To observe or to bet? Investigating purely exploratory and purely exploitative actions in children, adults, and computational models.

Eunice Yiu (ey242@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704 USA

Kai Sandbrink (kai.sandbrink@lmh.ox.ac.uk)
Department of Experimental Psychology, University of Oxford,
Oxford, United Kingdom

Alison Gopnik (gopnik@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704 USA

Abstract

Autonomous agents often need to decide between choosing actions that are familiar and have previously yielded positive results (exploitation) and seeking new information that could help uncover more effective actions (exploration). We present an "observe or bet" task that separates "pure exploration" from "pure exploitation": 75 five-to-seven-year-old children, 60 adults and computational agents have to decide either to observe an outcome without reward, or to bet on an action without immediate feedback at varying probability levels. Their performances were measured against solutions from the partially observable Markov decision process and meta-RL models. Children and adults tended to choose observation more than both algorithm classes would suggest. Children also modulated their betting policy based on the probability structure and amount of evidence, exhibiting "hedging behavior" a strategy not evident in standard bandit tasks. The results provide a benchmark for reasoning about reward and information in humans and neural network models.

Keywords: decision-making, exploration, probabilistic learning, reinforcement learning

Introduction

From hunting and foraging to achieving complex skills and tasks, agents need to autonomously search through a vast space of possible actions. As a result, an agent must strike a fine balance between the exploration of different options or opportunities and the exploitation of rewards (Cohen et al., 2017; Cook et al., 2013). This balance is commonly referred to as the exploration-exploitation trade-off. Understanding the specific kinds of heuristics and strategies that humans employ to solve this problem over the course of their development remains an open question in cognitive science.

Researchers have long argued that children are active and exploratory information seekers (e.g., Gopnik, 2020; Schulz, 2012; Piaget, 2013). However, previous studies used environments in which the reward and information that participants receive on each trial are confounded (e.g., Giron et al., 2023; Liquin & Gopnik, 2022; Meder et al., 2021; Schulz et al., 2019). In these experiments, exploratory actions lead to reward and exploitative moves result in information gain. There is no clear test that investigates how children select between reward and information when they are presented as independent options across varying probabilities. We investigated the behavior of children in a setting where "pure exploration" (i.e. actions that do supply any reward at all) was juxtaposed with "pure exploitation" (i.e. actions that do not supply any information at all). Previous work has shown that adults in similar versions of this task initially also observe more than is optimal (Tversky & Edwards, 1966), but can learn near-optimal exploratory behavior over several repetitions in the task (Navarro et al., 2016).

In the current studies we add child participants to investigate whether children will be particularly exploratory in these tasks. We also differentiate between two kinds of betting actions, those that correspond to the option that has received the strongest evidence so far (which would be chosen to maximize the expected value of reward) and the alternative option that goes against the current evidence which we call "hedging": occasionally choosing this option on some trials will lead to a pattern more like "probability matching" than "maximizing expected value," which can act as a "hedge" against future changes in the outcomes (Siegel and Goldstein, 1959; Gassmaier and Schooler, 2008; Rivas, 2013).

The study is the first to disambiguate the motives underlying exploratory and exploitative behavior in human children, in an observe vs bet task, and to make direct comparisons with human adults and state-of-the-art computational models on a level playing field.

Methods: Human Experiments

Participants. 75 child participants aged between 5 and 7 years old (M_{age} = 6.05 years, SD = .85, 43 females) were recruited and tested on Zoom. Four additional children were tested but excluded from the sample analysis as they either did not pass the comprehension checks (n = 2) or did not complete all the test trials (n = 1). This could be due to inattention or inability to understand the task. We

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

target this age group as our pilot study suggested that only children aged 5 and above could reliably comprehend the complexity of the task. In addition, we also recruited 60 adult participants aged 20 to 36 years old (M_{age} = 28.43 years, SD = 4.79, 29 females) on Prolific. The study was preregistered on AsPredicted [https://aspredicted.org/TG5_Q8B for children and https://aspredicted.org/BC8_YJM for adults].

Stimuli and Procedure. Following the structure of the observe or bet task (Tversky & Edwards, 1966), we presented participants with a computer game featuring a rewarding character and a non-rewarding character. Each character hid behind a separate door. Participants received one virtual coin if they found the rewarding character (a "kind elf" or a "kind princess") and did not gain or lose anything if they found the other non-rewarding character (a "mean monster" or a "mean thief"). They were explicitly given the goal of winning as many coins as possible. In every trial, participants were then given the option of either observing which doors the characters were hiding behind, or placing a bet on one of two probabilistically rewarding doors, without receiving feedback until the end of the experiment (Figure 1). Exactly one of the two doors paid out on every trial. Throughout the trials, the underlying payout probabilities remained constant. However, the payout probabilities varied between participants. We randomly assigned 25 children and 20 adults to the setting where the payout probability of the higher-paying door ρ was 1.0, 25 children and 20 adults to ρ = 0.75, and 25 children and 20 adults to ρ = 0.5. Although participants were not given the exact quantitative probability, they received a verbal description of their assigned environment: the environment was (i) "always the same" (ρ = 1.0), (ii) having a preferably higher-paying door even though it might "sometimes change" (ρ = 0.75), or (iii) "always changing, no one can tell" which door was higher-paying (ρ = 0.5). Prior to the experiment, participants watched narrated videos that explained the instructions of the game. Next, they had to answer ten questions from four rounds of comprehension checks correctly before they could proceed. If the participant failed to answer any one of the ten questions correctly, they would not be able to move on.

Practice Trials. Participants played a practice game consisting of 4 trials to familiarize themselves with the setup. These practice trials had the same setup as the subsequent test trials except that they involved visually different characters (e.g., a kind elf and an evil monster). Participants received feedback on the actual outcomes and reward they had accumulated at the end of the practice. The practice trials ensured that participants' decision-making was informed and reflective of the verbally described probabilistic scenarios.

Test Trials. Participants then proceeded to play the actual game, consisting of 12 test trials, which included two new characters who were visually different than the practice trials. They were told that this game had the *same* probability structure as the practice game. At the end of the test trials, they were asked to quantitatively express their perceived probability of receiving a reward from the left door versus the right using a slider. Feedback regarding the number of coins participants accumulated in the test trials was provided only after the participants stated their estimated probabilities. This ensures that participants' estimations were based on their exploratory behavior rather than the ground-truth results.

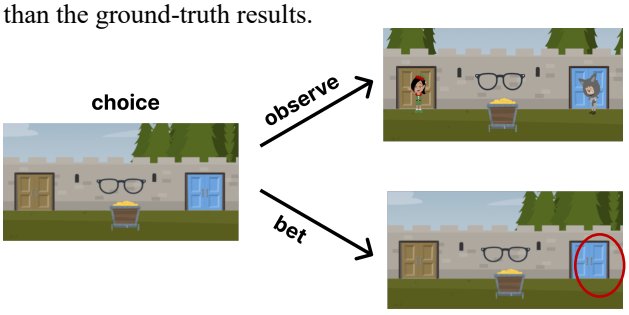

Figure 1. An example of a trial. Participants choose between "betting" with feedback delayed (outcome revealed at the end of the game) or "observing." Betting circles the chosen door without further feedback; observing reveals the respective locations of the kind character (left door) and the mean character (right door).

Methods: Computational Modelling

In this study, we employ a partially observable Markov decision process solver, which operates with precise probability levels, juxtaposed with a meta-reinforcement learning (meta-RL) model that does not have such explicit probabilistic information. These computational models serve as benchmarks for human behavior. Human participants were not provided with exact probability values, but they were informed in qualitative terms that conditions may either never change, sometimes change, or always change. We hypothesize that human performance will surpass that of the meta-RL model due to this inferred knowledge, but human estimations may not achieve the numerical exactitude that characterizes the POMDP solver's performance.

Formulation of the partially observable Markov decision process. To quantify optimal performance on the task, we first formulated the problem as a partially observable Markov decision process (POMDP, Åström, 1965; Kaelbling et al., 1998). A POMDP is a 7-tuple $\langle S, A, \Omega, O, T, r, \gamma \rangle$, where S is the (finite) non-empty state space, A is the (finite) non-empty action space, Ω is the (finite) non-empty observation space, $O : S \rightarrow P(\Omega)$ is the observation function, $T : S \times A \rightarrow P(S)$ is the probabilistic state-transition function, $r : S \times A \rightarrow P(R)$ is a bounded reward function, and $0 \leq \gamma \leq 1$ is the discount factor. We formalize the observe or bet task by defining the set of states as the product space given by the number of steps along with the two possibilities for which is the high-paying door. The set of actions is to observe, to bet on the left door, or to bet on the right door. The set of observations are given by the product of the set of the number of steps and the possible observations per step, no observation, observing a payout on the left door, and

observing a payout on the right door. We set the discount factor $\gamma := 1$.

As an upper bound of performance, we calculated the reward-maximizing policy for an agent aware of the probability structure of the task by using the JuliaPOMDP framework (Egorov et al., 2017) to calculate successive approximations of the reachable state under optimal policies (SARSOP, Kurniawati et al, 2008), a state-of-the-art solver for problems that require active information gathering (Ma & Pineau, 2015; Silver & Veness, 2010). We calculated for every trial the belief threshold at which one should switch from observing to betting for each probability setting.

Neural network architecture and training procedure. We also trained deep RL meta-agents on different probability levels of the same task and thus were unsure of the payout probability (Sandbrink & Summerfield, 2023). The neural networks had a standard architecture with an input layer, followed by an LSTM layer of 48 units, a fully connected layer of 24 units, and a softmax output layer of 3 units that correspond to the three possible actions. We used ReLU activation functions for the hidden layers. The state encoding at the input contained the following elements: one-shot encoding of the action chosen on the previous time step, the time remaining in the trial (scaled between 1 and 0, with 1 corresponding to the first time step in an episode), zero-to-one-shot feedback corresponding to the observation on the two doors (1 if the rewarding character was observed at the door on the previous time step; 0 if either that the agent did not observe or that the agent observed but this door did not contain the rewarding character).

We trained the neural network using the REINFORCE algorithm (Sutton et al., 1999) with a baseline of 1/3 (corresponding to the expected value of a random action) following the meta-reinforcement learning procedure (Duan et al., 2016; Wang et al., 2016; Wang et al., 2018). We meta-train the networks across the distribution of POMDPs defined by sampling $\rho \sim U[0.5, 1]$. To avoid biasing in a particular direction for comparing with the human data, we did not hold out any area of the training region. We trained the networks for 500000 episodes using a batch size of 50. The recurrent units of the LSTM layer were reset to 0 at the start of a new episode. We used the Adam optimizer with a learning rate of 1e-3. We started training with entropy regularization with coefficient 5, which we annealed to 0 geometrically over the course of 150000 episodes. We ran five instantiations of the RL neural networks, which learned to perform near-optimally on the task (Figure 2).

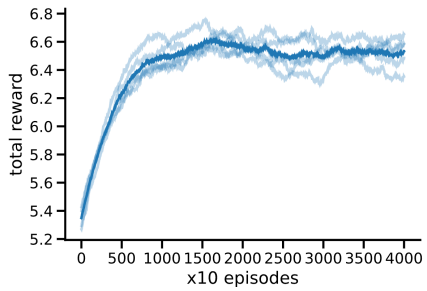

Figure 2: Learning curves for the task-driven five RL neural network models for (thick: aggregated), smoothed with a moving average window over 1000 episodes.

Fitting human behavior to computational process models. To characterize differences in behavior, we fitted the computational process model that Navarro et al. (2016) found explains adult human behavior best out of four possible candidates. This model posits that people use heuristics to approximate the ideal solution to the task, by keeping track of how much evidence relatively they have accumulated for the two doors, and switching from observing to betting once this evidence tally crosses a threshold that depends on the number of trials left in the episode. Specifically, on each trial, participants receive observations $x_t \in \{+1, 0, -1\}$ for observing the left door, betting, and observing the right door. They then update their evidence tally for trial t for a particular door based on $e_t = x_t + (1 - \alpha) \times e_{t-1}$ based on a forgetfulness factor α. The decision threshold at each trial is a piecewise linear function with initial value d_0 that is constant initially but beginning in trial c decreases linearly to a terminal value d_1 at the final trial. Finally, participants are modelled as selecting an action stochastically based on the interaction of these different terms. The probability of betting on the left door is given by:

$$P(\text{bet left}) = \Phi\left(\frac{e_t - d_t}{\sigma}\right)$$

where σ is a response stochasticity parameter and Φ is the cumulative distribution function of the standard normal distribution (Blanchard & Gershman, 2018). Following previous works (Navarro et al., 2016; Blanchard & Gershman, 2018), we fit five parameters $\alpha, d_0, d_1, c,$ and σ using hierarchical Bayesian probabilistic programming. Since we use the model in a stationary task, the normative decay parameter is zero, while the threshold depends on the probability structure of the environment, leading to the same normative predictions of front-loaded observations.

Results

Humans over-explore and do not modulate observations according to probability levels. High probability values for the high-paying door (i.e.., ρ closer to 1.0) in the environment mean that one should be more certain about which is the correct door before switching from observing to betting. Since the difference in payout rates between the two doors is bigger, the benefit from choosing the correct door is greater. However, because a higher probability level also corresponds to a greater belief update when observing, the optimal behavior in this task across probability levels, calculated using SARSOP (Figure 3A), is to make one observation at the beginning when ρ = 1.0 or ρ = 0.75, and to not make any observations when the probability is evenly split, i.e., ρ = 0.5. Neural networks that were meta-trained across all probability levels (and therefore simulate an agent that does not start out with any information about probability levels) observed exactly once and only did so at the beginning of every episode (Figure 3B).

Children tended to observe more than the optimal computational model would suggest. Out of 12 trials, children in the ρ = 1.0 condition made an average of 1.88 observations (*SE* = 0.25) which was not more the optimal solution of 1 observation, $t(24) = 1.32$, $p = 0.20$; those in the ρ = 0.75 condition made an average of 1.8 observations (*SE* = 0.25) which was marginally more than the optimal solution of 1 observation, $t(24) = 1.92$, $p = 0.067$; the rest in the ρ = 0.5 condition made an average of 2.8 observations (*SE* = 0.29) which was also significantly more than the optimal solution of 0 observation, $t(24) = 4.80$, $p < 0.01$. 30 out of 75 children strictly chose to bet and did not observe at all (n = 12 in ρ = 1.0, n = 10 in ρ = 0.75, n = 8 in ρ = 0.5). 2 children in the ρ = 1.0 condition only chose to observe and did not bet at all.

Similarly, out of 12 trials, adults in the ρ = 1.0 condition made an average of 1.15 observations (*SE* = 0.32) when the optimal solution was 1 observation, $t(19) = 0.47$, $p = 0.64$; those in the ρ = 0.75 condition made an average of 1.70 observations (*SE* = 0.34) when the optimal solution was 1 observation, $t(19) = 2.05$, $p = 0.054$; the rest in the ρ = 0.5 condition made an average of 1.35 observations (*SE* = 0.44) when the optimal solution was 0 observation, $t(19) = 2.90$, $p < 0.01$. 25 out of 60 adults strictly chose to bet and did not observe at all (n = 8 in ρ = 1.0, n = 6 in ρ = 0.75, n = 11 in ρ = 0.5). No adults strictly chose to observe and did not bet at all. Overall, children observed (μ = 2.37, *SE* = 0.33) *marginally more* than adults did (μ = 1.40, *SE* = 0.21), $t(144) = 1.49$, $p = 0.07$. This is consistent with other findings where children explored more than adults, though in very different tasks (e.g., Giron et al, 2023; Liquin & Gopnik, 2022).

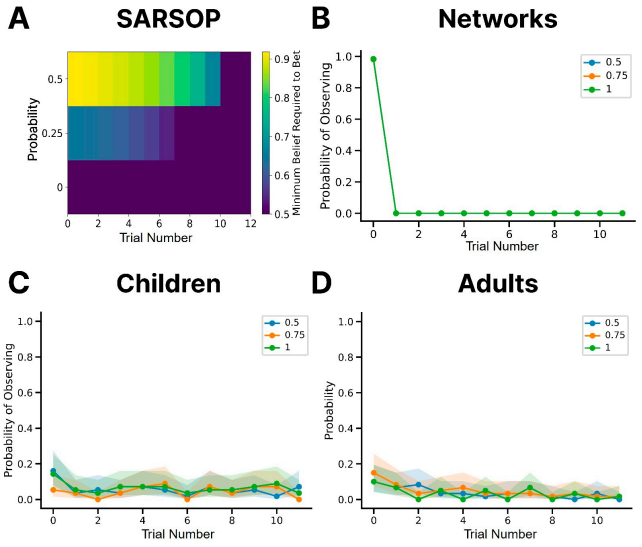

Figure 3: (A) Solutions for a partially observable Markov decision process approximated via SARSOP at ρ = 0.5, 0.75, 1.0, color bar indicates minimum belief for optimal betting. (B) Observing probability set by RL neural network agents. (C-D) Identical plots for children (C) and adults (D). Colored regions indicate the 95% Bayesian credible intervals under a Jeffreys prior.

Like the neural networks (which started the task without information on the task), both children and adults did not significantly modulate their observation rates based on the probability structure. In a one-way ANOVA test ($F(2, 72) = 0.97$, $p = 0.39$ for children; $F(2, 57) = 0.57$, $p = 0.57$ for adults); a generalized mixed-effect model with observation as a binary outcome did not yield a main effect of probability - this is further supported by the three pairwise comparisons of estimated marginal means in the three probability structures (ρ = 0.5 vs. ρ = 0.75, ρ = 0.5 vs. ρ = 1.0, ρ = 0.75 vs. ρ = 1.0 respectively) via Tukey's HSD test ($p = 0.50$, $p = 0.40$, $p = 0.98$ respectively for children; $p = 0.62$, $p = 0.98$, $p = 0.50$ respectively for adults).

Contrary to the optimal solution determined by the SARSOP model and the neural network solutions, children sampled their observations throughout the episode and not just on the first trial (Figure 3C). Nonetheless, trial number had a significant effect on their likelihood of observing in a generalized linear mixed-effects model, $\beta = -0.082$, $z = -2.65$, $p < 0.01$. In other words, as the likelihood of observing declined by 8.2% with every increasing trial number. Participants could keep track of which trial they were on out of the 12 trials. 35 out of 75 children (46.7%) did not choose to observe after the first trial. For adults, although they also observed at greater-than-optimal rates overall (Figure 3D), they attenuated their observation rate more strongly across the course of the game. As the trial number increased, adults' tendency to observe, unlike children's, declined significantly $\beta = -0.24$, $z = -5.79$, $p < 0.001$. 32 out of 60 adults (53.3%) did not choose to observe after the first trial.

Children modulate their betting policy based on the probability structure of the environment. We calculated the arm with the most evidence of reward as the arm that the agent had observed paying out more often, with a tie going to the most-recently-observed arm (following at least two observations) to account for recency effects. RL neural networks that were meta-trained across all probability levels were 100% likely to bet on the door with the most rewarding evidence in all the trials and across all probability levels (Figure 4A).

Children's observations did not differ significantly depending on the probability of reward. They were also not systematically distributed throughout the episode. However, betting behavior was sensitive to the payout structure that children observed. This analysis focuses on the 43 out of 75 children who chose to observe at least once before placing a bet (30 chose to strictly bet and 2 chose to strictly observe as previously discussed). More specifically, children were most likely to place their bets on the door with most rewarding evidence in the deterministic setting ρ = 1.0. On average, they were μ = 0.93 likely (*SE* = 0.046) to bet on the door that had the strongest evidence of reward (Figure 4B). However, in the indeterministic settings, children distributed their bets more evenly across the two doors (μ = 0.49, *SE* = 0.063 for ρ = 0.75 and μ = 0.50, *SE* = 0.046 for ρ = 0.5). A generalized linear mixed-effects model revealed a statistically significant

effect of ρ = 1.0 relative to ρ = 0.5, β = 2.70, z = 5.20, p < 0.001. Further pairwise comparisons via Tukey's HSD test showed that children were significantly more likely to bet on the most-rewarding-evidence door in the ρ = 1.0 condition than in the ρ = 0.5 condition (z = 5.20, p < 0.001) and the ρ = 0.75 condition (z = 5.03, p < 0.001).

By contrast, adults did not significantly modulate their bets across the different probability conditions (Figure 4C). This analysis focuses on the 35 out of 60 adults who chose to observe at least once before placing a bet (25 chose to strictly bet as previously discussed). They did not place absolute bets on the same door in the way the meta-RL did either. On average, adults placed their bet on the door with most rewarding evidence 82% of the time (SE = 0.08) in ρ = 1.0, 88% of the time (SE = 0.05) in ρ = 0.75, and 67% of the time (SE = 0.05) in ρ = 0.5. Pairwise comparisons via Tukey's HSD test showed that adults were not more likely to bet on the most-rewarding-evidence door in one probability condition compared to another (z = 1.85, p = 0.16 for ρ = 0.75 vs. ρ = 0.5; z = 1.75, p = 0.19 for ρ = 1.0 vs. ρ = 0.5; z = 0.017, p = 0.99 for ρ = 1 vs. ρ = 0.75).

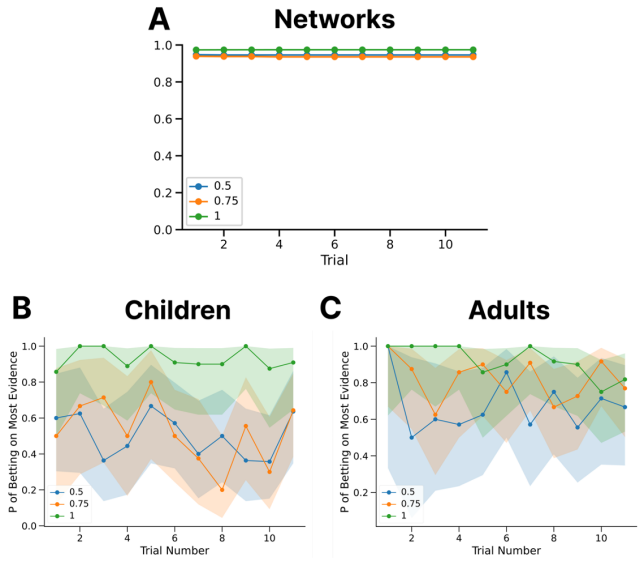

Figure 4: (A) Betting probability on the door with most rewarding evidence observed, set by RL neural network agents for ρ = 0.5, 0.75, 1.0. (B) Same probability for children (left) and adults (right). Colored regions represent the 95% Bayesian credible intervals under a Jeffreys prior.

Children's earnings approach that of adults.

Overall, across the 12 test trials and all probability conditions, children won an average of 5.84 coins (SE = 0.36) and adults won an average of 6.50 coins (SE = 0.30). There was no significant difference in the number of coins attained between the children and adults both in aggregate (75 children vs. 60 adults) and by probability level (25 children vs. 20 adults per level). While there was no effect of probability level on the number of coins won by adults, children were found to win significantly fewer coins in the ρ = 0.5 condition than in the ρ = 1.0 condition, t(40) = 2.25, p < 0.05.

Children distinguished between probability levels better than adults did.

Moreover, children generally differentiated the probability levels better than adults (Figure 5). Pairwise comparisons via Tukey's HSD test revealed that children estimated that the likelihood of receiving a reward from the rewarding door was significantly higher in the ρ = 1.0 condition (M = 73.7%, SE = 7.52%) than the ρ = 0.5 condition (M = 34.9%, SE = 5.62%), z = 4.18, p < 0.01; they also estimated a significantly higher likelihood in the ρ = 1.0 condition (M = 73.7%, SE = 7.52%) than the ρ = 0.5 condition (M = 56.7%, SE = 6.43%), z = 3.27, p < 0.01. However, adults did not significantly distinguish the likelihood of receiving a reward from the rewarding door between the different probability conditions.

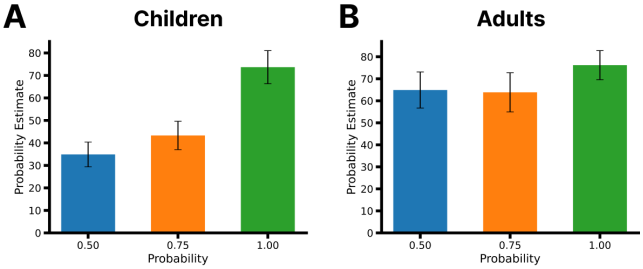

Figure 5: (A) Children's perceived reward likelihood from the rewarding door after the 12 test trials. (B) Same measure for adults. Error bars represent standard errors of the mean.

A computational process model explains components of child and adult behavior.

Finally, we found that the full computational process model proposed by Navarro et al. (2016) explains the adult data better overall than it does the child data, t(133) = -2.25, p = < 0.05. In particular, it explained the child data best in the ρ = 1.0 setting (Figure 6A), possibly reflecting inconsistencies between the behavior of children in the lower-probability setting and the expected behavior of the model which we will address in the Discussion. The fitted forgetfulness parameter α was higher for children than for adults, t(133) = 3.10, p < 0.05 (Figure 6B), indicating that participants preserved evidence most strongly in those cases. On the other hand, the only parameter determining the decision threshold that was different between children and adults was the final decision threshold, which was higher for children, t(133) = 3.12, p < 0.05 (Figures 6C-E). This reflects the fact that their observe probability does not decay as much over time. Children's choices generally did not exhibit higher stochasticity σ than adults', t(110) = 1.47, p = 0.14; Figure 6F), except for their choices in the 0.75 probability level, t(35) = -2.61, p < 0.05, suggesting that their betting behavior is particularly more volatile in this setting.

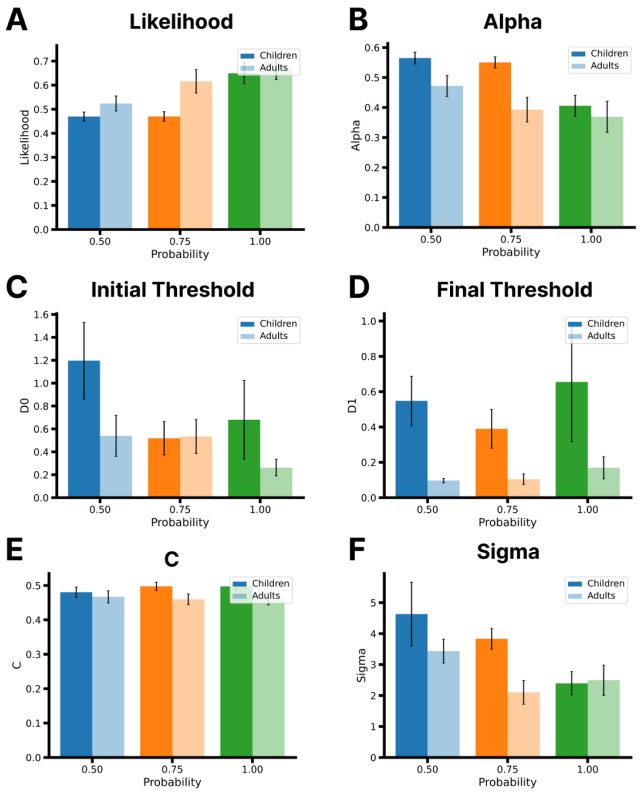

Figure 6: (A) Likelihood of participant data as determined by the fitted process model from Navarro et al. (2016) across various probability levels, with error bars representing the standard error of the mean. (B) Mean forgetfulness parameters α fitted by the model for children and adults. Error bars as before. (C) Initial decision threshold for switching from observing to betting. (D) Final decision threshold. (E) Threshold-change parameter c. (F) Stochasticity parameter σ. All plots from (B) to (F) follow the format of (B) with respective parameter changes.

Discussion

The current study corroborates the conclusion of older observe or bet tasks that adults seek more information than is computationally optimal, (Navarro et al., 2016; Tversky & Edwards, 1966); We also show a similar pattern in children, in fact children appear to be even more likely to choose to observe than to bet compared to adults. Children also distinguished the ground-truth probability levels better than adults did – again consistent with other finding where children's tendency to explore enabled them to learn more effectively than adults. Our findings also echo existing work demonstrating that children are generally sensitive to uncertainty (e.g., Redshaw & Suddendorf, 2016; Robinson et al., 2006).

While our findings demonstrate that even children tend to observe beyond computationally optimal, our study leaves open the question of what drives this behavior: whether it is driven by an intent to gain information or a desire to gain validation through repeated positive testing (Lapidow & Walker, 2020). Further research is warranted to tease apart these motivations.

To accommodate the limited attention spans of children, we designed our experimental horizon to consist of 12 trials. This is significantly fewer than the trial counts used in the experiments conducted by Tversky & Edwards (1966) as well as Navarro et al. (2016). In our task, the optimal strategy or both $\rho = 1.0$ or $\rho = 0.75$ involved a single early observation. Future research will aim to create distinct optimal strategies for each probability level ($\rho = 1.0$, $\rho = 0.75$ and $\rho = 0.5$). Specifically, we will investigate conditions under which $\rho = 0.75$ elicits more exploratory behavior compared to $\rho = 1.0$ and $\rho = 0.5$. Moreover, we are considering the integration of restless bandits, which would predict more value in exploration as the experiment progresses.

More critically, 5-to-7-year-old children behaved differently than adults in distinctive ways. Like adults, they do not modulate how much they observe, based on the probability structure of the environment. But unlike adults and computational models, they do modulate how much they bet based on the probability condition of the environment, diversifying their bets when uncertainty is high. Thus, given that children only hedge their bets when uncertainty in the environment is high and that they explicitly distinguish between probability levels even better than adults can, we argue that children are not merely behaving more noisily or randomly.

Children's "hedging behavior" is similar to "probability matching", and other studies have shown that children are more likely to probability match than adults (Denison et al., 2013). One alternative explanation for this behavior could be that children are trading off the resulting reduction in variance of rewards against both information gain and the expected value of reward. This strategy is effective in contexts in which it is necessary to make decisions under uncertainty as diverse as evolution (e.g., Philippi & Seger, 1989; Starrfelt & Kokko, 2012) and financial markets (e.g., Axén & Cortis, 2020).

Further studies may explore why children hedge, but not adults. In addition, both human children and adults do not behave like POMDP and meta-RL models. Further studies might also explore whether this apparently suboptimal behavior might be beneficial in more subtle ways. These differences between human and child behavior mean that the process model introduced by Navarro et al. is not as good a model for child behavior, and future studies should consider how to update the model to address this. In general, however, our results provide yet more support for the idea that children prioritize information over reward, and that they prefer to hedge their bets, probability matching rather than maximizing.

Acknowledgement

E.Y. was supported by the DOD ONR MURI Co-PI Self-Learning Perception through Real World Interaction funding. K.J.S. was funded by a Cusanuswerk Doctoral Scholarship. The authors are grateful to the members of the Cognitive

Development and Learning Lab at UC Berkeley, especially Eileen Liu and Megan Lui, for their help with data collection. The authors would also like to thank the participating children and families on ChildrenHelpingScience.com.

References

Åström, K. J. (1965). Optimal control of Markov processes with incomplete state information I. *Journal of mathematical analysis and applications*, *10*, 174-205.

Axén, G., & Cortis, D. (2020). Hedging on betting markets. *Risks*, *8*(3), 88.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child development*, *77*(2), 413-426.

Bellemare, M. G., Dabney, W., & Rowland, M. (2023). *Distributional reinforcement learning*. MIT Press.

Blanchard, T. C., & Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective, & Behavioral Neuroscience*, *18*, 117-126.

Cohen, J. D., McClure, S. M., & Yu, A. J. (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 B: Biological Sciences*, *362*(1481), 933-942.

Cook, Z., Franks, D. W., & Robinson, E. J. (2013). Exploration versus exploitation in polydomous ant colonies. *Journal of theoretical biology*, *323*, 49-56.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, *126*(2), 285-300.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Egorov, M., Sunberg, Z. N., Balaban, E., Wheeler, T. A., Gupta, J. K., & Kochenderfer, M. J. (2017). POMDPs. jl: A framework for sequential decision making under uncertainty. *The Journal of Machine Learning Research*, *18*(1), 831-835.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422. https://doi.org/10.1016/j.cognition.2008.09.007

Giron, A. P., Ciranka, S., Schulz, E., van den Bos, W., Ruggeri, A., Meder, B., & Wu, C. M. (2023). Developmental changes in exploration resemble stochastic optimization. *Nature Human Behaviour*, 1-13.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, *375*(1803), 20190502.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*(1-2), 99-134.

Kurniawati, H., Hsu, D., & Lee, W. S. (2008, June). Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems* (Vol. 2008).

Lapidow, E., & Walker, C. M. (2020). *The search for invariance: repeated positive testing serves the goals of causal learning* (pp. 197-219). Springer International Publishing.

Liquin, E. G., & Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, *218*, 104940.

Ma, H., & Pineau, J. (2015, March). Information gathering and reward exploitation of subgoals for POMDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1).

Meder, B., Wu, C. M., Schulz, E., & Ruggeri, A. (2021). Development of directed and random exploration in children. *Developmental science*, *24*(4), e13095.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive psychology*, *85*, 43-77.

Philippi, T., & Seger, J. (1989). Hedging one's evolutionary bets, revisited. *Trends in ecology & evolution*, *4*(2), 41-44.

Piaget, J. (2013). *The construction of reality in the child* (Vol. 82). Routledge.

Poupart, P., Kim, K. E., & Kim, D. (2011, March). Closing the gap: Improved bounds on optimal POMDP solutions. In *Proceedings of the International Conference on Automated Planning and Scheduling* (Vol. 21, pp. 194-201).

Redshaw, J., & Suddendorf, T. (2016). Children's and apes' preparatory responses to two mutually exclusive possibilities. *Current Biology*, *26*(13), 1758-1762.

Rivas, J. (2013). Probability matching and reinforcement learning. *Journal of Mathematical Economics*, *49*(1), 17–21. https://doi.org/10.1016/j.jmateco.2012.09.004

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., & Apperly, I. A. (2006). Children's sensitivity to their own relative ignorance: Handling of possibilities under epistemic and physical uncertainty. *Child development*, *77*(6), 1642-1655.

Sandbrink, K., & Summerfield, C. (2023, August). *Learning the Value of Control with Deep RL.* Proceedings of the Conference on Cognitive Computational Neuroscience, Oxford, UK.

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*(11), 1561-1572.

Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in cognitive sciences*, *16*(7), 382-389.

Siegel, S., & Goldstein, D. A. (1959). Decision-making behavior in a two-choice uncertain outcome situation. *Journal of Experimental Psychology*, *57*(1), 37–42. https://doi.org/10.1037/h0045959

Silver, D., & Veness, J. (2010). Monte-Carlo planning in large POMDPs. *Advances in neural information processing systems*, *23*.

Sophian, C., & Somerville, S. C. (1988). Early developments in logical reasoning: Considering alternative possibilities. *Cognitive Development*, *3*(2), 183-222.

Starrfelt, J., & Kokko, H. (2012). Bet-hedging—a triple trade-off between means, variances and correlations. *Biological Reviews*, *87*(3), 742-755.

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, *12*.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., ... & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, *21*(6), 860-868.