nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0526-x

Corrected: Publisher Correction

Sources of suboptimality in a minimalistic explore–exploit task

Mingyu Song[1,2,3,4], Zahy Bnaya[2,3,4] and Wei Ji Ma[2,3*]

People often choose between sticking with an available good option (exploitation) and trying out a new option that is uncertain but potentially more rewarding (exploration)[1,2]. Laboratory studies on explore–exploit decisions often contain real-world complexities such as non-stationary environments, stochasticity under exploitation and unknown reward distributions[3-7]. However, such factors might limit the researcher's ability to understand the essence of people's explore–exploit decisions. For this reason, we introduce a minimalistic task in which the optimal policy is to start off exploring and to switch to exploitation at most once in each sequence of decisions. The behaviour of 49 laboratory and 143 online participants deviated both qualitatively and quantitatively from the optimal policy, even when allowing for bias and decision noise. Instead, people seem to follow a suboptimal rule in which they switch from exploration to exploitation when the highest reward so far exceeds a certain threshold. Moreover, we show that this threshold decreases approximately linearly with the proportion of the sequence that remains, suggesting a temporal ratio law. Finally, we find evidence for 'sequence-level' variability that is shared across all decisions in the same sequence. Our results emphasize the importance of examining sequence-level strategies and their variability when studying sequential decision-making.

Many daily-life decisions involve a tradeoff between exploration and exploitation[1,2]. Should you stick with the same brand of breakfast cereal or try a new one? Should you stay in your current job or explore new opportunities? Exploitation generally means choosing the action believed to have the maximum expected reward, while exploration is choosing any other action, which may be beneficial in the long run[8]. It is not well understood how, and by how much, people deviate from optimality in explore–exploit decisions. Studying such optimality is difficult because the pure explore–exploit problem is often intertwined with multiple cognitive processes: the reward distributions may be unknown to participants and require learning[3,4]; to evaluate options, participants need to remember and aggregate a series of past actions and rewards[5]; sometimes, the outcome of an exploitation action is also uncertain[3,5,7]; the environment might be non-stationary, either because it changes over time[5,6] or because the agent's decisions affect the environment[4]; and the number of consecutive decisions might be stochastic (indefinite decision horizon), which requires the estimation of remaining opportunities to exploit any new information[3,9].

To investigate optimality in the absence of these complexities, we introduce a paradigm in which the participant has full information about the task structure and the only form of stochasticity is the one intrinsically associated with the outcome of an explorative action. The participant was told that they were a tourist in a foreign city. Restaurant ratings in this city ranged from 1.0 to 5.0 and followed a truncated Gaussian distribution, which was visualized on the screen throughout the experiment (Fig. 1a). Each trial consisted of 5 to 10 virtual 'days'. On each of the 'days', the participant decided to go either to a random new restaurant (exploration) or to the best restaurant so far (exploitation), and received a reward equal to the rating of the restaurant. Their goal was to maximize the total reward of the entire trial.

We denote the trial length by T, the reward by r, the action by a (0 for exploitation and 1 for exploration), the highest reward received so far in a trial by r^*, and the number of days left by t_{left}. The optimal policy in this task depends only on r^* and t_{left}. For a given state (r^*, t_{left}), the optimal agent compares the expected future reward $Q(r^*, t_{\mathrm{left}}; a)$ between $a=0$ and $a=1$. This Q function is specified by the Bellman equation, which we calculate using the value iteration algorithm[10] (see Supplementary Methods 1 for the derivation of the optimal policy). The optimal agent explores more when r^* is lower and when there are more days left (Fig. 1b). Since, over the course of a trial, r^* increases and t_{left} decreases, the optimal policy switches from exploration to exploitation at most once in a trial (see Supplementary Methods 2 for formal proof of this property for any deterministic reward distribution, and Supplementary Methods 3 for the calculation of the expected number of switches under the current reward distribution). A similar single-switch policy is optimal in a multi-armed bandit task[11] where each decision is inferred to be under latent 'exploration' or 'exploitation' state.

We tested 49 laboratory participants and 143 online participants using Amazon Mechanical Turk (see Methods for details). All participants passed a task comprehension test. Each laboratory participant completed 180 trials (1,170 choices), and each online participant completed 60 trials (390 choices). At the end, laboratory participants were asked to write about their strategy. In the main text, we report the results from laboratory participants; the results and conclusions are consistent for online participants (see Supplementary Figs. 10 and 11). We characterize the choice data using a set of summary statistics (Fig. 1c–g). Participants' choices (Fig. 1c) resembled a stochastic version of the optimal policy (Fig. 1b). Participants explored more in the beginning of a trial than towards the end, and when the highest reward received so far was lower (Fig. 1c,d and Supplementary Fig. 1a). Logistic regression on the choice against r^* and t_{left} returned coefficients of -5.64 ± 0.61 (mean ± s.e.m.) for r^* (two-sided t-test with zero: $t_{48} = -9.20$; $P < 0.001$; effect size $d = -1.31$; 95% CI on d: -1.69 to -0.93; here and elsewhere, t-tests are all two-tailed) and 0.545 ± 0.038 for t_{left} ($t_{48} = 14.4$; $P < 0.001$; $d = 2.06$; 95% CI: 1.56 to 2.56).

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [2]Center for Neural Science, New York University, New York, NY, USA. [3]Department of Psychology, New York University, New York, NY, USA. [4]These authors contributed equally: Mingyu Song, Zahy Bnaya. *e-mail: weijima@nyu.edu

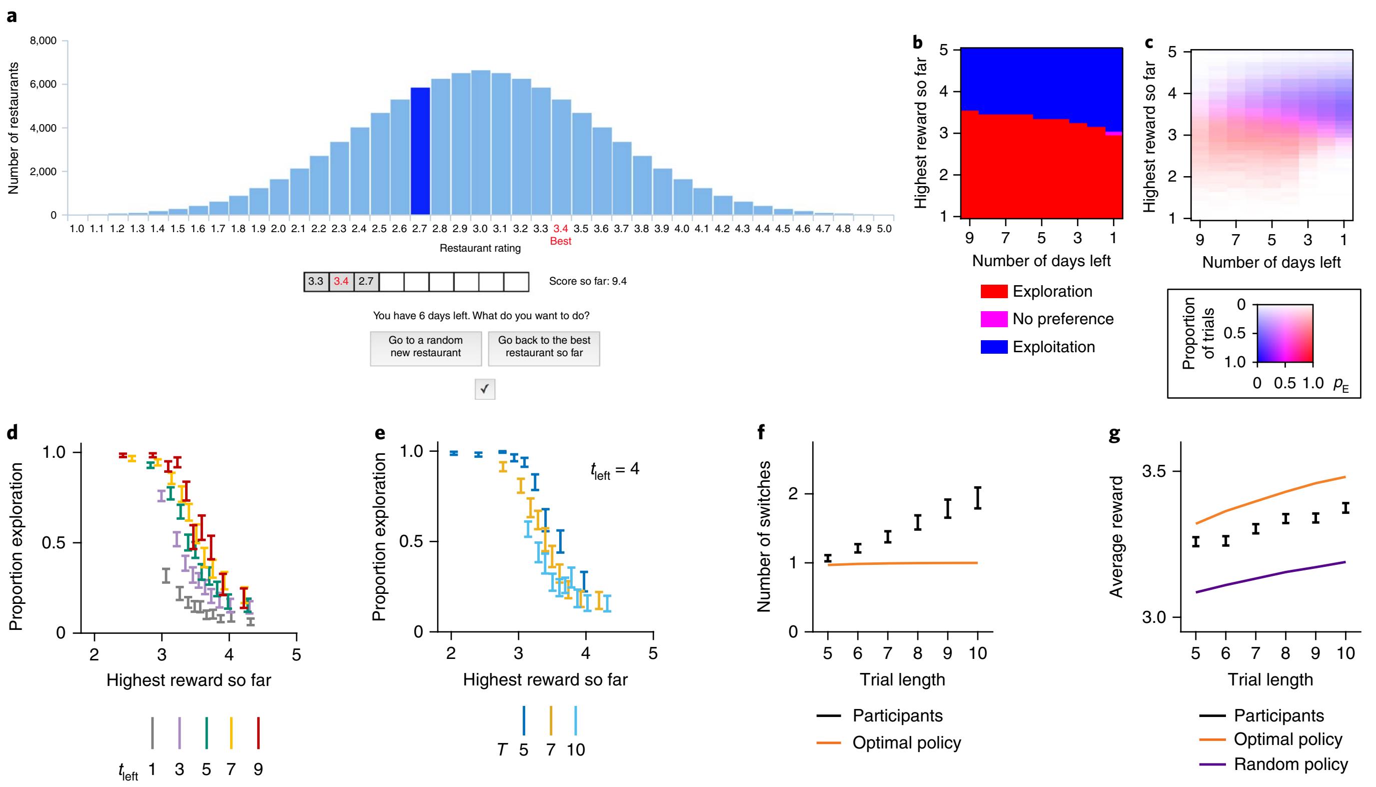

Fig. 1 | Experimental design, optimal policy and summary statistics. a, Example of a response screen (on day 4 of a 9-day trial). The histogram shows the distribution of restaurant ratings. The dark blue bar shows the most recent reward (2.7). The red text indicates the highest reward so far (3.4). The history of rewards and 'score so far' (that is, the accumulated reward: 9.4) are shown below the histogram. **b**, Under the optimal policy, the decision on whether to explore only depends on the highest reward so far (r^*) and the number of days left (t_{left}). The optimal agent explores when r^* is low and t_{left} is high. **c–g**, Summary statistics of data from laboratory participants. Error bars indicate ±1 s.e.m. across participants. (**c**) Proportion of decisions for which participants explored, as a function of the highest reward so far and the number of days left, averaged across participants. The colour code is two-dimensional: the hue represents the proportion of exploration (p_E) and saturation = log(1 + proportion of trials)/log(2). (**d**) Slices from the plot in **c**. For each participant and each t_{left}, we divided the r^* values from all decisions into ten quantiles; within each quantile, we calculated the proportion of decisions for which the participant explored. We plotted the mean and s.e.m. of that proportion against the mean across participants of the median r^* in that quantile. (**e**) Proportion of exploration as a function of the highest reward so far for $t_{\text{left}} = 4$, broken down by trial length (that is, the total number of days (T)). In the optimal policy, T would be irrelevant. Similar effects for other t_{left} values (Supplementary Fig. 2). (**f**) Number of switches between exploration and exploitation, averaged across trials, as a function of the trial length, for participants (black) and the optimal policy (orange) (see Supplementary Method 3). (**g**) Average reward as a function of trial length for participants (black), the optimal policy (orange) and the random agent (purple). In **d** and **e**, we only show part of the data (see Supplementary Figs. 1 and 2 for the full data). Panels **c–g** show qualitative and quantitative deviations of human behaviour from the optimal policy.

Participants' behaviour, however, also showed clear deviations from the optimal policy. Their choices were influenced not only by r^* and t_{left}, but also by the trial length T. For a given t_{left}, participants explored more in shorter than in longer trials (Fig. 1e and Supplementary Figs. 1b and 2). Adding T as a regressor in the logistic regression returned a coefficient of −0.276 ± 0.036 for T ($t_{48} = -7.73$; $P < 0.001$; $d = -1.10$; 95% CI: −1.46 to −0.74); the coefficients for r^* and t_{left} remained significantly different from zero (both $P < 0.001$). Furthermore, we noted before that the optimal agent switches at most once per trial. Participants switched more frequently than was optimal for any given trial length (Fig. 1f; $P = 0.0058$ for $T = 5$; $P < 0.001$ for $T = 6$–10; all after Bonferroni–Holm correction). The number of switches also increased with trial length (linear regression slope: 0.180 ± 0.024; $t_{48} = 7.44$; $P < 0.001$; $d = 1.06$; 95% CI: 0.71 to 1.41). Participants' average reward per 'day' was significantly higher than the average reward of an agent who randomly explores or exploits, but also significantly lower than an optimal agent (Fig. 1g; in both comparisons, $P < 0.001$ for every trial length, after Bonferroni–Holm correction). Participants earned a higher average reward per day on longer trials (linear regression slope: 0.0242 ± 0.0029; $t_{48} = 8.36$; $P < 0.001$; $d = 1.19$; 95% CI: 0.82 to 1.56; Fig. 1g). Taken together, these results show that participants adopted a reasonable but suboptimal strategy. Similar to the optimal agent, participants tended to switch to exploitation towards the end of a trial, and when the highest reward received so far was high. However, their choices were more stochastic and were influenced by trial length; on average, they switched more often between exploration and exploitation and earned less reward than the optimal agent.

Some of the observed deviations from optimality could simply be due to decision noise. Therefore, we first considered a model obtained by adding softmax decision noise[12] to the optimal policy. The probability of exploring is then:

$$P(a=1) = f(\beta_0 + \beta \Delta Q(r^*, t_{\text{left}})) \tag{1}$$

where $\Delta Q(r^*, t_{\text{left}}) \equiv Q(r^*, t_{\text{left}}; 1) - Q(r^*, t_{\text{left}}; 0)$ and $f(x) = \frac{1}{1+e^{-x}}$ is the logistic function. When the inverse temperature β is higher, behaviour is closer to deterministic. The parameter β_0 captures a bias towards exploration or exploitation. We call this model the Opt model. We fitted the Opt model to individual choices for each individual participant using maximum-likelihood estimation. The fits to the summary statistics are qualitatively similar to the data

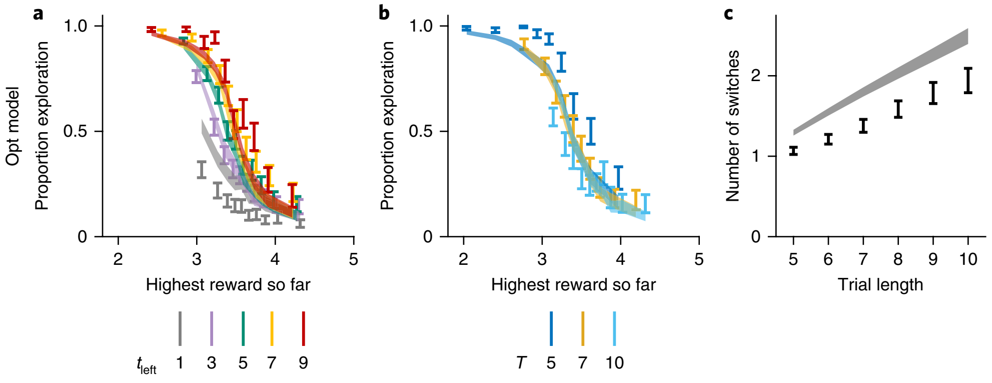

Fig. 2 | Fits of the Opt model to selected summary statistics. a–c, The Opt model fits poorly despite allowing for bias and decision noise. Graphs in **a–c** correspond to the summary statistics in Fig. 1d–f, respectively. Error bars represent data; shaded areas represent model fits (both are ±1 s.e.m).

(the three most diagnostic summary statistics are shown in Fig. 2; see Supplementary Fig. 3 for model fits to the other summary statistics), but quantitatively, the Opt model does not fully capture the influence of the number of days left on the proportion of exploration (Fig. 2a), nor could it predict the effect of trial length on choices (Fig. 2b). The Opt model also overestimates the number of switches (Fig. 2c). These results indicate that the Opt model is not a good description of participants' choices, and that the deviation of human behaviour from the optimal policy is not merely due to bias and softmax decision noise.

Next, we considered the possibility that deviations from optimality are not only due to bias and noise, but also due to systematic suboptimalities in the policy. Since calculating optimal future values could be computationally demanding, people might instead use simple heuristics[13,14]. Indeed, participants reported using heuristic strategies (see Supplementary Results 1): most participants (30 out of 43 valid responses) reported to have explored in the beginning of a trial and switched to exploitation when the best restaurant rating reached a threshold. Some reported that the threshold was related to trial length and where they were in a trial. Therefore, we considered threshold rules, in which the agent starts out exploring and once r^* exceeds a time-dependent threshold, switches to exploitation and keeps exploiting until the end of the trial. We still allow for bias (inherent in the threshold function) and softmax decision noise. Denoting the threshold by θ, the probability of exploring is then:

$$P(a=1)=f(\beta(\theta-r^{\star})) \tag{2}$$

Inspired by the optimal policy, we first considered a model whose threshold function also depends on the number of days left (t_{left}), but in a linear way: $\theta=kt_{\text{left}}+b$ (k and b are model parameters). We call this the Num model.

We compared model fits using the Akaike information criterion[15] corrected for small sample sizes (AICc)[16], where a lower AICc value means a better fit. AICc is lower for the Num model by 70 (95% bootstrapped confidence interval (BCI): 50 to 95) compared with the Opt model. For all model comparisons reported in this paper, using the Bayesian information criterion[17] gives consistent results (see Supplementary Table 1 and Supplementary Figs. 7b, 9d, 12d, 14d and 16d).

The Num model fits better to the proportion of exploration as a function of t_{left} (Fig. 3a). However, the dependence of participants' choice on trial length (T) is not well accounted for (Fig. 3b); this is expected because the Num agent does not use information about T.

To explore the specific threshold strategy people used, we fitted a flexible-threshold model where $\theta(t_{\text{left}}, T)$ is a discrete function of t_{left} and T (Fig. 3g; 40 free parameters, including 39 thresholds and a softmax noise). The fitted threshold depends both on t_{left} and T: it decreases over the course of a trial, and is lower on longer trials. When we change the independent variable from the number of days left (t_{left}) to the proportion of days left (t_{left}/T), the thresholds at different trial lengths collapse onto each other, suggesting that participants used a threshold rule that almost linearly depended on the proportion of days left (Fig. 3h). Thus, we consider a model where $\theta=kt_{\text{left}}/T+b$, which we will call the Prop model. The Prop model provides a better fit to the data than the Num model (AICc difference: 19.7; 95% BCI: 7.2 to 32.2) and than the Opt model (AICc difference: 90; 95% BCI: 67 to 120). Specifically, it fits well to the effect of T on the proportion of exploration (Fig. 3e). It also preserves the good fit on the influence of the number of days left (Fig. 3d). These findings suggest that the relevant quantity in deciding when to switch to exploitation is the relative rather than absolute time in the trial. This perhaps reflects an organism's need to plan across multiple different time scales[18].

All models that we have considered so far systematically overestimate the number of switches (Figs. 2c and 3c,f), which could result from the overestimation of softmax decision noise. This motivates us to consider another type of variability: a common random shift of the thresholds across an entire trial (a sequence of 5 to 10 choices) (Fig. 4a). Sequence-level variability could be interpreted as a form of planning: before any choices are made in a trial, the agent might set the decision threshold for all choices within that trial based on the trial length, but with a random shift. Different from choice-level softmax noise, which changes the slope of the psychometric curve, sequence-level variability randomly shifts the psychometric curve horizontally on each trial (Fig. 4b). If an agent's thresholds are variable at the sequence level, but we fit a fixed-threshold model (such as Num or Prop) to their data, the choice-level softmax noise would be overestimated to account for the sequence-level variability, and as a result the number of switches would be overestimated.

We implemented sequence-level variability as a Gaussian random variable η with a mean of 0 and a variance of σ^2. Thus, the probability of exploration for the variable-threshold models is:

$$P(a=1)=f(\beta(\theta+\eta-r^{\star})) \tag{3}$$

where $\eta\sim\mathcal{N}(0,\sigma^2)$. We added this sequence-level variability to the Prop and Num models to obtain the Prop-V and Num-V models, respectively. When fitting these models, we had to take into account that sequence-level variability introduces dependencies between choices within a trial (see Methods). Adding sequence-level variability greatly improved the fits of both the Prop and Num models (Fig. 4c): AICc decreased by 61 (95% BCI: 43 to 86) from Prop to Prop-V, and by 67 (95% BCI: 48 to 93) from Num to Num-V, and both models fitted well to the number of switches (Fig. 4f and Supplementary Fig. 4c). The good fits to other summary statistics were preserved

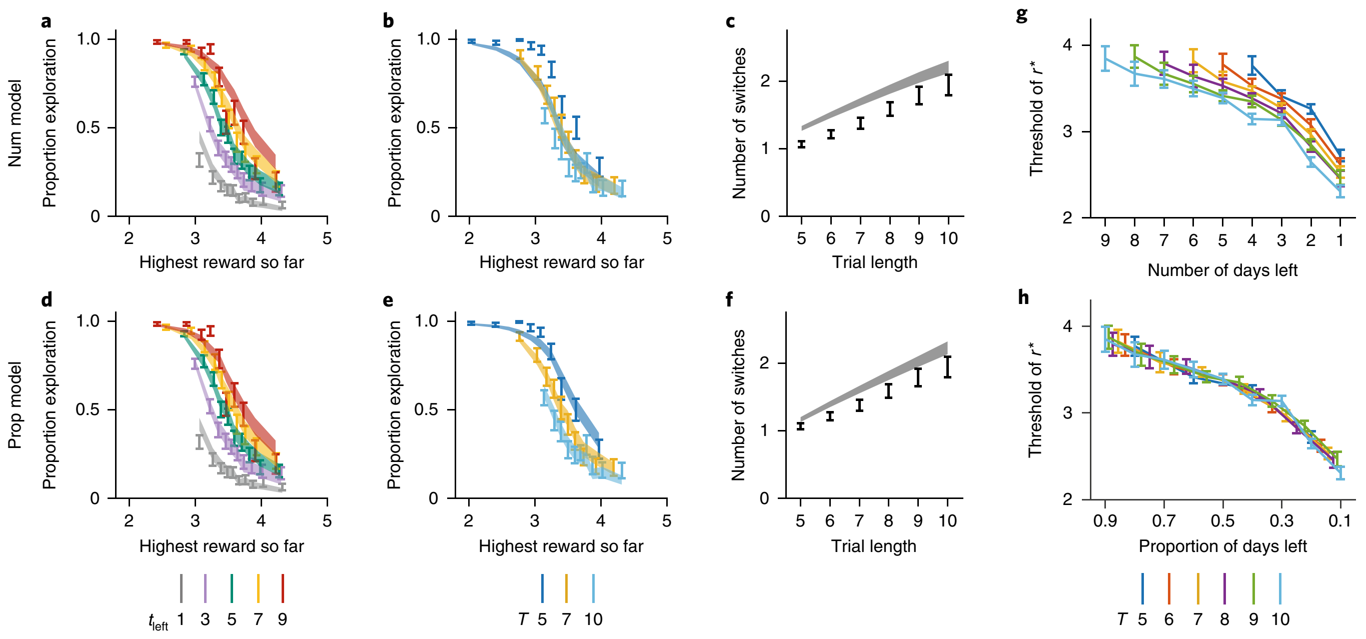

Fig. 3 | Evidence of a threshold rule depending on the proportion of days left. Participants' behaviour is better characterized by a threshold rule that depends on the proportion of days left than by one that depends on the number of days left. This suggests that relative time is more important than absolute time. **a–f**, Model fits of the Num (**a**–**c**) and Prop models (**d**–**f**) to the summary statistics in Fig. 1d–f, respectively. **g**, Fitted threshold of r^* as a discrete function of t_{left} and T. **h**, Same curves as in **g**, but with the independent variable changed to the proportion of days left (each curve is stretched along its x axis). Error bars (data) and shaded areas (model fits) indicate ±1 s.e.m.

for the Prop-V model (Fig. 4d,e). The Num-V model still could not fit to the effect of trial length on the proportion of exploration (Supplementary Fig. 4b). Thus, the Prop-V model fitted best to data among all the models we tested: the AICc difference between Prop-V and the second-best model Num-V is 13.8 (95% BCI: 2.5 to 23.0). Bayesian model selection for group studies[19,20] shows there is some evidence for heterogeneity between participants: the estimated model frequency is 0.716 for the Prop-V model and 0.255 for the Num-V model; the protected exceedance probability (the probability of one model being more frequent in the group of participants than the other models) is 0.9996 for the Prop-V model. Adding a risk attitude parameter to the Prop-V model (see Methods) does not improve the fit much: the value of AICc decreases by 4.1; 95% BCI: −8.6 to −1.9 (Supplementary Fig. 7).

Next, we examined the possibility that the Prop-V model fitted better than the Prop model, not because of sequence-level variability but because the model tries to compensate for an incorrect linear assumption for the threshold. We compared the Prop-V model with the flexible-threshold model without sequence-level variability (Fig. 3g), which can account for threshold functions of arbitrary shapes. The Prop-V model fitted much better (AICc difference: 155; 95% BCI: 103 to 246), indicating that the good fit of the Prop-V model cannot be attributed solely to threshold mismatch.

Given the evidence for sequence-level threshold variability, what is the origin of this variability? One possibility is that the threshold changes systematically over the course of the experiment due to learning. To examine this, we compared the first and second halves of the task. The only difference we found was fewer switches between exploration and exploitation in the second half of the task than in the first half ($t_{48} = 2.61$; $P = 0.012$). However, this effect was weak (effect size $d = 0.17$; 95% CI: 0.038 to 0.31). We did not find evidence for a change in the average reward per trial ($t_{48} = -0.31$; $P = 0.76$; $d = -0.025$; 95% CI: −0.19 to 0.14). We also fitted the Prop-V model to the first and second halves of the task separately and found no difference in the parameter estimates (paired t-test: $P > 0.27$ for each of the four parameters). None of these results changed qualitatively when we split the data into thirds and performed analyses of variance (Supplementary Tables 2 and 3). Other possibilities are that sequence-level variability is due to inter-trial dependencies or a wrong belief that the reward distribution changes over trials. Thoroughly testing for any of these possibilities would require targeted follow-up experiments.

Now, we consider a more remote alternative class of models. As we saw, the optimal policy only depends on r^* and t_{left}. However, people might consider information that is irrelevant for maximizing future reward. Specifically, they might use the history of choices and rewards within the same trial in their current choice[21]. Therefore, an alternative approach to characterizing deviations from optimality would be to look for dependencies of choices on intra-trial history-dependent factors. We tested for eight history-dependent factors: preceding action, number of days past, proportion of days past, trial length, preceding reward, average reward, minimum reward, and regret (the difference between the preceding reward and the guaranteed reward of exploitation). We split the data based on each of these factors separately, and tested whether the proportion of exploration choice differed between the splits (Supplementary Fig. 5a,b); we controlled for the values of r^* and t_{left}. We found significant effects of the preceding action, number of days past, proportion of days past, and trial length ($P < 0.001$ after Bonferroni–Holm correction; $P > 0.05$ for the other four factors). Next, we built a logistic regression model containing the eight history factors as well as r^* and t_{left}; we call this model, which has 11 free parameters, the History model. This model fitted well to all summary statistics in Fig. 1 and to the history effects in Supplementary Fig. 5b. To examine the importance of the individual history factors, we compared model fits by adding one factor at a time to the null model (with only r^* and t_{left} as regressors; Supplementary Fig. 5c), and by leaving one factor out at a time from the History model (Supplementary Fig. 5d), as well as computing the posterior for each factor using Bayesian model selection[19,20] (Supplementary Fig. 5e). According to all three metrics, there was

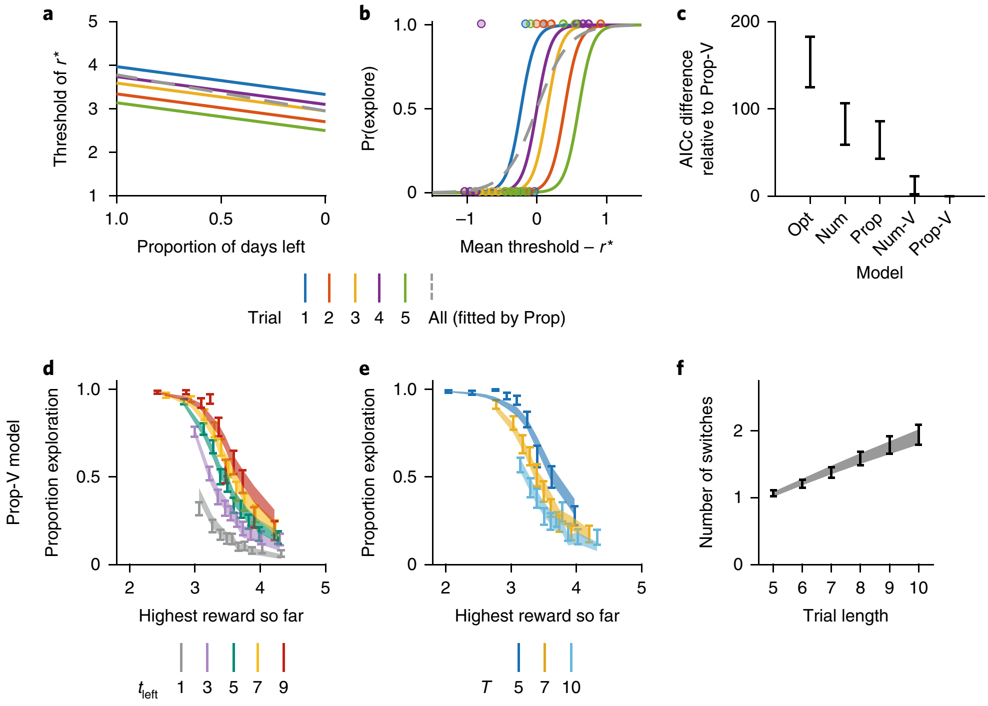

Fig. 4 | Sequence-level variability as implemented through variable-threshold models (Num-V and Prop-V). Models with sequence-level variability account better for participants' choices. **a**, We postulate that choices are subject not only to softmax noise, but also to variability at the level of an entire sequence of choices (a trial), as implemented through random trial-to-trial shifts in the threshold function (five example trials shown). Here we use the Prop-V model as an example, but the same idea applies to the Num-V model. **b**, Corresponding random shifts in the psychometric curves of the probability of exploration as a function of the difference between the mean threshold and r^*. Dots are data simulated using the Prop-V model for the five example trials; the simulated data are then fitted using the Prop model and the resulting psychometric curve is shown in dashed grey. If Prop-V (or Num-V) is the true model, fitting the Prop (or Num) model will overestimate the choice-level softmax noise (the dashed grey line is shallower than the coloured lines), and therefore the number of switches per trial. **c**, Comparison of all five models. The AICc values of the Prop-V model are used as a baseline. Positive values mean worse fits compared with Prop-V. Error bars represent 95% BCIs. Both variable-threshold models fit better than their fixed-threshold counterparts. **d**–**f**, Model fits of the Prop-V model to the summary statistics in Fig. 1d–f, respectively. Error bars (data) and shaded areas (model fits) indicate ±1 s.e.m.

the strongest evidence for 'preceding action' (AICc decreases by 82 (95% BCI: 75 to 92) if adding 'preceding action' into the null model and increases by 33 (95% BCI: 30 to 37) if dropping it out of the history model; factor posterior: 0.947 ± 0.019), suggesting a tendency to repeat the preceding action—a phenomenon commonly found in valued-based decision-making tasks[21,22].

However, the History model has the following problems: (1) there are no fewer than ten regressors in total, each with a corresponding free parameter; and (2) the choice of these history factors and how they appeared in the model is ad hoc. Perhaps surprisingly, the parsimonious Prop-V model, with only one explicit history-dependent factor (the proportion of days left) and many fewer free parameters (four in total), also accounts for all eight intra-trial history-dependent effects (Supplementary Fig. 6). This suggests that the evidence for many of the history factors was due to them being correlated with 'the proportion of days left'. Specifically, the Prop-V model can account for the apparent effect of preceding action. This is because the sequence-level variability causes choices within a trial to be correlated, which the History model fits as the agent's tendency to repeat the preceding action. This result might provide a different perspective on the commonly found choice inertia effects in value-based decision-making: they might be due to sequence-level mechanisms. Altogether, these results illustrate the value of looking for principled and parsimonious models[23,24].

We designed a simplified explore–exploit task that allowed us to quantitatively characterize the deviations of human behaviour from optimality. We found that human participants behaved qualitatively similarly to the optimal policy, but that they adopted a heuristic decision rule instead of calculating optimal values. In particular, they scaled their decision threshold to the trial length. This suggests a temporal ratio law for planning that might be easier and more intuitive to implement than a rule that uses the absolute amount of time left. It is broadly consistent with the notion that ratios of temporal durations, not absolute values, are relevant in interval timing[25,26] and long-term memory[27]; however, it is not clear whether the similarity is more than superficial.

In addition to choice-level softmax noise, which is commonly used to model value-based decisions, participants also seemed to exhibit sequence-level variability. Even though we only considered a very simple form of such variability, the finding suggests that when studying other sequential decisions (such as multi-armed bandit problems[28], foraging tasks[29], the secretary problem[30], the four-in-a-row game[31], and the travelling salesman problem[32]), one has to take into account variability at the level of the entire sequence rather than only at the level of individual decisions.

A similar simplified task was used by Sang and colleagues[33–35]. Participants were asked to maximize the total score across a sequence of 20 decisions, each of which consisted of either flipping a card from a deck of 100 cards numbered from 1 to 100, or picking one of the cards they had already flipped. Like in the present study, participants explored more when the highest number so far was lower and when more decisions were left in the trial. Participants also switched more between exploration and exploitation than the optimal policy. One of these studies[33] also manipulated the trial length (uniformly from 5 to 35 decisions), but this did not affect performance. This was potentially because the participant did not know the trial length ahead of time, so they would not be able to plan accordingly. The authors tested several heuristic models, including a random-policy model, an ε-greedy model, threshold models and sampling models inspired by the secretary problem literature that switch from exploration to exploitation based on an initial sample of options. Consistent with our study, they found evidence for threshold models; specifically, a model with a linearly decreasing threshold (similar to our Num model) fitted best. However, because the participant did not know the trial length ahead of time (in the study in which trial length varied[33]), it was not possible to investigate whether the threshold depended on the number or proportion of decisions left. Additionally, it would be interesting to see whether, in their data, the threshold exhibits variability at the sequence level. Compared with the modest learning effect on the number of switches in the current study, Sang and colleagues found strong learning effects both on performance (for example, the average reward and the number of switches) and on the estimated decision thresholds and decision noise. This was potentially because Sang et al. informed the participant at the end of each trial of the points that the optimal strategy would have earned; this could have served as an explicit error signal. Trials were also longer in Sang's task, so if participants tried out different strategies, they could see a greater change in performance compared with our task, where trials were shorter. Both factors could have motivated the participants to improve their strategies.

In this paper, we have adopted the definitions of 'exploitation' (choosing the option with the maximum expected outcome) and 'exploration' (choosing any other option at random) that are common in the reinforcement learning literature[8]. However, other definitions have also been widely used[2], which are based on (1) behavioural patterns[36] (exploration is alternating between options while exploitation is focusing on one option); (2) uncertainty of the options[11] (exploration is choosing an option with higher uncertainty); and (3) outcomes obtained from a choice[1,37] (exploration is obtaining information while exploitation is obtaining reward). These definitions are not mutually exclusive; in our task, definitions (1) and (2) also apply, but not definition (3), as we provided participants with full information and do not consider 'information' as an outcome.

Methodologically, the sources of suboptimality that we found would have been difficult to identify with a more traditional design, as they might have been confounded with failures of learning or memorization. The minimalistic features of our task—in particular the ability to perform an exploitation action without much stochasticity, and a fixed, known horizon—not only are a modelling convenience, but potentially also approximate some real-world decisions: the problem was originally motivated by author W.J.M. having to decide whether to go to the same breakfast place in Beijing during his few days of visit there. Nevertheless, many real-life scenarios involve factors that we did not consider here, such as unknown reward distributions[38], non-stationary environments[39], or a fixed goal for cumulative reward[40]. These complexities could affect behaviour. For example, a non-stationary environment could provoke information seeking[41]; a changing discrepancy between the accumulated reward and the fixed goal could drive risk-taking or risk-averse behaviour that may appear as exploration or exploitation, respectively. Future studies could incorporate these factors into the current task; the current approach might still be useful as a starting point for process models that explicitly dissociate various sources of suboptimality.

Our work might help to guide the search for the neural basis of explore–exploit behaviour. Previous work has found involvement of the rostrolateral prefrontal cortex in exploration actions[5,42,43]. Our paradigm might be useful to break down this activation into 'pure' exploration (as studied here) and learning about the reward distributions. Moreover, the dorsolateral prefrontal cortex has been found to encode task variables in value-based decision-making[44–47]. Thus, we would expect that this area contains signals correlated with the trial-to-trial decision threshold, as well as with the decision variable (the difference between the threshold and the highest reward so far).

Methods

Experiment 1. *Task.* Participants were placed in a computerized task environment with a backstory in which they were a tourist in a foreign city. A trial consisted of T 'days' (T varied pseudo-randomly between 5 and 10, with an equal number of each). On each of the T days, they had to decide where to go for dinner. The rating of the restaurants in the city varied between 1.0 (worst) to 5.0 (best); the distribution was a renormalized truncated Gaussian with a mean of 3.0 and standard deviation of 0.6. The number of restaurants was assumed to be much greater than ten. The histogram of restaurant's rating was shown on the screen. On each 'day' except for the first one, the participant would choose to click one of two buttons: 'Go to a random new restaurant' or 'Go to the best restaurant so far'. On the first 'day', only the first button could be clicked. After clicking on either button, a check mark button would appear in the bottom centre of the screen, and the button that did not get chosen would be disabled (not shown in Fig. 1a). Participants needed to click on the check mark button to confirm their choice. After a choice was confirmed, the resulting rating would be shown (for example, '2.7') and the corresponding bar would be highlighted in the distribution. If the exploration button was clicked, the rating would be drawn randomly from the Gaussian distribution. If the exploitation button was clicked, the rating would be the highest rating obtained so far in the trial. The screen showed the number of 'days' left, highest rating so far (marked in red), numerical history of rewards (with the highest in the series highlighted) and score (sum of ratings) so far in the trial. At the end of the trial, the participant was told their trial score and the next trial commenced. The task was self-paced with no time limit.

Procedure. Participants were explained the task and the distribution of possible rewards using a series of self-paced instruction screens. To test their understanding of the distribution, they were shown the histogram with a simple two-alternative question of the type 'If you randomly go to a restaurant in this city, which rating is more probable? A. 2.5 B. 4.5' (the correct answer would be A). Participants had to correctly answer at least two out of three such questions to be allowed to continue; participants who did not meet this criterion would be sent home. The remaining participants were shown the example screen in Fig. 1a with the different elements annotated to familiarize them with it. After the instruction phase, participants completed 180 trials in a single session, each consisting of 5 to 10 decisions. After completing all 180 trials, we asked participants, 'what kind of strategy did you use in the task?', to which they typed an answer (no length limits). See Supplementary Results 1 for full answers to the question. At the end of the session, we randomly chose one trial and determined a bonus payout based on the score. Participants were told this procedure, and we emphasized to them that each trial was equally important in determining their payout.

Participants. In total, 49 participants (10 male, 24 female and 15 unknown; aged 18–57 years) participated in the experiment. All of them passed the qualifying test as described in the section 'Procedure'. Participants received US$10 for completing the experiment (about 45 min), plus a performance bonus up to US$5. The experiments were approved by the University Committee on Activities Involving Human Participants of New York University. Each participant gave informed consent before the experiment. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to or larger than those reported in previous publications with laboratory participants[4,11,38] and online participants[48]. The trial sequence was randomized for each participant.

Experiment 2: online experiment. Experiment 2 was conducted on Amazon Mechanical Turk, programmed using the psiTurk interface[49]. We obtained separate Institutional Review Board approval for this project from the University Committee on Activities Involving Human Participants of New York University. Each participant gave informed consent before accepting the task. The procedure was the same as in experiment 1, except for the following differences: (1) participants completed only 60 trials; (2) participants were paid nothing if they did not pass the task-comprehension test, US$1.50 if they passed the test and completed the experiment, and a performance bonus of up to US$5; (3) participants were recruited through the general Amazon Mechanical Turk task listing; and (4) due to technical problems, participants' reported strategies were not recorded. A total of 143 participants participated (no demographic information provided).

Experiments 3–5. Experiment 3 was the same as experiment 2, except that most history information (trial length, previous rewards and the accumulated reward so far) was hidden from participants (Supplementary Fig. 10). In total, 131 Amazon Mechanical Turk participants participated (no demographic information provided).

Experiments 4 and 5 were conducted before the other three experiments. They were the same as experiments 1 and 2 except that we did not require participants to click on a 'confirm' button after each choice. The old design was thought to bias participants towards continuing the same choice. However, the results turned out to be consistent with or without the confirm button. A total of 16 laboratory participants (7 male and 9 female; aged 20–43 years) and 108 Amazon Mechanical Turk participants (no demographic information provided) participated in experiments 4 and 5, respectively.

The results from experiments 3–5 (Supplementary Figs. 11–16) were consistent with the main experiments (1 and 2), so we take them as replicates of the main experiments.

Models and model fitting. *Likelihood function.* The Opt model and fixed-threshold models (the Num and Prop models) assume that all choices within a trial are independent, so the likelihood function is simply the multiplication of the likelihood of individual choices. However, in variable-threshold models (the Num-V and Prop-V models), choices within a trial are not independent; instead, they are conditionally independent given the random shift of thresholds on that particular trial. Denote the total number of trials by N, the action, reward and best reward so far of the tth action in the ith trial by a_{it}, r_{it} and r^*_{it}, respectively, and the random shift of threshold on trial i by η_i. Denote by $\mathbf{a}_i$ the vector of actions a_{ij}, where $j=1,...,T_i$. Then, the likelihood function of the parameters k, b, σ and β is:

$$\begin{aligned} p(\mathbf{a}_1,\ldots,\mathbf{a}_N|k,b,\sigma,\beta) &= \prod_{i=1}^{N} p(\mathbf{a}_i|k,b,\sigma,\beta) \\ &= \prod_{i=1}^{N} \int p(\mathbf{a}_i|k,b,\eta_i,\beta)p(\eta_i|\sigma)\mathrm{d}\eta_i \\ &= \prod_{i=1}^{N} \int \left[\prod_{t=1}^{T_i} p(a_{it}|k,b,\eta_i,\beta)\right] p(\eta_i|\sigma)\mathrm{d}\eta_i \end{aligned}$$

where:

$$p\,(a_{it}=1|k,b,\eta_i,\beta) = \frac{1}{1+e^{-\beta(\theta_{it}+\eta_i-r^*_{it})}}$$

$$\theta_{it}=k\,(T_i-t+1)+b\,(\text{Num-V}) \;\text{ or }\; \theta_{it}=k\frac{T_i-t+1}{T_i}+b\,(\text{Prop-V})$$

$$p\,(\eta_i|\sigma) \sim \mathcal{N}(0,\sigma^2)$$

Constraints on thresholds. In the threshold models (including the Num, Prop, Num-V and Prop-V models), we did not constrain the parameters k, b and σ, so the threshold θ was unbounded. For the model thresholds to be cognitively meaningful, we constrained θ to between 1 and 5 by setting values <1 to 1 and values >5 to 5; not doing so did not qualitatively change any of our results.

Risk attitude parameter. We added a risk attitude parameter α to the Prop-V model to get the Prop-V-risk model. We assume that subjective utility is a power-law function of reward r[50], denoted by r^α, and use this instead of r when solving the Bellman equation. A risk-seeking agent will have $\alpha>1$, and a risk-averse agent will have $\alpha<1$.

Algorithms for parameter fitting. To fit the Opt model, we used the mnrfit function in MATLAB. To fit the Num, Prop, Num-V and Prop-V models, we used the fminunc function in MATLAB, in which we ran each fitting 100 times from random starting points to reduce the risk of ending up in a local maximum.

Parameter and model recovery. We conducted parameter recovery and model recovery tests on all models, using synthetic data generated from all models.

Fits to the summary statistics. The fits to summary statistics (Fig. 1c–g) were obtained by simulating data from the model using the fitted parameters and extracting summary statistics in the same way as for the real data.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

All experimental and analysis codes used in this paper are available at https://github.com/mingyus/explore-exploit.

Data availability

All data that support the findings of this paper are available at https://github.com/mingyus/explore-exploit.

Received: 15 June 2018; Accepted: 19 December 2018;
Published online: 11 February 2019

References

1. Cohen, J. D., McClure, S. M. & Angela, J. Yu Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Phil. Trans. R. Soc. Lond. B* **362**, 933–942 (2007).
2. Mehlhorn, K. et al. Unpacking the exploration–exploitation tradeoff: a synthesis of human and animal literatures. *Decision* **2**, 191–215 (2015).
3. Acuna. D. & Schrater. P. Bayesian modeling of human sequential decision-making on the multi-armed bandit problem. In *Proc. 30th Annual Conference of the Cognitive Science Society* 2065–2070 (Cognitive Science Society, 2008).
4. Constantino, S. M. & Daw, N. D. Learning the opportunity cost of time in a patch-foraging task. *Cogn. Affect. Behav. Neurosci.* **15**, 837–853 (2015).
5. 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).
6. Knox, W. B., Otto, A. R., Stone, P. & Love, B. The nature of belief-directed exploratory choice in human decision-making. *Front. Psychol.* **2**, 398 (2012).
7. Steyvers, M., Lee, M. D. & Wagenmakers, E.-J. A Bayesian analysis of human decision-making on bandit problems. *J. Math. Psychol.* **53**, 168–179 (2009).
8. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).
9. Seale, D. A. & Rapoport, A. Optimal stopping behavior with relative ranks: the secretary problem with unknown population size. *J. Behav. Decis. Mak.* **13**, 391–411 (2000).
10. Bellman, R. *Dynamic Programming* 1st edn (Princeton Univ. Press, Princeton, 1957).
11. Lee, M. D., Zhang, S., Munro, M. & Steyvers, M. Psychological models of human and optimal performance in bandit problems. *Cogn. Syst. Res.* **12**, 164–174 (2011).
12. McFadden, D. et al. in *Frontiers in Econometrics* (ed. Zarembka, P.) 105–142 (Academic Press, New York, 1973).
13. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).
14. Simon, H. A. Rational choice and the structure of the environment. *Psychol. Rev.* **63**, 129–138 (1956).
15. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
16. Cavanaugh, J. E. et al. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Stat. Probabil. Lett.* **33**, 201–208 (1997).
17. Schwarz, G. et al. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
18. Kello, C. T. et al. Scaling laws in cognitive sciences. *Trends Cogn. Sci.* **14**, 223–232 (2010).
19. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—revisited. *NeuroImage* **84**, 971–985 (2014).
20. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *NeuroImage* **46**, 1004–1017 (2009).
21. Lau, B. & Glimcher, P. W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* **84**, 555–579 (2005).
22. Ito, M. & Doya, K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J. Neurosci.* **29**, 9861–9874 (2009).
23. Boehner, P. *Ockham: Philosophical Writings* (Nelson, Canada, 1957).
24. Chater, N. & Vitányi, P. Simplicity: a unifying principle in cognitive science? *Trends Cogn. Sci.* **7**, 19–22 (2003).
25. Buhusi, C. V. & Meck, W. H. What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.* **6**, 755–765 (2005).
26. Gibbon, J. Scalar expectancy theory and Weber's law in animal timing. *Psychol. Rev.* **84**, 279–325 (1977).
27. Brown, G. D. A., Neath, I. & Chater, N. A temporal ratio model of memory. *Psychol. Rev.* **114**, 539–576 (2007).
28. Robbins, H. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **58**, 527–535 (1952).
29. Charnov, E. Optimal foraging: the marginal value theorem. *Theor. Popul. Biol.* **9**, 129–136 (1976).
30. Seale, D. A. & Rapoport, A. Sequential decision making with relative ranks: an experimental investigation of the "secretary problem". *Organ. Behav. Hum. Decis. Process.* **69**, 221–236 (1997).
31. Van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y. & Ma, W. J. A computational model for decision tree search. (2017). In *Proc. 39th Annual Conference of the Cognitive Science Society* 1254–1259 (Cognitive Science Society, 2017).
32. MacGregor, J. N. & Ormerod, T. Human performance on the traveling salesman problem. *Percept. Psychophys.* **58**, 527–539 (1996).
33. Sang, K. *Modeling Exploration/Exploitation Behavior and the Effect of Individual Differences*. PhD thesis, Indiana Univ. (2017).
34. Sang, K., Todd, P. & Goldstone, R. Learning near-optimal search in a minimal explore/exploit task. In *Proc. 33rd Annual Conference of the Cognitive Science Society* 2800–2805 (Cognitive Science Society, 2011).

35. Sang, K., Todd, P. M., Goldstone, R. & Hills, T. T. Explore/exploit tradeoff strategies in a resource accumulation search task. Preprint at https://psyarxiv.com/zw3s8 (2018).
36. Hills, T. T., Todd, P. M. & Goldstone, R. L. The central executive as a search process: priming exploration and exploitation across domains. *J. Exp. Psychol. Gen.* **139**, 590–609 (2010).
37. Navarro, D. J., Newell, B. R. & Schulze, C. Learning and choosing in an uncertain world: an investigation of the explore–exploit dilemma in static and dynamic environments. *Cogn. Psychol.* **85**, 43–77 (2016).
38. 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**, 2074–2081 (2014).
39. Stoll, F. M., Fontanier, V. & Procyk, E. Specific frontal neural dynamics contribute to decisions to check. *Nat. Commun.* **7**, 11990 (2016).
40. Kolling, N., Wittmann, M. & Rushworth, M. F. S. Multiple neural mechanisms of decision making and their competition under changing risk pressure. *Neuron* **81**, 1190–1202 (2014).
41. Mai, J.-E. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior* (Emerald Group Publishing, UK, 2016).
42. Badre, D., Doll, B. B., Long, N. M. & Frank, M. J. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron* **73**, 595–607 (2012).
43. Boorman, E. D., Behrens, T. E. J., Woolrich, M. W. & Rushworth, M. F. S. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* **62**, 733–743 (2009).
44. Barraclough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* **7**, 404–410 (2004).
45. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
46. Wallis, J. D. & Miller, E. K. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *Eur. J. Neurosci.* **18**, 2069–2081 (2003).
47. Watanabe, M. Reward expectancy in primate prefrontal neurons. *Nature* **382**, 629–632 (1996).
48. Rich, A. S. & Gureckis, T. M. Exploratory choice reflects the future value of information.*Decision* **5**, 177–192 (2018).
49. Gureckis, T. M. et al. psiTurk: an open-source framework for conducting replicable behavioral experiments online. *Behav. Res. Methods* **48**, 829–842 (2016).
50. Glimcher, P. & Fehr, E. *Neuroeconomics* 2nd edn (Academic Press, 2014).

Acknowledgements

The authors thank R. Polonia for helpful comments on the manuscript, and people in W.J.M.'s laboratory for helpful discussions. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

All authors designed the study, developed the models, interpreted the results, and wrote the paper. M.S. and Z.B. collected the data and performed the analyses.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0526-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to W.J.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Wei Ji Ma

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	PsiTurk, JavaScript
Data analysis	MATLAB

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data in this paper are available at https://github.com/mingyus/explore-exploit (under folder data/).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	quantitative experimental
Research sample	Experiment 1: 49 laboratory participants (10 male, 24 female and 15 unknown, aged 18-57) Experiment 2: 143 Amazon Mechanical Turk participants (no demographic information collected) Experiment 3: 131 Amazon Mechanical Turk participants (no demographic information collected) Experiment 4: 16 laboratory participants (7 male and 9 female, aged 20-43) Experiment 5: 108 Amazon Mechanical Turk participants (no demographic information collected)
Sampling strategy	random sampling; no sample-size calculation was performed
Data collection	Experiments 1 and 4 were conducted with computers. Participants in experiments 2, 3 and 5 performed the task with their own computers via Internet. In all experiments participants went through self-paced instructions screens.
Timing	Experiment 1: 12/16/2016 - 1/23/2017 Experiment 2: 4/19/2017 - 5/1/2017 Experiment 3: 5/10/2017 - 5/21/2017 Experiment 4: 8/13/2015 - 9/7/2015 Experiment 5: 8/31/2015 - 9/16/2015
Data exclusions	No data were excluded from analyses.
Non-participation	In experiments 1 and 4, no participants dropped the study; In experiments 2, 3 and 5 (Amazon Mechanical Turk experiments), 9, 6 and 6 participants respectively did not pass the qualifying test (a test on their understanding of the instructions) and thus did not proceed to the main task.
Randomization	Not applicable.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	Laboratory participants were recruited from the New York University participant pool. Amazon Mechanical Turk participants were recruited through the general Amazon Mechanical Turk task listing. No self-selection bias was aware to the authors.