Children prioritize purely exploratory actions in observe or bet tasks

Eunice Yiu[*]
Department of Psychology
University of California, Berkeley
ey242@berkeley.edu

Kai Sandbrink[*]
Department of Experimental Psychology
University of Oxford
kai.sandbrink@lmh.ox.ac.uk

Eileen Liu
Department of Psychology
University of California, Berkeley
eileen-liu@berkeley.edu

Alison Gopnik
Department of Psychology
University of California, Berkeley
gopnik@berkeley.edu

Abstract

In reinforcement learning, agents often need to decide between selecting actions that are familiar and have previously yielded positive results (exploitation), and seeking new information that could allow them to uncover more effective actions (exploration). 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 and AI. In this study we develop an "observe or bet" task that separates "pure exploration" from "pure exploitation." Participants have the option to either observe an instance of an outcome and receive no reward, or to bet on an action that is eventually rewarding, but offers no immediate feedback. We collected data from 56 five-to-seven-year-old children who completed the task at one of three different probability levels. We compared how children performed against both approximate solutions to the partially-observable Markov decision process and meta-RL models that were meta trained on the same decision making task across different probability levels. We found that the children observe significantly more than the two classes of algorithms. We then quantified how children's policies differ between the different probability levels by fitting probabilistic programming models and by calculating the likelihood of the children's actions under the task-driven model. The fitted parameters of the behavioral model as well as the direction of the deviation from neural network policies demonstrate that the primary way children change the frequency with which they bet on the door for which they have less evidence. This suggests both that children model the causal structure of the environment and that they produce a "hedging behavior" that would be impossible to detect in standard bandit tasks, and that reduces variance in overall rewards. The results shed light on how children reason about reward and information, providing a developmental benchmark that can help shape our understanding of both human behavior and RL neural network models.

1 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

[*] These authors contributed equally to this work

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

exploration of different options or opportunities and the exploitation of rewards (1, 2). This balance is commonly referred to as the exploration-exploitation trade-off.

Researchers have long argued that children are active and exploratory information seekers (e.g., 3, 4, 5). However, previous studies used environments in which the reward and information that participants receive on each trial are confounded (e.g., 6, 7, 8, 9). In these experiments, exploratory actions lead to reward and exploitative moves result in information gain, which obscures interpretation of the two concepts. In our study, 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 10, but can learn near-optimal exploratory behavior over several repetitions in the task (11). Moving beyond these studies, we use child participants and we also consider the two kinds of betting actions, those that correspond to the one that the participant has received the strongest evidence for so far (which would be chosen to maximize the expected value of reward) and the one that goes against the current evidence (which we call “hedging"; if this arm has been chosen less frequently so far, it reduces the variance in rewards). The study is the first of its kind to disambiguate the motives underlying exploratory and exploitative behavior in children.

2 Methods

Child Experiment Following the structure of the observe or bet task (10), we presented 56 five- to seven-year-old children with a game featuring a rewarding character and a non-rewarding character. We target this age group as a large amount of existing literature suggests that it is not until the age of 4 that children begin to reliably understand and reasonably act on uncertainty and counterfactual possibilities in the physical world where feedback is not immediately provided (as in our case of "bet" actions) (e.g., 12, 13, 14, 15, 16, 17). Moreover, our pilot study suggested that only children aged 5 and above could reliably comprehend the complexity of the task. In this game, each character hid behind separate doors. Children received one coin if they found the rewarding character and did not gain or lose anything if they found the other non-rewarding character. In every trial, children 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 1A). Exactly one of the two actions paid out on every trial. Children played games of 12 trials, during which the underlying payout probabilities remained constant. However, we varied the payout probabilities between participants. In this ongoing study, we randomly assigned 19 children to the setting where the payout probability ρ of the higher-paying door was $\rho = 1.0$, 19 to $\rho = 0.75$, and 18 to $\rho = 0.5$. While children were not given the exact probability of the environment, the experimenter gave a clue by describing their assigned environment as "always the same" ($\rho = 1.0$), having a preferably higher-paying door even though it might "sometimes change" ($\rho = 0.75$), and "always changing, no one can tell" which option was higher-paying ($\rho = 0.5$). Children had to pass all of the comprehension checks before they could proceed in the study. 2 additional participants who failed to pass were excluded from the final participant pool ($n = 56$). Next, children played four practice trials to familiarize themselves with the setup and received feedback on the reward they had accumulated at the end of the practice. They then proceeded to play the actual game which they were told had the same probability structure as the practice game. The game at test involved characters (a kind princess and a mean thief) that were visually different from the practice games (a kind elf and a mean monster). At the end of the test trials, children were asked to quantitatively express their perceived probability of receiving a reward from the left door versus the right using a slider. The results indicated that they recognized the differences between levels (see Figure 1B). The study was preregistered here: `https://aspredicted.org/blind.php?x=TG5_Q8B`.

Computational Modelling In order to quantify optimal performance on the task, we first formulated the problem as a partially-observable Markov decision process (POMDP, 18, 19). This POMDP is defined in Supplementary Section S1.1. 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 (20) to calculate successive approximations of the reachable state under optimal policies (SARSOP, 21), a state-of-the-art solver for problems that require active information gathering (22, 23). This allowed us to calculate for every trial the belief threshold at which one should switch from observing to betting for each probability setting (Figure 1C). We then compared children's behavior across different probability levels with the SARSOP solutions. We also compared both the children and the SARSOP solution to the solution of deep RL meta-agents which were trained on different probability levels of the same task and thus were unsure of the payout probability (following 24, see Supplementary Section S1.2 for architecture and training procedure).We ran five instantiations of the RL neural networks, which learned to perform near-optimally on the task (see Supplementary Figure S1 for the learning curve). Finally, we fitted parameters from the model for human behavior from 11 to the child data in order to analyze differences in child behavior between the different probability settings.

3 Results

3.1 Children 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 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, is to make one observation at the beginning when $\rho = 1.0$ or $\rho = 0.75$, and to not make any observations when the probability is evenly split, i.e. $\rho = 0.5$.Neural networks that are meta-trained across all probability levels (and therefore simulate an agent which does not start out with any information as to which probability level it is operating under) observe exactly once at the beginning of every episode (Figure 1D). While children on average did not observe significantly more than once in $\rho = 1.0$ and $\rho = 0.75$ conditions, they observed significantly more than the optimal solution in the $\rho = 0.5$ condition (Figure 1D): out of 12 trials, participants in the $p = 1.0$ condition made an average of 2.32 observations ($SE = 0.31$) when the optimal solution was 1 observation, $t(9) = 1.54$, $p = 0.14$; those in the $p = 0.75$ condition made an average of 1.58 observations ($SE = 0.27$) when the optimal solution was 1 observation, $t(18) = 1.78$, $p = 0.25$; the rest in the $p = 0.5$ condition made an average of 2.11 observations ($SE = 0.31$) when the optimal solution was 0 observation, $t(17) = 3.43$, $p < 0.01$. 24 out of 56 children strictly chose to bet and did not observe at all (n=8 in $\rho = 1.0$, n=9 in $\rho = 0.75$, n=7 in $\rho = 0.5$). 2 children in $\rho = 1.0$ strictly chose to observe and did not bet at all.

Like the neural networks (which start the task without information on the task), children did not significantly modulate their observation rates based on the probability structure in a one-way ANOVA test ($F(2, 53) = 0.33$, $p = 0.72$); a generalized mixed-effect model with observation as a binary outcome did not yield a main effect of probability - this is further supported in the three pairwise comparisons of estimated marginal means in the three probability structures via Tukey's HSD test ($p = 0.81$, $p = .99$, $p = 0.74$ respectively). Aggregating across all probabilities, we found that children on average make 2.0 observations ($SE = 0.38$), which is significantly more than 1 observation, $t(55) = 2.61$, $p < 0.05$.

Contrary to the optimal solution determined by the SARSOP model and the neural network solutions, children sampled their observations more throughout the episode (Figure 1C). This is different than adult behavior in similar versions of the task: Although adults also observe at greater-than-optimal rates (11), they attenuate their observation rate strongly across the course of an episode, whereas children continue to observe until the end of the episode. This means that only few children front-load their observations (i.e., only choosing to observe at the start of a task, see Figure 1E). That said, children generally reduced their observation behavior as trial number increased in a generalized linear mixed-effects model ($\beta = -0.07$,$z = -1.88$, $p = 0.06$).

3.2 Children modulate their betting policy based on the probability structure of the environment

While children neither significantly differ in how much they observe between probability structures nor systematically alter how much they observe throughout an episode, we find evidence of betting behavior that is sensitive to the payout structure of the environment in children's observe actions. Children were most likely to place their bets on the most-recently-observed rewarding door in high-probability settings ($p = 1.0$), i.e., they preferentially bet on the action that they had accumulated the most evidence for (see Figure 1F), $\mu = 0.91$, $SE = 0.06$. We quantify the arm that they had

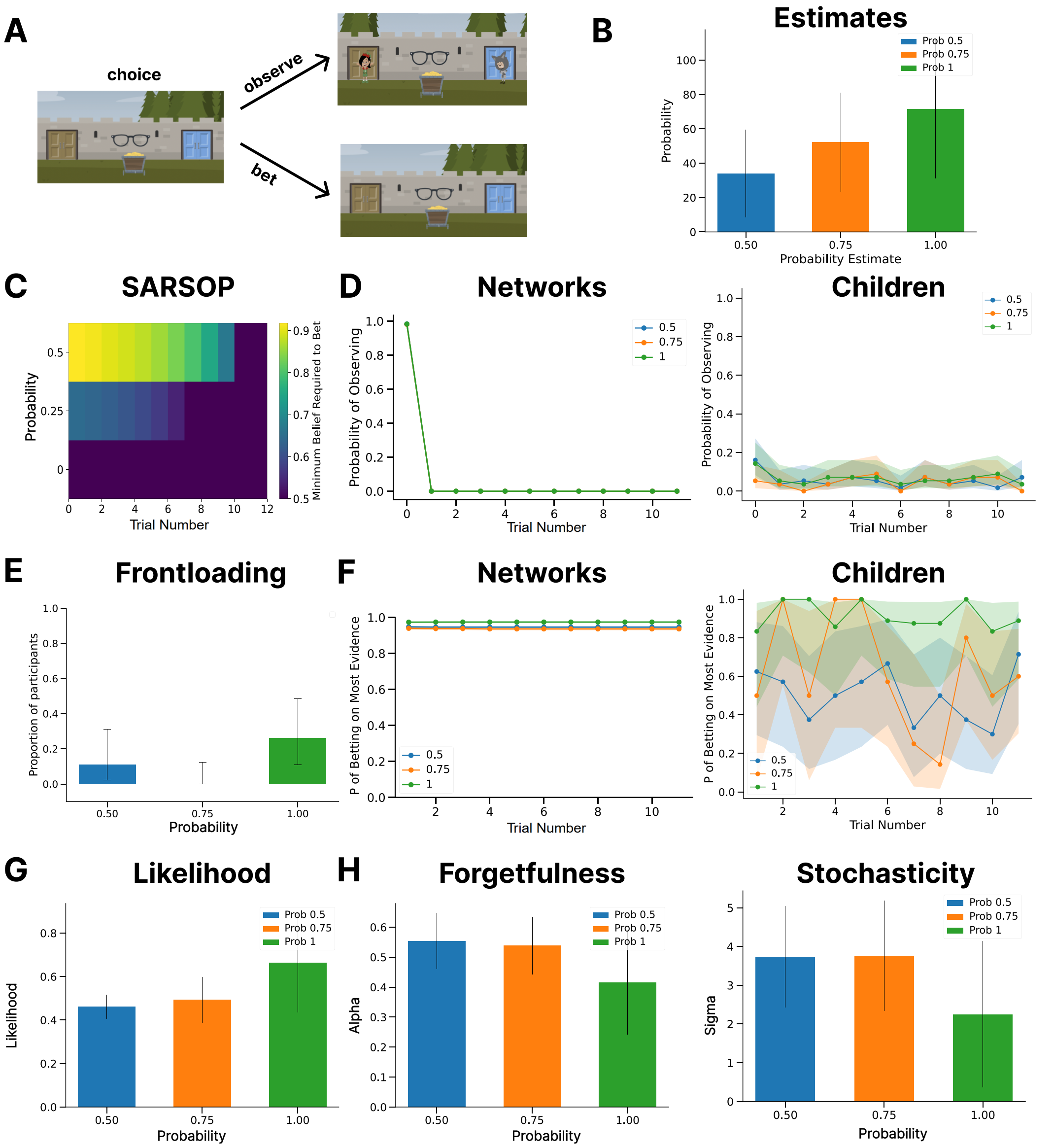

Figure 1: (**A**) On every trial, participants have the choice between "observing," when they see which door contains a princess and which contains a thief, or "betting," in which case receive reward (with feedback delayed) if they pick the door with the princess. (**B**) Children's perceived likelihood that the rewarding door provides reward measured after they completed all 12 test trials. Error bars signify SD. (**C**) The solution to the partially-observable Markov decision process approximated using SARSOP for the three different probability levels. The colorbar indicates the minimum belief in the doors for which betting is optimal (with observing optimal below that threshold). For the 1 and 0.75 probability levels, the threshold corresponds to one observation; for the 0.5 probability level, to zero. (**D**) (*left*) the reinforcement-learning neural network agents and (*right*) the children for upper probabilities of $(green)$ 1, $(orange)$ 0.75 and $(blue)$ 0.5. Error bars for neural networks signify SEMand for children represent the 95% Bayesian credible intervals for an unknown proportion under a Jeffreys prior. (**E**) Proportion of children who front-load their observations exactly for three different probability settings. Error bars represent the 95% Bayesian credible intervals for an unknown proportion under a Jeffreys prior. (**F**) Probability of betting on the recently-observed rewarding door (as opposed to the other one) for (*left*) RL neural network models and (*right*) children. Error bars for neural networks signify SEMand for children the 95% Bayesian credible intervals for an unknown proportion under a Jeffreys prior. (**G**) Likelihood of the fitted process model (from Navarro et al., 2016) for different probability levels. Error bars represent SD. (**H**) Mean posterior parameter values for (*left*) forgetfulness and (*right*) stochasticity parameters for the fitted process model. Error bars represent SD.

accumulated the *most evidence* for as the one that they have observed paying out more often, with a tie (following at least two observations) going to the most-recently-observed arm to account for recency effects. In episodes where the payout probabilities are more even, however, children distributed their bets more evenly across the two doors, ($\mu = 0.51$, $SE = 0.07$ for $\rho = 0.75$ structure and $\mu = 0.52$, $SE = 0.05$ for $\rho = 0.5$ structure). A generalized linear mixed-effects model with bet based on the most-recently-observed door as a binary outcome reveals a statistically significant effect of $\rho = 1.0$ relative to $\rho = 0.5$, $\beta = 2.28$, $z = 5.05$, $p < 0.001$. Further pairwise comparisons via Tukey's HSD test show that children are significantly more likely to be on the most-recently-observed door in the $\rho = 1.0$ condition compared to the $\rho = 0.5$ condition ($z = 2.28$, $p < 0.001$) and the $\rho = 0.75$ condition ($z = 2.06$, $p < 0.001$).

To validate this difference in behavior, we fit the computational process model from Navarro et al. to behavior (11). This model has been shown to be the best out of 4 candidate models for an adult version of the task and has been established in subsequent work (25). In this model, behavior can vary based on an evidence decay parameter, a decision threshold, and a stochasticity term. The model performs better than chance at predicting child performance, although it explains the child data best in the $\rho = 1.0$ setting (Figure 1G). The parameters related to decision threshold do not vary systematically between probability levels (see Supplementary Figure 3). However, both the evidence decay parameter α and the stochasticity term σ are higher for more even probability levels, corresponding to higher probabilities for taking the unobserved door (Figure 1G).

4 Discussion

This study provides a developmental benchmark for integrating reinforcement learning, observation (exploration) and betting (exploitation) behaviors in both the cognitive science and AI communities. Our findings corroborate with existing work that children are sensitive to and make anticipatory responses to uncertainties in the physical world (e.g., 13, 15). Furthermore, we show that when they are given the opportunity to get information about the uncertainty by observing outcomes, 5-to-7-year-olds opt to observe at similar frequencies across all uncertainty levels. They do not modulate how much they observe, but rather diversify their bets to address high uncertainties. Supplementary Section S1.4 describes how this reduction in variance in the reward space can be described as trading off against both information gain of the transition structure as well as 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 [26, 27] and financial markets [28]. This could be one explanation for (or realization of) the counterfactual behavior that has previously been described by others. To our knowledge, this study is the first to show that from a young age, humans' desire to "hedge bets" in the face of uncertain environments could be a driving force for modulations of exploratory behavior in standard bandit tasks, and could play a wider role in the use of higher-order moments in RL [29]. In next steps, we plan to compare children's performance with that of adults performing the same task and to model the results with RL models of developmental learning.

References

[1] Jonathan D Cohen, Samuel M McClure, and Angela J Yu. 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, 2007.

[2] Zoe Cook, Daniel W Franks, and Elva JH Robinson. Exploration versus exploitation in polydomous ant colonies. *Journal of theoretical biology*, 323:49–56, 2013.

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

[4] Laura Schulz. The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in cognitive sciences*, 16(7):382–389, 2012.

[5] Jean Piaget. *The construction of reality in the child*, volume 82. Routledge, 2013.

[6] Anna P Giron, Simon Ciranka, Eric Schulz, Wouter van den Bos, Azzurra Ruggeri, Björn Meder, and Charley M Wu. Developmental changes in exploration resemble stochastic optimization. *Nature Human Behaviour*, pages 1–13, 2023.

[7] Emily G Liquin and Alison Gopnik. Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, 218:104940, 2022.

[8] Björn Meder, Charley M Wu, Eric Schulz, and Azzurra Ruggeri. Development of directed and random exploration in children. *Developmental science*, 24(4):e13095, 2021.

[9] Eric Schulz, Charley M Wu, Azzurra Ruggeri, and Björn Meder. Searching for rewards like a child means less generalization and more directed exploration. *Psychological science*, 30(11): 1561–1572, 2019.

[10] Amos Tversky and Ward Edwards. Information versus reward in binary choices. *Journal of Experimental Psychology*, 71(5):680, 1966.

[11] Daniel J Navarro, Ben R Newell, and Christin Schulze. Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive psychology*, 85:43–77, 2016.

[12] Brian Leahy, Michael Huemer, Matt Steele, Stephanie Alderete, and Susan Carey. Minimal representations of possibility at age 3. *Proceedings of the National Academy of Sciences*, 119 (52):e2207499119, 2022.

[13] Jonathan Redshaw and Thomas Suddendorf. Children's and apes' preparatory responses to two mutually exclusive possibilities. *Current Biology*, 26(13):1758–1762, 2016.

[14] Sarah R Beck, Elizabeth J Robinson, Daniel J Carroll, and Ian A Apperly. Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child development*, 77(2): 413–426, 2006.

[15] Elizabeth J Robinson, Martin G Rowley, Sarah R Beck, Dan J Carroll, and Ian A Apperly. Children's sensitivity to their own relative ignorance: Handling of possibilities under epistemic and physical uncertainty. *Child development*, 77(6):1642–1655, 2006.

[16] Catherine Sophian and Susan C Somerville. Early developments in logical reasoning: Considering alternative possibilities. *Cognitive Development*, 3(2):183–222, 1988.

[17] Marjorie Taylor. Conceptual perspective taking: Children's ability to distinguish what they know from what they see. *Child development*, pages 703–718, 1988.

[18] K. J Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, February 1965. ISSN 0022-247X. doi: 10.1016/0022-247X(65)90154-X.

[19] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, May 1998. ISSN 0004-3702. doi: 10.1016/S0004-3702(98)00023-X.

[20] Maxim Egorov, Zachary N Sunberg, Edward Balaban, Tim A Wheeler, Jayesh K Gupta, and Mykel J Kochenderfer. Pomdps.jl: A framework for sequential decision making under uncertainty. *The Journal of Machine Learning Research*, 18(1):831–835, 2017.

[21] Hanna Kurniawati, David Hsu, and Wee Sun Lee. SARSOP: Efficient Point-Based POMDP Planning by Approximating Optimally Reachable Belief Spaces. 2008.

[22] Hang Ma and Joelle Pineau. Information Gathering and Reward Exploitation of Subgoals for POMDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), March 2015. ISSN 2374-3468. doi: 10.1609/aaai.v29i1.9659.

[23] David Silver and Joel Veness. Monte-Carlo Planning in Large POMDPs. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[24] Kai Sandbrink and Christopher Summerfield. Learning the value of control with deep rl. In *Proceedings of the Conference on Cognitive Computational Neuroscience*, Oxford, UK, August 2023.

[25] Tommy C. Blanchard and Samuel J. Gershman. Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective & Behavioral Neuroscience*, 18(1):117–126, February 2018. ISSN 1531-135X. doi: 10.3758/s13415-017-0556-2.

[26] Tom Philippi and Jon Seger. Hedging one's evolutionary bets, revisited. *Trends in Ecology & Evolution*, 4(2):41–44, February 1989. ISSN 0169-5347. doi: 10.1016/0169-5347(89)90138-9.

[27] Jostein Starrfelt and Hanna Kokko. Bet-hedging—a triple trade-off between means, variances and correlations. *Biological Reviews*, 87(3):742–755, 2012. ISSN 1469-185X. doi: 10.1111/j.1469-185X.2012.00225.x.

[28] Gustav Axén and Dominic Cortis. Hedging on Betting Markets. *Risks*, 8(3):88, September 2020. ISSN 2227-9091. doi: 10.3390/risks8030088.

[29] Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023.

[30] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. 1999.

[31] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL\$$\hat{2}$\$: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv:1611.02779 [cs, stat]*, November 2016.

[32] Jane X. Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn, 2016.

[33] Jane X. Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6):860–868, June 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0147-8.

S1 Supplementary Material

S1.1 Formulation of the partially-observable Markov decision process

A POMDP is a 7-tuple $\langle \mathcal{S}, \mathcal{A}, \Omega, \mathcal{O}, T, r, \gamma \rangle$, where $\mathcal{S}$ is the (finite) non-empty state space, $\mathcal{A}$ is the (finite) non-empty action space, Ω is the (finite) non-empty observation space, $\mathcal{O} \colon \mathcal{S} \to \mathbb{P}(\Omega)$ is the observation function, $T \colon \mathcal{S} \times \mathcal{A} \to \mathbb{P}(\mathcal{S})$ is the probabilistic state-transition function, $r \colon \mathcal{S} \times \mathcal{A} \to \mathbb{P}(R)$ is a bounded reward function, and $0 \leq \gamma \leq 1$ is the discount factor. We formalize the observe or bet task for use with the solver 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. The observation, transition, and reward functions that correspond to the regular observe-or-bet rules, and we set $\gamma := 1$.

This formulation of the POMDP takes the perspective of an agent who knows what the overall probability level is, but does not know the assignment to a particular door. This is the formulation that we use for the POMDP solver. In contrast, the neural networks are meta-trained across all probability levels, which means they start an episode with a uniform prior about the probability level. Because we only give participants general indication of probability ranges rather than an exact estimate, we can assume that they are and in any case start the trial without information about which arm is the higher-paying one, we can assume that the participants operate with some uncertainty as to the exact transition structure of the environment and therefore are operating under the POMDP described above.

S1.2 Neural network architecture and training procedure

The neural networks we trained 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-hot 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-hot feedback corresponding to the observation on the two doors (1 indicating that the princess was observed at the door on the previous time step and 0 either that the agent did not observe or that the agent observed but this door did not contain the princess), and then a flag indicating the start of a new episode as well as that showed the feedback tally for the reward received on the previous episode on the first step of a new episode, and that showed 0 otherwise.

We train the neural network using the REINFORCE algorithm [30] with a baseline of 1/3 (corresponding to the expected value of a random action) following the meta-reinforcement learning procedure [31, 32, 33] in which we train the not only on a single payout probability level but vary the probability level between different episodes. (In essence, we meta-train the networks across the distribution of POMDPs defined by sampling $\rho \sim \mathcal{U}[0.5, 1]$. In order to avoid biasing in a particular direction for comparing with the human data, we do not hold out any area of the training region.) We train the networks for 500000 episodes using a batch size of 50. The recurrent units of the LSTM layer are reset to 0 at the start of a new episode. We use the Adam optimizer with a learning rate of 1e-3. We start training with entropy regularization with coefficient 5, which we annealed to 0 geometrically over the course of 150000 episodes. These values were determined based on being reasonable as well as a very limited amount of manual trial and error. No further hyperparameter optimization was conducted. The remaining parameters were left at their default values.

S1.3 Behavioral analyses

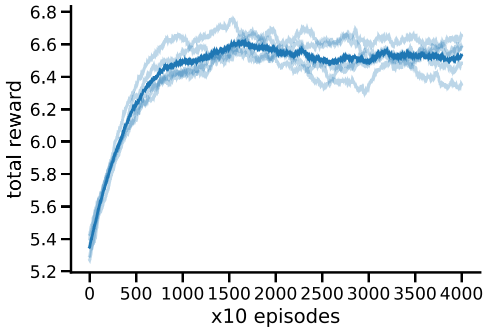

Figure S1: Learning curves for the task-driven RL neural network models for (*faint*) each of the individual models and (*thick*) aggregated over all five, smoothed using a moving average window over 1000 episodes.

S1.4 Explanation of information gain and reduction in variance

An actor in the POMDP described in Supplementary Section S1.1 will on every step face the choice between three different actions, each of which maximizes a different quantity. Here we assume that participants are operating under the same POMDP as the solver and interpret the qualitative labels of different probability levels given at the start of a trial to indicate the probability values of 0.5, 0.75, and 1 exactly.

Information gain Because we assume that the participant starts knowing the probability level but not the exact reward allocation to a particular door, we can describe the information state of the participant at every point of time as a Bernoulli distribution with parameter b describing the belief that the dominant door is the left one. Every time the agent chooses to observe, b is updated corresponding to the Bayesian update rule.

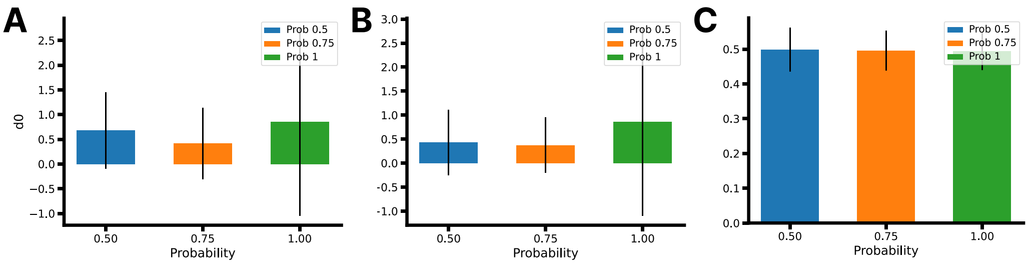

Figure S2: Fitted parameters for the Navarro model corresponding to the decision threshold. (**A**) The initial decision threshold d_0 for different payout probability levels. Error bars represent standard deviation. (**B**) Same as A, except showing the final decision threshold d_1. (**C**) Same as B, except showing the fraction of trials after which the decision threshold starts to decrease linearly from d_0 to d_1.

$$b_{t+1} = \frac{\rho \times b_t}{\rho \times b_t + (1-\rho) \times (1-b_t)} \tag{1}$$

where b_t corresponds to the previous belief on the step at time t and b_{t+1} corresponds to the belief after the observation has been taken into account on the next step at time $t+1$.

If B_t corresponds to the Bernoulli distribution describing the participants' belief state at time t and B_{t+1} at time $t+1$, the information gain is then given by the KL-divergence between the two states:

$$D_{\mathrm{KL}}(B_t||B_{t+1}) = b_t \times \log\left(\frac{b_t}{b_{t+1}}\right) + (1-b_t) \times \log\left(\frac{1-b_t}{1-b_{t+1}}\right) \tag{2}$$

This quantity is maximized by taking an observe action.

Reward Exploitation Betting on one of the doors yields the opportunity of receiving reward. We define $R_t(a)$ as the random variable corresponding to the reward yielded by taking action a at time step t. If b_t corresponds to the belief that the chosen door is the higher-paying one (corresponding to the notation above, we can assume without loss of generality that the participant is choosing the left door as we can temporarily set $b \leftarrow 1-b$ otherwise), then from the perspective of the participant the expected value of betting on that door is given by

$$\mathbb{E}[R_t(a)] = b_t \times \rho + (1-b_t) \times (1-\rho) \tag{3}$$

since we have two sources of uncertainty, the epistemic uncertainty described by the agent's belief B_t at time t and the aleatoric uncertainty given by uncertainty which door will reveal the princess under the assumption that a certain door is the dominant one given by the payout probability level ρ. This quantity is maximized by betting on the door that has been observed most frequently so far (since b_t is greater than 0.5 for this one and $\rho \geq 0.5$.)

Reduction of Variance Continuing this formulation of the random variable $R_t(a_t)$ describing the reward for taking a given action a at time t based on the two sources of uncertainty, the total variance for a series of actions $(a_t)_{t=1}^{T}$ is described by

$$\mathrm{Var}\left(\sum_{t=1}^{T} R_t\right) = \sum_{t=1}^{T} \mathrm{Var}(R_t) + 2\sum_{t<s}^{T} \mathrm{Cov}(R_t, R_s) \tag{4}$$

where T is the total number of time steps (12 in this experiment) and we write R_t to describe the reward for the chosen action a_t for simplicity. We consider cases where $p \neq 0.5$ since otherwise the agent's choices won't have any impact on its expected reward or reward variance. To illustrate how the variance in rewards is minimized by spreading bets between the different doors, consider a sequence of n betting actions uninterrupted by observe actions $(a_t)_{t=t_0}^{t_0+n}$. Since the participants' belief does not change over the course of the sequence, the R_t all have the same expected value and variance equal to $\mathbb{E}[R_{t_0}]$ and $\mathrm{Var}(R_{t_0})$ respectively. However, because the participants' belief state would have updated between taking different actions if the participant had been able to observe the outcome, the events are not independent and therefore the covariance for payouts of actions at any two times t and s $\mathrm{Cov}(R_t, R_s)$ is not zero. Specifically, betting twice on the same door yields $\mathrm{Cov}(R_t, R_s) > 0$ since receiving reward for betting on a door once increases the likelihood of receiving it a second time, and betting on two different doors yields $\mathrm{Cov}(R_t, R_s) < 0$ by the same logic. (As an example, consider the case of $\rho = 1$ where the door that contains the princess is unknown, where betting on two different doors yields an expected value of 1 with variance 0.) Therefore, assuming that the majority of the participant's betting actions up to a given point have been placed on the door for which it has the highest belief, the participant will maximize their reduction of variance by placing the subsequent bet on the other arm. Furthermore, the participant will maximize their reduction of variance over a course of a series of actions by spreading their bets evenly between the two doors.