Human Exploration Strategically Balances Approaching and Avoiding Uncertainty

Yaniv Abir[1*], Michael N. Shadlen[2,3], Daphna Shohamy[1,2]

[1]Department of Psychology, Columbia University, New York, NY, USA.

[2]Zuckerman Mind Brain Behavior Institute, and Kavli Institute for Brain Science, Columbia University, New York, NY, USA.

[3]Department of Neuroscience and Howard Hughes Medical Institute, Columbia University, New York, NY, USA.

Author Note

Correspondence concerning this article should be addressed to Yaniv Abir, Columbia University, 3227 Broadway, New York, NY 10027, USA. Email: yaniv.abir@columbia.edu

Abstract

The purpose of exploration is to reduce goal-relevant uncertainty. This can be achieved by choosing to explore the parts of the environment one is most uncertain about. Humans, however, often choose to avoid uncertainty. How do humans balance approaching and avoiding uncertainty during exploration? To answer this question, we developed a task requiring participants to explore a simulated environment towards a clear goal. We compared human choices to the predictions of the optimal exploration policy and a hierarchy of simpler strategies. We found that participants generally explored the object they were more uncertain about. However, when overall uncertainty about choice options was high, participants avoided objects they were more uncertain about, learning instead about better known objects. We examined reaction times and individual differences to understand the costs and benefits of this strategy. We conclude that balancing approaching and avoiding uncertainty ameliorates the costs of exploration in a resource-rational manner.

Introduction

The purpose of exploration is to reduce uncertainty about the aspects of one's environment that are goal relevant or otherwise important. Yet, devising an optimal strategy to reduce uncertainty is very difficult[1–3], especially for agents with limited memory and processing capacities. A simple heuristic that is often efficient for exploration is focusing on the parts of the environment that one is most uncertain about. This principle of approaching uncertainty has been found useful by statisticians[4,5], by researchers developing artificial intelligence[6–9], and by cognitive scientists interested in understanding the exploratory behaviour of humans and other animals[2,10]. Indeed, humans have been shown to approach uncertainty when learning about rewards in the environment through trial and error[2,11–13]. However, there are also many examples of uncertainty avoidance in the decision making of humans and animals. Uncertainty avoidance has been documented in situations where resolving uncertainty may reveal negative outcomes[14–18] or news[19,20], or make overcoming a conflict in motivation more difficult[20,21].

Given prior evidence of both approaching and avoiding uncertainty, we asked how these two conflicting tendencies manifest when individuals choose what to explore. We addressed this question in two parts. First, we asked whether humans tend to approach uncertainty when exploring, and if so, what decision rule do they use to choose options they are more uncertain about. While the theory of exploration makes a strong case for approaching uncertainty as an efficient principle for exploration[2,4,10], how humans implement this principle in their choices has yet to be determined, with several algorithms of varying sophistication suggested in the literature[2].

Second, we asked what conditions might compel individuals to avoid uncertainty instead of approaching it. We were interested in whether uncertainty avoidance would arise even when participants were clearly incentivised to resolve uncertainty about the environment. We hypothesized that the difficulty of making an exploratory choice might encourage avoiding uncertainty. Always approaching uncertainty may well be an efficient policy for an agent with unlimited cognitive resources. Since humans have finite memory systems, inference bandwidth, and time, it stands to reason that they would try to conserve these resources by regulating their exploration[22]. We reasoned that avoiding uncertainty could help reduce the strain on limited resources when making difficult exploration choices by lowering the amount of information that needs to be processed.

We developed a task requiring participants to make multiple exploratory choices, incrementally building knowledge in the service of a distant goal. Importantly, participants were given reward feedback only at the end of a round and not after every trial, allowing us to focus on choices made to accumulate knowledge, rather than choices driven by the need to exploit available rewards. We designed a task that posed a challenging exploration problem for participants, taxing their memory and learning systems, as is typical in real-life exploration[22,23]. The task could nonetheless be captured by a few mathematical expressions, which permit the derivation of the optimal exploration policy. This optimal policy served as a basis for a quantitative analysis of participants' behaviour with the aim of identifying the algorithm driving their exploratory choices[24–26].

Participants explored an environment comprising four tables, each with two decks of cards (**Fig. 1**). The face of each card, which was hidden from view, could be either orange or blue. Participants chose to reveal a card on one of two tables presented to them as exploratory choice options. They made these exploratory choices with the aim of learning the difference in the distributions of colours between the two decks on each table. After a random number of exploration trials, participants' learning was assessed on a separate test. Participants were rewarded solely for their performance on this test. We sought to explain how participants chose which table to explore given the sequence of cards they had already observed.

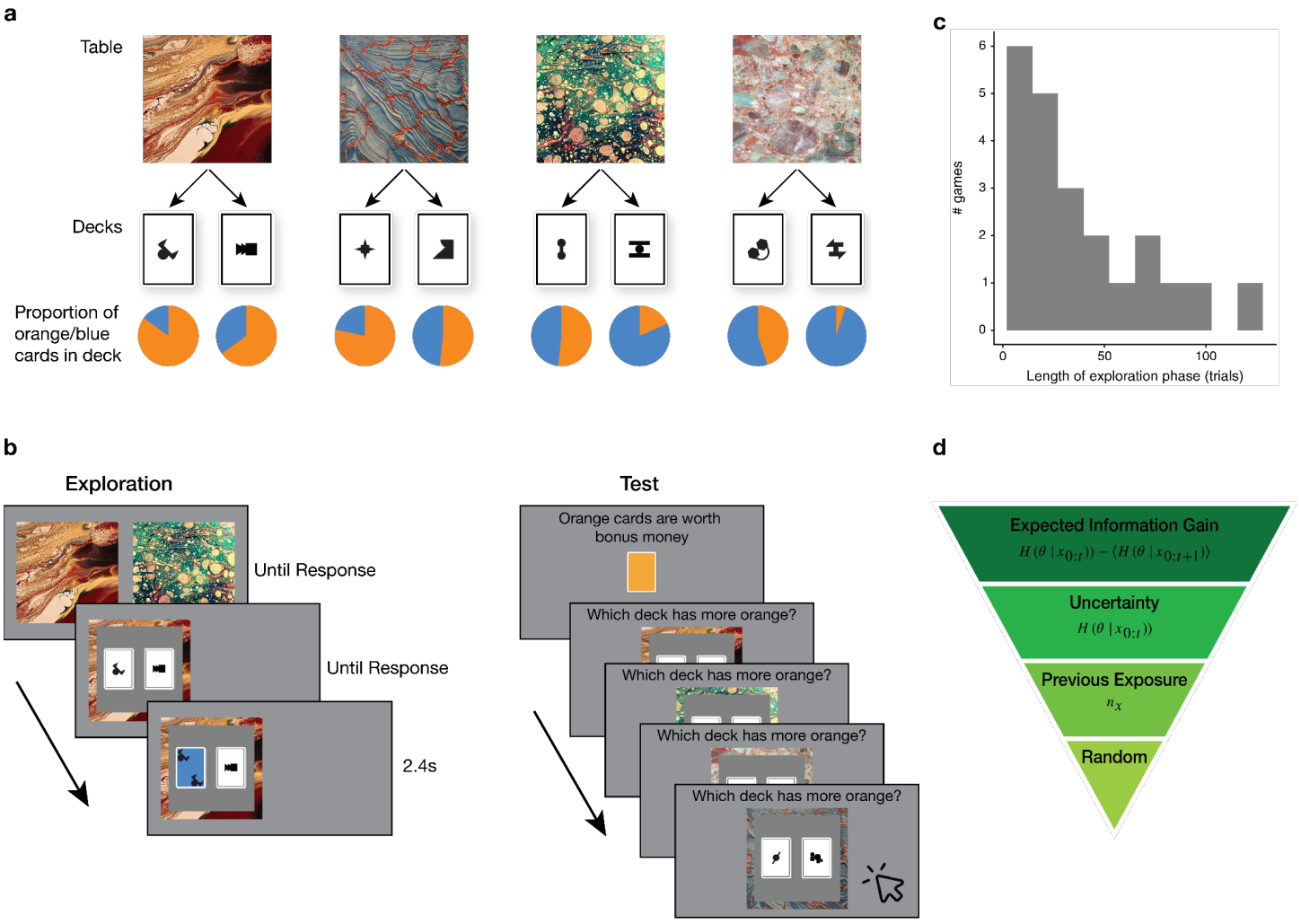

Figure 1. Exploration with no immediate reward in an incremental learning task. **a**, Statistical structure of the task. Participants explored four tables, each containing two decks with different proportions of blue/orange cards. The goal was to learn the difference in proportions of the decks on each table. **b**, On a single exploration trial (left), participants chose between two tables, and then sampled a card from one of the decks on that table, observing its colour. After a random number of exploration trials, participants were tested on their knowledge (right). A colour was designated as rewarding, and participants then chose the deck with the highest proportion of the rewarding colour on each table. They were rewarded for correct test-phase choices, and received no reward during exploration. **c**, Participants played 18 rounds. The length of exploration in each round followed a shifted geometric distribution, such that the test was equally likely to occur following any trial after the first 10. **d**, We considered a hierarchy of strategies for choosing which table to explore. The normatively prescribed strategy is to choose the table affording maximal expected information gain. This is the table for which the next card is expected to maximally decrease uncertainty (measured as entropy H) about the value of the goal-relevant latent parameter θ, given observations thus far x. A simpler strategy is to choose the table with the maximum uncertainty, as it does not necessitate computing an expectation over the next observation. An even simpler heuristic is to equate previous exposure and choose the table with the least previous observations n_x. Even though these three strategies vary considerably in complexity, they are all uncertainty-approaching on average. Lastly, people may be random explorers.

We compared participants' choices to a hierarchy of plausible strategies, differing in computational complexity (**Fig. 1e**). We found that on the majority of trials participants chose to explore the more uncertain table. However, when we examined behaviour as a function of difficulty we observed a systematic bias in their choices. Specifically, when choice difficulty was high, participants tended to avoid exploring the table they were more uncertain about. Surprisingly, participants who had a stronger tendency to avoid uncertainty learned no worse than other participants. Furthermore, reaction time data revealed that participants spent less time deliberating when making choices that avoided uncertainty. Altogether, these data are consistent with the idea that strategically balancing approaching and avoiding uncertainty serves to manage the use of cognitive resources during exploration.

Results

194 participants from an online pre-registered[27] sample were recruited to complete up to 18 rounds of the exploration task over four sessions. The task simulated a room with four tables, with two decks of cards on each table (**Fig. 1a-b**). If a card was flipped, it was revealed to be, for example, either orange or blue (each round used a different pair of colours). The proportion of orange vs. blue cards, π, differed between the two decks on each table. Participants' goal was to learn $\text{sgn}(\pi_1-\pi_2)$, or which deck had more orange (blue) cards on each table. We will denote this term, which serves as the learning desideratum for participants, as θ.

The task began with an exploration phase, followed by a test phase. On each trial of the exploration phase participants chose which of two tables to explore, and then revealed one card from a deck on that table (**Fig. 1b**). Participants were instructed that the exploration phase would be followed by a test phase after a random number of trials (drawn from a geometric distribution to discourage pre-planning, **Fig. 1c**). They were further instructed that one of the colours would be designated as rewarding at the beginning of the test phase. During the test phase, participants were asked to indicate which deck had more of the rewarding colour on each table (**Fig. 1b**). They also rated their confidence in the choice. For every correct test-phase choice they received \$0.25. Crucially, they received no reward during exploration.

Three Hypothetical Strategies Derived by Rational Analysis

To explain how participants chose between tables in the exploration phase, we employed rational analysis[24,25]: we asked how an optimal agent might solve the problem of choosing which table to explore on each trial of the task. We limited our consideration to strategies that optimise learning only for the next trial, since a globally optimal strategy is intractable in our case[2,3]. We started by deriving the optimal strategy and progressively simplified it to generate two additional hypotheses. Despite varying complexity, all three strategies direct an agent using them to approach the options they are more uncertain about.

The optimal strategy, given by the expression at the top of **Fig. 1d**, is choosing the table affording maximal expected information gain (EIG)[4,28,29]. EIG is the difference between the uncertainty in the value of the learning desideratum, θ, given observed cards $x_{0:t}$, and the expected uncertainty after observing the next card on trial $t + 1$. In other words, EIG is the amount of uncertainty resolvable on the next trial.

Computing the second term in the EIG expression requires averaging over future unseen outcomes, which may be beyond the ability of participants. As an alternative, they might avoid computing this term by simply choosing the table they were more uncertain about at the moment of making the

choice (**Fig. 1d**, second tier)[2]. While this strategy has intuitive appeal, computing uncertainties may still be too complicated for human participants. An even simpler heuristic is given on the third tier of figure 1d: choosing the table with the least prior exposure[2,30], measured as the number of already observed cards n_x. Since on average uncertainty is lower with more observations, this strategy is an approximate way to approach uncertainty. Finally, participants might explore at random, rather than in a directed manner[2,31,32].

Test Phase Performance Validates Observation Model

To relate the three hypothesized strategies to participants' behaviour, we first had to assume a model of participants' beliefs about θ and the mechanism by which they updated these beliefs. We used a Bayesian observer model which forms beliefs about θ based on the actual card sequence each participant observed, and updates these beliefs according to Bayes' rule (**Fig. 2**). The Bayesian observer is only a model of how participants are making inferences from observations, and is agnostic regarding which table or deck should be chosen for exploration.

Before evaluating the hypothesised exploration strategies, we sought to validate the assumptions of the Bayesian observer model. To this end, we related the predictions of the Bayesian observer model to participants' choices during the test phase. We predicted that test accuracy should be greater when the Bayesian observer model had low uncertainty about θ at the end of the learning phase. Using a multilevel logistic regression model, we confirmed that test accuracy was strongly related to the Bayesian observer's uncertainty b=-5.54, 95% posterior interval (PI)=[-6.18, -4.93] (**Fig. 3a**, Table S2). Participants' reports of confidence after making a correct choice also followed the Bayesian observer's uncertainty b=-4.03, 95% PI=[-4.51, -3.57]. After committing errors, participants' reported confidence was lower overall b=-1.07, 95% PI=[-1.25, -0.90], and considerably less dependent on Bayesian observer uncertainty, interaction b=-3.20, 95% PI=[-4.60,-1.93] (**Fig. 3b**, Table S3).

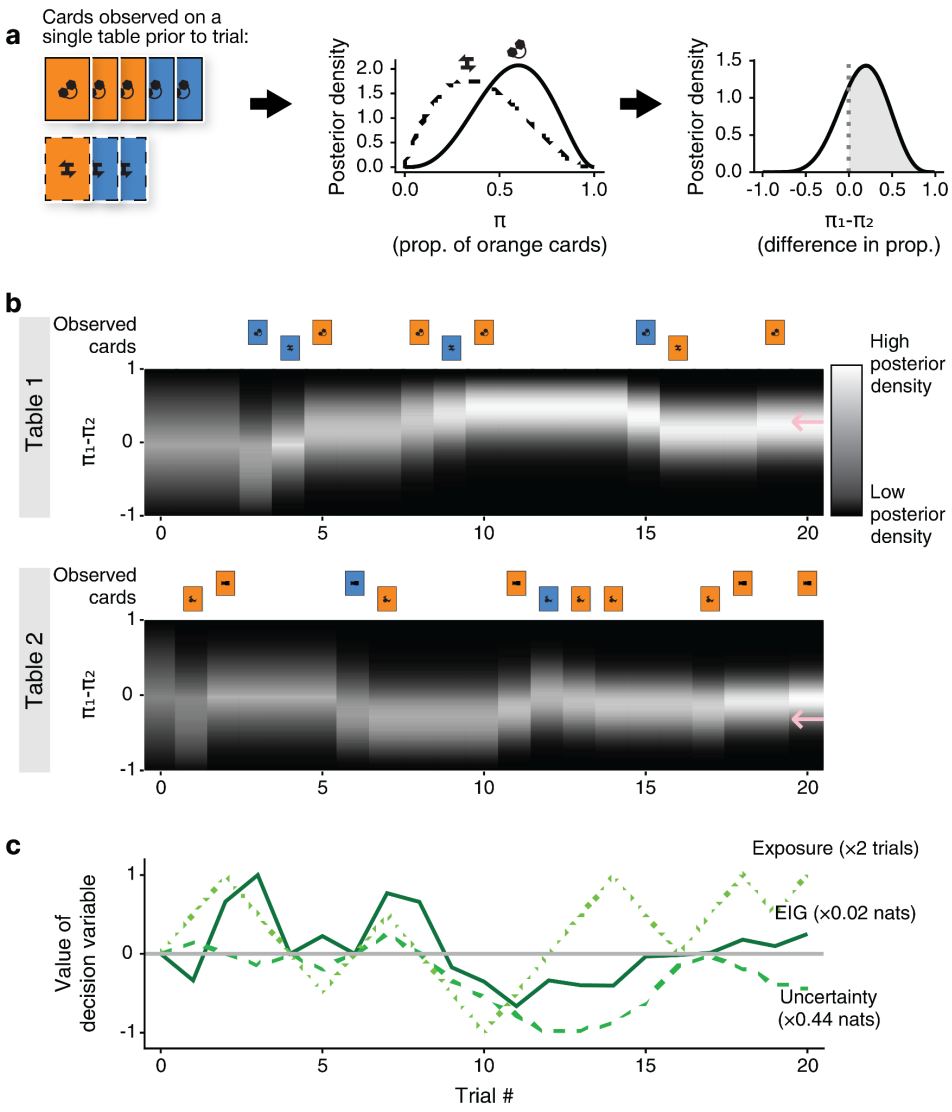

Figure 2. Hypothetical strategies make differing predictions for exploratory choice behaviour. We computed the three quantities hypothesized to drive exploratory choices using a Bayesian observer model. To illustrate this process, we plot the derivation of Bayesian belief on a single trial (**a**) and across multiple trials (**b, c**). For visualization, we use a simplified version with two tables only. **a** depicts the Bayesian observer's belief about a single table on a single trial. Given a sequence of previously observed cards (left), the Bayesian observer forms posterior beliefs about the proportion of orange cards in each deck (centre). These beliefs are expressed as Beta distributions. From these, it is possible to derive a belief about the difference in the proportion of orange cards between the two decks π_1-π_2 (right). The probability that $\pi_1 > \pi_2$ is given by the proportional size of the area marked in grey (0.74 in this example). **b** Depicts the same process over a series of 20 trials. The observed card sequence for each table is presented at the top of each panel. The matching belief state about π_1-π_2 is plotted below it as an evolving posterior density in white (high) and black (low). The pink arrows mark the true value of π_1-π_2 for that round. As the round progresses, belief converges towards the true value, and becomes more certain. **c,** The three choice strategies prescribe different table choices on most trials. The difference between table 1 and table 2 in each of the three quantities (EIG, uncertainty and exposure) is plotted for each trial. This difference is the hypothesized decision variable for choosing between tables 1 and 2. A positive value indicates a preference for exploring table 1, and a negative value a preference for table 2. The three variables are normalized to facilitate visual comparison.

Uncertainty is the Best Predictor of Exploratory Choice

To evaluate the three exploration strategies, we tested whether participants' exploration-phase choices could be predicted from the difference between the two tables that were presented as choice options in each of the hypothesized quantities. We fit the data with a multilevel logistic regression model for each strategy (Tables S4-6). In a formal comparison of the three models we found that uncertainty was the best predictor of exploratory choices, as indicated by a reliably better prediction metric (**Fig. 3c**).

Accordingly, as shown in **Figure 3d**, the uncertainty for the table presented on the right relative to the table presented on the left (Δ uncertainty) predicted participants' choices. Δ EIG provides a poorer fit to choices, and Δ exposure was anti-correlated with choice, in contradiction of the exposure hypothesis.

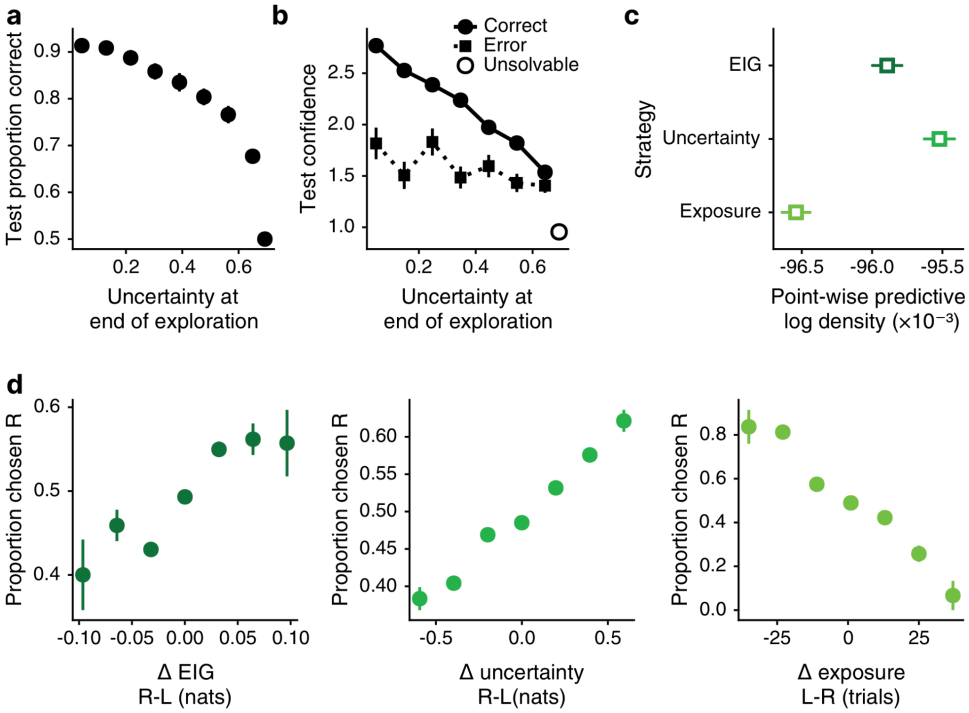

Figure 3. Uncertainty is the best predictor of choice. The Bayesian observer model is validated by participants' accuracy and confidence on the test phase. **a,** Participants were accurate when an exploration phase ended with low uncertainty, and performed at chance level when the phase ended with high uncertainty. **b,** Participant's confidence on correct choices fell with rising uncertainty. Confidence on errors did not depend as much on Bayesian observer uncertainty. **c**, Uncertainty is the best predictor of exploration-phase choices. Out of the three hypothesized strategies it has the highest approximate expected log predictive density. **d**, This result can be assessed qualitatively by plotting the difference in each hypothesized quantity between the two tables presented on each trial against actual choices of the table presented on the right. For each plot, the relevant hypothesis predicts a positive smooth curve. Δ uncertainty, plotted in the centre, matches this prediction better than Δ EIG (left). The relationship between Δ exposure (right) and choice is negative, rather than the hypothesized positive correlation. Data presented as mean values ±1 SE.

Participants Approach or Avoid Uncertainty According to Overall Uncertainty

Having identified uncertainty as the strategy that best accounts for participants' choices, we asked whether that strategy is modulated by the difficulty of making an exploratory choice. Our main index of difficulty was participants' overall uncertainty about the two options they could choose to explore on a

given trial (**Fig. 4a**). Since table choice options were presented at random, participants sometimes had to choose between tables they already knew a lot about, and sometimes between tables they were very uncertain about. When overall uncertainty was high, the choice between tables had to be made with very little evidence, and so was more difficult[2].

We found a systematic deviation in exploration strategy in relation to choice difficulty, as indexed by overall uncertainty. When overall uncertainty for the two choice options was below a certain threshold, participants chose the more uncertain table, as expected. However, when overall uncertainty was above the threshold, they chose the less uncertain table, thereby slowing the rate of information-intake (**Fig. 4b,c**).

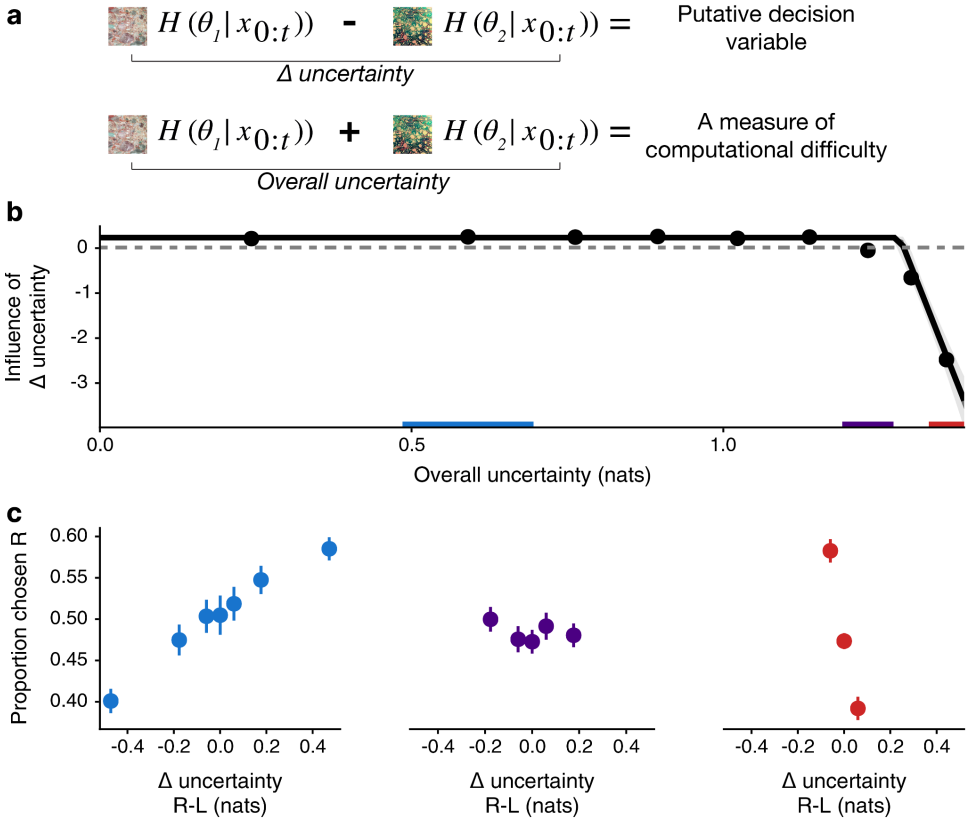

Figure 4. Participants approach vs. avoid Δ uncertainty as a function of overall uncertainty. a, While the Δ uncertainty is the decision variable identified above, overall uncertainty, defined as the sum of uncertainty for both tables, is a measure of decision difficulty. **b**, The influence of Δ uncertainty on choice differed markedly below and above a threshold of overall uncertainty. Below a certain estimated threshold of overall uncertainty, Δ uncertainty had a significant positive effect on choice. Above this threshold of overall uncertainty, the influence of Δ uncertainty decreased significantly. Points denote mean posterior estimate from regression models fitted to binned data, error bars mark 50% PI. The solid line depicts the prediction from a breakpoint regression model capturing the non-linear relationship and estimating the threshold, with darker ribbon marking 50% PI and light ribbon marking 95% PI. Data from three regions of overall uncertainty marked in colour are plotted in **c**. For low overall uncertainty (blue) participants tend to choose the table they are more uncertain about, as normatively prescribed. But that relationship is broken for medium levels of overall uncertainty (purple). For high overall uncertainty (red), participants strongly prefer to choose the table they are less uncertain about, thereby slowing down the rate of information intake. Data plotted as mean ±SE.

We validated this observation using a multilevel piecewise-regression model, allowing for the influence of Δ uncertainty on choice to differ below and above a fitted threshold of overall uncertainty. We observed a positive relationship between Δ uncertainty and choice below the threshold b=0.97, 95% PI=[0.84, 1.11], but above the threshold we found that the influence of Δ uncertainty on choice became strongly negative, interaction b=-4.3e+02, 95% PI=[-5.3e+02, -3.5e+02]. The value of the threshold was estimated to be 1.28 nats of overall uncertainty (95% PI=[1.27, 1.29]; Table S7). This bias in exploration cannot be viewed merely as a noisier version of optimal performance. Rather, it constitutes a systematic modulation of exploration strategy on difficult trials.

Approaching but not Avoiding Uncertainty is Associated with Test Performance

So far, we identified two aspects describing participants' exploration strategy – a baseline tendency to approach uncertainty and a tendency to avoid uncertainty when overall uncertainty is high. Since efficient learning is the purpose of exploration, we asked how each of these tendencies affected learning as reflected in performance at test. If the most important determinant of successful exploration is the ability to maximise information intake, then participants who tend to approach uncertainty to a greater degree should learn more and perform better at test, while participants with a strong tendency to avoid uncertainty should learn less and perform worse at test.

To test these predictions we examined individual differences in exploration strategy. We correlated each participant's test performance with the parameters from the piecewise regression model describing their tendencies to approach and avoid uncertainty. We found that participants' baseline tendency to approach uncertainty indeed predicted better performance at test b=2.98, 95% PI=[2.70, 3.26] (**Fig. 5b**; Table S8).

In contrast, we found no evidence that participants with a strong tendency to avoid uncertainty performed worse at test. First, participants who started avoiding Δ uncertainty at a lower overall uncertainty threshold actually had a weak tendency to do better at test b=-0.05, 95% PI=[-0.10, -0.01]. Secondly, the magnitude of uncertainty avoidance under high overall uncertainty was not associated with test performance b=0.05, 95% PI=[-0.05, 0.14] (**Fig. 5,** Tables S9-10). Thus, modulating exploration according to overall uncertainty was not maladaptive, resulting in no decrement to learning. This result suggests that the rate of information intake is not the limiting factor for the efficiency of exploration and learning.

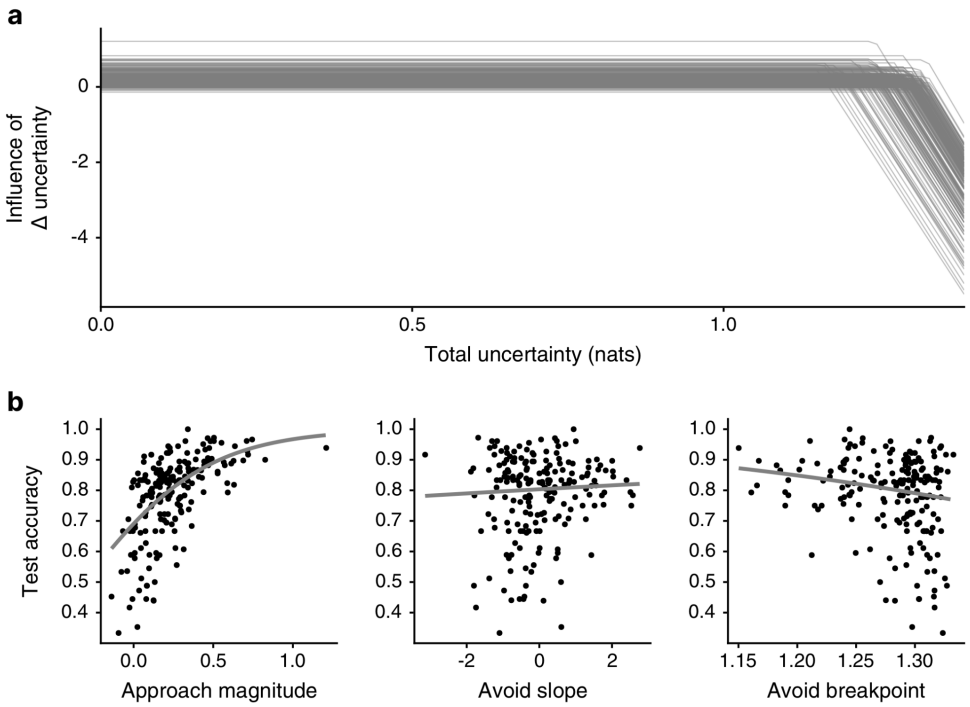

Figure 5. Individual differences in exploration strategy are correlated with test performance. a, We observe substantial individual differences in strategy. Replotting Fig. 4e for each individual reveals differences in the baseline influence of Δ uncertainty on choice, and the interaction with overall uncertainty. **b**, Associations between test performance and the parameters describing approaching and avoiding uncertainty. The baseline tendency to approach uncertainty (left) is strongly associated with performance at test, such that participants who are unable to approach uncertainty also perform poorly at test. The two parameters describing uncertainty avoidance at high overall uncertainty - the slope of the interaction (middle) and position of the breakpoint (right) are not strongly correlated with test performance, indicating that uncertainty avoidance does not hinder learning.

Individual Differences in Exploration Reaction Times Associated with Test Performance

Examining reaction times (RTs) further illuminated the link between exploration strategy and successful learning (**Fig. 6**). We focused on RTs measured on exploration-phase trials with overall uncertainty below the estimated threshold, since we found choice strategy on these trials to be tightly linked to successful test performance. We expected RTs to depend on the absolute value of Δ uncertainty, since this value captures the amount of available evidence upon which participants could base their choice. Specifically, when the absolute value of Δ uncertainty is large, the decision should be easy and RTs should be short, while when the absolute value is small, the decision should be difficult and RTs long. To test for this possible relationship between Δ uncertainty and RTs, we fit the data with a generative model of choice and RTs. We used a sequential sampling model, which explains decisions as the outcome of a process of sequential sampling that stops when the accumulation of evidence satisfies a bound. This model explains RTs as jointly influenced by participant's efficacy in deliberating about Δ uncertainty, and their tendency to deliberate longer vs. make quick responses[33–35]. The basic predictions of sequential sampling models are that greater deliberation efficacy should be manifested as a greater dependence of RT on the strength of the evidence (here, absolute Δ uncertainty), and that a stronger tendency to deliberate manifests in longer RTs when the evidence is weak[36].

We found that RTs indeed varied in relation to the absolute value of Δ uncertainty as expected b=0.67, 95% PI=[0.57, 0.77] (Table S11). Crucially, a strong dependence of RT on the absolute value of

Δ uncertainty predicted better performance at test b=0.83, 95% PI=[0.58, 1.09]. We further found that participants who tended to deliberate longer for the sake of accuracy also tended to perform better at test b=1.42, 95% PI=[0.55, 2.31] (**Fig. 6c**, Table S12). In summary, participants who were better at deliberating about uncertainty during exploration, and who deliberated for longer, performed better at test.

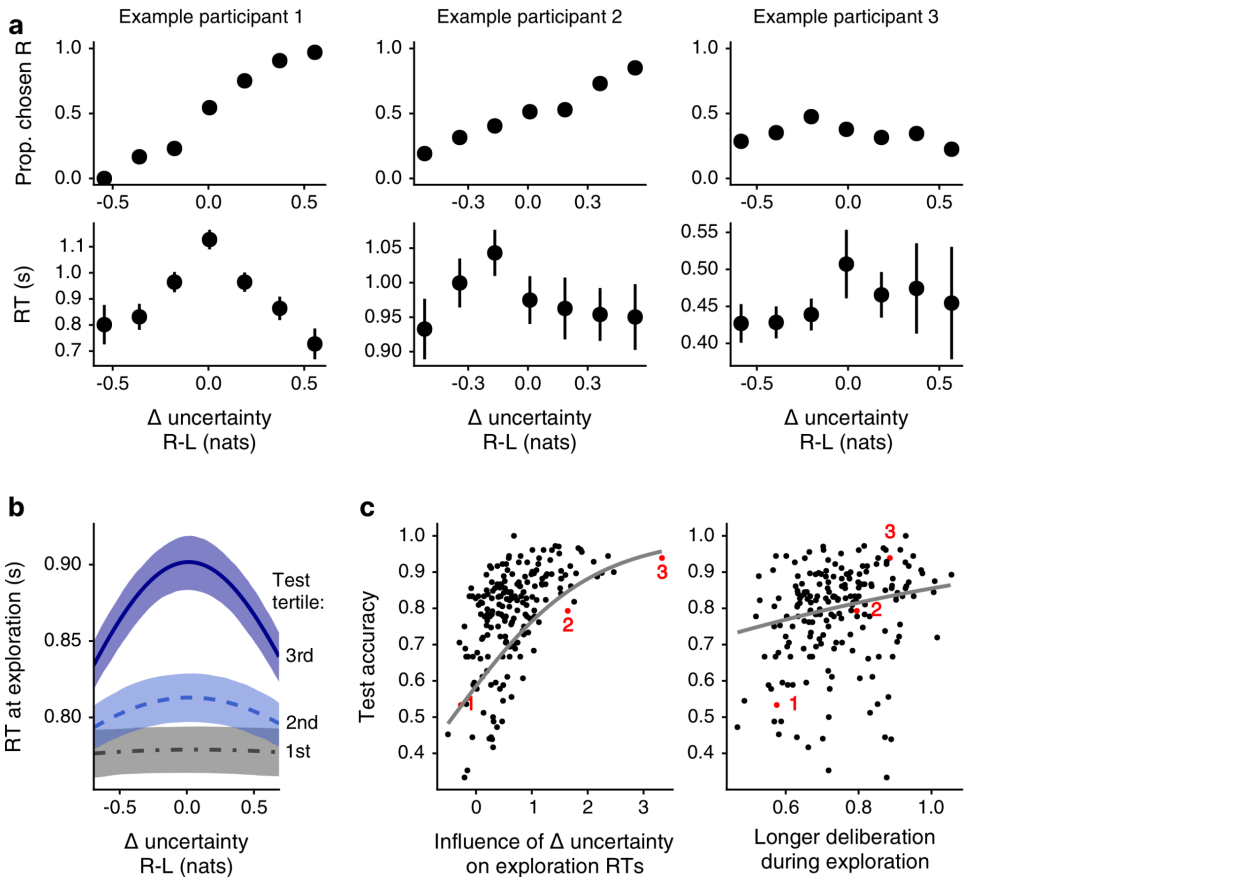

Figure 6. Individual differences in choice reaction times explained by a sequential sampling model. Participants varied not only in the pattern of their choices, but also in their RTs. **a**, Data from three example participants. The relationship of choice and RTs with Δ uncertainty weakens from left to right. Data plotted as mean ±SE. **b**, These individual differences were captured by a sequential sampling model, explaining choices and RTs as the interaction between participant's efficacy of deliberating about Δ uncertainty and their tendency to deliberate longer vs. make quick responses. Plotting model predictions, we observe a u-shaped dependence of RTs on Δ uncertainty for participants whose performance at test was in the top accuracy tertile. This characteristic u-shape is commensurate with Δ uncertainty being the decision variable used to guide exploration. This relationship is weaker for participants in the bottom two test accuracy tertiles. Such participants also exhibit shorter RTs overall. Lines mark mean predictions from a sequential sampling model fit by tertiles for visualization, ribbons denote 50% PIs. **c,** Correlating the sequential sampling model parameters with test performance confirms these observations. Participants with a stronger dependence of RT on Δ uncertainty perform better at test, as do participants who deliberate longer for the sake of accuracy. Example participants from **a** are marked in red. Lines are mean predictions from a logistic regression model.

Participants Display a General Tendency to Repeat Their Previous Choice

The observed change in strategy in association with overall uncertainty can also be described as a tendency to revisit the better-learnt tables when overall uncertainty is high. We also identified an

independent pattern of revisiting previous choices that held across all levels of Δ uncertainty or overall uncertainty. We found that participants generally preferred to re-choose the table they had last chosen b=0.50, 95% PI=[0.42,0.59] (**Fig. 7b**, Table S13). This tendency to repeat choices was also reflected in RTs, which for repeat choices were less related to Δ uncertainty (b=-0.34, 95% PI=[-0.44,-0.24]). We also found that participants tended to make repeat choices more quickly rather than deliberate longer (b=-0.05, 95% PI=[-0.06,-0.04]; **Fig. 7c**, Table S14). Hence, repeating a choice seems to be an additional heuristic participants employed to avoid the deliberation involved in choosing according to Δ uncertainty.

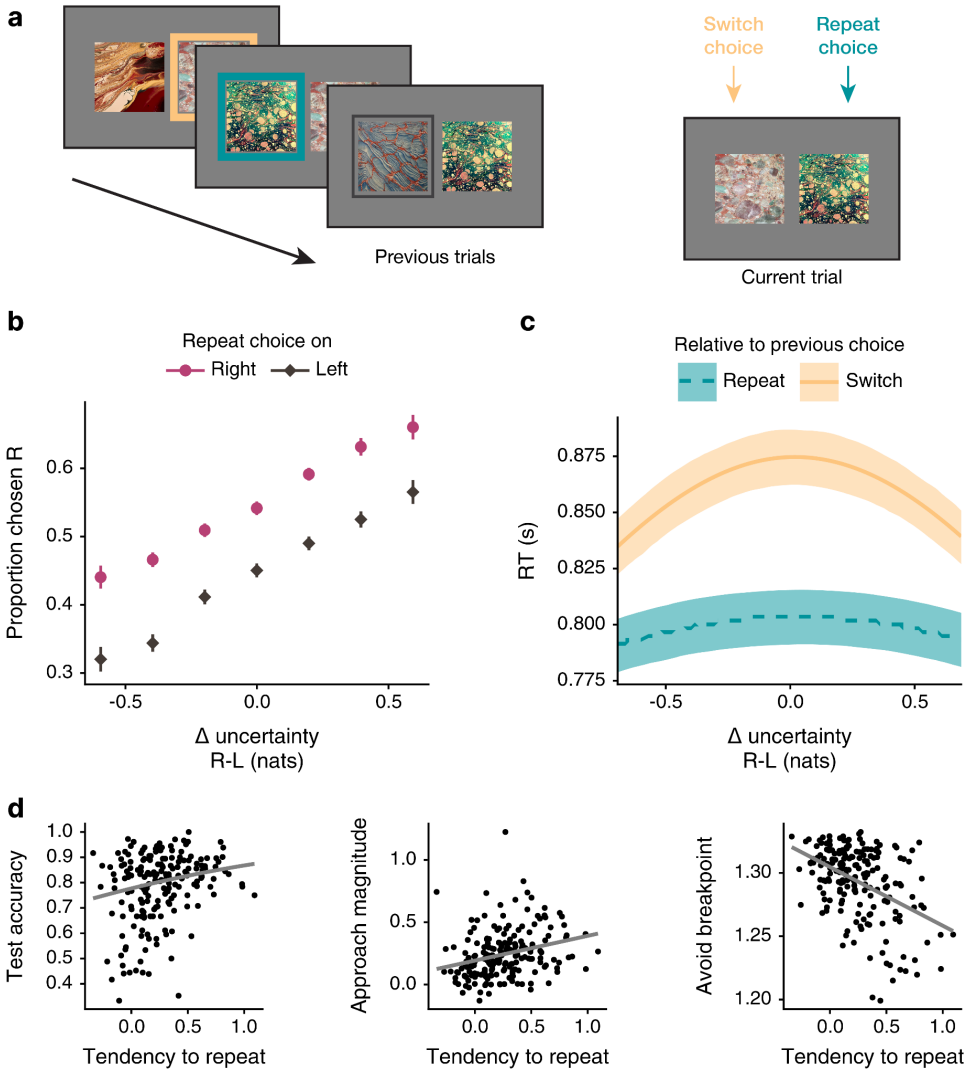

Figure 7. Participants tend to repeat previous choices instead of deliberating over uncertainty. a, On a given trial one table has been chosen more recently than the other (frames denote previous choices). In the example the green table had been chosen more recently, hence it is designated the repeat option and the other table the switch option. **b**, Participants tend to choose the table displayed on the right more often when it is the repeat option than when it is the switch option. Data plotted as mean ±SE. **c**, When choosing a repeat option, participants' RTs are shorter and less dependent on Δ uncertainty. Lines mark mean predictions from a sequential sampling model, ribbons denote 50% PIs. **d**, Participants who tended to repeat their previous choice also tended to perform better at test (left), were more likely to have a stronger baseline tendency to approach uncertainty (middle), and were more likely to start avoiding uncertainty at a lower overall uncertainty threshold (right). Regression lines are plotted for visualization.

As in other aspects of exploration strategy, we observed considerable individual differences in the tendency to repeat previous choices. These differences were associated with the uncertainty-based aspects of exploration discussed above (**Fig. 7d**). Participants with a general tendency to repeat choices show stronger uncertainty avoidance when overall uncertainty is high, indicating that these two conceptually related strategies also co-occur in the population r=-0.61, 95% PI=[-0.75,-0.44] (Table S13). Furthermore, the tendency to repeat previous choices is associated with better test performance, logistic regression b=0.09, 95% PI=[0.07,0.12] (Table S15). The tendency to repeat is also correlated with a stronger baseline tendency to approach uncertainty r=0.32, 95% PI=[0.17,0.46] (Table S13), which was shown above to be correlated with test performance. Thus, while from a normative point of view repeating the previous choice appears to be a context-insensitive strategy, in practice participants who use this strategy do not learn any worse.

Forgetting as a Conceptual Control

We have found that participants deviate from the optimal exploration policy, reducing the rate of information intake to strategically conserve cognitive resources. However, explaining a deviation from optimality as strategic is interesting only to the extent that the null hypothesis – that participants are simply failing at making good decisions – is also a-priori plausible. We turned to forgetting as a second source of difficulty in our task to make sure that this interpretation of behaviour as strategic was not a forgone conclusion. Since only two out of the four tables are randomly presented on each trial as choice options, there was variability in the number of trials since either of the presented tables was last explored – a number we refer to as memory lag. We assumed that choosing between tables that had not been explored for a long time is more difficult than between tables for which evidence has been recently observed. Indeed, we found that RTs were longer with large memory lag, indicating greater difficulty of making a choice (lognormal regression b=0.02, 95% PI=[0.02, 0.03]; **Fig. 8a**, Table S16). Furthermore, we observed that exploration choices on trials with a greater lag depended less on Δ uncertainty b=-0.08, 95% PI=[-0.11, -0.05], and that the tendency to repeat the last chosen table on these trials was also diminished b=-0.13, 95% PI=[-0.15, -0.11] (**Fig. 8b**, Table S17). Finally, on trials with a large lag the difference in RTs between making a repeat and a switch choice disappeared, interaction b=0.02, 95% PI=[0.02,0.03] (**Fig. 8a**, Table S16). These patterns suggest that prior evidence is forgotten with increasing memory lag and that as a consequence exploration becomes more random. Hence, in contrast to the systematic effect of overall uncertainty, forgetting results in a failure to make principled exploratory choices.

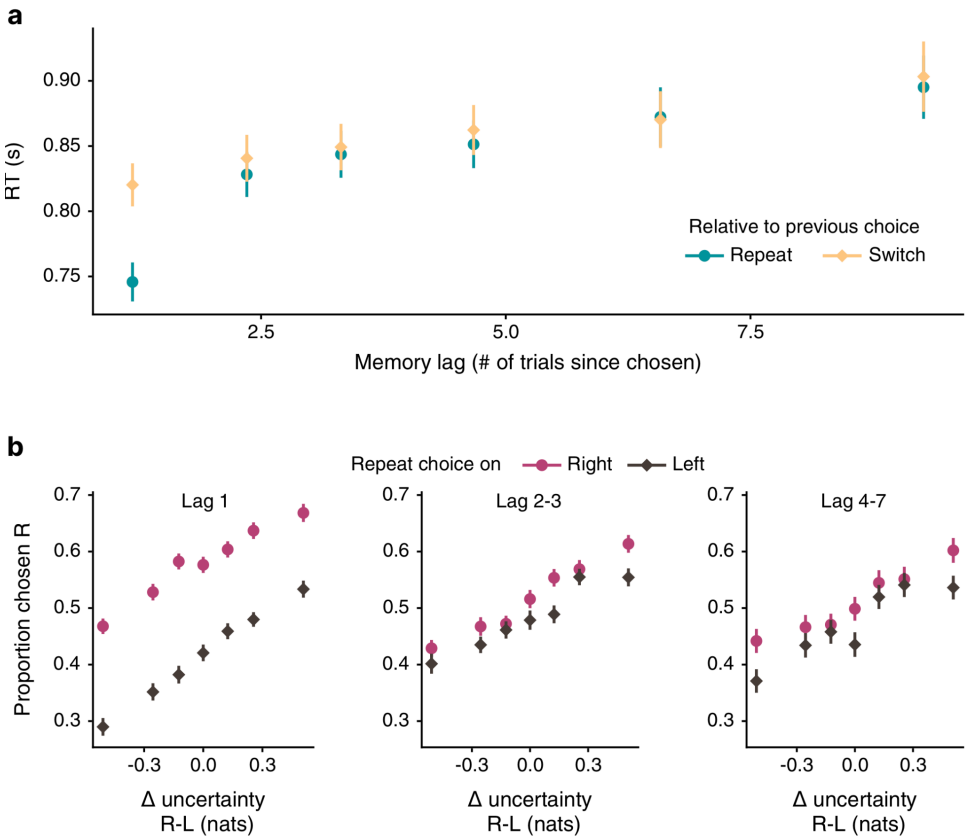

Figure 8. Forgetting is associated with random choice rather than a systematic bias. a, Memory lag, defined as trials since last choice, serves as a proxy for forgetting and contributes to the difficulty of making an exploratory choice. RTs rise with memory lag. The repeat choice RT advantage disappears with rising memory lag. Data plotted as mean ±SE. **b**, With higher memory lag choices become less dependent on Δ uncertainty, as indicated by flatter curves. The tendency to repeat the last choice is also diminished with memory lag. Both effects amount to choice becoming more random due to forgetting. Data plotted as mean ±SE.

Discussion

We examined the cognitive computations behind exploratory choices in a setting allowing for incremental learning in the service of a distant goal. We found that uncertainty played an important role in guiding participants' choices about how to sample their environment for learning. In general, participants chose to learn more about the options they were more uncertain about. However, when overall uncertainty was especially high, participants instead avoided the more uncertain options and sampled the options they already knew more about. In addition, we found that participants tended to repeat previous choices. Together, this pattern suggests that participants systematically balance approaching and avoiding uncertainty while exploring.

We further examined this pattern by assessing individual differences in exploration and test performance. We found that strategically avoiding uncertainty is not associated with a detriment to learning, even though it slows down the rate of information intake. Reaction time modelling revealed a tradeoff between deliberation effort and learning efficiency. Participants who deliberated longer learned better, but deliberation time could be shortened by repeating previous choices. Based on these results, we conclude that balancing approaching and avoiding uncertainty is a way to manage cognitive resources by

regulating deliberation costs. In this sense, our results serve as an example of how human cognition is adapted to the inherent constraints of the human mind, as predicted by the resource rationality framework[22].

This work extends the treatment of exploration in two established literatures. Researchers of reinforcement learning have previously examined how exploration manifests when agents learn incrementally about their environment. Crucially, this literature has focused on cases where reward can be gained on each trial[1,2,13,31,32,37–39]. In contrast, our task was designed to remove the impetus to exploit current knowledge immediately, a motivation that predominates exploration in tasks with immediate reward. Accordingly, we were able to observe many exploratory choices and had greater experimental power to describe in detail how participants approach uncertainty and when they avoid it instead. Exploration has also been studied in the information search literature[28,40–44]. In most studies of this field participants make decisions without relying on their memory - as the entire history of learning is displayed to them on screen (cf. related work in active sensing[45]). This differs from our task, which places heavy demands on memory. Rather than treating capacity limitations as sources of noise and a nuisance to measurement, we find that the rational use of limited resources is a central computational goal of exploration.

We observed considerable individual differences in exploration strategy, as would be expected in a complex task requiring memory-based learning and inference over a hierarchical environment. In the face of such variability, one may question the prudence of drawing conclusions about the population, since the average might be a poor summary of a plurality of idiosyncratic strategies. However, the strong correlation we observed between individual differences in exploration and test performance mitigates this concern. The correlation suggests that participants who were engaged with this task and able to learn from observation can well be described as exploring by balancing approaching and avoiding uncertainty. The relationship between test performance and RTs lends additional mechanistic support to this idea. Our results underline the importance of acknowledging and interpreting individual differences in cognition as strong tests of behavioural patterns and cognitive mechanisms.

Our theoretical analysis and experiments leave several open questions. First, overall uncertainty in our task was correlated with the number of cards observed. While our results hold when trial number is added as a covariate to the regression models (see Table S18), future work orthogonalizing overall uncertainty and time on task would help to fully disentangle the contribution of each factor to uncertainty avoidance.

Another open question is the source of difficulty engendered by overall uncertainty. Decisions based on high overall uncertainty may be more difficult due to limitations in committing prior experiences to memory, in inferring latent parameters from disparate experiences, in retrieving prior knowledge, or in estimating the uncertainty of existent knowledge. While the idea that decisions based on high overall uncertainty are more difficult has been previously motivated on computational grounds[2,46], an explanation grounded in cognitive mechanisms is still needed. Accordingly, the mechanism by which uncertainty avoidance ameliorates choice difficulty remains unknown.

One intriguing explanation for the source of difficulty and the way it is managed lies in the distinction between learning strategies dependent on remembering single experiences and those dependent on slower incremental learning of summary statistics[47–52]. Both strategies could contribute to performance in tasks such as ours. A participant may be encoding prior observations as single instances, or summarize them into a central tendency with a margin of uncertainty around it. Crucially, each strategy is associated with a different profile of cognitive resource use. Keeping track of individual experiences is

much costlier than tracking a single expectation and a confidence interval around it[52,53] and more likely to incur costs when switching between exploring different tables. Prior work suggests individuals switch between using single experiences and summary statistics according to the reliability of each strategy, and the cost of using it[52,53]. In our case, summary statistics may be perceived as unreliable when overall uncertainty is high, compelling participants to rely on committing individual experiences to working memory[47,50,52,54,55]. Furthermore, recent work examining how humans make a series of related decisions demonstrates that the tension between remembering single experiences and discarding them in favour of summary statistics is accompanied by a tendency to revisit previous choices instead of switching to new alternatives[56].

The idea of a balance between approaching and avoiding uncertainty has conceptual parallels in other literatures. A group of relevant findings concern how animals explore their proximal environment. A classic finding in rats is that when placed in a novel open arena, they alternate between the exploration strategy of walking around the arena (uncertainty approaching) and a strategy of returning to their initial position and pausing there (termed "home base" behaviour, which is uncertainty avoiding)[14]. Relatedly, by using computational models to understand how rats use their whiskers to explore near objects, researchers have identified an alternation between uncertainty approaching and avoiding strategies[15]. Recent work in mice and primates has successfully uncovered neural circuits driving exploration by framing the problem of exploration as striking a balance between exploration and avoidance[16,17,57]. Our findings highlight the shared computational principles between human exploration in symbolic space and animal exploration of the physical environment and suggest that mechanisms involved in avoidance responses may also play a part in epistemic knowledge building.

The questions we addressed here were partly motivated by the well-established observation that humans and animals often avoid uncertainty in various situations. Two broad categories of explanation for such avoidance have been proposed[20]. First, individuals avoid resolving uncertainty when it could lead to negative news, for example by avoiding ambiguous prospects when making economic choices[58,59]. An extension of this idea is dread avoidance[19,20]. One might avoid resolving the uncertainty about a medical diagnosis to avoid the unpleasant affective response to the news, even if the information could be very useful in determining treatment. A second reason to avoid uncertainty is the strategic management of conflict between different motivations, or different mechanisms of action selection[20,21]. For example, to maintain their diet, an individual might choose to avoid resolving the uncertainty about what snacks can be found in the office kitchen. Our findings highlight another kind of strategic uncertainty avoidance. In our tasks there were no negative consequences to learning about the colour proportions of card decks, and no conflicting motivations. Rather, we explain participants' tendency to avoid uncertainty in terms of managing their limited memory and learning resources.

Finally, human planning[60], learning[28], and sensing[15,45] are increasingly studied as active processes, situated within and interacting with our environment. Understanding the complicated dynamics between agent and environment has been greatly facilitated by comparing behaviour against the computational ideal of maximising the amount of information observed[2,10,40]. The findings we present here suggest a modification to this computational premise. Rather than trying to uncover as much information as possible, the goal of human exploration is to maximize the amount of information retained in memory, by modulating the rate and order of observed information.

Methods

Data Collection and Participants

A sample of 298 participants was recruited via Amazon MTurk to participate in four sessions of the exploration task. They were paid $3.60 for each session and earned a bonus contingent on their test phase performance, adding up to $4.50 for the first session and $6 for later sessions. Additionally, a $2 bonus was paid out for completion of the fourth session. Participants were asked to complete the four sessions over the course of a week and were invited by email to each session after the first, as long as the data from their last session was not excluded according to the criteria we had specified (see below).

The first session was terminated early for 89 participants due to recorded interactions with other applications during the experiment or failure to comply with instructions (see Supplementary Information). An additional 32 sessions played by participants who had successfully completed the first session were excluded for the same reasons. One participant was excluded after reporting technical problems with stimulus presentation in the second session. Twenty-seven further sessions were excluded for failure to sample cards from both decks, a prerequisite for learning on which participants were instructed as part of the training. Altogether, data from 194 participants was included in the analysed sample (120 female, 72 male, 2 other gender, average age 29.63, range 20-48). This sample included 194 first sessions, 156 second sessions, 129 third sessions, and 116 fourth sessions.

Before running this experiment, we pre-registered a sample size of 190 participants satisfying our exclusion criteria. We chose this number to be three times larger than a preliminary sample of N=62 participants, which provided the dataset we used to develop our analysis approach and pipeline, and first identify exploration strategies as described above. Results for the preliminary sample are provided in Supplementary Information.

Task Design and Procedure

On each round of the exploration task participants were presented with a simple environment of four tables with two decks of cards on each table. Tables were distinguished by unique colourful patterns and decks by geometric symbols that did not repeat within an experimental session. The hidden side of each card was painted in one of two colours, with a unique colour pair for each round. The proportion of colours in each deck were determined pseudo-randomly (see Supplementary Information), resulting in variability in the difference in proportion between each deck pair - the learning desideratum of this task.

At the beginning of each round, participants were first presented with the colour pair for the round, and then with the table-deck assignments. Participants then had to pass a multiple-choice test on the table-deck assignment, making sure they remembered the structure of the task before proceeding to explore. Failing to get a perfect score on this test resulted in repeating this phase.

The exploration phase then commenced. Trial structure for the exploration phase is depicted in **Fig. 1b**. The lengths of the exploration phases varied from round to round. They were sampled from a geometric distribution with rate 1/44, shifted by 10 trials. The same list of round lengths was used for all participants, but their order was randomised.

Following the exploration phase, participants were tested on their learning. They were presented with the rewarding colour for this round, and then had to indicate which deck had a greater proportion of that colour on each table (**Fig. 1c**). After answering this question for each of the four tables, they rated their confidence in each of the four choices on a 1-5 Likert scale. Participants were then told whether each of the test choices were correct, and the true colour proportions for the two decks on each table were presented to them as 10 open cards.

The first session started with extensive instructions explaining the structure of each of the two phases of the task and clearly stating the learning goal. Participants were also instructed on the independence of colour proportion within each deck pair, necessitating sampling from both decks to succeed in the task. The instructions also included training on how to make the relevant choices in each of the two stages. A quiz followed the instruction phase, and participants had to repeat reading the instructions if they had given the wrong response to any question on this quiz.

Each session started with a short practice round (12-19 trials). Data from this round was excluded from analysis. In the first session participants then played three more rounds and in later sessions five more rounds, for a total of 18 experimental rounds.

Data Analysis

Bayesian Observer

Each of the three hypothesized strategies for exploration postulates a different summary statistic of prior learning as the driver of exploratory choice. To derive these summary statistics, we first had to construct a model of prior learning. We chose a simple Bayesian observer model[45,61]. Like our participants, this model's goal was to learn $\theta=\text{sgn}(\pi_1-\pi_2)$ from observed outcomes $x_{0:t}$. It did so by placing a probabilistic prior over the value of each π_i, updating it after every observation according to Bayes' rule, and solving for θ using the rules of probability. The result is a posterior distribution capturing the agent's expectation of the value of θ, and their uncertainty about the expectation. This process is depicted in **Fig. 2** for two tables and their matching pairs of decks.

This computation can be put into formulaic form as follows. At the beginning of a round, the Bayesian observer places a flat Beta distribution prior on the proportion of colours in each of the eight decks:

$$\pi_i \sim Beta(1,1)$$

After observing a card, this prior would be updated to form a posterior distribution. Since the posterior of a Beta prior and a Bernoulli observation likelihood is also a Beta distribution, the posterior has a simple analytic form: after completing t trials, observing c_i cards of one colour and $t - c_i$ cards of the other colour, the posterior would be:

$$\pi_i|x_{0:t} \sim Beta(1 + c_i, 1 + t\ - c_i)$$

We can then find the probability that $\theta=1$, i.e. that $\pi_1>\pi_2$, by calculative the probability that π_2 is smaller than a given $\pi_1=z$, and integrating over z, the possible values of π_1:

$$P(\theta = 1|x_{0:t})\ =\ \int_0^1 \quad f_{\pi_1|x_{0:t}}(z)F_{\pi_2|x_{0:t}}(z)\text{dz}$$

Where f is the Beta probability density function, F is the Beta cumulative density function, and $x_{0:t}$ are observations thus far. We computed the value of this integral numerically using the Julia package QuadGK.jl. Finally, θ can only take two values, and so

$$P(\theta = -1|x_{0:t})\ =\ 1 - P(\theta = 1|x_{0:t})$$

Computing Hypothesized Decision Variables

The theory of decision making defines a decision variable as the quantity evaluated by the decision maker in order to choose between two choice options[33]. The difficulty of the decision should scale with the absolute value of the decision variable. Each of the three hypothesized strategies is defined by a specific summary statistic of prior learning that might serve as the decision variable for an exploratory choice. The three summary statistics are given in figure 1e.

Both EIG and uncertainty are derived from the uncertainty of the posterior for θ as defined above. We quantified uncertainty as the entropy of the posterior belief[4,29,40]:

$$H(\theta|x_{0:t}) = -\sum_{\theta=-1,1} P(\theta|x_{0:t}) ln P(\theta|x_{0:t})$$

Entropy takes the unit of nats, ranging from 0 should the participant be absolutely sure about the value of θ for both table choice options, to 0.69 when they know nothing about a table. This is the equivalent of 1 bit of information, were we to replace the natural logarithm with a base 2 logarithm.

Sequential Sampling Model of Reaction Times

To draw inference from participants' RTs we turned to the sequential sampling theory of deliberation and choice. This theory encompasses a family of models in which decisions arise through a process of sequential sampling that stops when the accumulation of evidence satisfies a threshold or bound[33,36]. From this family of models we chose to use the drift diffusion model (DDM) to fit our data, as it is very well described and extensively studied[33,35]. The DDM explains RTs as the culmination of three interpretable terms. The first is the efficacy of a participant's thought process in furnishing relevant evidence for the decision - in our case the efficacy of calculating Δ uncertainty (the *drift rate* in DDM parlance). The second term governs the participant's speed-accuracy tradeoff by determining how much evidence they require to commit to a decision. This can also be thought of as how long a participant is willing to deliberate when a decision is difficult (*bound height*). Finally, the portion of the RT not linked to the deliberation process is captured by a third term (*non-decision time*). Since behaviour was considerably different when overall uncertainty was high, DDM models were fit excluding trials with total uncertainty above the participant's estimated threshold. See Supplementary Information for a full specification of the model.

Estimating Multilevel Bayesian Models for Inference

The regression coefficients and PIs reported here were all estimated using multilevel regression models accounting for individual differences in behaviour. We used regularising priors to facilitate robust estimation (Table S1). For predicting choices, we used logistic regression, for confidence ratings we used ordinal-logistic regression, and for average RTs we used lognormal regression. We estimated these models with Hamiltonian Monte Carlo implemented in the Stan probabilistic programming language using the R package brms. Three Monte-Carlo chains were run for each model, collecting 1000 samples each after a warm up period of at least 1000 samples (warm up was extended if convergence had not been reached). Sequential sampling models were estimated using slice sampling, implemented in the python package HDDM. Four Monte-Carlo chains were run for each model, collecting 2000 samples each after a warm up period of at least 2000 samples. Convergence for both model types was assessed using the $\hat{R}$ metric, and visual inspection of trace plots. R syntax formulae and coefficients for covariates for all models mentioned in the main text are reported in Supplementary Information.

Model Evaluation

We compared the models of choice and RTs to alternative models, either reduced or expanded (see Supplementary Information). We used the LOO R package to perform approximate leave-one-out cross validation for models implemented in Stan. This method uses pareto-smoothed importance sampling to approximate cross validation in an efficient manner[62]. Models implemented in HDDM were compared

using the DIC metric. We also performed posterior predictive checks for our models, making sure they capture the theoretically important qualitative features of the data.

Data Availability

The entire dataset discussed here has been deposited on the OSF database https://osf.io/6zyev/. Source data for figures are provided in this paper.

Code Availability

The code used to run the experiment as well as the scripts and the software environment we used for analysis have been deposited on the OSF database https://osf.io/6zyev/.

References

1. Cohen, J. D., McClure, S. M. & Yu, A. J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. B Biol. Sci.* **362**, 933–942 (2007).
2. Schulz, E. & Gershman, S. J. The algorithmic architecture of exploration in the human brain. *Curr. Opin. Neurobiol.* **55**, 7–14 (2019).
3. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction, 2nd ed*. xxii, 526 (The MIT Press, 2018).
4. MacKay, D. J. C. Information-based objective functions for active data selection. *Neural Comput.* **4**, 590–604 (1992).
5. Sebastiani, P. & Wynn, H. P. Maximum entropy sampling and optimal Bayesian experimental design. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62**, 145–157 (2000).
6. Badia, A. P. *et al.* Agent57: Outperforming the Atari Human Benchmark. in *Proceedings of the 37th International Conference on Machine Learning* 507–517 (PMLR, 2020).
7. Raposo, D. *et al.* Synthetic Returns for Long-Term Credit Assignment. Preprint at http://arxiv.org/abs/2102.12425 (2021).
8. Bellemare, M. *et al.* Unifying Count-Based Exploration and Intrinsic Motivation. 9.
9. Pathak, D., Agrawal, P., Efros, A. A. & Darrell, T. Curiosity-driven Exploration by Self-supervised

Prediction. in *Proceedings of the 34th International Conference on Machine Learning* 2778–2787 (PMLR, 2017).

10. Schwartenbeck, P. *et al.* Computational mechanisms of curiosity and goal-directed exploration. *eLife* **8**, 1–45 (2019).

11. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074 (2014).

12. Speekenbrink, M. & Konstantinidis, E. Uncertainty and Exploration in a Restless Bandit Problem. *Top. Cogn. Sci.* **7**, 351–367 (2015).

13. Wu, C. M., Schulz, E., Pleskac, T. J. & Speekenbrink, M. Time pressure changes how people explore and respond to uncertainty. *Sci. Rep.* **12**, 4122 (2022).

14. Eilam, D. & Golani, I. Home base behavior of rats (Rattus norvegicus) exploring a novel environment. *Behav. Brain Res.* **34**, 199–211 (1989).

15. Gordon, G., Fonio, E. & Ahissar, E. Emergent Exploration via Novelty Management. *J. Neurosci.* **34**, 12646–12661 (2014).

16. Botta, P. *et al.* An Amygdala Circuit Mediates Experience-Dependent Momentary Arrests during Exploration. *Cell* **183**, 605-619.e22 (2020).

17. Ahmadlou, M. *et al.* A cell type–specific cortico-subcortical brain circuit for investigatory and novelty-seeking behavior. *Science* **372**, eabe9681 (2021).

18. Glickman, S. E. & Sroges, R. W. Curiosity in zoo animals. *Behaviour* **26**, 151–187 (1966).

19. Gigerenzer, G. & Garcia-Retamero, R. Cassandra's regret: The psychology of not wanting to know. *Psychol. Rev.* **124**, 179 (2017).

20. Golman, R., Hagmann, D. & Loewenstein, G. Information avoidance. *J. Econ. Lit.* **55**, 96–135 (2017).

21. Carrillo, J. D. & Mariotti, T. Strategic ignorance as a self-disciplining device. *Rev. Econ. Stud.* **67**, 529–544 (2000).

22. Lieder, F. & Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the

optimal use of limited computational resources. *Behav. Brain Sci.* **43**, (2020).

23. Hartley, C. A. How do natural environments shape adaptive cognition across the lifespan? *Trends Cogn. Sci.* **26**, 1029–1030 (2022).

24. Anderson, J. R. *The adaptive character of thought*. (Psychology Press, 1990).

25. Chater, N. & Oaksford, M. Ten years of the rational analysis of cognition. *Trends Cogn. Sci.* **3**, 57–65 (1999).

26. Waskom, M. L., Okazawa, G. & Kiani, R. Designing and Interpreting Psychophysical Investigations of Cognition. *Neuron* **104**, 100–112 (2019).

27. Abir, Yaniv, Shadlen, M. N. & Shohamy, D. Memory-based incremental exploration in a stochastic environment. *AsPredicted* https://aspredicted.org/hx6gj.pdf (2021).

28. Gureckis, T. M. & Markant, D. B. Self-directed learning: A cognitive and computational perspective. *Perspect. Psychol. Sci.* **7**, 464–481 (2012).

29. Yang, S. C. H., Wolpert, D. M. & Lengyel, M. Theoretical perspectives on active sensing. *Curr. Opin. Behav. Sci.* **11**, 100–108 (2016).

30. Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3**, 397–422 (2002).

31. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore--exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074 (2014).

32. Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).

33. Shadlen, M. N. & Kiani, R. Decision making as a window on cognition. *Neuron* **80**, 791–806 (2013).

34. Shadlen, M. N. & Shohamy, D. Decision Making and Sequential Sampling from Memory. *Neuron* **90**, 927–939 (2016).

35. Ratcliff, R. & McKoon, G. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Comput.* **20**, 873–922 (2008).

36. Palmer, J., Huk, A. C. & Shadlen, M. N. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J. Vis.* **5**, 1 (2005).

37. Tversky, A. & Edwards, W. Information versus reward in binary choices. *J. Exp. Psychol.* **71**, 680–683 (1966).

38. Song, M., Bnaya, Z. & Ma, W. J. Sources of suboptimality in a minimalistic explore–exploit task. *Nat. Hum. Behav.* **3**, 361–368 (2019).

39. Brown, V. M., Hallquist, M. N., Frank, M. J. & Dombrovski, A. Y. Humans adaptively resolve the explore-exploit dilemma under cognitive constraints: Evidence from a multi-armed bandit task. *Cognition* **229**, 105233 (2022).

40. Oaksford, M. & Chater, N. A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychol. Rev.* **101**, 608–631 (1994).

41. Markant, D. B. & Gureckis, T. M. Is it better to select or to receive? Learning via active and passive hypothesis testing. *J. Exp. Psychol. Gen.* **143**, 94–122 (2014).

42. Rothe, A., Lake, B. M. & Gureckis, T. M. Do People Ask Good Questions? *Comput. Brain Behav.* **1**, 69–89 (2018).

43. Ruggeri, A., Sim, Z. L. & Xu, F. "Why is Toma late to school again?" Preschoolers identify the most informative questions. *Dev. Psychol.* **53**, 1620 (2017).

44. Petitet, P., Attaallah, B., Manohar, S. G. & Husain, M. The computational cost of active information sampling before decision-making under uncertainty. *Nat. Hum. Behav.* **5**, 935–946 (2021).

45. Yang, S. C. H., Lengyel, M. & Wolpert, D. M. Active sensing in the categorization of visual patterns. *eLife* **5**, 1–22 (2016).

46. Shafir, E. Uncertainty and the difficulty of thinking through disjunctions. *Cognition* **50**, 403–430 (1994).

47. Poldrack, R. A. *et al.* Interactive memory systems in the human brain. *Nature* **414**, 546–550 (2001).

48. Knowlton, B. J., Mangels, J. A. & Squire, L. R. A Neostriatal Habit Learning System in Humans.

Science **273**, 1399–1402 (1996).

49. Collins, A. G. E. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* **35**, 1024–1035 (2012).

50. Collins, A. G. E., Ciullo, B., Frank, M. J. & Badre, D. Working memory load strengthens reward prediction errors. *J. Neurosci.* **37**, 4332–4342 (2017).

51. Plonsky, O., Teodorescu, K. & Erev, I. Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychol. Rev.* **122**, 621 (20150615).

52. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).

53. Nicholas, J., Daw, N. D. & Shohamy, D. Uncertainty alters the balance between incremental learning and episodic memory. *eLife* **11**, e81679 (2022).

54. Duncan, K., Semmler, A. & Shohamy, D. Modulating the Use of Multiple Memory Systems in Value-based Decisions with Contextual Novelty. *J. Cogn. Neurosci.* **31**, 1455–1467 (2019).

55. Bavard, S., Rustichini, A. & Palminteri, S. Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. *Sci. Adv.* **7**, eabe0340 (2021).

56. Zylberberg, A. Decision prioritization and causal reasoning in decision hierarchies. *PLOS Comput. Biol.* **17**, e1009688 (2021).

57. Ogasawara, T. *et al.* A primate temporal cortex–zona incerta pathway for novelty seeking. *Nat. Neurosci.* **25**, 50–60 (2022).

58. Ellsberg, D. Risk, ambiguity, and the Savage axioms. *Q. J. Econ.* 643–669 (1961).

59. Fox, C. R. & Tversky, A. Ambiguity aversion and comparative ignorance. *Q. J. Econ.* **110**, 585–603 (1995).

60. Hunt, L. T. *et al.* Formalizing planning and information search in naturalistic decision-making. *Nat. Neurosci.* **24**, 1051–1064 (2021).

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

62. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017).

Acknowledgements

We thank the Shohamy lab, Christopher A. Baldassano, and Gabriel M. Stine for their insightful discussion of the project. We are thankful for the support of the Stan user community, especially Matti Vuorre and Paul Bürkner. We are grateful for funding support from the NSF (award #1822619 to D.S.), NIMH/NIH (#MH121093 to D.S.) and the Templeton Foundation (#60844 to D.S.).

Author Information

Contributions

Y.A., M.N.S, and D.S. designed research; Y.A. collected and analyzed data; M.N.S and D.S supervised.