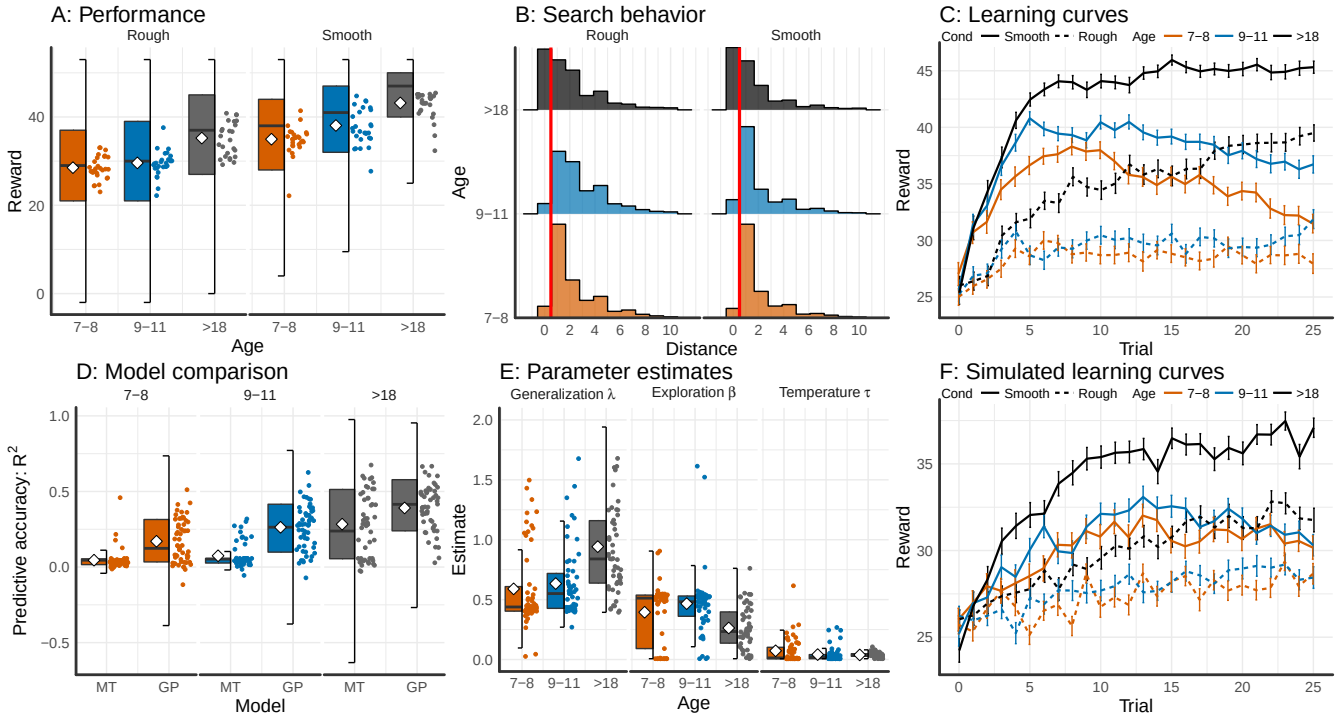

Figure 2. Main results. **A:** Tukey box plots of rewards, showing the distribution of all choices for all participants. Each dot is the participant-wise mean and diamonds indicate group means. **B:** Histograms of distances between consecutive choices by age group and condition, with a distance of zero 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.

ference 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 adults ($t(49) = 5.98$,

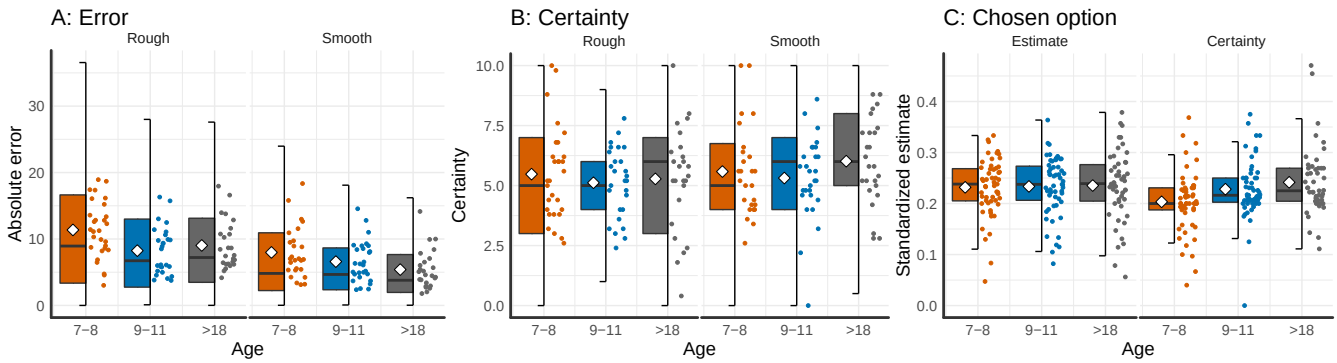

Figure 3. Bonus round results. **A:** Absolute error of participant predictions about the rewards of unobserved tiles. **B:** Certainty judgments, where 0 is least certain and 10 is most certain. **C:** Standardized predictions and certainty estimates, where we divided judgments about the chosen option (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.

$p < .001$, $d = 0.85$, $BF > 100$), 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 di-

viding 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 (Stadie, Levine, & Abbeel, 2015).

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.

References

Acuna, D., & Schrater, P. R. (2009). Structure learning in human sequential decision-making. In *Advances in neural information processing systems* (pp. 1–8).

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

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*(3), 284–297.

Bonawitz, E. B., van Schijndel, T. J., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234.

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*(1), 193–207.

Davidson, D. (1996). The effects of decision characteristics on children's selective search of predecisional information. *Acta Psychologica*, *92*(3), 263–281.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual

differences in exploration and exploitation. *Nature neuroscience*, *12*(8), 1062–1068.

Gittins, J. C., & Jones, D. M. (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, *66*(3), 561–565.

Gopnik, A. (2017). Making AI more human. *Scientific American*, *316*(6), 60–65.

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*(2), 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*(30), 7892–7899.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Hagen, J. W., & Hale, G. A. (1973). The development of attention in children. *ETS Research Report Series*, *1973*(1).

Hartley, C. A., & Somerville, L. H. (2015). The neuroscience of adolescent decision-making. *Current Opinion in Behavioral Sciences*, *5*(9), 108–115.

Klahr, D. (1982). Nonmonotone assessment of monotone development: An information processing analysis. In S. Strauss & R. Stavy (Eds.), *U-shaped behavioral growth* (pp. 63–86). New York: Academic Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Lalonde, R. (2002). The neurobiological basis of spontaneous alternation. *Neuroscience & Biobehavioral Reviews*, *26*(1), 91–104.

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*(2), 284–299. doi: 10.1016/j.cognition.2013.12.010

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

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*(3), 191–215.

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. Retrieved from `http://www.jstatsoft.org/v40/i06/`

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Palminteri, S., Kilford, E. J., Coricelli, G., & Blakemore, S.-J. (2016). The computational development of reinforcement learning during adolescence. *PLoS Computational Biology*, *12*(6), e1004953.

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, *35*(4), 1.

Piaget, J. (1964). Part i: Cognitive development in children: Piaget development and learning. *Journal of Research in Science Teaching*, *2*(3), 176–186.

Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning*. Cambridge, MA, USA: MIT Press.

Riedmiller, M., Hafner, R., Lampe, T., Neunert, M., Degrave, J., Van de Wiele, T., ... Springenberg, J. T. (2018). Learning by playing-solving sparse reward tasks from scratch. *arXiv preprint arXiv:1802.10567*.

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*(2), 225–237.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*(1), 203–216. doi: 10.1016/j.cognition.2015.07.004

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.*.

Schulz, L. E. (2015). Infants explore the unexpected. *Science*, *348*(6230), 42–43.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 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*(2), 155–164.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Stadie, B. C., Levine, S., & Abbeel, P. (2015). Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*(4), 1004–1017.

Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in neural information processing systems* (pp. 1038–1044).

Thiel, A., Driessen, G., & Hoffmeister, T. S. (2006). Different habitats, different habits? response to foraging information in the parasitic wasp venturia canescens. *Behavioral Ecology and Sociobiology*, *59*(5), 614–623.

Turing, A. (1950). Computing intelligence and machinery. *Mind*, *59*(2236), 433–460.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian latent-normal inference for the rank sum test, the signed rank test, and Spearman's ρ. *arXiv preprint*

arXiv:1712.06941.

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*(6), 2074–2081.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2017). Mapping the unknown: The spatially correlated multi-armed bandit. In *Proceedings of the 39th annual meeting of the cognitive science society* (p. 1357-1362).

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Exploration and generalization in vast spaces. *bioRxiv*, 171371.

Statistical tests

We report both frequentist and Bayesian test statistics throughout the paper. Whereas frequentist tests are reported as either Student's t-tests (for the behavioral data and model comparisons) or Mann-Whitney-U tests (for the parameter comparisons), we rely on Bayes factors to quantify the relative evidence the data provide in favor of the alternative hypothesis (H_A) over the null (H_0). The Bayes factor quantifies the likelihood of the data under H_A relative to the likelihood of the data under H_0. We denote such a Bayes factor as BF. For example, a BF of 10 indicates that the data are 10 times more likely under the H_A than under the H_0, while a BF of 0.1 indicates that the data are 10 times more likely under the H_0 than under the H_A.

For testing hypotheses regarding the behavioral data and the model comparison, we use the default two-sided Bayesian t-test for independent samples using a Jeffreys-Zellner-Siow prior with its scale set to $\sqrt{2}/2$ Rouder, Speckman, Sun, Morey, and Iverson (2009). The prior is truncated below 0 for the directional tests performed to create Figure 4, which shows pairwise comparisons between the different models. All other statistical test are non-directional tests for which we use a symmetric prior.

For testing hypotheses regarding the model parameters, we use the frequentist Mann-Whitney-U test and report Kendall's r_τ as an effect size for the difference between two groups. The Bayesian test is based on performing posterior inference over the test statistics and assigning a prior by means of a parametric yoking procedure van Doorn, Ly, Marsman, and Wagenmakers (2017). This then leads to a posterior distribution for Kendall's r_τ, and via the Savage-Dickey density ratio test, also yields an interpretable Bayes factor. The null hypothesis posits that the parameters between two groups do not differ and the alternative hypothesis posits the presence of an effect and assigns an effect size using a Cauchy distribution with the scale parameter set to $1/\sqrt{2}$.

Full modeling results

We report a full model comparison of two models each combined with three different sampling strategies (see also Fig. 10-11). Different *models of learning* (i.e., Gaussian Process regression and Mean Tracker) are combined with a *sampling strategy*, in order to make predictions about where a participant will search next, given the history of previous observations. Table 1 contains the predictive accuracy, the number of participants best described, and the parameter estimates of each combination of a learning model and a sampling strategy. In total, cross-validated model comparisons for both models required approximately two days of computation time distributed on a cluster of 160 Dell C6220 nodes.

Models of Learning

Gaussian Process. We use Gaussian Process (GP) regression as a Bayesian model of generalization. A GP is defined as a collection of points, any subset of which is multivariate Gaussian. Let $f : \mathcal{X} \rightarrow \mathbb{R}^n$ denote a function over input space $\mathcal{X}$ that maps to real-valued scalar outputs. This function can be modeled as a random draw from a GP:

$$f \sim \mathcal{GP}(m, k), \tag{4}$$

where m is a mean function specifying the expected output of the function given input $\mathbf{x}$, and k is a kernel (or covariance) function specifying the covariance between outputs:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{5}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right] \tag{6}$$

Here, we fix the prior mean to the median value of unscaled payoffs, $m(\mathbf{x}) = 25$, and use the kernel function to encode an inductive bias about the expected spatial correlations between rewards (see Radial Basis Function kernel below).

Conditional on observed data $\mathcal{D}_t = \{\mathbf{x}_j, y_j\}_{j=1}^t$, where $y_j \sim \mathcal{N}(f(\mathbf{x}_j), \sigma_j^2)$ is drawn from the underlying function with added noise $\sigma_j^2 = 1$, we can calculate the posterior predictive distribution for a new input $\mathbf{x}_*$ as a Gaussian with mean and variance given by:

$$\mathbb{E}[f(\mathbf{x}_*)|\mathcal{D}_t] = m_t(\mathbf{x}_*) = \mathbf{k}_*^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}_t \tag{7}$$

$$\mathbb{V}[f(\mathbf{x}_*)|\mathcal{D}_t] = v_t(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_\star^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}_*, \tag{8}$$

where $\mathbf{y} = [y_1, \ldots, y_t]^\top$, $\mathbf{K}$ is the $t \times t$ covariance matrix evaluated at each pair of observed inputs, and $\mathbf{k}_* = [k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_t, \mathbf{x}_*)]$ is the covariance between each observed input and the new input $\mathbf{x}_*$.

Radial Basis Function kernel. We use the Radial Basis Function (RBF) kernel as a component of the $\mathcal{GP}$ algorithm of generalization. The RBF kernel specifies the correlation between inputs $\mathbf{x}$ and $\mathbf{x}'$ as

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\lambda}\right). \tag{9}$$

This kernel defines a universal function learning engine based on the principles of Bayesian regression and can model any stationary function. Note that sometimes the RBF kernel is specified as $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2l^2}\right)$ whereas we use $\lambda = 2l^2$ as a more psychologically interpretable formulation. Intuitively, the RBF kernel models the correlation between points as an exponentially decreasing function of their distance. Here, λ modifies the rate of correlation decay, with larger λ-values corresponding to slower decays, stronger spatial correlations, and smoother functions. As $\lambda \to \infty$, the RBF kernel assumes functions approaching linearity, whereas as $\lambda \to 0$, there ceases to be any spatial correlation, with the implication that learning happens independently for each discrete input without generalization (similar to the assumption of the Mean Tracker model described below). We treat λ as a free parameter, and use cross-validated estimates to make inferences about the extent to which participants generalize.

Mean Tracker. The Mean Tracker model is implemented as a Bayesian updating model, which assumes the average reward associated with each option is constant over time (i.e., no temporal dynamics), as is the case in our experiment. In contrast to the GP regression model (which also assumes constant means over time), the Mean Tracker learns the rewards of each option independently, by computing an independent posterior distribution for the mean μ_j for each option j. We implemented a version that assumes rewards are normally distributed (as in the GP model), with a known variance but unknown mean, where the prior distribution of the mean is a normal distribution. This implies that the posterior distribution for each mean is also a normal distribution:

$$p(\mu_{j,t}|\mathcal{D}_{t-1}) = \mathcal{N}(m_{j,t}, v_{j,t}) \tag{10}$$

For a given option j, the posterior mean $m_{j,t}$ and variance $v_{j,t}$ are only updated when it has been selected at trial t:

$$m_{j,t} = m_{j,t-1} + \delta_{j,t} G_{j,t}\left[y_t - m_{j,t-1}\right] \tag{11}$$

$$v_{j,t} = \left[1 - \delta_{j,t} G_{j,t}\right] v_{j,t-1} \tag{12}$$

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

$$G_{j,t} = \frac{v_{j,t-1}}{v_{j,t-1} + \theta_\epsilon^2} \tag{13}$$

where θ_ϵ^2 is the error variance, which is estimated 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 $m_{j,t-1}$, multiplied by $G_{j,t}$. At the same time, the estimated variance $v_{j,t}$ is reduced by a factor of $1 - G_{j,t}$, which is in the range $[0, 1]$. The error variance (θ_ϵ^2) can be interpreted as an inverse sensitivity, where smaller values result in more substantial updates to the mean $m_{j,t}$, and larger reductions of uncertainty $v_{j,t}$. We set the prior mean to the median value of payoffs $m_{j,0} = 25$ and the prior variance to $v_{j,0} = 250$.

Sampling strategies

Given the normally distributed posteriors of the expected rewards, which have mean $\mu(\mathbf{x})$ and uncertainty (formalized here as standard deviation) $\sigma(\mathbf{x})$, for each search option $\mathbf{x}$ (for the Mean Tracker, we let $\mu(\mathbf{x}) = m_{j,t}$ and $\sigma(\mathbf{x}) = \sqrt{v_{j,t}}$, where j is the index of the option characterized by $\mathbf{x}$), we assess different sampling strategies that (combined with a softmax choice rule, Eq. 14) make probabilistic predictions about where participants would search next.

Upper Confidence Bound sampling. Given the posterior predictive mean $\mu(\mathbf{x})$ and its attached standard deviation $\sigma(\mathbf{x}) = \sqrt{\sigma^2(\mathbf{x})}$, we calculate the upper confidence bound using a weighted sum

$$\mathrm{UCB}(\mathbf{x}) = \mu(\mathbf{x}) + \beta\sigma(\mathbf{x}), \tag{14}$$

where the exploration factor β determines how much reduction of uncertainty is valued (relative to exploiting known high-value options). We estimate β as a free parameter indicating participants' tendency towards directed exploration.

Mean Greedy Exploitation. A special case of the Upper Confidence Bound sampling strategy (with $\beta = 0$) is a greedy exploitation component that only evaluates points based on their expected rewards

$$\mathrm{M}(\mathbf{x}) = \mu(\mathbf{x}), \tag{15}$$

This sampling strategy only samples options with high expected rewards, i.e. greedily exploits the environment.

Variance Greedy Exploration. Another special case of the Upper Confidence Bound sampling strategy (with $\beta \to \infty$) is a greedy exploration component which only samples points based on their predictive variance

$$\mathrm{V}(\mathbf{x}) = \sigma(\mathbf{x}). \tag{16}$$

This sampling strategy only cares about reducing uncertainty without attempting to generate high rewards.

Model comparison

We use maximum likelihood estimation (MLE) for parameter estimation, and cross-validation to measure out-of-sample predictive accuracy. A softmax choice rule transforms each model's prediction into a probability distribution over options:

$$p(\mathbf{x}) = \frac{\exp(q(\mathbf{x})/\tau)}{\sum_{j=1}^{N} \exp(q(\mathbf{x}_j)/\tau)}, \tag{17}$$

where $q(\mathbf{x})$ is the predicted value of each option $\mathbf{x}$ for a given model (e.g., $q(\mathbf{x}) = \mathrm{UCB}(\mathbf{x})$ for the UCB model), and τ is the temperature parameter. Lower values of τ indicate more concentrated probability distributions, corresponding to more precise predictions and therefore less random sampling. All models include τ as a free parameter. Additionally, we estimate λ (generalization parameter) for the Gaussian Process regression model and θ_ϵ^2 (error variance) for the Mean Tracker model. Finally, we estimate β (exploration bonus) for the Upper Confidence Bound sampling strategy.

Cross validation. We fit all combinations of models and sampling strategies—per participant—using cross-validated MLE implemented via a Differential Evolution algorithm for optimization Mullen, Ardia, Gil, Windover, and Cline (2011). Parameter estimates are constrained to positive values in the range $[\exp(-5), \exp(5)]$. We use leave-one-out cross-validation to iteratively form a training set by leaving out a single round, computing a MLE on the training set, and then generating out-of-sample predictions on the remaining round. This is repeated for all combinations of training set and test set. This cross-validation procedure yields one set of parameter estimates per round, per participant, and out-of-sample predictions for 200 choices per participant overall (rounds 2-9 $\times$ 25 choices).

Predictive accuracy. Prediction error (computed as log loss) is summed up over all rounds, and is reported as *predictive accuracy*, using a pseudo-R^2 measure that compares the total log loss prediction error for each model to that of a random model:

$$R^2 = 1 - \frac{\log \mathcal{L}(\mathcal{M}_k)}{\log \mathcal{L}(\mathcal{M}_{\text{rand}})}, \tag{18}$$

where $\log \mathcal{L}(\mathcal{M}_{\text{rand}})$ is the log loss of a random model (i.e., picking options with equal probability) and $\log \mathcal{L}(\mathcal{M}_k)$ is the log loss of model k's out-of-sample prediction error. Intuitively, $R^2 = 0$ corresponds to prediction accuracy equivalent to chance, while $R^2 = 1$ corresponds to theoretical perfect predictive accuracy, since $\log \mathcal{L}(\mathcal{M}_k)/\log \mathcal{L}(\mathcal{M}_{\text{rand}}) \to 0$ when $\log \mathcal{L}(\mathcal{M}_k) \ll \log \mathcal{L}(\mathcal{M}_{\text{rand}})$.

Parameter comparison. We report the mean parameter estimates per participant, averaged over all 8 rounds in Table 1. Because the optimization routine does not always converge perfectly, Figures 1E, 7, 10, and 11 visualizing the parameter estimates exclude outliers higher than 5. However, all parameter estimates were included in the statistical analyses. The presence of these outliers motivated us to use non-parametric tests (without removing the outliers) to compare the different parameters across age groups, in order to achieve more robustness. For the GP-UCB estimates, 1.1% of the λ-estimates, for 3.4% of the β-estimates, and 2.7% of the τ-estimates have been removed this way. However, the significance of our test results do not change even if we use parametric t-tests, but remove outliers before performing these tests. Specifically, after removing values higher than 5, we find that adults show higher λ-estimates than older children ($t(103) = 3.88$, $p < .001$, $d = 0.76$, $BF > 100$), who do not differ in their λ-estimates from younger children ($t(108) = 1.00$, $p = .32$, $d = 0.19$, $BF = 0.3$). Moreover, the β-estimates are higher for older children than for adults ($t(103) = 4.41$, $p < .001$, $d = 0.87$, $BF > 100$), but do not differ between the two groups of children $t(108) = 1.41$, $p = .160$, $d = 0.27$, $BF = 0.5$). Crucially, the random exploration parameters τ do neither differ between older children and adults ($t(103) = .55$, $p = .58$, $d = 0.11$, $BF = 0.2$), nor do they differ between the two groups of children ($t(108) = 1.74$, $p = .08$, $d = 0.33$, $BF = 0.79$).

Notice that our criterion of outlier removal is less strict than other criteria. For example, Tukey proposed removing values which are higher than 1.5×IQR. In our data set, this would lead to a removal of 1.8% of the λ-estimates, 5.1% of the β-estimates, and 11.2% of the τ-estimates. For completeness, we also check if the parameter differences change when applying this more aggressive criterion. After performing a Tukey outlier removal, adults still show a higher λ-estimate than older children ($t(102) = 3.87$, $p < .001$, $d = 0.76$, $BF > 100$), who do not differ from younger children in their λ-estimates ($t(108) = 1.04$, $p = .29$, $d = 0.20$, $BF = 0.3$). The β-estimates are again higher for older children than for adults ($t(100) = 4.69$, $p < .001$, $d = 0.93$, $BF > 100$), and do not differ between the two groups of children ($t(103) = 1.06$, $p = .29$, $d = 0.21$, $BF = 0.3$). Finally, the estimates of the τ-parameter do neither differ between adults and older children ($t(103) = 1.36$, $p = .17$, $d = 0.26$, $BF = 0.5$), nor between the two groups of children ($t(108) = 0.96$, $p = .34$, $d = 0.18$, $BF = 0.3$). Taken together, we believe that our results robustly speak for the existence of strong developmental differences in generalization and directed exploration, but lack reliable evidence in support of differences in random exploration. These results hold no matter which method of outlier removal is applied.

Full model comparison results

Figure 4 shows the log-scaled Bayes Factor of every combination of each learning model with each sampling strategy, compared in terms of predictive accuracy to every other combination, and separated by the different age groups. Results

are based on one-sided Bayesian t-tests. We find that the GP-UCB model wins against every other combination, resulting in large Bayes factors overall, but also for every individual age group.

In addition to these Bayes factors for comparisons between individual models, we also compare the cross-validated log-loss at the group level using (predictive) Bayesian model selection, estimating each model's protected probability of exceedance, which is the likelihood that the proportion of participants generated using a given combination exceeds the proportion of participants generated using all other combinations, while controlling for the chance rate Stephan, Penny, Daunizeau, Moran, and Friston (2009). This analysis revealed that the overall hypothesis of all models performing equally well was rejected by a Bayesian omnibus test at $p < .001$. Moreover, the protected probability of exceedance was virtually 1 for the GP-UCB model, where 1 is the upper theoretical limit. This result is also obtained if breaking down the analysis by the different age groups, again always leading to a protected probability of exceedance of 1 for the GP-UCB model.

Finally, we report how many participants are best predicted by each combination of a model and a sampling strategy. The results of this comparison are shown in Figure 5 (see also Table 1). As before, the GP-UCB model is the best overall model and also predicts most participants best for every age group. Overall, the GP-UCB model predicted 106 out of 160 participant best. In the different age categories, the GP-UCB model best predicted 32 out of 55 younger children, 44 out of 55 older children, and 30 out of 50 adult participants. Therefore, the GP-UCB model was by far the best model in our comparison.

Other forms of random exploration

A softmax function with a temperature parameter τ is only one way to define random exploration. Another approach towards assessing random exploration is ϵ-greedy exploration. Given k number of arms (64 in our experiment), epsilon-greedy exploration chooses

$$p(\mathbf{x}) = \begin{cases} 1 - \epsilon, \text{if} \arg\max UCB(\mathbf{x}), \\ \epsilon/(k-1), \text{otherwise.} \end{cases} \tag{19}$$

We test the ϵ-greedy method of exploration by using it instead of a softmax function in combination with the GP regression model and a UCB-sampling strategy. The results of this comparison show that the ϵ-greedy exploration model was systematically worse at predicting behavior than the softmax model reported in the main text (mean predictive accuracy: $R^2 = 0.21$, $t(159) = 6.67$, $p < .001$, $d = 0.53$, $BF > 100$). Additionally the softmax model also had better predictive accuracy than the ϵ-greedy exploration model for adults (mean predictive accuracy: $R^2 = 0.26$, $t(49) = 9.29$ $p < .001$, $d = 1.31$, $BF > 100$), and for older children (mean predictive accuracy: $R^2 = 0.21$, $t(54) = 3.60$, $p < .001$, $d = 0.49$, $BF = 38.5$), but not for younger children (mean predictive accuracy: $R^2 = 0.17$, $t(54) = 0.33$, $p = .74$, $d = 0.04$, $BF = 0.2$).

Next, we looked for age-related differences in the parameter estimates of the ϵ-greedy model, specifically the directed exploration parameter β and the alternative random exploration parameter ϵ. Notice, however, that interpreting estimates of inferior computational models can be problematic and should only be done with caution.

As in the softmax-parameterized models, we find larger λ-estimates for adults than for older children (λ-estimates: 0.99 vs. 0.24, $U = 1975$, $r_\tau = 0.31$, $p < .001$, $BF > 100$), whereas the two children groups do not differ in their λ-estimates (λ-estimates: 0.31 vs. 0.24, $U = 1299$, $r_\tau = 0.10$, $p < .20$, $BF = 0.4$). Furthermore, we find more directed exploration (i.e., larger β parameters) for older children than for adults (β-estimates: 17.30 vs. 5.38, $U = 555$, $r_\tau = 0.42$, $p < .001$, $BF > 100$), but found no difference between the two groups of children (β-estimates: 17.20 vs. 17.30, $U = 1684$,$r_\tau = 0.0$, $p = .3$, $BF = 0.27$). We also found a smaller difference in ϵ-greedy exploration between adults and older children (ϵ-estimates: 0.00012 vs. 0.00014, $U = 960$,$r_\tau = 0.21$ $p = .007$, $BF = 6.95$), but not between the two groups of children (ϵ-estimates: 0.00014 vs. 0.00016, $U = 1774$, $r_\tau = 0.12$, $p = .11$, $BF = 0.46$). Notice that the relative proportion of random exploration decisions according to the ϵ-parameter estimates is so small, that within the 200 choices we are modeling this accounts for a difference of exactly 1 in every 250 choices, which means that there is no actual difference in participants' ϵ-parameters for most practical matters. The overall age-related effect in this analysis was larger for directed exploration than for ϵ-greedy exploration ($r_\tau = 0.40$ vs. $r_\tau = 0.25$). Thus, there are two reasons to believe that children are driven more strongly by directed than by random exploration. Firstly, the GP-UCB model combined with a softmax formulation of random exploration predicted participants better than an ϵ-greedy model, and finds no age-related difference in terms of random exploration described by the temperature parameter τ. Secondly, parameter estimates of the ϵ-greedy model find only small and practically meaningless age-related difference in ϵ-exploration, but again large age-related differences in the directed exploration parameter β.

Model recovery

We present model recovery results that assess whether or not our predictive model comparison procedure allows us to correctly identify the true underlying model. To assess this, we generated data based on each individual participant's mean parameter estimates (excluding outliers larger than 5 as before). More specifically, for each participant and round,

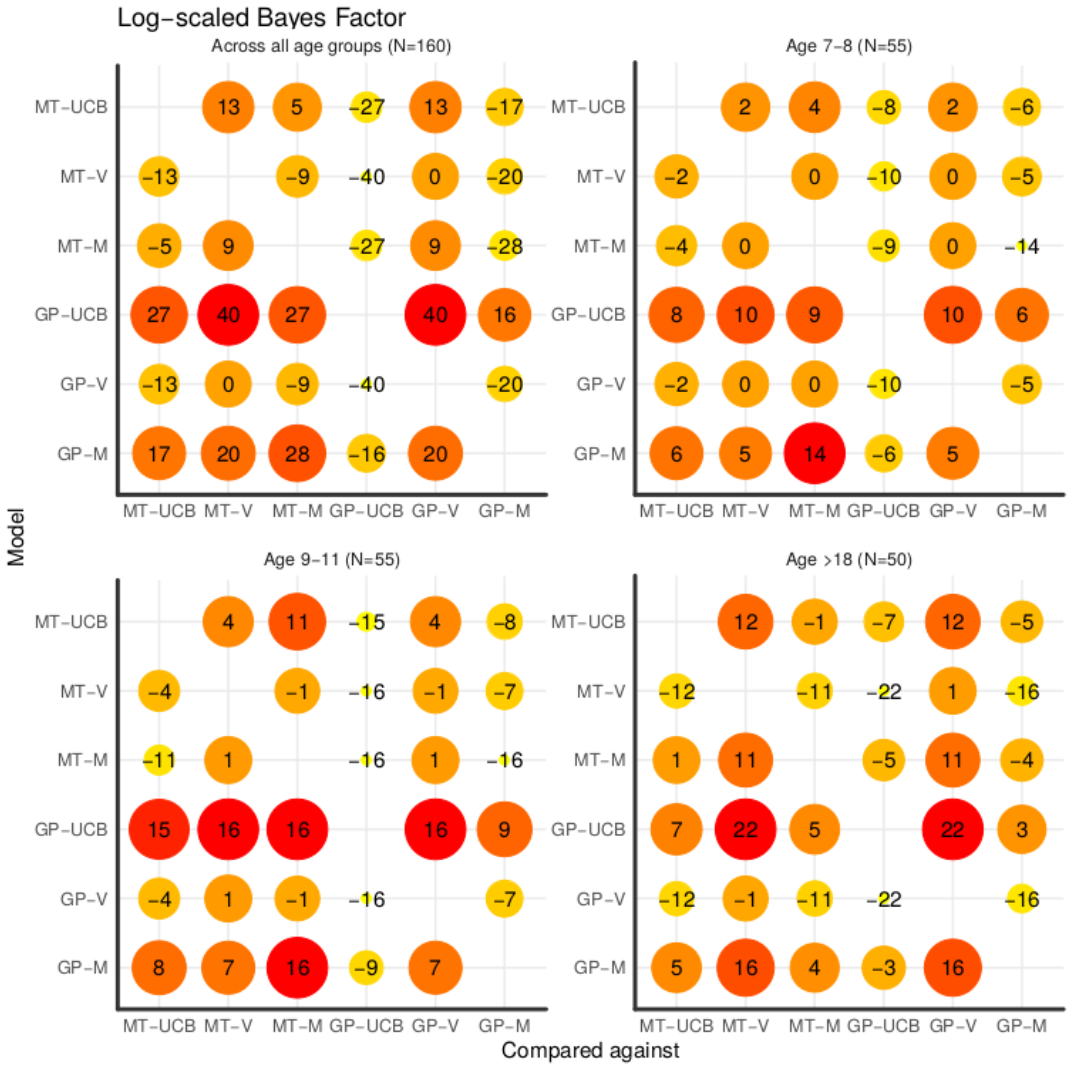

Figure 4. Log-scaled Bayes factor for every pair-wise model comparison, between a model on the y-axis compared against a model on the x-axis. Numbers indicate $\log_{10}$ of the Bayes Factor, with a value of 1 indicating that the alternative hypothesis is around 10-times more likely than the null hypothesis, whereas a value of -1 means that the null hypothesis is around 10-times more likely than the alternative hypothesis. Larger Bayes Factors are accompanied by larger circles and darker shades of red. Comparisons show both learning models—the Gaussian Process (GP) and the Mean Tracker (MT)—combined with every sampling strategy, Upper Confidence Bound sampling (UCB), Variance-greedy sampling (V), and Mean-greedy sampling (M).

we use the cross-validated parameter estimates to specify a given model, and then generate new data in the attempt to mimic participant data. We generate data using the Mean Tracker and the GP regression model. In all cases, we use the UCB sampling strategy in conjunction with the specified learning model. We then utilize the same cross-validation method as before in order to determine if we can successfully identify which model has generated the underlying data. Figure 6 shows the cross-validated predictive performance for the simulated data.

Recovery Results

Our predictive model comparison procedure shows that the Gaussian Process model is a better predictor for data generated from the same underlying model, whereas the Mean Tracker model is only marginally (if at all) better at predicting data generated from the same underlying model. This suggests that our main model comparison results are robust to Type II errors, and provides evidence that the better predictive accuracy of the GP model on participant data is unlikely

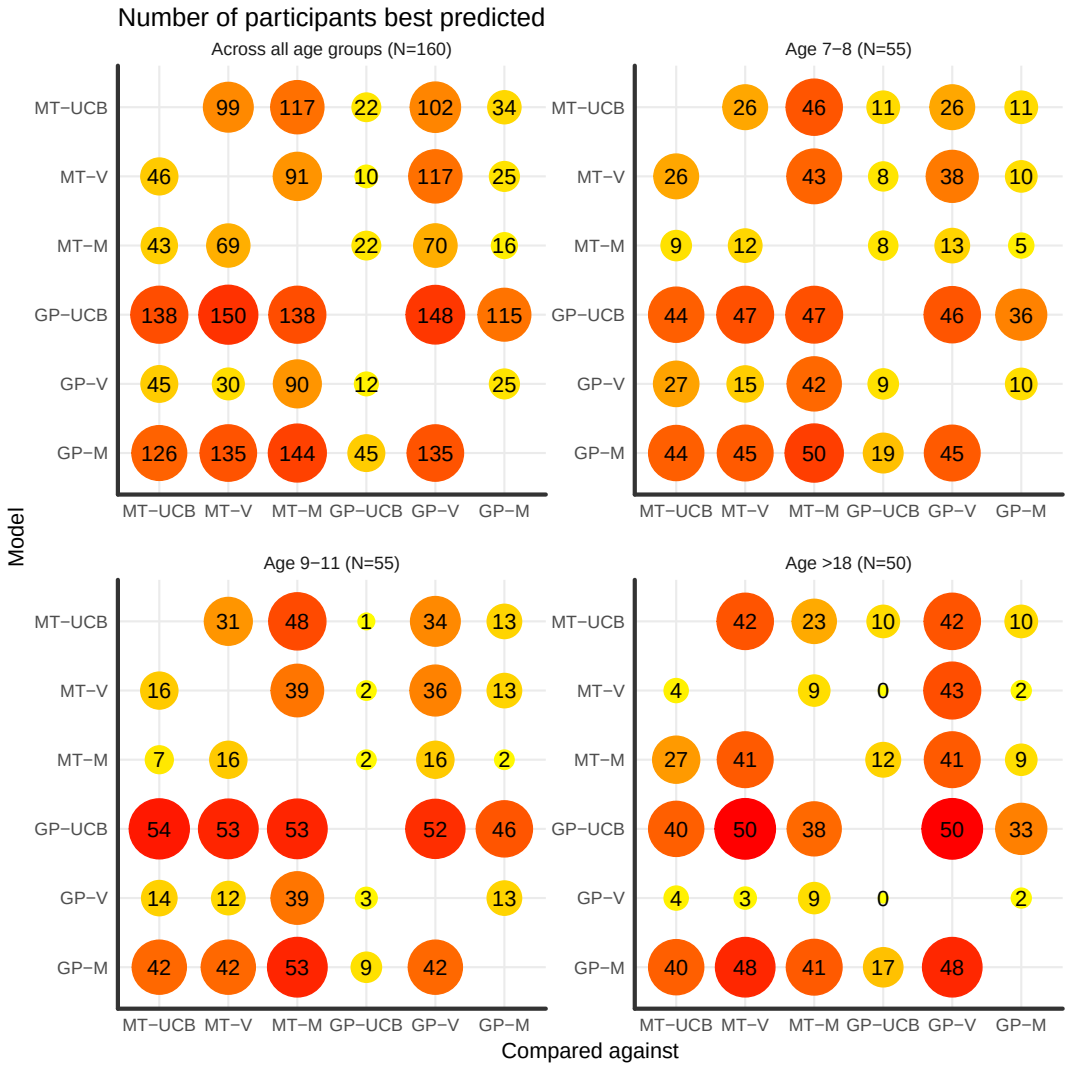

Figure 5. Number of participants best predicted for every pair-wise model comparison, comparing a model on the y-axis with a model on the x-axis. Numbers show absolute counts and can only be interpreted in relation to the overall number of participants per group (shown in brackets at the top of each panel). A larger proportion of participants best predicted are accompanied by larger circles and darker shades of red. Comparisons show both models—the Gaussian Process (GP) and the Mean Tracker (MT)— combined with every sampling strategy, Upper Confidence Bound sampling (UCB), Variance-greedy sampling (V), and Mean-greedy sampling (M).

due to differences in model mimicry.

When the Mean Tracker model generates data using participant parameter estimates, the same Mean Tracker model performs better than the GP model ($t(159) = 4.55$, $p < .001$, $d = 0.37$, $BF > 100$) and predicts 112 out of 160 simulated participants best. Notice, however, that both models perform poorly in this case, with both models achieving an average pseudo-r-squared of around $R^2 = 0.02$. This also shows that the Mean Tracker is not a good generative model of human-like behavior in our task.

When the Gaussian Process regression model has generated the underlying data, the same model performs significantly better than the Mean Tracker model ($t(159) = 18.8$, $p < .001$, $d = 1.48$, $BF > 100$) and predicts 153 of the 160 simulated participants best. In general, both models perform better when the GP has generated the data with the GP reaching an average pseudo-r-squared of $R^2 = 0.20$ and the MT of 0.12. This makes our finding of the Gaussian Process regression model as the best predictive model even stronger as –technically– the Mean Tracker model can mimic parts of its behavior. Moreover, the resulting values of predictive accuracy are similar to the ones we found using participant data. This indicates that our empirical values of predictive accuracy were as good as possible under the assumption that

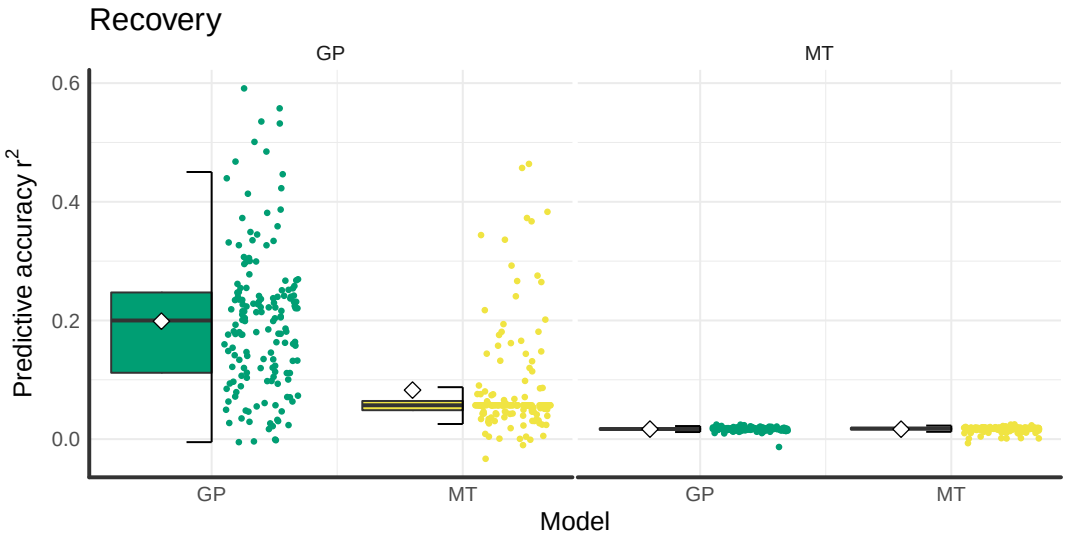

Figure 6. Tukey box plots of model recovery results including individual data points (points) and overall means (diamonds). **Left:** Model performance based on data generated by the GP-UCB model specified with participant parameter estimates. The matching GP-UCB model makes better predictions than a mismatched MT-UCB model. **Right:** Model performance based on data generated by the MT-UCB model specified with participant parameter estimates. Both GP and MT models predict this data poorly, and perform similar to a random model.

a GP model generated the data.

Parameter Recovery

Another important question is whether the reported parameter estimates of the GP-UCB model are reliable and recoverable. We address this question by assessing the recoverability of the three underlying parameters, the length-scale λ, the directed exploration factor β, and the random exploration (temperature) parameter τ of the softmax choice rule. We use the results from the model recovery simulation described above, and correlate the empirically estimated parameters used to generate data (i.e., the estimates based on participants' data), with the parameter estimates of the recovering model (i.e., the MLE from the cross-validation procedure on the simulated data). We assess whether the recovered parameter estimates are similar to the parameters that were used to generated the underlying data. We present parameter recovery results for the Gaussian Process regression model using the UCB sampling strategy. We report the results in Figure 7, with the generating parameter estimate on the x-axis and the recovered parameter estimate on the y-axis.

The correlation between the generating and the recovered length-scale λ is $r_\tau = .85$, $p < .001$, the correlation between the generating and the recovered exploration factor β is $r_\tau = 0.75$, $p < .001$, and the correlation between the generating and the recovered softmax temperature parameter τ is $r_\tau = 0.81$, $p < .001$.

These results show that the correlation between the generating and the recovered parameters is very high for all parameters. Thus, we have strong evidence to support the claim that the reported parameter estimates of the GP-UCB model are recoverable, reliable, and therefore interpretable. Importantly, we find that estimates for β (exploration bonus) and τ (softmax temperature) are indeed recoverable, providing evidence for the existence of a *directed* exploration bonus, as a separate phenomena from *random* exploration in our behavioral data.

Next, we analyze whether or not the same differences between the parameter estimates that we found for the experimental data can also be found for the simulated data. Thus, we compare the recovered parameter estimates from the data generated by the estimated parameters for the different age groups. This comparison shows that the recovered data exhibits the same characteristics as the empirical data. The recovered λ-estimates for simulated data from adults was again larger than the recovered lambda estimates for older children ($U = 2021$, $r_\tau = 0.33$, $p < .001$, $BF > 100$), whereas there was no difference between the recovered parameters for the two simulated groups of children ($U = 1800$, $p = .08$, $r_\tau = 0.14$, $BF = 1$). As in the empirical data, the recovered estimates also showed a difference between age groups in their directed exploration behavior such that the recovered β was higher for simulated older children than for simulated adults ($U = 730$, $p < .001$, $r_\tau = 0.33$, $BF > 100$), whereas there was no difference between the two simulated groups of children ($U = 1730$, $p = .19$, $r_\tau = 0.10$, $BF = 0.6$). There was no difference between the different recovered τ-parameters (max-$BF = 0.5$). Thus, our model can also reproduce similar group differences between generalization and directed exploration as found in the empirical data.

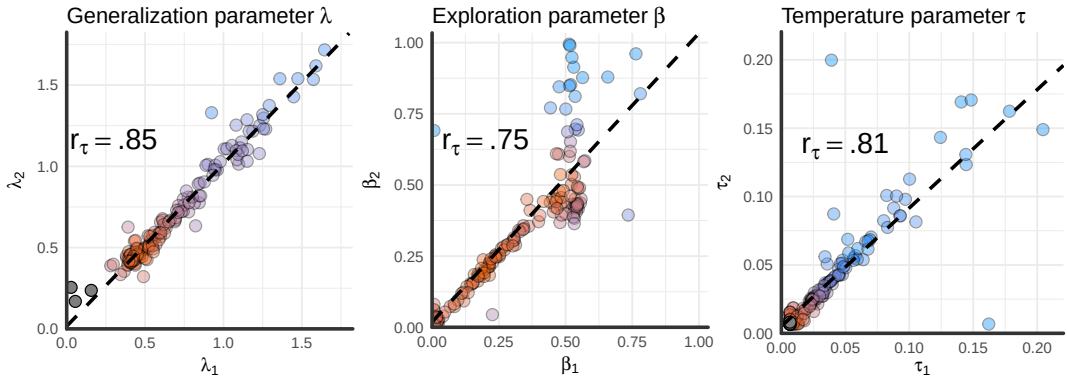

Figure 7. Parameter recovery results. The generating parameter estimate is on the x-axis and the recovered parameter estimate is on the y-axis. The generating parameter estimates are from the cross-validated participant parameter estimates, which were used to simulate data (see Model recovery). Recovered parameter estimates are the result of the cross-validated model comparison (see Model comparison) on the simulated data. While the cross-validation procedure yielded 8-estimates per participant, one for each round, we show the mean estimate per (simulated) participant. The dashed line shows a linear regression on the data, while Kendall's rank correlation r_τ is shown in the plot. For readability, colors represent the bivariate kernel density estimate, with red indicating higher density.

Counter-factual parameter recovery

Another explanation of the finding that children differ from adult participants in their directed exploration parameter β but not in their random exploration parameter τ could be that the softmax temperature parameter τ can sometimes track more of the random behavioral difference between participants than the directed exploration parameter β. If the random exploration parameter τ indeed tends to absorb more of the random variance in the data than the directed exploration parameter β, then perhaps one is always more likely to find differences in β rather than differences in τ. To assess this claim, we simulate data using our GP-UCB model as before but swap participants' parameter estimates of β with their estimates of τ and vice versa. Ideally, this simulation can reveal whether it is possible for our method to find differences in τ but not β in a counter-factual parameter recovery where the age groups differ in their random but not their directed exploration behavior. Thus, we generate data from the swapped GP-UCB model and then use our model fitting procedure to assess the GP-UCB-model's parameters from this generated data. The results of this simulation reveal that simulated adults do not differ from simulated older children in their estimated directed exploration parameters β ($U = 1170$, $r_\tau = 0.11$, $p = .18$, $BF = 0.7$). Furthermore, the two simulated children groups also do not differ in terms of their directed exploration parameter β ($U = 1696$,$r_\tau = 0.09$, $p = .27$, $BF = 0.4$). However, the random exploration parameter τ is estimated to be somewhat higher for simulated older children than for simulated adults ($U = 1771$, $r_\tau = 0.21$, $p = .01$, $BF = 2.5$) and shows no difference between the two simulated children groups ($U = 1631$, $p = .48$, $r_\tau = 0.06$, $BF = 0.4$). This means that the GP-UCB model can pick up on differences in random exploration as well and therefore that our findings are unlikely due to a false positive.

Further Behavioral Analysis

Learning over trials and rounds

We analyze participants learning over trials and rounds using a hierarchical Bayesian regression approach. Formally, we assume the regression weight parameters θ_x for $x \in \{0, 1, 2\}$ are hierarchically distributed as

$$\theta_x \sim \mathcal{N}(\mu_x, \sigma_x); \tag{20}$$

we further assume a weakly informative prior of the means and standard deviation over the regression equation, defined as:

$$\mu_x \sim \mathcal{N}(0, 100) \tag{21}$$

$$\sigma_x \sim \text{Half-Cauchy}(0, 100) \tag{22}$$

We fit one hierarchical model of means and standard deviations over all participants using Hamiltonian Markov chain Monte Carlo sampling as implemented in the PyMC-environment (Patil, Huard, & Fonnesbeck, 2010). This yields hierarchical estimates for each regression coefficient overall as well as individual estimates for each participant. Doing so for a model containing an intercept (mean performance), a standardized coefficient of the effect of trials on rewards, as well as a standardized coefficient of the effect of rounds on rewards results into the posterior distributions shown in Figure 8.

The overall posterior mean of participants' rewards is estimated to be 34.4 with a 95%-credible set of $[33.5, 35.4]$. The standardized effect of trials onto rewards is estimated to be 1.48 with a 95%-credible set of $[1.1, 1.9]$, indicating

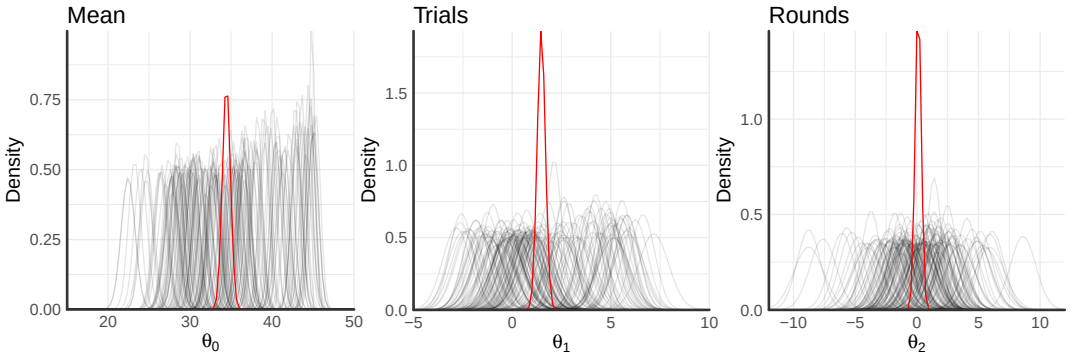

Figure 8. Posterior regression coefficients of mean performance and the effects of both trials and rounds onto participants' rewards. Individual lines densities correspond to participant-wise estimates, whereas red lines show hierarchical estimates.

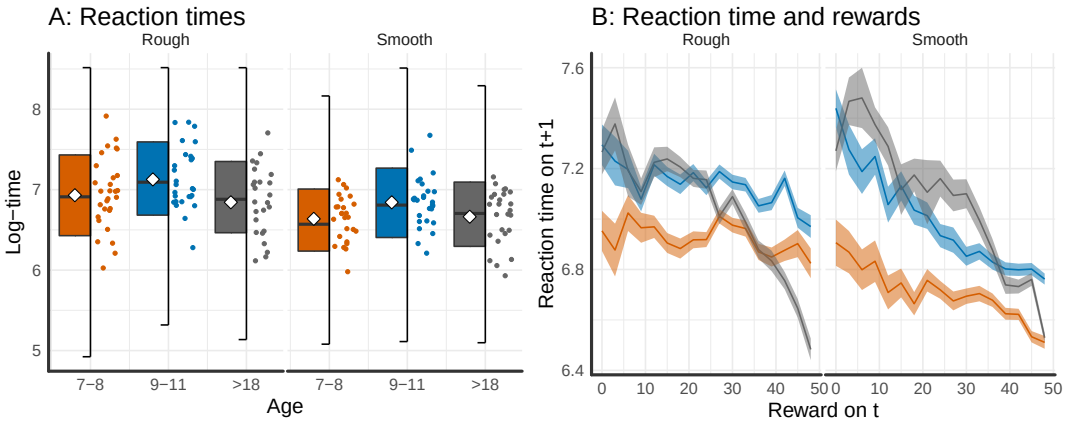

Figure 9. Log-reaction times. **A:** Reaction time by age group and condition. **B:** Reaction times as a function of previous reward.

a strong effect of learning over trials. The standardized effect of rounds onto participants' rewards is estimated to be 0.11 with a 95%-credible set of $[-0.4, 0.6]$, indicating no effect. Taken together, these results indicate that participants performed well above the chance-level of 25 overall and improved their score greatly over trials. They did not, however, learn or adapt their strategies across rounds.

Reaction times

We analyze participants log-reaction times as function of previous reward, age group, and condition. Reaction times are filtered to be smaller than 5000ms and larger than 100ms (3.05% removed in total). We find that reaction times are larger for the rough as compared to the smooth condition (see Fig. 9A; $t(158) = 4.44$, $p < .001$, $d = 0.70$, $BF > 10$). Moreover, adults are somewhat faster than children of age 9-11 ($t(103) = -2.60$, $p = .01$, $d = 0.51$, $BF = 4$). The difference in reaction times between children of age 9-11 and children of age 7-8 is only small ($t(108) = 2.27$, $p = .03$, $d = 0.43$, $BF = 1.97$).

We also analyze how much a previously found reward influences participants' reaction times, i.e. whether participants slow down after a bad outcome and/or speed up after a good outcome (see Fig. 9B). The correlation between the previous reward and reaction times is negative overall with $r = -0.18$, $t(159) = -15$, $p < .001$, $BF > 100$, indicating that larger rewards lead to faster reaction times, whereas participants might slow down after having experienced low rewards. This effect is even stronger for the smooth as compared to the rough condition ($t(158) = -4.43$, $p < .001$, $d = 0.70$, $BF > 100$). Moreover, this effect is also stronger for adults than for older children ($t(103) = -5.51$, $p < .001$, $d = 1.08$, $BF > 100$) and does not differ between the two groups of children ($t(108) = 1.93$, $p = .06$, $d = 0.37$, $BF = 1.05$).

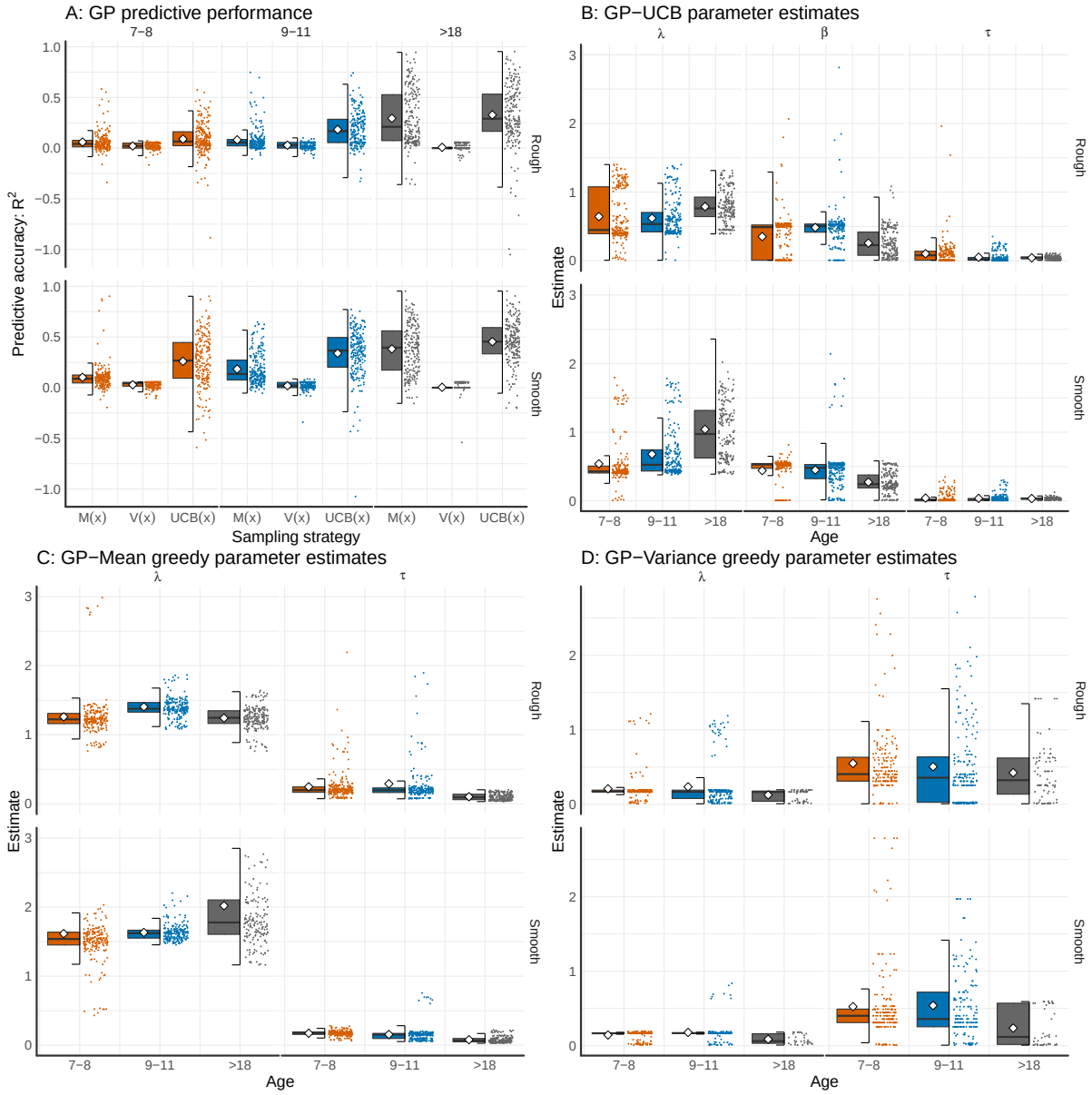

Figure 10. Pseudo-r-squared and parameter estimates for the different Gaussian Process-models by age group and condition. All Tukey box plots include raw data points and mean shown as diamond.

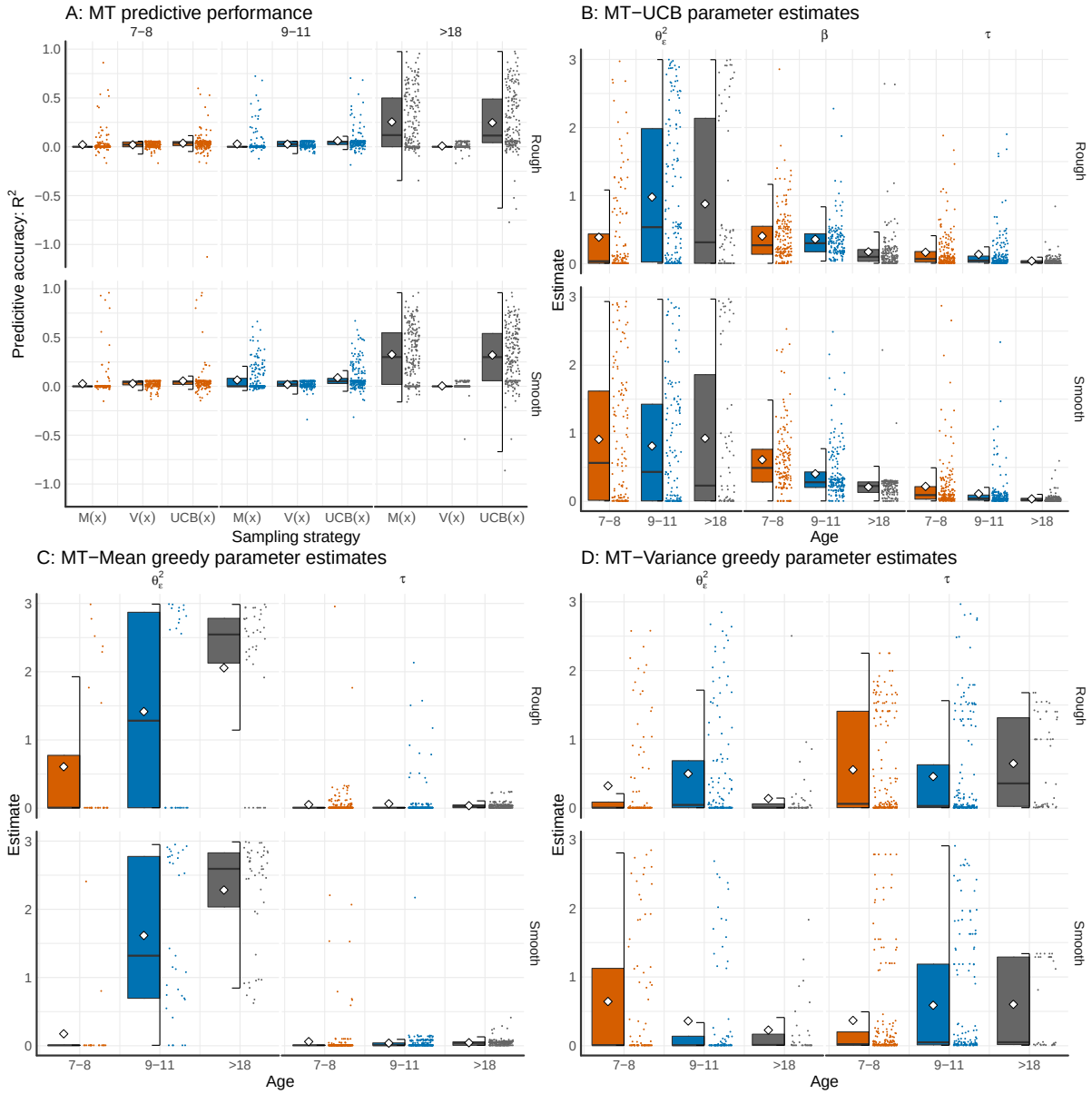

Figure 11. Pseudo-r-squared and parameter estimates for the different Mean Tracker-models by age group and condition. All Tukey box plots include raw data points and mean shown as diamond.

Table 1

Modeling Results. Parameter estimates higher than 5 have been excluded. Estimates are based on medians over means of participants' parameter estimates. Columns indicate (from left to right) the average predictive performance of each model (R^2), how many participants each model predicted best (#best predicted), the average log-loss over all 8 rounds of predictions (Log-loss), the protected probability of exceedance (Pr. of exceed.), the GP's generalization parameter (λ), the directed exploration parameter (β) for all models involving a UCB sampling strategy, the MT's error variance ($\sqrt{\theta_\epsilon^2}$), and the random exploration (softmax temperature) parameter (τ).

Model	Pseudo-R^2	#best predicted	Log-loss	Pr. of exceed.	Generalization λ	Exploration β	Error variance $\sqrt{\theta_\epsilon^2}$	Softmax τ
Overall:								
MT-Mean greedy	0.02	10	103.9	0	–	–	2.98	0.01
MT-Variance greedy	0.01	3	102.7	0	–	–	0.58	0.08
MT-UCB	0.05	5	99.3	0	–	0.41	1.60	0.16
GP-Mean greedy	0.09	34	94.4	0	1.46	–	–	0.16
GP-Variance greedy	0.02	2	102.1	0	0.18	–	–	0.56
GP-UCB	0.28	106	74.8	1	0.59	0.45	–	0.03
Age 7-8:								
MT-Mean greedy	0.00	3	103.9	0	–	–	0.53	0.01
MT-Variance greedy	0.00	3	100.8	0	–	–	1.22	0.08
MT-UCB	0.03	3	100.5	0	–	0.47	1.20	0.11
GP-Mean greedy	0.07	14	96.9	0	1.37	–	–	0.18
GP-Variance greedy	0.01	0	102.7	0	0.17	–	–	0.41
GP-UCB	0.14	32	89.2	1	0.46	0.48	–	0.03
Age 9-11:								
MT-Mean greedy	0.01	0	103.9	0	–	–	2.86	0.01
MT-Variance greedy	0.01	0	101.7	0	–	–	0.65	0.08
MT-UCB	0.05	0	99.3	0	–	0.33	1.34	0.07
GP-Mean greedy	0.08	9	95.1	0	1.52	–	–	0.17
GP-Variance greedy	0.01	2	101.5	0	0.17	–	–	0.38
GP-UCB	0.26	44	76.6	1	0.74	0.53	–	0.02
Adults:								
MT-Mean greedy	0.29	7	72.8	0	–	–	3.38	0.03
MT-Variance greedy	0.00	0	103.4	0	–	–	0.13	0.87
MT-UCB	0.29	2	74.2	0	–	0.18	2.94	0.03
GP-Mean greedy	0.33	11	69.1	0	1.38	–	–	0.07
GP-Variance greedy	0.01	0	103.9	0	0.10	–	–	0.23
GP-UCB	0.40	30	62.7	1	0.85	0.24	–	0.03