nature human behaviour

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

Sources of suboptimality in a minimalistic explore–exploit task

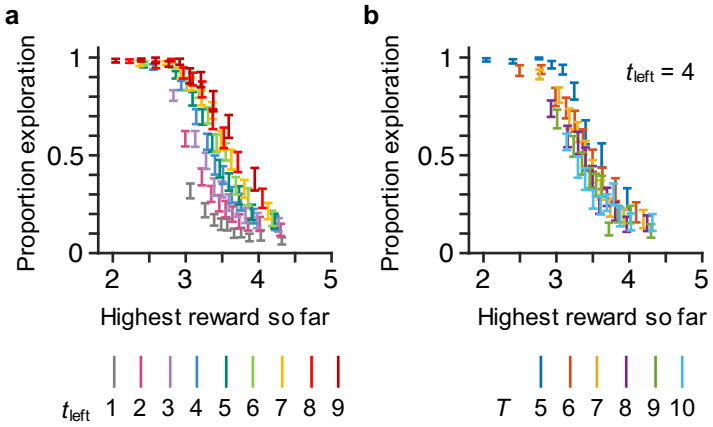
Mingyu Song[1,2,3,4], Zahy Bnaya[2,3,4] and Wei Ji Ma[2,3]*

[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

SUPPLEMENTARY INFORMATION

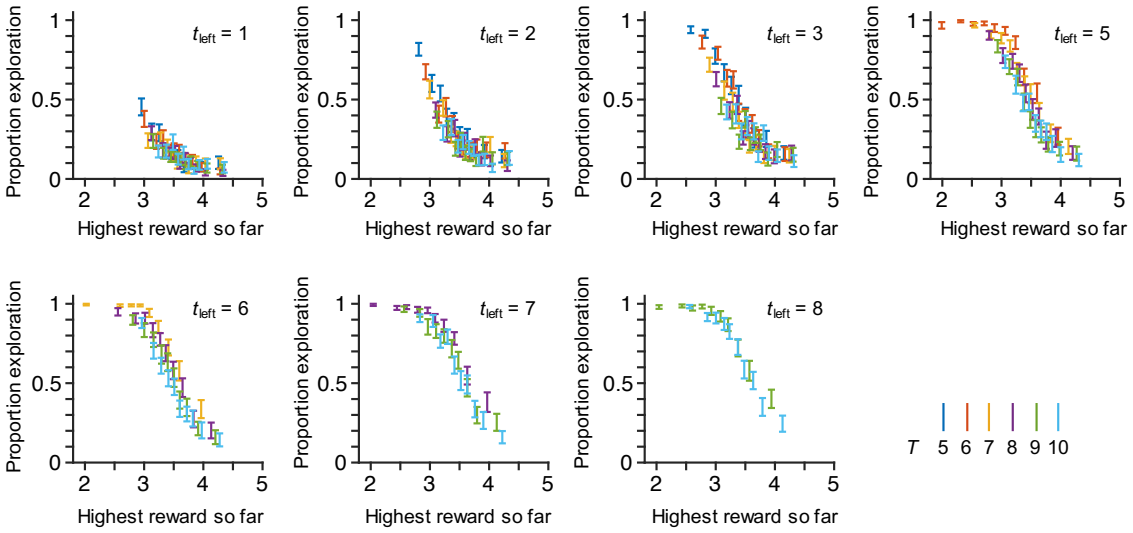
Sources of suboptimality in a minimalistic explore-exploit task
Mingyu Song, Zahy Bnaya, Wei Ji Ma

Supplementary Figure 1



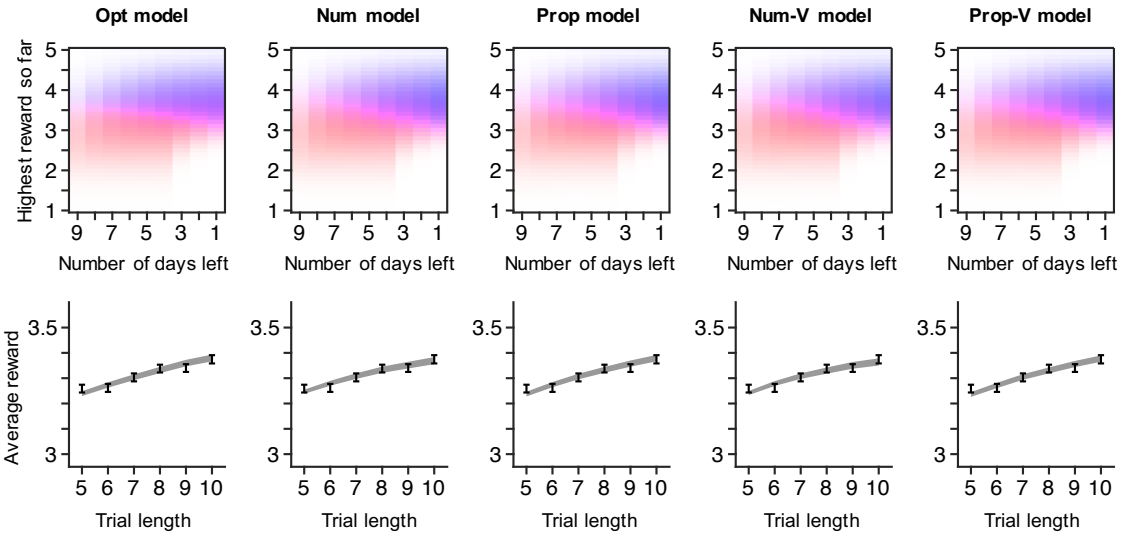
Full version of Figure 1d and 1e. (a) Proportion of decisions in which participants explored as a function of the highest reward so far and the number of days left, averaged across participants. For each participant and each t_{left}, we divided the r^* values from all decisions into 10 quantiles; within each quantile, we calculated the proportion of decisions in 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. **(b)** Proportion of exploration as a function of the highest reward so far for $t_{left} = 4$, broken down by trial length, aka total number of days (T). Error bars represent 1 s.e.m.

Supplementary Figure 2



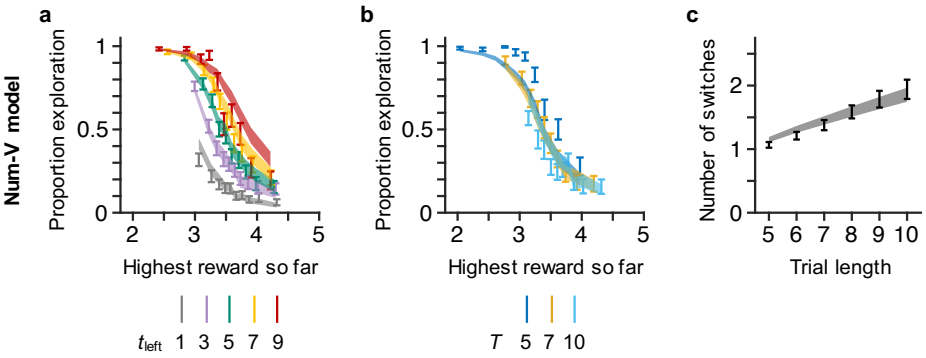
Same as Supplementary Figure 1b (proportion of exploration as a function of the highest reward, broken down by trial length, aka total number of days *T*), but for t_{left} = 1, 2, 3, 5, 6, 7 and 8. Error bars represent 1 s.e.m.

Supplementary Figure 3



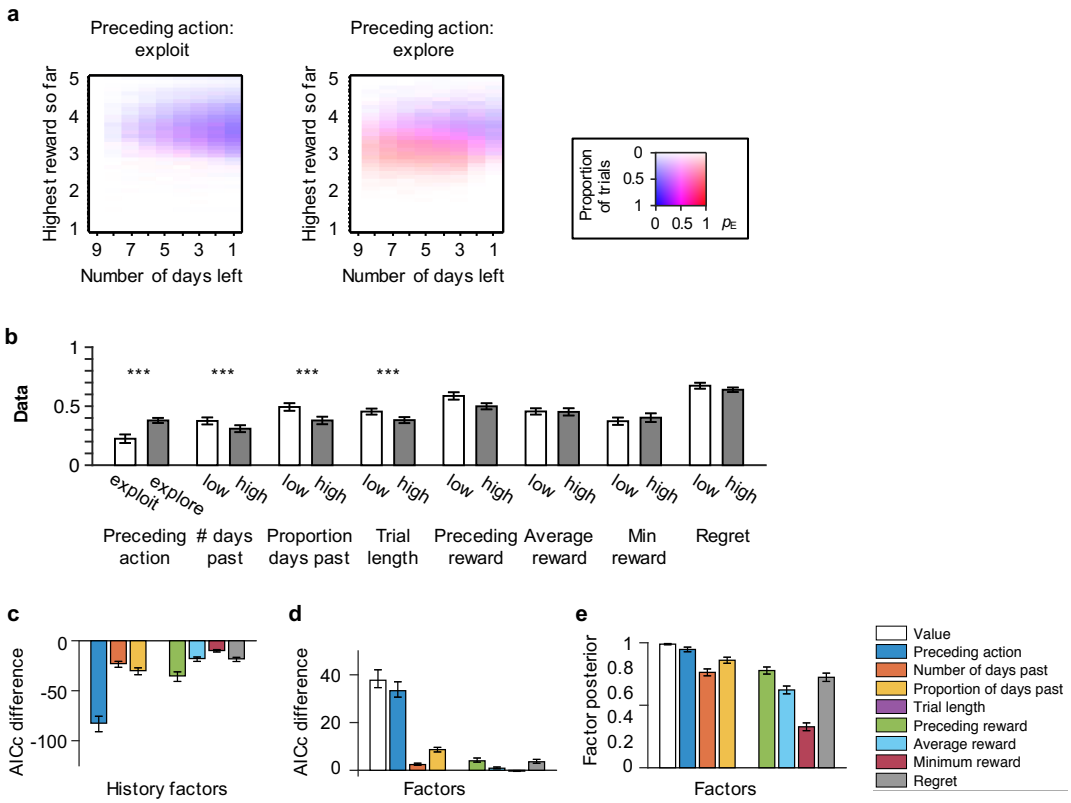
Top row: model fits to Figure 1c (proportion of decisions in which participants explored as a function of the highest reward so far and the number of days left, averaged across participants). Bottom row: model fits to Figure 1g (average reward as a function of trial length). Error bars (data) and shaded areas (model fits) represent 1 s.e.m.

Supplementary Figure 4



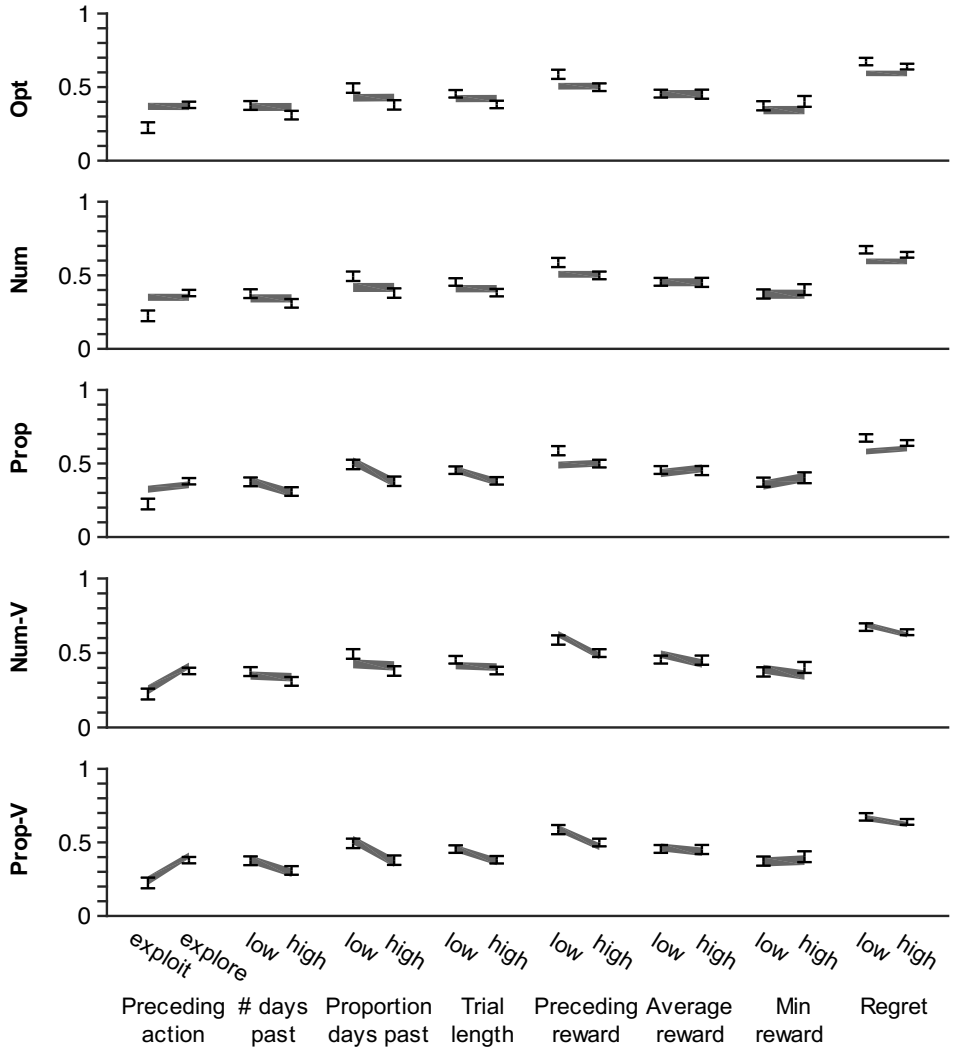
Fits of the Num-V model to the summary statistics (besides the ones already shown in Supplementary Figure 2). **(a)** Fits of the Num-V model to Figure 1d (proportion of decisions in which participants explored as a function of the highest reward so far and the number of days left, averaged across participants); **(b)** Fits of the Num-V model to Figure 1e (proportion of exploration as a function of the highest reward so far for $t_{left} = 4$, broken down by trial length, aka total number of days T); **(c)** Fits of the Num-V model to Figure 1f (the number of switches between exploration and exploitation, averaged across trials, as a function of trial length). Error bars (data) and shaded areas (model fits) represent 1 s.e.m.

Supplementary Figure 5



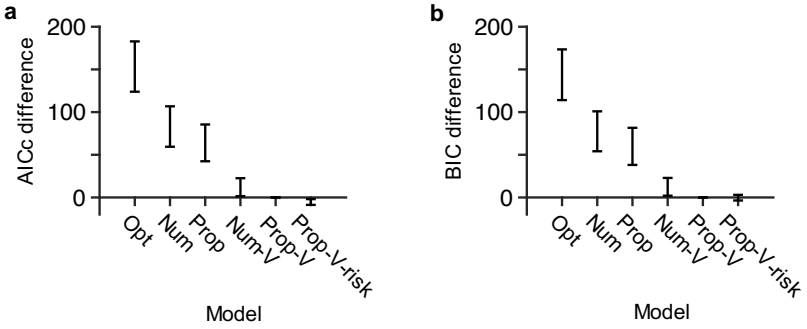
(a) Proportion of exploration conditioned on the preceding action. For each participant, we calculated the proportion of exploration for each combination of best reward so far, r^* and number of days left, t_{left}, under the two conditions. The average over lab participants are shown here as an example. Same legend as Figure 1c. **(b)** The first two bars: average of the values in (a) over the combinations of r^* and t_{left} that are common between both conditions. The other bars: as the first two bars but conditioned on the number of days past (low: $t \leq 4$; high: $t>4$), the proportion of days past (low: <0.5, high: >0.5), trial length (low: ≤7, high: >7), the preceding reward (low: <3; high: >3), average reward (low: average reward<3; high: average reward>3), minimum reward (low: <3; high: >3), and regret (low: <0; high: >0). ***: $p<0.001$ in a two-tailed paired t-test between two conditions after Boferroni-Holm correction. **(c)** AICc difference when adding in one history factor to the null model (with only r^* and t_{left} as regressors). **(d)** AICc difference when dropping out one factor from the History model. Error bars represent 95% bootstrapped confidence intervals in (c) and (d). **(e)** Factor posterior based on Bayesian model selection analyses [1,2] on models with all possible combination of the eight factors. Error bars represent s.e.m. The "value" factor refers to r^* and t_{left}. Results for the "value" factor is provided as a reference for the history factors. There is strong evidence for preceding action. Other history factors are identifiable by group but not individually. Error bars represent 1 s.e.m.

Supplementary Figure 6



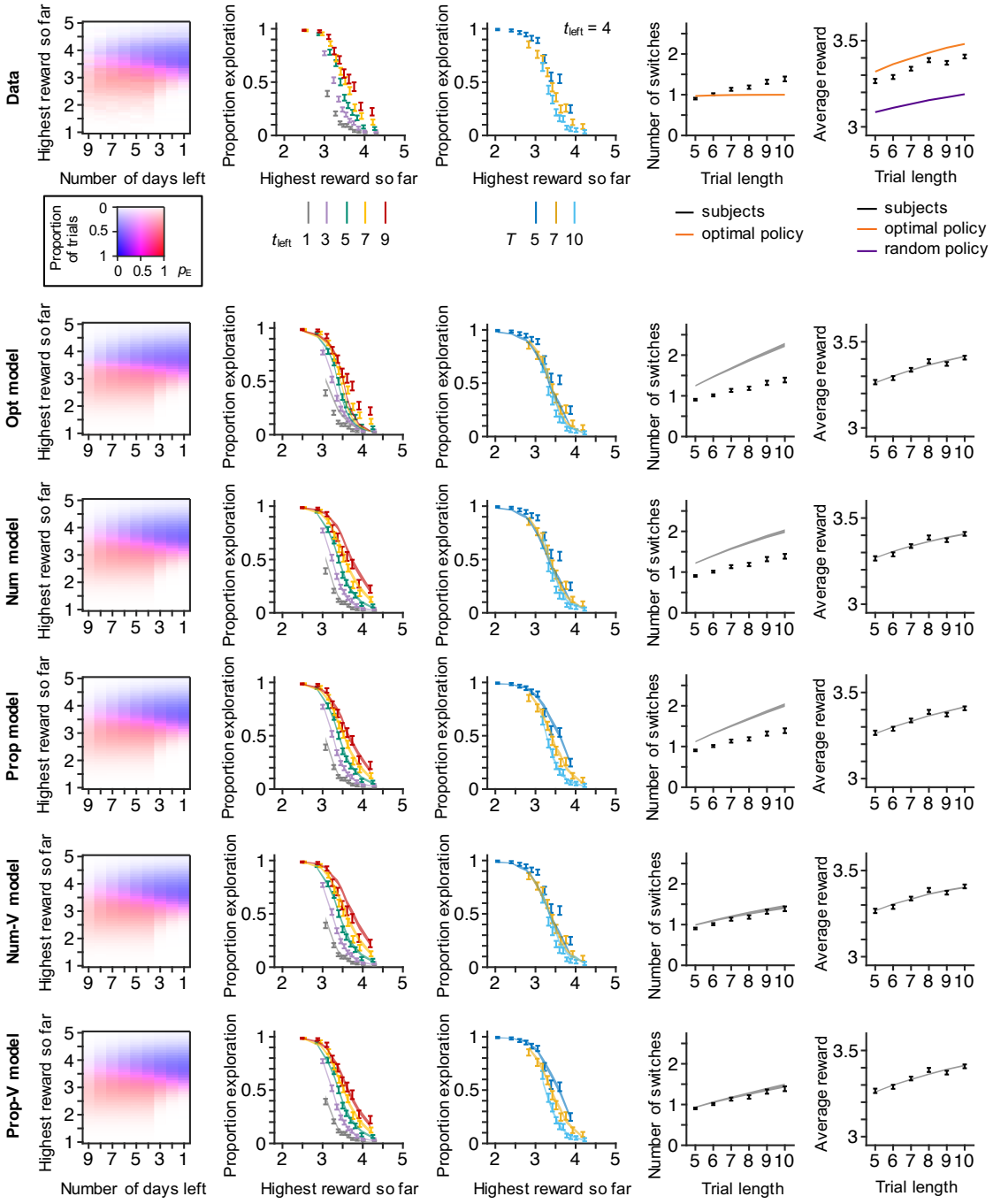
The model fits to effects of history factor in Supplementary Figure 5b. The error bars represent data (same as Supplementary Figure 5b); the shaded areas represent model fits. The Prop-V model accounts well for all eight history factors. Other four models can't account for all eight factors, indicating that both the threshold rule on the proportion of days past and the sequence-level variability are necessary for the good fits to intra-trial history effects. Error bars (data) and shaded areas (model fits) represent 1 s.e.m.

Supplementary Figure 7



(a) AICc and **(b)** BIC comparisons, including the Prop-V-risk model (the Prop-V model with an additional risk attitude factor, implemented as an exponent on the r^* term). Adding the risk factor into the Prop-V model doesn't improve the fit much. Error bars represent 95% bootstrapped confidence intervals.

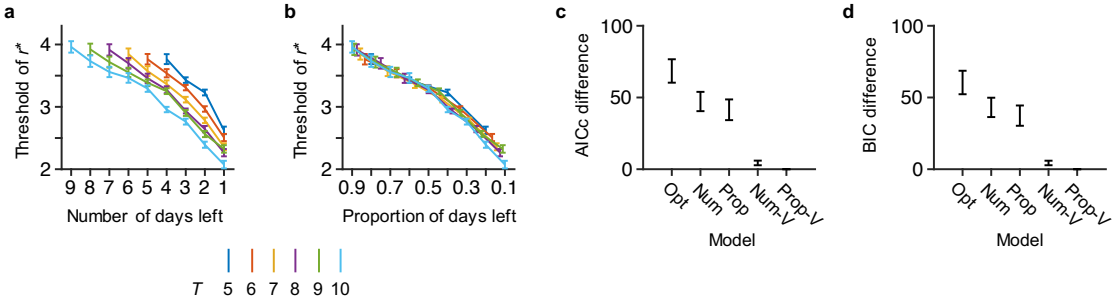
Supplementary Figure 8



Summary statistics (first row) and model fits (second to last rows) in Experiment 2 (143 Mturk participants). First column: Proportion of decisions in which participants explored as a function of the highest reward so far and the number of days left, averaged across participants. Second column: Slices from the plot in the first column. Third column: Proportion of exploration as a function of the highest reward so far for $t_{left} = 4$, broken down by trial length, aka total number of days (T). Forth column: The number of

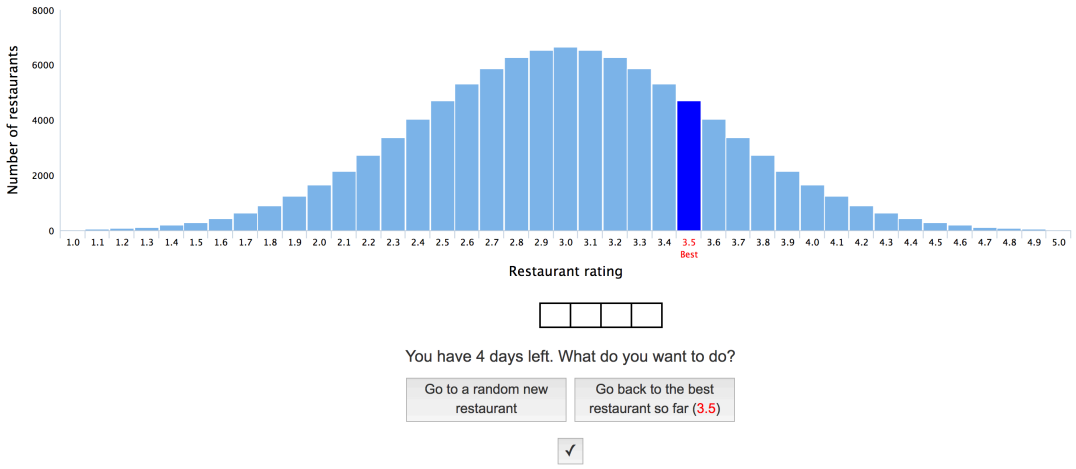
switches between exploration and exploitation, averaged across trials, as a function of trial length. Fifth column: Average reward as a function of trial length. Error bars (data) and shaded areas (model fits) represent 1 s.e.m.

Supplementary Figure 9



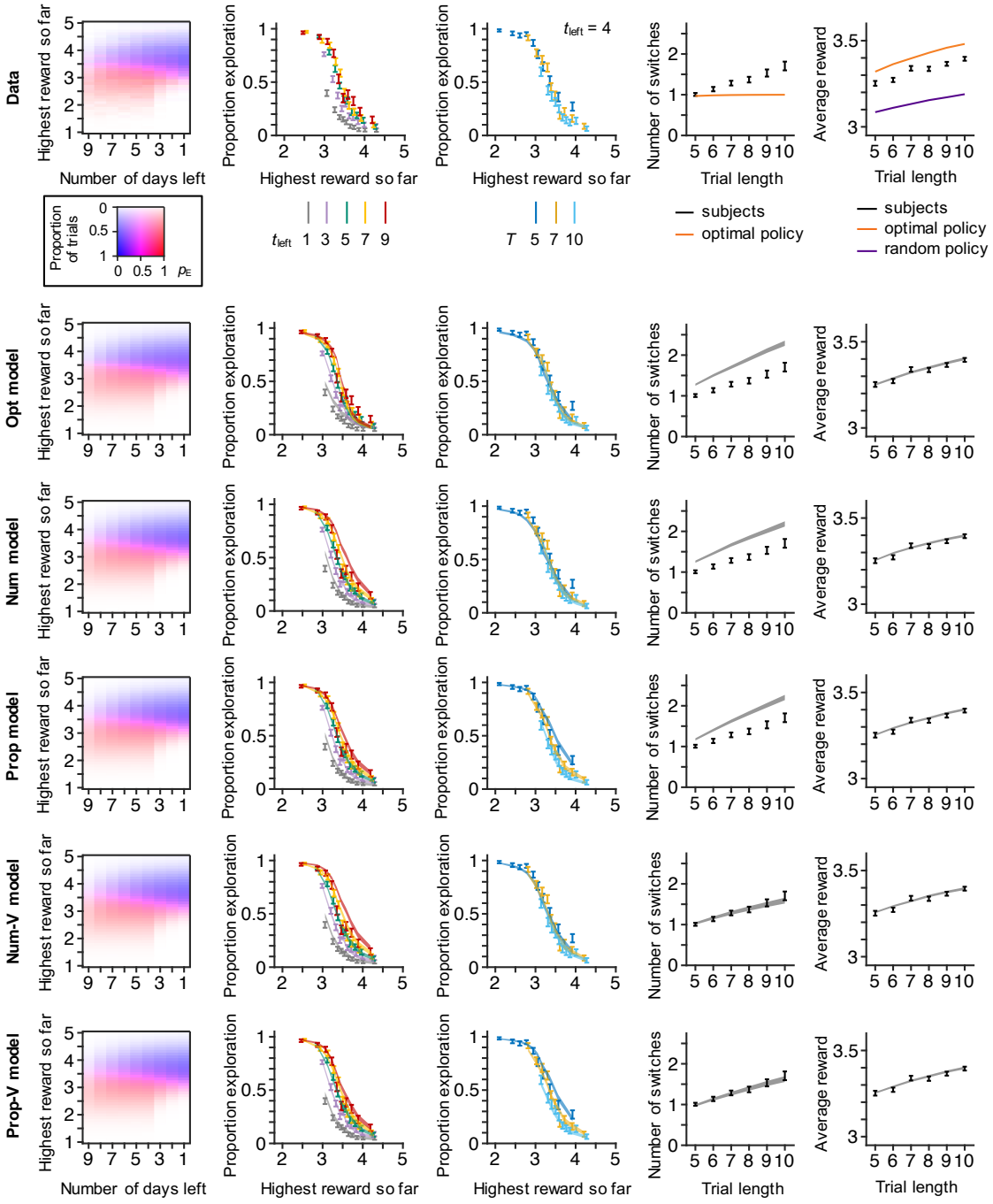
Counterparts of Figure 3d, 3e and 4c in Experiment 2 (n = 143). (a) The fitted threshold of r^* as a discrete function of t_{left} and T. **(b)** The same curves as in (a) with the independent variable changed to proportion of days left (each curve is stretched along the x axis respectively). In (a) and (b), error bars represent 1 s.e.m. **(c)** AICc and **(d)** BIC comparisons. In (c) and (d), error bars represent 95% bootstrapped confidence intervals.

Supplementary Figure 10



Design of Experiment 3. Most history information (trial length, previous rewards and the accumulated reward so far) was hidden from participants.

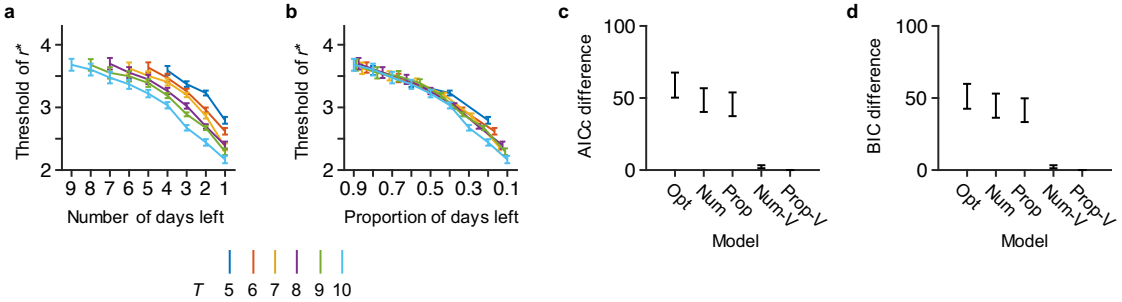
Supplementary Figure 11



Summary statistics (first row) and model fits (second to last rows) in Experiment 3 (131 Mturk participants). First column: Proportion of decisions in which participants explored as a function of the highest reward so far and the number of days left, averaged across participants. Second column: Slices from the plot in the first column. Third column: Proportion of exploration as a function of the highest reward so far for $t_{left} = 4$, broken down by trial length, aka total number of days (T). Forth column: The number of

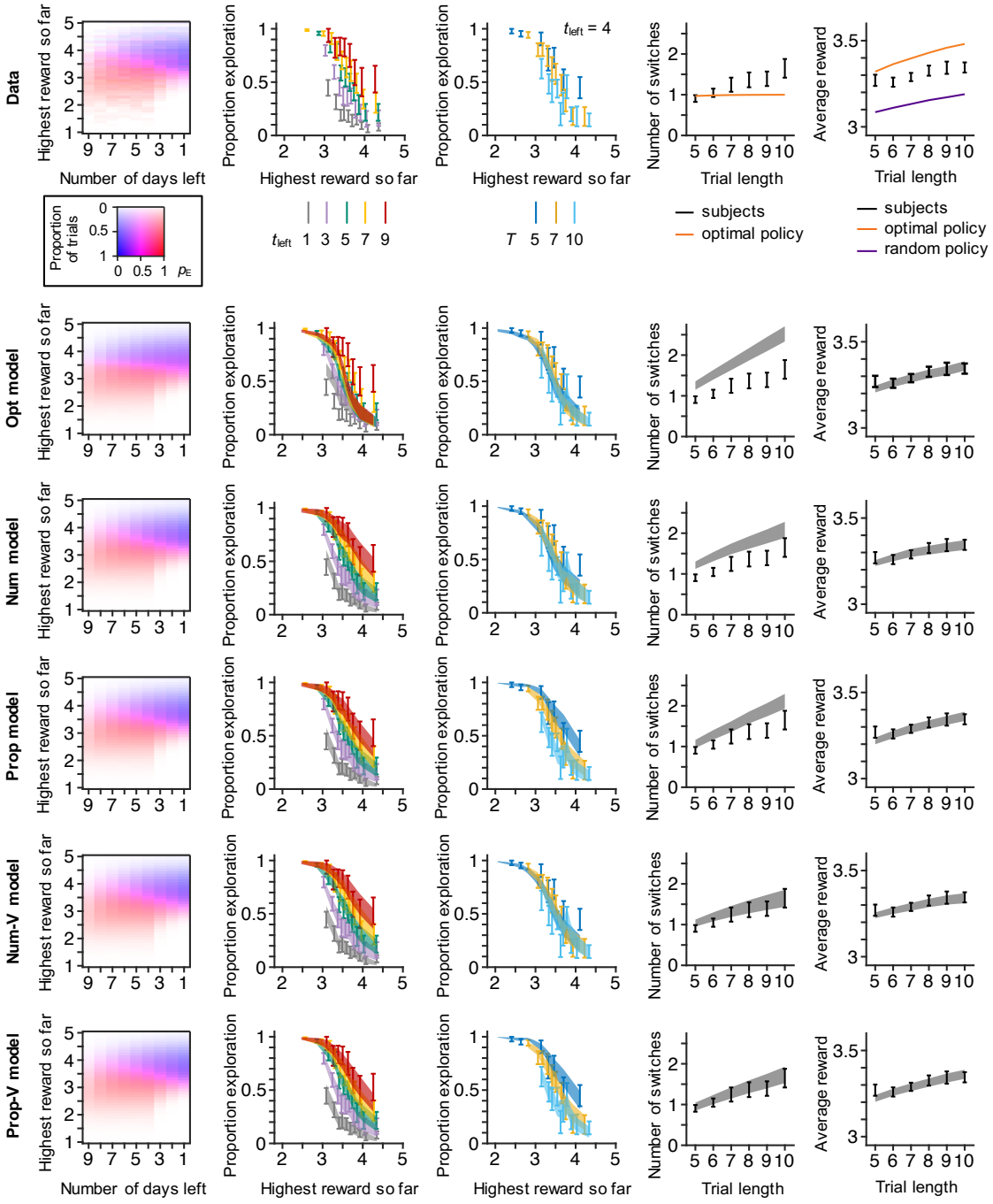
switches between exploration and exploitation, averaged across trials, as a function of trial length. Fifth column: Average reward as a function of trial length. Error bars (data) and shaded areas (model fits) represent 1 s.e.m.

Supplementary Figure 12



Counterparts of Figure 3d, 3e and 4c in Experiment 3 (*n* = 131). (a) The fitted threshold of r^* as a discrete function of t_{left} and T. **(b)** The same curves as in (a) with the independent variable changed to proportion of days left (each curve is stretched along the x axis respectively). In (a) and (b), error bars represent 1 s.e.m. **(c)** AICc and **(d)** BIC comparisons. In (c) and (d), error bars represent 95% bootstrapped confidence intervals.

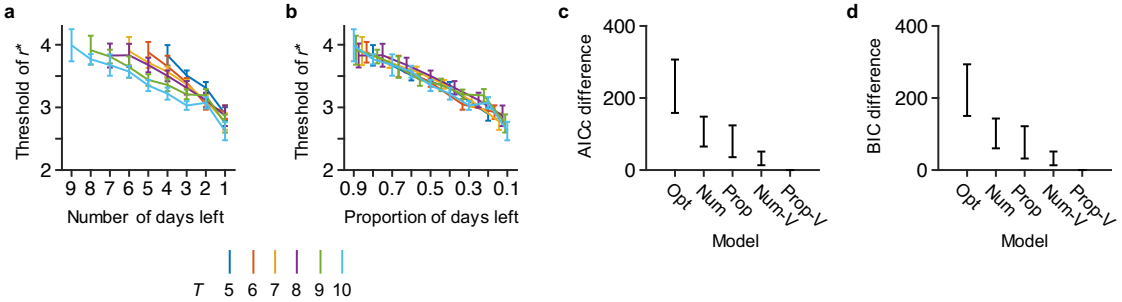
Supplementary Figure 13



Summary statistics (first row) and model fits (second to last rows) in Experiment 4 (16 lab participants). First column: Proportion of decisions in which participants explored as a function of the highest reward so far and the number of days left, averaged across participants. Second column: Slices from the plot in the first column. Third column: Proportion of exploration as a function of the highest reward so far for $t_{left} = 4$, broken down by trial length, aka total number of days (T). Forth column: The number of

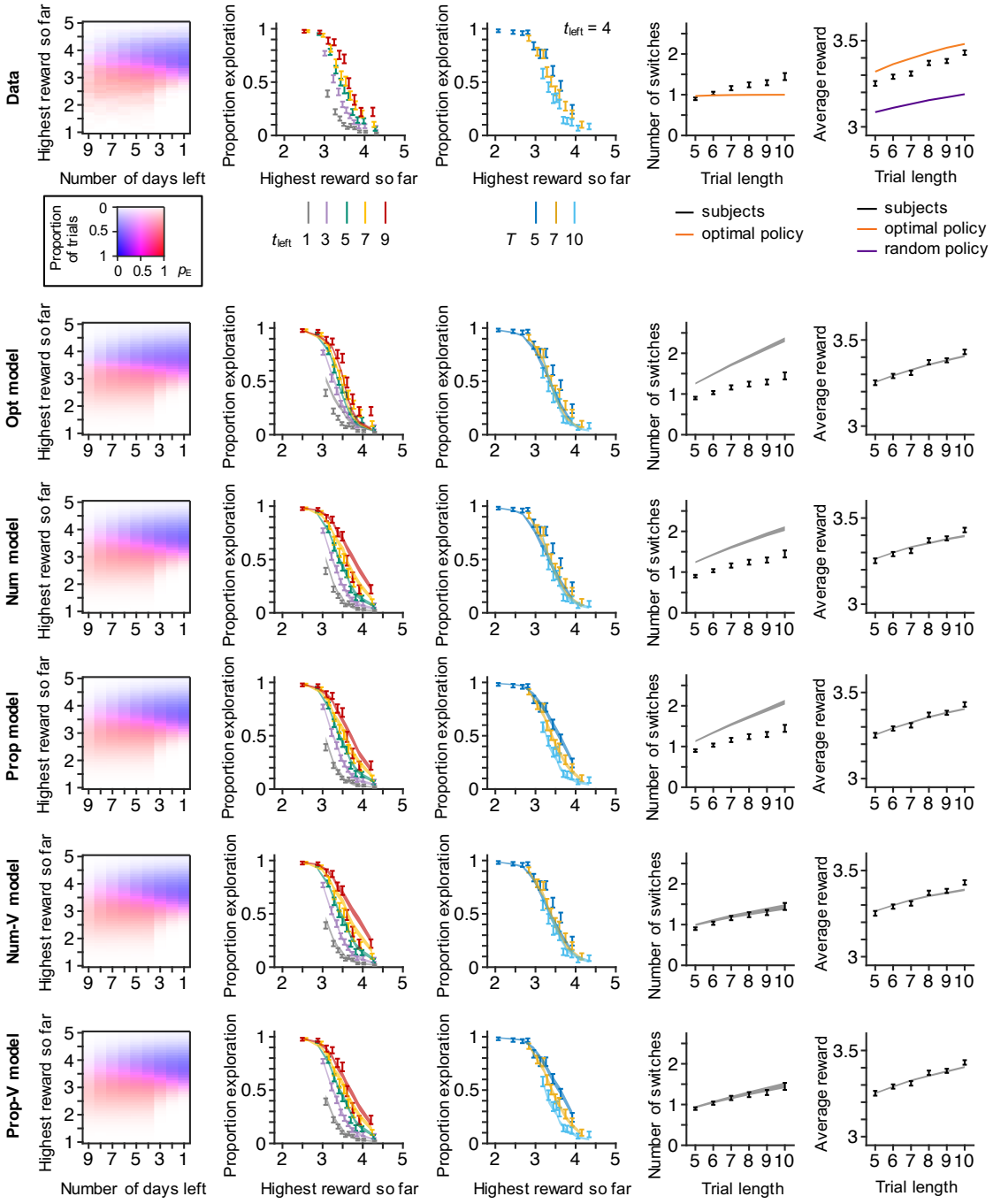
switches between exploration and exploitation, averaged across trials, as a function of trial length. Fifth column: Average reward as a function of trial length. Error bars (data) and shaded areas (model fits) represent 1 s.e.m.

Supplementary Figure 14



Counterparts of Figure 3d, 3e and 4c in Experiment 4 (n = 16). **(a)** The fitted threshold of r^* as a discrete function of t_{left} and T. **(b)** The same curves as in (a) with the independent variable changed to proportion of days left (each curve is stretched along the x axis respectively). In (a) and (b), error bars represent 1 s.e.m. **(c)** AICc and **(d)** BIC comparisons. In (c) and (d), error bars represent 95% bootstrapped confidence intervals.

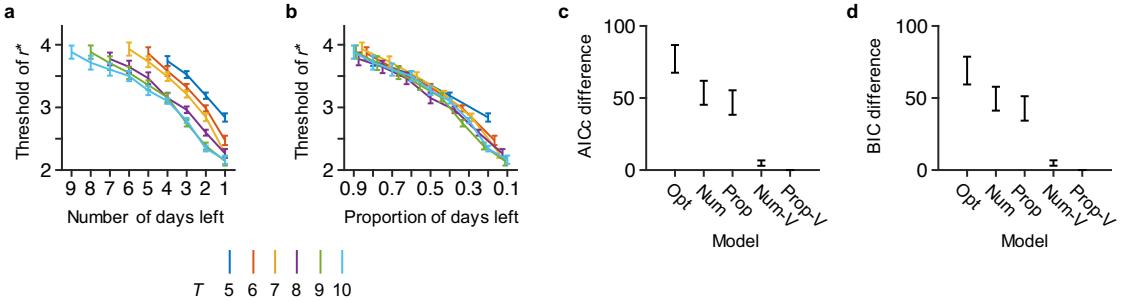
Supplementary Figure 15



Summary statistics (first row) and model fits (second to last rows) in Experiment 5 (108 Mturk participants). First column: Proportion of decisions in which participants explored as a function of the highest reward so far and the number of days left, averaged across participants. Second column: Slices from the plot in the first column. Third column: Proportion of exploration as a function of the highest reward so far for $t_{left} = 4$, broken down by trial length, aka total number of days (T). Forth column: The number of

switches between exploration and exploitation, averaged across trials, as a function of trial length. Fifth column: Average reward as a function of trial length. Error bars (data) and shaded areas (model fits) represent 1 s.e.m.

Supplementary Figure 16



Counterparts of Figure 3d, 3e and 4c in Experiment 5 (*n* = 108). **(a)** The fitted threshold of r^* as a discrete function of t_{left} and T. **(b)** The same curves as in (a) with the independent variable changed to proportion of days left (each curve is stretched along the x axis respectively). In (a) and (b), error bars represent 1 s.e.m. **(c)** AICc and **(d)** BIC comparisons. In (c) and (d), error bars represent 95% bootstrapped confidence intervals.

Supplementary Table 1: Bayesian Information Criterion results

Model	Bayesian Information Criterion values (relative to the Prop-V model; 95% bootstrapped confidence interval in brackets)
Opt	141 [114, 172]
Num	76 [54, 101]
Prop	56 [37, 81]
Num-V	13.8 [2.3, 22.8]
Prop-V	0

Supplementary Table 2: Repeated-measure two-way ANOVA on average reward and number of switches per trial

	df	*F*	*p*	Partial η_2
Average reward	Fold: 2 Trial length: 5 Fold * trial length: 10	Fold: 0.39 Trial length: 18.18 Fold * trial length: 0.51	Fold: 0.68 Trial length: < 0.001 Fold * trial length: 0.88	Fold: 0.008 Trial length: 0.275 Fold * trial length: 0.011
Number of switches	Fold: 2 Trial length: 5 Fold * trial length: 10	Fold: 6.86 Trial length: 43.27 Fold * trial length: 1.03	Fold: 0.016 Trial length: < 0.001 Fold * trial length: 0.41	Fold: 0.13 Trial length: 0.47 Fold * trial length: 0.02

Supplementary Table 3: **Repeated-measures one-way ANOVA on the estimated parameters in the Prop-V model**

Parameter	df	*F*	*p*	Partial η_2
k	2	1.15	0.32	0.023
b	2	0.92	0.40	0.018
σ	2	1.01	0.37	0.020
$1/\beta$	2	1.08	0.35	0.022

Supplementary Method 1: Problem formulation as a Markov Decision Process

We denote the restaurant rating by r and its probability distribution (as displayed on the screen) by $p(r)$. Our task can be modeled as a deterministic Markov Decision Process represented by the tuple (S, A, P, R). S is the set of all possible states s, each of which is defined by a pair (r^*, t_{left}), which are the highest reward received so far (initialized as $r^*=1$ on the first day) and the number of days left, respectively. A is the set of possible actions in each state: 0 (exploitation) and 1 (exploration), except that on the first day, only exploration is possible.

The transition function $P(s, a, s')$ describes the probability of reaching a new state $s'=\left(r^{*\prime}, t_{\text{left}}{}'\right)$ when applying action a in state $s=(r^*, t_{\text{left}})$. For exploitation, the transition function is

$$P\left(s,0,s'\right)=\begin{cases}1 & \text{when } t'_{\text{left}}=t_{\text{left}}-1 \text{ and } r^{*\prime}=r^*;\\ 0 & \text{otherwise.}\end{cases}$$

For exploration, the transition function is

$$P\left(s,1,s'\right)=\begin{cases}p\left(r^{*\prime}\right) & \text{when } t'_{\text{left}}=t_{\text{left}}-1 \text{ and } r^{*\prime}>r^*;\\ \Pr\left(r\le r^{*\prime}\right) & \text{when } t'_{\text{left}}=t_{\text{left}}-1 \text{ and } r^{*\prime}=r^*;\\ 0 & \text{otherwise.}\end{cases}$$

The reward function $R\left(s,a,s'\right)$ describes the expected immediate reward received after choosing action a. In our case, it is independent of s', and equal to r^* for a=0 and 3.0 for a=1.

The optimal value of each state and state-action pair is then specified by the Bellman equations [3],

$$V\left(s\right)=\max_a Q\left(s,a\right)$$
$$Q\left(s,a\right)=\sum_{s'\in S}P\left(s,a,s'\right)\left(R\left(s,a,s'\right)+V\left(s'\right)\right)$$

The optimal policy can be directly calculated from $Q(s,a)$ as $\pi\left(s\right)=\operatorname{argmax}_a Q(s,a)$.

Plugging in the specific functions of our problem, the optimal policy takes the form

$$\pi\left(r^*, t_{\text{left}}\right) = \operatorname*{argmax}_a Q\left(r^*, t_{\text{left}}; a\right)$$
$$V\left(r^*, t_{\text{left}}\right) = \max_a Q\left(r^*, t_{\text{left}}; a\right)$$
$$Q\left(r^*, t_{\text{left}}; a=0\right) = r^* + V\left(r^*, t_{\text{left}} - 1\right)$$
$$Q\left(r^*, t_{\text{left}}; a=1\right) = 3.0 + \sum_r p(r) V\left(\max\left(r^*, r\right), t_{\text{left}} - 1\right)$$

Supplementary Method 2: Proof that it is never optimal to switch more than once

Definition: A *single-switch* policy is a policy in which explorations never follow exploitations. We call policies that do not follow this rule *multi-switch* policies.

Theorem: For any instance of our task, the optimal policy is a *single-switch* policy.

Proof: Let $\tilde{V}_\pi(l)$ be the expected partial reward received by applying policy π on the first l days, starting from the initial state. Let $\phi_\pi(l)$ be the number of exploration actions performed during the first l days by policy π. Let $\rho(l)$ be the expected highest reward received after exactly l explorations.

Assume by contradiction a multi-switch optimal policy π^* that switches to exploitation on day $t > 1$ for $l \geq 1$ days and then switches back to exploration for $l_2 \geq 1$ days. The expected value of policy π^* is:

$$V_{\pi^*} = \tilde{V}_{\pi^*}\left(t-1\right) + \rho\left(\phi_{\pi^*}\left(t-1\right)\right) \cdot l + \left\langle r \right\rangle \cdot l_2 + V_{\pi^*}\left(s\right).$$

Here, we used $V_{\pi^*}(s)$ to denote the expected reward received from the resulting state with $T - (t - 1 + l + l_2)$ days left and an expected highest reward of $\rho\left(\phi_{\pi^*}\left(t-1\right) + l_2\right)$ $\rho(\phi_{\pi^*}(t-1) + l_2)$.

We define the "swap" operator on policy π^* which swaps the l_2 exploration decisions with the l exploitation decisions performed by π^*. We refer to the resulting policy as $\hat{\pi}^*$. The expected reward of $\hat{\pi}^*$ is

$$V_{\hat{\pi}^*} = \tilde{V}_{\hat{\pi}^*}\left(t-1\right) + \rho\left(\phi_{\hat{\pi}^*}\left(t-1\right) + l_2\right) \cdot l + \left\langle r \right\rangle \cdot l_2 + V_{\hat{\pi}^*}\left(s\right).$$

Since ρ is a monotonically increasing function, $\rho(\phi_{\hat{\pi}^*}(t-1) + l_2) > \rho\left(\phi_{\hat{\pi}^*}(t-1)\right)$ for any $l_2 \geq 1$. And thus:

$$V_{\hat{\pi}^*} > V_{\pi^*}$$

which contradicts the assumption that π^* is an optimal policy. In other words, for any multi-switch policy π there exists a "swap" policy $\hat{\pi}$ with a larger expected reward. Thus, the optimal policy must be single-switch.

Supplementary Method 3: Expected number of switches for the optimal policy

We denote the cumulative reward distribution by Φ. Under the optimal policy with random tiebreaking, the number of switches per trial is 0 when *r** on day *T* is strictly less than 3.0, 0.5 when r* on day T equals 3.0, and 1 when *r** on day *T* is greater than 3.0. Thus, the expected number of switches per trial of the optimal agent is

$$\Phi\left(2.9\right)^{T-1}\cdot 0+\left(\Phi\left(3.0\right)^{T-1}-\Phi\left(2.9\right)^{T-1}\right)\cdot 0.5+\left(1-\Phi\left(3.0\right)^{T-1}\right)\cdot 1$$
$$=1-0.5\cdot\left(\Phi\left(3.0\right)^{T-1}+\Phi\left(2.9\right)^{T-1}\right)$$

Supplementary Result 1: Reported strategies from laboratory participants in Experiment 1

Note:

- **Each paragraph was from one participant (49 in total)**
- **N/A means no response**
- **Responses in gray are considered as invalid responses that cannot be interpreted as strategies**
- **Responses with a tick mark are considered as having "explored in the beginning of a trial and switched to exploitation when the best restaurant rating reached a threshold" (see main text)**

1. I tried to set a rating threshold below which I tended to take the risk of trying new restaurants. Say, I hit an above 4.0 rating restaurant on the first day, then I just kept returning to it, rather than risking entering a lower rating restaurant the next day! √

2. I was more willing to gamble to get a random higher score if the average was 3.6 or lower. However, if the random score I pulled was higher than 3.6, then I tended to stick with that. For all instances, once I'd made it halfway through the number of days, I preferred to not gamble and stick with the highest rated average that I'd gotten in that experiment. √

3. Used the idea of probability. Kept trying for a new score if it was very low and the chances of it increasing were high. Once I got a score that if I changed restaurant, would fall lower I stuck with it. √

4. If the initial rating (score) was higher than 3.3, I stayed. Otherwise, I clicked for new rating (score). √

5. I tried to gauge how probable it was to get a better rated restaurant based on how high the best score was and how many days were left. If I started on a relatively high score (3.5 or more) and there were a lot of days (more than 6) I would fish around a little. But if less days, I would likely just return to the best restaurant, or at the most try one more random one to see what came up. If only a couple days were left, I usually reverted to the best score to finish out unless it was low (below 3.0). √

6. i used the lowest rating mostly the new reataurant

7. Whenever the rating started low (1-3.4), I went with a random restaurant because the rating could only go up from there. If it started moderately or very high (3.5-5) I would take a chance on a random restaurant to see if I could improve. If it went up I

took a chance again, if not, I went back to the highest rated restaurant. Sometimes I went to the same restaurant back to back if I saw that the average ratings were all pretty low. √

8. I aimed to end each trip with a return to the highest rated restaurant unless I saw a trend of finding ratings under 3.5, in that case, I continued to go to random restaurants. For trips longer than 5 days, I am more willing to explore different restaurants; however, I would tend to visit restaurants with a rating greater than or equal to 3.8 twice in a row. If I previously encountered a restaurant with a rating of 3.5 or greater and found a rating of 2, I will go back to the highest rated restaurant for a "palette cleanser."

9. Dependent on length of stay, I'd select a random restaurant until I got a higher rating (up to 3/4 selects). If satisfied with the highest rating of these first 3 to 4 selects, I'd frequent the highest rated throughout for remainder of stay. Which I thought would give me the best score possible. √

10. If the curve was higher to the left of the current best, I stayed with the current best (more probability of loosing than gaining). If the curve was higher to the right, I picked a new one at random. √

11. Usually after 3 attempts to get the highest rating possible, I would stick with the best restaurant I had at the beginning of my trip. In my head, early risk was better than late risk. I possibly could have missed out on big rewards at the end of my "trips", however they wouldn't have made much of a difference if I got all 3 ratings and below throughout the rest.

12. I deemed "3.5" as the score to look for. So, I would choose a random restaurant until I got this score or above OR until I used up about half of the days I had in the session. Some sessions may not fit this pattern, especially in the first section. In these sessions, if I had mostly low scores, I would "go back" to the highest scoring restaurant for the final day for a small boost of points. √

13. I tested usually the first two to three restaurants if the number of days was greater than 6/7, and then would stick with the highest rated restaurant if the difference between the highest and lowest rating was great. I only tested two restaurants for options of 5 days or less, and stuck with the higher rating regardless. At times if I had a particularly good score in the upper 3s or lower 4s, and a greater number of days, I would try out a new restaurant at least once before sticking with the highest rating. Anything over 4.2 I usually stayed with for the duration of the trip. √

14. N/A

15. When the restaurant had a rating that was 3 or higher, I was more likely to try different places. However, when the food was nasty, I stuck to the highest ranking restaurant.

16. My first step was to always pick a random number (obviously) then from there, based on the number generated, I either clicked random number again or returned the same best number. Sometimes to maximize the sum of ratings I experimented in where I would get a high number from the random button, then I would click on random again to test if the number was larger, as a result I clicked on the best button after to get the new larger number or if not I would just return the (former)best number.

17. I tried to get a good restaurant in the first few clicks and then tried to stay with it. I had a cutoff around 4 points (I stayed with it when I got a restaurant around the cutoff) √

18. Randomize until I got a rating of at least 3.0, then stick with that rating for the rest of each and every assignment √

19. To choose the best Restaurant.

20. chance

21. I kept rolling until I got a >3.2 best rating, unless there were only a few days left, in which case I used whatever best rating I had. √

22. I used my beginner's level knowledge of probability and statistics to roughly gauge the likelihood of getting a higher rating earlier in each set of days, switching to returning to the highest restaurant so far when I got a very high number, or when I was around halfway through. I also tried to return to the highest rated restaurant for the last day if it was over 3.0; in that case, the probability of getting something higher would be less than 50%, so it wasn't worth trying. But I didn't use this strategy throughout. √

23. If I got a 4 or above, I would choose to return to that same restaurant rather than roll the dice and risk getting a lower score. If I got anything below 3.3, I always would choose another random restaurant. If it was a longer sequence and several of the numbers were low, then I'd pick the highest number for the final selection; it gave me a feeling of control because I was guaranteed to know what that final number would be. I took more chances the longer the sequence was. For example, in a 10-restaurant sequence, if the first number was a 3.8, I'd be more willing to roll the dice and see if I

got a higher number. But if there were only 5 restaurants, then I was more likely to return to that same restaurant multiple times. √

24. I would see if I could get a high percentage in the task early on, and then I would continue going back to the same restaurant until the task was completed. √

25. tried to calculate statistics based on the distribution and on the number of steps left tried to maximized the total score for instance, when had 10 days, tried to play with more options, when i had treshold sat to around 3.7 - if equal or above, i would choose repeat same restureant then when had 9 days, the trehsold went down to around 3.6 - and again with same decision mechanism √

26. If the rating was equal or lower than 3 I would attempt again. If it was greater than 4 I would most likely keep going to the same restaurant. √

27. I would explore different restaurants until i found a restaurant i liked with a high rating. once i found a restaurant with a high rating i would stick with that restaurant for the rest of the trip and maybe try a new place out once or twice for the remainder of the trip. √

28. always duplicate numbers 3.4 and up anything lower can be risk

29. My strategy was to get the highest rated restaurant and continue going to that restaurant for the rest of my days. √

30. My strategy changed in the middle of the game. After a while, I decided that if I got 3.0 or above, then I would try once more for a higher rating and if I did not get it, then opt for returning to that original 3.0+ restaurant. I also decided that if I got a 3.4 or above, I would just keep going back. At the beginning I wasn't mindful of the idea that it is best to try different places early on, rather than try a random one at the end because the earlier you hit a high rating, the more often you can return to the restaurant with a high rating. √

31. I tried to maximize my total assignment scores. I created a competition with myself to beat my previous best score (which changed from 30, later 35, finally 40). I tried to quickly assess probabilities of getting a better score with a random pick vs. best score. I have some familiarity and experience with statistics, so I think that helped make quick decisions.

32. I would click new restaurant until I got a high enough rating that I was satisfied with √

33. standard deviation-- statistics percent probability

34. I don't think I really had a strategy, but I knew that if I was going to several bad restaurants in a row (below 3 rating) at the start of the trip, for the last 2-3 days of the vacation I almost always decided to return to the best rated restaurant rather than choose a new one. If I was actually traveling, I would want to ensure that I would enjoy the food for my last days of the vacation rather than take a chance on a restaurant that might be terrible. For some of the assignments I think that I was more focused on getting the highest score (even if that meant going to the same restaurant every day on the vacation, which I would never do in real life), but for others I definitely tried to choose my answers as if I were truly in that situation.

35. If I had 1 3.7 or higher I tended to stay with that for the rest of the trip, I did not take many risks If my score was above a 3.7. If it was below 3.0 I always randomized to get a greater rating. The last rating I selected tended to be my highest rating from the previous days. √

36. I wanted to visit multiple places, regardless of the ratings, but if I saw that I was visiting mostly low-rated restaurants I rewarded myself with my highest rating. Also, for the most part, I found that I liked to end with a higher-rated restaurant that I'd already visited, rather than a random one.

37. Obviously, the higher scores are less likely, but in order to ensure you get a higher score, you want to optimize your guessing to go-back ratio by making sure your guesses don't jeopardize that round's score. so, halfway through the round, if you're not able to score high, just go with the best so far. but if you get a high score to start off you're start just sticking to the best so far for the entire round. likewise, if you get a very low score in the start, you can continue randomizing for a few more times because another low score is unlikely. √

38. As a rule I like to try different restaurants, so in the beginning of my time in a new city I will usually keep trying new places, even if I really like one of the first ones. When I saw a pattern in the beginning to end of the time that the restaurants were between 1-3, usually I felt I would rather take a risk and try a new place. In this case I usually went back to the best meal. Unless I was feeling like taking a risk and trying something new, which sometimes happens too :)

39. I attempted to obtain values greater than or equal to 3.5. If the majority of my scores were greater than or equal to 3.5, then I chose a random restaurant for the final "day." √

40. aimed for the highest rating and used it when necessary

41. Unless I received a 3.5 or higher in the first half of my visit, I would visit a random restaurant. If I was in the second half of my stay in the foreign city, I would stick with the highest rated restaurant for the remainder of my visit. For example, if there were 10 days I would visit a random restaurant for the first five days (unless I went to a restaurant with a high rating) and stay with my highest rated restaurant for the remaining 5 days. √

42. For shorter days, hit new restaurant till I reach 3.3. For more than 7 days, choose new restaurant till I reach 3.6. Last 2 days are generally always previous restaurant unless highest rating is below 3. √

43. The strategy I developed over the 180 assignments was that if I were to obtain the highest rated place I would then press random and if the new value were to be a lower number I would return to the old place to still maintain the high rating and repeat to see if I could get a higher one in the next day. If I were to get a higher number I would press go to a new place again to see if I could get a higher rated place

44. The experiment tried to find out risk taking behavior. If I began to get very high scores, I kept trying new restaurants; if not, I went back to the restaurant with the highest/best score. If I began to see my scores sinking, I went back to the restaurant with the best score for the last days. If the restaurant ratings were not fluctuating that much, I kept trying new restaurants.

45. I have took my basis as 3.2 , if any number came above that, i used the same number for that assignment. √

46. if i received a number in the first half of the trials that was very high i stuck with that choice throughout. if my first few choices were too low i sought out other higher numbers randomly in the first half of all choices. after that point i just repeatedly chose the highest rated option no matter how low it was. i mainly stuck with any number above the 60th percentile of ratings. √

47. I was trying to keep average score of each trip higher than 3.0 If I got score relatively high and rare, then I will go to this restaurant until the end of the trip If I got scores are less than 3.0 at the beginning, then I will continue visit random restaurants

48. Make the average rating at lease 3.0. The longer the days, the more you could try for random restaurants. It the trip is short, stick with the best.

49. first try to see the trend and after I have less than half of the days left I just keep choosing the best one. Sometimes if the first one is really good I will just keep choosing that one. If it's really bad then I'll just try a few more times. √

Supplementary Result 2: Bayesian statistics results

For all the frequentist statistics in the paper, we supplement results from Bayesian statistics obtained using JASP (JASP Team, 2018). All our conclusions are robust to changes in the chosen prior distribution over r/δ.

1. Logistic regression on choice against r^* and t_{left}

Bayesian one sample t-test on the regression coefficients with 0:
r^* coefficient: δ = -1.281, 95%CI: [-1.661, -0.898], BF_{10} = 2.411e9, extreme evidence for H1
t_{left} coefficient: δ = 2.025, 95%CI: [1.646, 2.468], BF_{10} = 1.175e6, extreme evidence for H1

2. Logistic regression on choice against r^*, t_{left} and T

Bayesian one sample t-test on the regression coefficients with 0:
r^* coefficient: δ = -1.230, 95%CI: [-1.618, -0.706], BF_{10} = 9.797e8, extreme evidence for H1
t_{left} coefficient: δ = 1.920, 95%CI: [1.380, 2.454], BF_{10} = 1.709e15, extreme evidence for H1
T coefficient: δ = -1.069, 95%CI: [-1.438, -0.725], BF_{10} = 1.926e7, extreme evidence for H1

3. Logistic regression on the number of switches against T

Bayesian one sample t-test on the regression coefficient with 0:
slope: δ = 1.023, 95%CI: [0.664, 1.375], BF_{10} = 7.251e6, extreme evidence for H1

4. Bayesian one sample t-test on the number of switches with the optimal policy:

T = 5: δ = 0.388, 95%CI: [0.107, 0.667], BF_{10} = 6.047, moderate evidence for H1
T = 6: δ = 0.548, 95%CI: [0.260, 0.850], BF_{10} = 123.515, extreme evidence for H1
T = 7: δ = 0.667, 95%CI: [0.347, 0.973], BF_{10} = 1728, extreme evidence for H1
T = 8: δ = 0.789, 95%CI: [0.445, 1.114], BF_{10} = 30970, extreme evidence for H1
T = 9: δ = 0.827, 95%CI: [0.492, 1.151], BF_{10} = 63741, extreme evidence for H1
T = 10: δ = 0.858, 95%CI: [0.520, 1.183], BF_{10} = 126033, extreme evidence for H1

5. Logistic regression on the average reward against T

Bayesian one sample t-test on the regression coefficient with 0:

slope: δ = 1.161, 95%CI: [0.786, 1.529], BF_{10} = 1.545e8, extreme evidence for H1

6. Bayesian one sample t-test on the average reward with the optimal policy (3.3223, 3.3606, 3.4005, 3.4294, 3.4569, 3.4834 for T = 5, 6, 7, 8, 9, and 10 respectively):

T = 5: δ = -0.563, 95%CI: [-0.862, -0.259], BF_{10} = 162.433, extreme evidence for H1
T = 6: δ = -0.853, 95%CI: [-1.191, -0.525], BF_{10} = 126510, extreme evidence for H1
T = 7: δ = -0.851, 95%CI: [-1.183, -0.530], BF_{10} = 112327, extreme evidence for H1
T = 8: δ = -0.815, 95%CI: [-1.151, -0.472], BF_{10} = 53412, extreme evidence for H1
T = 9: δ = -1.056, 95%CI: [-1.413, -0.728], BF_{10} = 1.417e7, extreme evidence for H1
T = 10: δ = -0.914, 95%CI: [-1.258, -0.574], BF_{10} = 547892, extreme evidence for H1

7. Bayesian one sample t-test on the average reward with the random policy (3.0856, 3.1105, 3.1319, 3.1533, 3.1726, 3.1898 for T = 5, 6, 7, 8, 9, and 10 respectively):

T = 5: δ = 1.566, 95%CI: [1.146, 1.987], BF_{10} = 1.526e12, extreme evidence for H1
T = 6: δ = 1.325, 95%CI: [0.964, 1.746], BF_{10} = 6.787e9, extreme evidence for H1
T = 7: δ = 1.515, 95%CI: [1.073, 1.932], BF_{10} = 4.844e11, extreme evidence for H1
T = 8: δ = 1.680, 95%CI: [1.286, 2.117], BF_{10} = 1.555e13, extreme evidence for H1
T = 9: δ = 1.520, 95%CI: [1.156, 1.919], BF_{10} = 5.827e11, extreme evidence for H1
T = 10: δ = 1.583, 95%CI: [1.228, 2.029], BF_{10} = 1.827e12, extreme evidence for H1

8. Learning effect

Bayesian paired t-test on the fitted parameters and average reward on the first vs. second half of the data:

k: δ = -0.061, 95%CI: [-0.440, 0.298], BF_{10} = 0.166, moderate evidence for H0
b: δ = 0.195, 95%CI: [-0.178, 0.574], BF_{10} = 0.274, moderate evidence for H0
σ: δ = 0.171, 95%CI: [-0.195, 0.562], BF_{10} = 0.245, moderate evidence for H0
β: δ = 0.178, 95%CI: [-0.195, 0.556], BF_{10} = 0.249, moderate evidence for H0
average reward: δ = -0.058, 95%CI: [-0.426, 0.313], BF_{10} = 0.163, moderate evidence for H0

Supplementary References

1. 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).
2. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies - Revisited. *Neuroimage* **84,** 971–985 (2014).
3. Bellman, R. *Dynamic programming*. (Courier Dover Publications, 1957).