Supplementary Material: Humans use directed and random exploration to solve the exploration-exploitation dilemma

Robert C. Wilson[1,*], Andra Geana[1,2], John M. White[2], Elliot A. Ludvig[1,3], and Jonathan D. Cohen[1,2]

[1]Princeton Neuroscience Institute, Princeton University, Princeton NJ 08540.

[2]Department of Psychology, Princeton University Princeton NJ 08540.

[3]Department of Psychology, University of Warwick, Coventry, UK, CV4 7AL.

Full instructions for the task

Before taking part in the experiment, participants read a set of illustrated onscreen instructions describing the task and its mechanics. Here we provide the full text for these instructions. Each bullet point corresponds to a single screen in the instructions. In the interests of space we have not included the illustrations themselves.

- Welcome! Thank you for volunteering for this experiment.
- In this experiment we would like you to choose between two one-armed bandits of the sort you might find in a casino.
- The one-armed bandits will be represented like this
- Every time you choose to play a particular bandit, the lever will be pulled like this ...
- ... and the payoff will be shown like this. For example, in this case, the left bandit has been played and is paying out 77 points.
- Each bandit tends to pay out about the same amount of reward on average, but there is variability in the reward on any given play.
- For example, the average reward for the bandit on the right might be 50 points, but on the first play we might see a reward of 52 points because of the variability ...
- ... on the second play we might get 56 points ...
- ... if we open a third box on the right we might get 45 points this time ...
- ... and so on, such that if we were to play the right bandit 10 times in a row we might see these rewards ...
- Both bandits will have the same kind of variability and this variability will stay constant throughout the experiment.

- One of the bandits will always have a higher average reward and hence is the better option to choose on average.
- To make your choice: Press < to play the left bandit. Press > play the right bandit
- On any trial you can only play one bandit and the number of trials in each game is determined by the height of the bandits. For example, when the bandits are 10 boxes high, there are 10 trials in each game ...
- ... when the bandits are 5 boxes high there are only 5 trials in the game.
- Finally, the first 4 choices in each game are instructed trials where we will tell you which option to play. This will give you some experience with each option before you make your first choice.
- These instructed trials will be indicated by a green square inside the box we want you to open and you must press the button to choose this option in order to move on to see the reward and move on the next trial. For example, if you are instructed to choose the left box on the first trial, you will see this:
- If you are instructed to choose the right box on the second trial, you will see this:
- Once these instructed trials are complete you will have a free choice between the two bandits that is indicated by two green squares inside the two boxes you are choosing between.
- Press space when you are ready to begin. Good luck!

Difference in behavior over blocks

To look for evidence of learning of the strategies over long timescales, we looked at the behavior in the different blocks. The results, plotted in figure S1, show no significant differences between behavior in any of the blocks.

An ANOVA of the information bonus with factors for horizon and block shows neither a main effect of block (F(3,239) = 0.48, p = 0.69), nor an interaction of block with horizon (F(3,239) = 0.41, p > 0.75). Likewise an ANOVA for decision noise with factors for horizon, block and information condition shows no main effect of block (F(3,479) = 0.88, p = 0.45) nor any interaction with the other factors (block x horizon: F(3,479) = 1.66, p = 0.18; block x information condition: F(3,479) = 0.39, p = 0.53).

Thus if subjects do change their exploratory behavior in this task, they do so on a timescale that is either much slower or much faster than the 80 games in a single block.

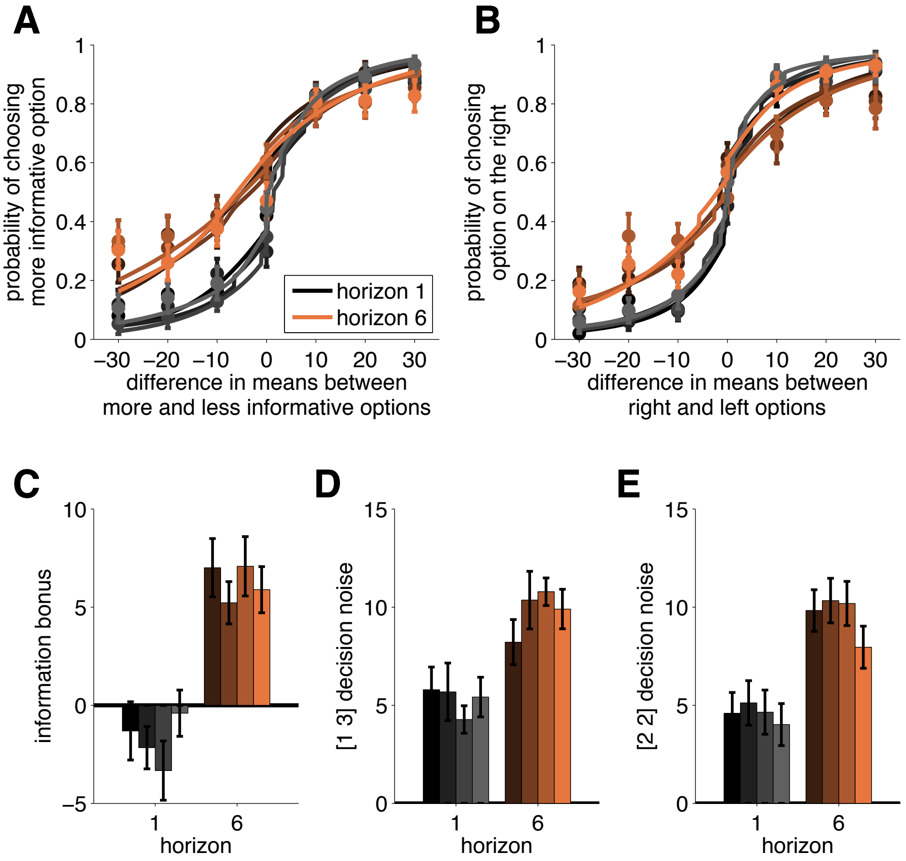

Figure S1. Behavior on the first free-choice trial across blocks. (**A**, **B**) Choice curves for the unequal (**A**) and equal (**B**) information conditions on the first free-choice trial as a function of the difference in mean between the two options. Earlier blocks correspond to darker colors. In all cases, there is no effect of block on behavior. (**C**, **D**, **E**) Mean parameter fits for the information bonus (**C**) and decision noise in the [1 3] condition (**D**) and decision noise in the [2 2] condition showing an increase in information bonus and decision noise between horizons 1 and 6 that is not modulated by block number. Error bars are s.e.m across participants.

Difference in behavior between different center-mean conditions

In every game in our experiment we set the mean of one of the options, the "center mean", to be either 40 or 60 and then determined the mean of the other option by sampling the difference uniformly from [-30, -20, -12, -8, -4, 4, 8, 12, 20, 30]. One possible question is whether the choice of center mean matters? To test this we performed the same analysis – fitting choice curves for each subject with a simple logistic and then averaging parameters – separately for the two center mean conditions.

Results of this analysis are shown in Figure S2. An ANOVA for the information bonus (with factors for horizon and center-mean) reveals a main effect of horizon ($F(1,119) = 24.38$, $p < 10^{-4}$) as well as center-mean ($F(1,119) = 11.14$, $p < 0.005$) but no interaction between them. Post hoc t tests showed a significant decrease in information bonus between the center-mean 40 and center-mean 60 conditions for both horizon 1 ($t(29) = 2.52$, $p < 0.02$) and horizon 6 ($t(29) = 2.58$, $p < 0.02$).

A similar ANOVA for decision noise (with horizon, center-mean and uncertainty condition as factors) showed only a main effect of horizon ($F(1,119) = 75.78$, $p < 10^{-8}$), with marginal support for main effects of center-mean ($F(1,119) = 3.56$, $p = 0.07$) and uncertainty condition ($F(1,119) = 3.32$, $p = 0.08$).

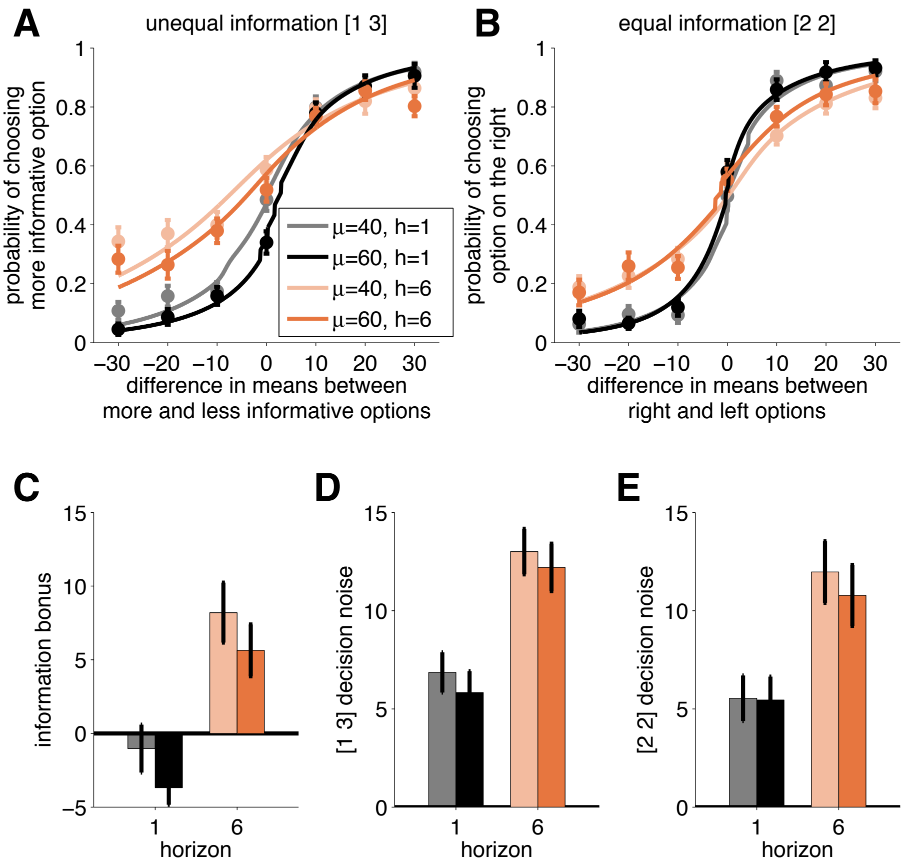

Figure S2. Behavior on the first free-choice trial across center-mean conditions. (**A**, **B**) Choice curves for the unequal (**A**) and equal (**B**) information conditions on the first free-choice trial as a function of the difference in mean between the two options. (**C**, **D**, **E**) Mean parameter fits for the information bonus (**C**) and decision noise in the [1 3] condition (**D**) and decision noise in the [2 2] condition showing an increase in information bonus and decision noise between horizons 1 and 6 that is not modulated by center mean. Error bars are s.e.m across participants.

Behavior on later trials

As mentioned in the main text, the reward-information confound makes it difficult to interpret the behavior on later trials in terms of directed exploration. Despite this qualification, it is still possible to look at later-trial behavior. In Figure S3A-D we plot the choice curves grouped by the difference, Δ, in the number of times each option is played. Here the equal information conditions, [2 2] on free-choice trial 1, [3 3] on trial 3 and [4 4] on trial 5, in panel A correspond to $\Delta = 0$ because each option has been played an equal number of times. $\Delta = 1$ encompasses the [2 3] on trial 2, [3 4] on trial 4 and [4 5] on trial 6 information conditions and so on for $\Delta = 2$ (C) and $\Delta = 3$ (D). All of these plots show an increase in the slope as the game progresses consistent with a decrease in random exploration. Similarly, for the unequal conditions, there appears to be a decrease in the indifference point on later trials consistent with a decrease in directed exploration.

Model fitting confirms these observations with a decrease in information bonus (E) and decision noise (F) across the game in all information conditions. In particular we fit the same logistic function to all trials using the observed mean as the mean for each option and again set the information, I_a equal to ±1/2, to interpret the information bonus as the indifference point of each curve. Separate ANOVAs on the information bonus for $\Delta = 1$, 2 and 3 reveals a significant main effect of trial number (for $\Delta = 1$: $F(2,89) = 7.93$, $p < 0.001$, for $\Delta = 2$: $F(2,89) = 22.42$, $p < 10^{-7}$, for $\Delta = 3$: $F(2,89) = 5.53$, $p < 0.01$). Likewise an ANOVA on the decision noise for $\Delta = 0$, 1, 2 and 3 yields similar results (for $\Delta = 0$: $F(2,89) = 35.38$, $p < 10^{-10}$, for $\Delta = 1$: $F(2,89) = 28.52$, $p < 10^{-8}$, for $\Delta = 2$: $F(2,89) = 84.25$, $p < 10^{-17}$, for $\Delta = 3$: $F(2,89) = 5.53$, $p < 0.01$)

Figure S3E clearly shows that the information bonus decreases rapidly and even becomes negative (consistent with ambiguity aversion) after the first trial. This highlights the difficulty of detecting directed exploration when fitting behavior to all trials and also suggests that directed exploration may decrease over time.

One difficulty with interpreting these results is that the difference in information between the two options is not only a function of Δ, but also depends on the total number of times the options have been played. Thus, the option played once in the [1 3] condition is much more informative than the option played three times in the [3 5] condition. This makes it difficult to disentangle changes due to trial number from changes due to available information without making strong assumptions about the functional form of the information. Thus to measure how directed exploration changes parametrically as a function of both trial number and information will require a different experiment.

The changes in decision noise are easier to interpret because information does not factor into the decision noise. Also, in the equal condition, the reward-information confound is not a factor because there is no information to confound with reward. Thus the simplest interpretation of the change in decision noise is that random exploration is decreasing over the course of the game. Intriguingly, the form of this decrease seems to be the same regardless of information condition suggesting a general purpose mechanism for random exploration, consistent with it being a simpler strategy than directed exploration which changes with information and trial number. Overall this analysis supports our conclusion that decision noise is used and adapted as a strategy for random exploration.

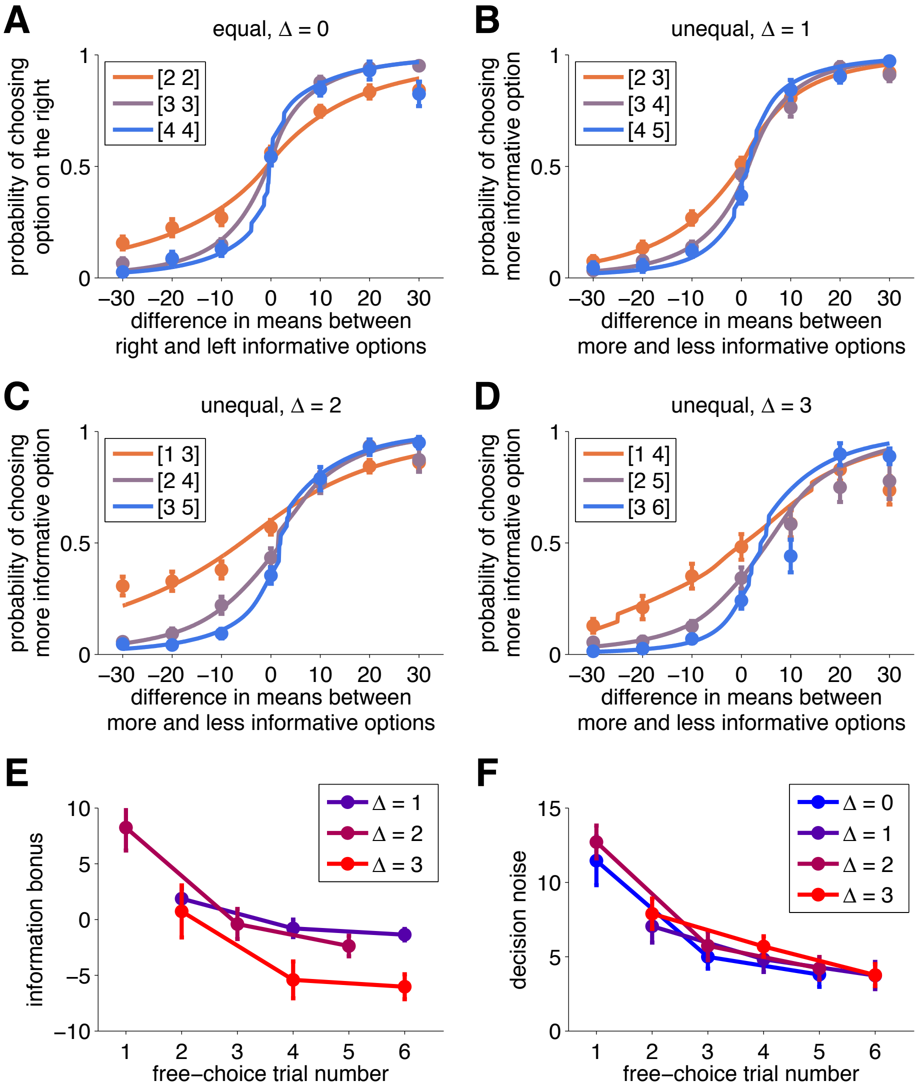

Figure S3: Behavior on later trials. (**A-D**) Choice curves on later trials grouped by the difference in the number of times each option has been played, Δ. (**A**) The equal conditions, Δ = 0, replicates Figure 3A in the main text. (**C-D**) unequal information conditions Δ = 1 (**B**), Δ = 2 (**C**) and Δ = 3 (**D**). (**E-F**) Fit information bonus (**E**) and decision noise (**F**) as a function of free trial number for the different Δs.

Reaction times

One potential explanation for the difference in decision noise between horizon 1 and horizon 6 is that participants change their speed-accuracy tradeoff between horizons. Specifically, that they favor speed over accuracy in horizon 6 – perhaps because they feel they can make up any loss of accuracy on later trials or because they simply wish to get the longer games over with. If this were the case, then we might expect to see evidence for this in their reaction times – with the first free choice in horizon 6 being faster than in horizon 1.

Contrary to this prediction we found no effect of horizon on reaction time on the first free choice trial (Figure S4A and C). In particular a repeated measures ANOVA on the z-scored reactions times showed no main effect of either horizon (F(1,119) = 0.82, p = 0.37) or information condition (F(1,119) = 1.8, p = 0.18). We did, however, find a small difference in z-scored reaction times on the last forced-choice trial. An ANOVA with factors for horizon and information condition showed a main effect of both horizon (F(1,119) = 16.61, p = 10^{-4}) and information condition (F(1,119) = 21.96, p =10^{-5}). While this effect is consistent with a speed-accuracy tradeoff, the overall effect size is small, with the difference in reaction times between horizons on the order of 50 ms versus the approximately 750 ms reaction time for the first free-choice trial.

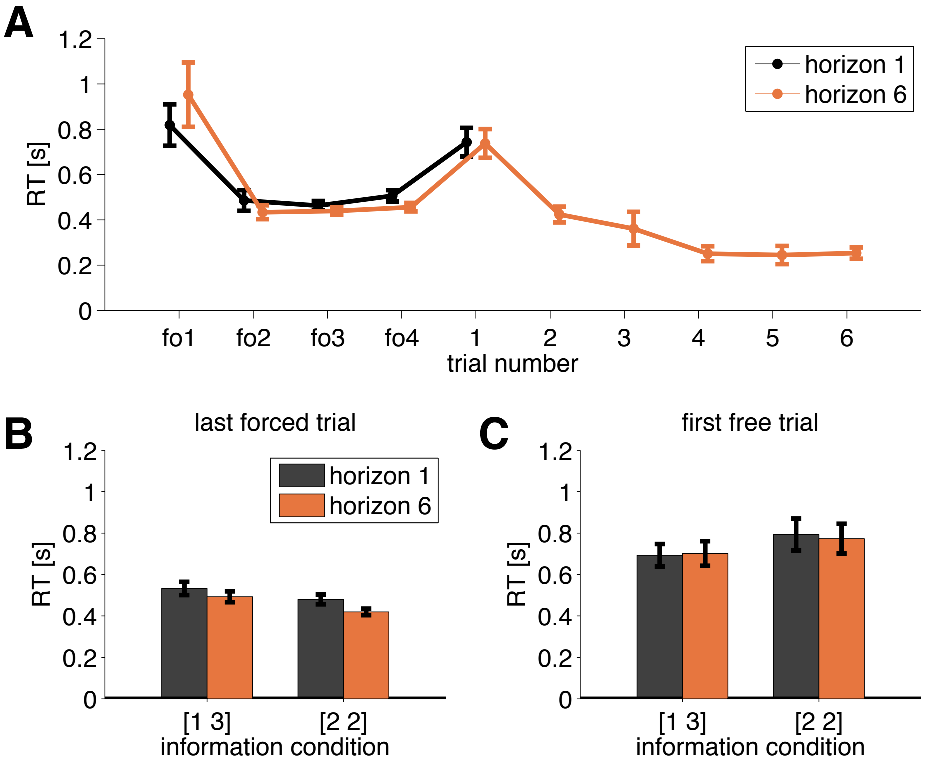

Figure S4. Reaction times. (**A**) Mean reaction times over the course of the entire game (averaged over both information conditions) for horizon 1 and horizon 6 games. (**B**, **C**) Reaction times for each information condition on the last forced-choice trial (**B**) and first free-choice trial (**C**).

Replication

In order to demonstrate the robustness of our findings, here we report the results of a near-replication of the study that was run before the task reported in the paper.

This replication differed in the following ways from the task reported in the main text. Most importantly, we used three horizon conditions instead of two: horizon 1, 6 and 11. Because the horizon 11 games were so long this greatly reduced the total number of games that subjects could play to 150 games in five blocks (versus 320 games in four blocks in the main task). The difference in means between the two options was also different and focused more on the indifference point around 0. Specifically, the true differences in means were 0, 5 and 10 points (as opposed to 4, 8, 12, 20 and 30 in the main task).

The participants were also paid in this version of the experiment. Payment was $12 for the one-hour session plus up to $3 'performance bonus' that scaled linearly with points earned and was rounded up to the nearest dollar. In practice, because of the relatively small spread in total points earned in this experiment, this meant that all participants received the full $3 bonus.

Finally, the visual layout of the task was different, Figure S5A. In particular, the one-armed bandits were only able to display one reward at a time and the history of rewards was instead conveyed with "reward history bars" at the side of the screen. These history bars behaved in much the same way as the bandits in the main task: after a particular option was played, the reward on that trial was added to the appropriate history bar, while the corresponding space for the unplayed option was filled with an 'XX'.

Despite the changes in experiment design basic performance was similar to the main experiment (Figure S5B) and the same reward-information confound developed on the later trials (Figure

S5C). We also found the same changes with horizon in both information bonus (ANOVA main effect of horizon, $F(2,227) = 12.54$, $p < 0.001$) and decision noise (ANOVA main effect of horizon, $F(2,227) = 14.15$, $p < 10^{-5}$) (Figure S6).

One new finding was that there was no difference in either the information bonus or decision noise on the first free-choice trial in the horizon 6 and horizon 11 conditions (difference in information bonus: t(37) = 0.97, p = 0.35, [1 3] decision noise: t(37) = 1.44, p = 0.16, [2 2] decision noise t(37) = 0.70, p = 0.49). Whether this result is due to our inability to resolve small differences in exploration between the longer horizons, or because exploration is categorically modulated by horizon (with low exploration at horizon 1 and high exploration at any other horizon) is difficult to tell.

Our analysis of behavior on the later trials of horizon 6 games (Figure 3 and Figure S3) sheds light on this for random exploration. Specifically, the result that decision noise decreases over the course of these games suggests that random exploration, at least, is parametrically modulated by horizon. It is not possible, however, to draw the same conclusion about directed exploration. Although we do see a similar decrease in information bonus over the course of the games (Figure S3E), interpretation of this is made more difficult because of the reward-information confound. Therefore we conclude that, while random exploration is modulated parametrically by horizon, directed exploration may be modulated categorically.

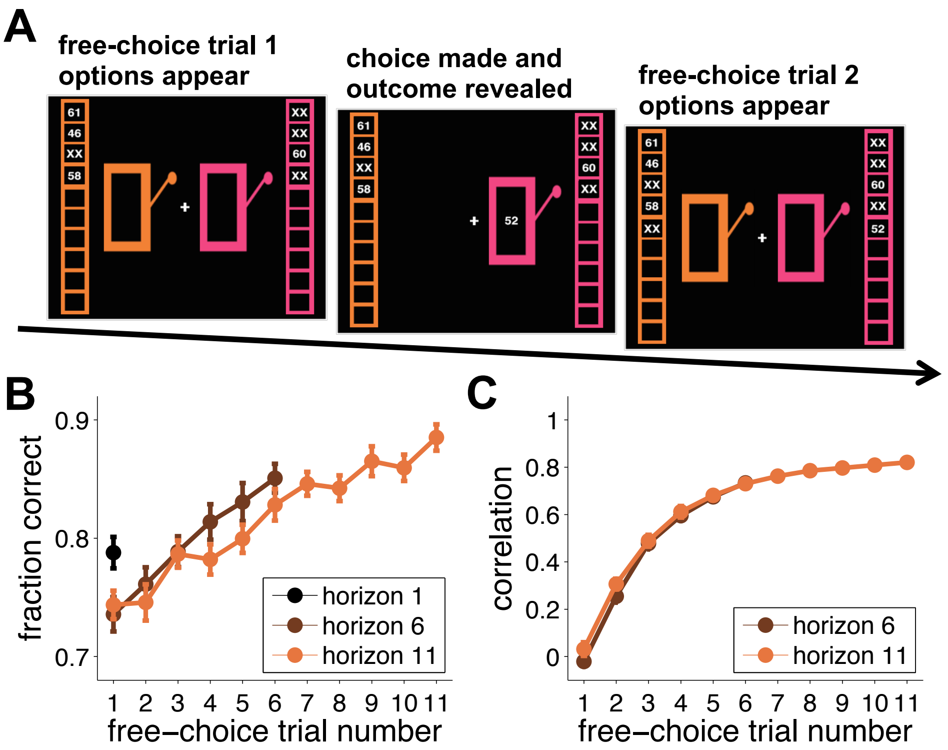

Figure S5. Repeat experiment. (**A**) Example of a free-choice trial showing the different visual layout in this version of the task. (**B**) Basic performance showing an increase in fraction correct as a function of free-choice trial number. (**C**) Correlation between the difference in means between the options and difference in the number of times each option has been played over the horizon 6 and 11 games.

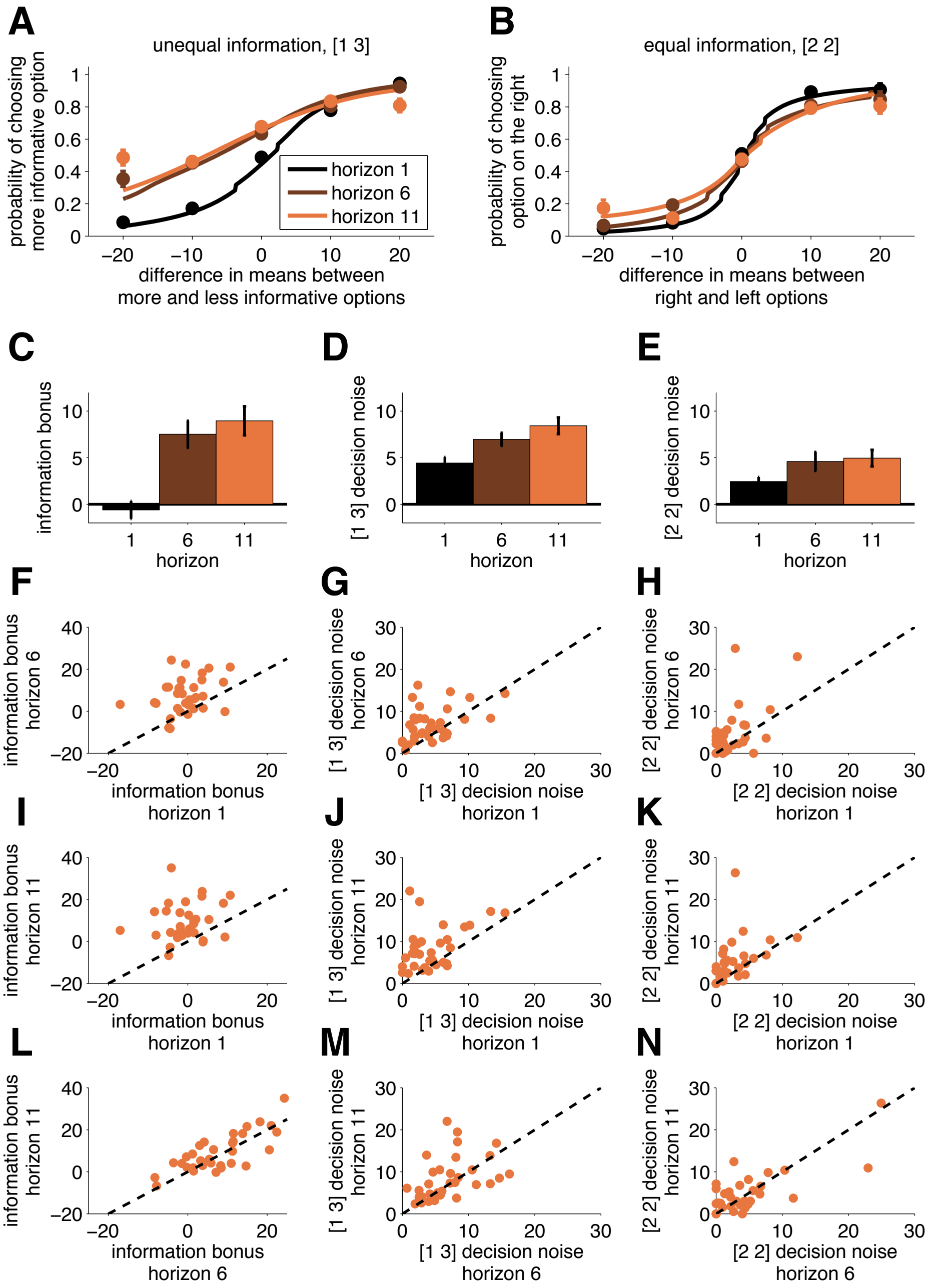

Figure S6. Behavior on the first free-choice trial in the replication experiment. (**A**, **B**) Choice curves for the unequal (**A**) and equal (**B**) information conditions. (**C**, **D**, **E**) Mean parameter fits for the information bonus (**C**) and decision noise in the [1 3] condition (**D**) and decision noise in the [2 2] condition (**E**). (**F-N**) Scatter plots comparing parameter fits for individual subjects in horizons 1 and horizon 6 (**F**-**H**), 1 and 11 (**I**-**K**) and 6 and 11 (**L**-**N**). The dashed lines denote equality.

Model-based analysis of behavior

To test whether our conclusions in the main paper are contingent on our modeling choices, we performed a more detailed model-based analysis. In particular, we considered models with a number of cognitive factors that are known to influence choice and asked whether including these changed our results.

The factors we considered were:

- Unequal weighting of the rewards
- Choice kernel
- Variance bias
- Trend bias
- Non-linear utility function

We incorporated these factors by changing the equation for the value of each option, Q_a, in the original model. As before we fit the models separately for each subject and for each information and horizon condition. Importantly, each of the models was parameterized in such a way that allowed us to extract an information bonus and decision noise from each of the fits. This allowed us to compare the information bonus and decision noise between models to assess whether adding a particular factor removed the effect of horizon.

Models

Unequal weighting of the rewards

Our original model makes decisions based on the means of the observed samples for each option, thus it implicitly assumes that each of the past outcomes has equal influence on the decision. However, it is well known that humans do not always treat all of the past equally and there is good evidence that recent and maximal outcomes are overweighted in decision making (Kahneman et al., 1993; Redelmeier & Kahneman, 1996; Do et al., 2008; Ludvig et al., 2014; Blanchard et al., 2014).

To take into account these possibilities we built three models. The first, termed the "reward order" model, allows arbitrary weighting of rewards based on their order, thus encompassing primacy and recency effects as special cases. The second and third models give special weight to maximal outcomes in slightly different ways. The "peak bias" model, finds one maximum reward, the global maximum over the four forced play trials, while the "local peak bias" model finds two maximum rewards, one for each option.

More concretely, the equations for the models are as follows:

Reward order

This model assumes that participants compute the value for each option as:

$$Q_a = \frac{\sum_{i=1}^{4} \omega_i r_i x_{ia}}{\sum_{i=1}^{4} x_{ia}} + \alpha I_a + B s_a$$

where r_i is the reward outcome on forced play *i*, ω_i is the weight given to that outcome and x_{ia} indicates whether option *a* has been played ($x_{ia} = 1$) on trial *i* or not ($x_{ia} = 0$). As before, I_a is the information, α is the information bonus, s_a is the side, and B is the spatial bias. To allow for comparison with our original model we set the weight of the first outcome to 1 ($\omega_i = 1$).

Peak bias

This model computes values as

$$Q_a = R_a + \omega_{max} r_{max} m_a + \alpha I_a + B s_a$$

where R_a is the mean of option, *a*, ω_{max} is the weight given to the maximum reward, r_{max}, and m_a is and indicator variable that is 1 if option *a* is the option with the maximum value and 0 otherwise.

Local peak bias

This model computes values as

$$Q_a = R_a + \omega_{max} r_{max}^{a} + \alpha I_a + B s_a$$

where r_{max}^{a} is the maximum reward seen on option *a*.

Choice kernel

An important suboptimality in human decision making is the tendency for past choices themselves, regardless of outcome, to influence future behavior (Lau & Glimcher 2005; Ito & Doya 2009; Akaishi et al. 2014). The best known of these biases is perseveration in which past choices tend to be repeated, but other possibilities such as alternation are known to occur, especially in animals. To model this effect we included a term for the influence of the past choices as follows

$$Q_a = R_a + \sum_{i=1}^{4} \kappa_i c_i + \alpha I_a + B s_a$$

where c_i is the choice on trial i ($c_i = +1$ for right and $c_i = -1$ for left) and κ_i is the choice kernel that determines the weight given to past choices. As with the reward order model, to allow comparison with the original model, we set the weight on the first choice as zero $\kappa_1 = 0$.

Variance bias

To test whether the variance of the outcomes on the forced plays affected choice, we included a term for the empirical standard deviation, $\hat{\sigma}_a$,

$$Q_a = R_a + V_a \hat{\sigma}_a + \alpha I_a + Bs_a$$

Trend bias

Trend bias describes the tendency for people to bias their choices towards options whose outcomes appear to be increasing over time (Loewenstein & Prelec, 1993). To test for this in our participants we included a term related to the linear trend, τ_a, of rewards for each option.

$$Q_a = R_a + D_a \tau_a + \alpha I_a + Bs_a$$

Non-linear utility function

Finally, we include the possibility that the utility of a reward is a non-linear function of its value. In particular we assume the following form for the utility

$$u(r) = r^{\eta}$$

and base decisions on the mean utility, U_a, by computing the value as

$$Q_a = U_a + \alpha I_a + Bs_a$$

Results

The design of our models allows us to extract an information bonus and decision noise from each of them that can be compared across models. Using this we can then test whether our main finding, that information bonus and decision noise increase between horizons 1 and 6, still holds when we include the additional factors in the model.

These results are shown in Figure S7. We plot stacked histograms of the change in information bonus and decision noise across horizon conditions. These histograms are all significantly shifted to the right of zero indicating that the increase in information bonus and decision noise is present in all of the models.

Furthermore, model comparison using the Bayes Information Criterion (BIC), showed that all participants were best fit using the original model.

Taken together, these results suggest that our original conclusions are valid and that neither the change in information bonus nor decision noise is due to any of these cognitive factors.

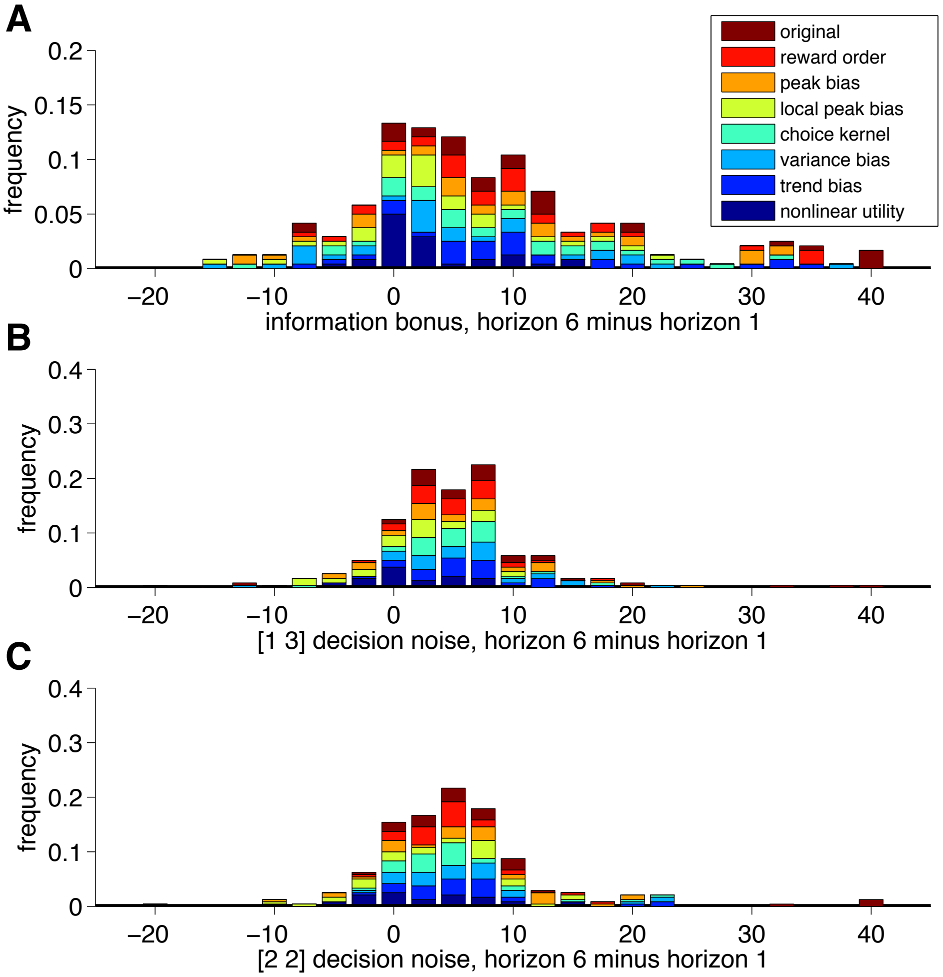

Figure S7. Histograms of the difference in information bonus (**A**), [1 3] decision noise (**B**) and [2 2] decision noise (**C**) for each participant as computed by the different models.

Model of optimal behavior

The optimal model solves a dynamic programming problem (Bellman, 1957; Duff, 2002) to compute the action that will maximize the expected total reward over the course of each game.

To do this the model first infers a distribution over the mean of each option given the observed rewards. We write r_t to denote the reward on trial *t* in the game, c_t to be the choice on trial *t* and D_t to be the set of choices and rewards up to and including time *t*. We assume that the model knows that the rewards are generated from a truncated Gaussian distribution and we further assume that it knows that the standard deviation of this distribution, σ_n.

In this case, the inferred distribution over the mean of option *a*, μ^a, given the history of choices and rewards is

$$p(\mu^a \mid D_t) \propto \sqrt{\frac{n_t^a}{2\pi}} \frac{1}{\sigma_n} \exp\left(-\frac{n_t^a(\mu^a - R_t^a / n_t^a)^2}{2\sigma_n^2} \right) p(\mu^a) \qquad (1)$$

where n_t^a is the number of times option *a* has been played, R_t^a is the cumulative sum of the rewards obtained from playing option *a* and $p(\mu^a)$ is the prior of the mean. In our model we assumed an improper, uniform prior on μ^a (although we should note that it is straightforward to include a Gaussian prior instead). With this prior, equation (1) shows that the model's state of knowledge about option *a* is summarized by the two numbers, n_t^a and R_t^a. We can thus define the *hyperstate* (Duff, 2002), S_t, the state of information that the model has about both options as

$$S_t = (n_t^A, R_t^A, n_t^B, R_t^B) \ . \qquad (2)$$

With the hyperstates defined in this way we can now specify a Markov decision process within this state space. In particular we can define a transition matrix, $T(S_{t+1} \mid S_t, a)$, which describes the probability of transitioning between states S_{t+1} and S_t given action a. To compute this we note that if action $a = A$ is chosen on trial t and reward r_t is observed, then new state on the next trial will be

$$S_{t+1} = (n_t^A + 1, R_t^A + r_t, n_t^B, R_t^B) \tag{3}$$

Further, given the distribution over the mean, using equation (1) we can predict that this outcome will occur with probability

$$\begin{aligned} p(r_t \mid S_t, A) &= \int d\mu^A p(r_t \mid \mu^A) p(\mu^A \mid S_t) \\ &= \sqrt{\frac{n_t^A}{2\pi(1+n_t^A)}} \frac{1}{\sigma_n} \exp\left(-\frac{(r_t - R_t^A / n_t^A)^2}{2\sigma_n^2} \right) \end{aligned} \tag{4}$$

Note that this result comes because both $p(r_t \mid \mu^a)$ and $p(\mu^a \mid D_t)$ are Gaussians, with $p(\mu^a \mid D_t)$ defined in equation (1) and

$$p(r_t \mid \mu^a) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(r_t - \mu^a)^2}{2\sigma_n^2} \right) \tag{5}$$

In practice, to make the algorithm tractable we only consider a subset of possible outcomes, focusing on a set of 51 possible outcomes between 0 and 100 for the horizon 6 case and 21 possible outcomes in the horizon 11 case. Given this approximation we can then compute the set of possible states encountered during the task and solve the dynamic program by iterating the equations for the state values

$$V(S_t) = \max_a Q(a, S_t) \tag{6}$$

and the action values

$$Q(a, S_t) = \sum_{S'_{t+1}} T(S_{t+1} \mid S_t, a)(r_t(S_{t+1}) + V(S_{t+1})) \tag{7}$$

In particular we start at the last trial, $t = H$, and work backwards in time to the first trial. Here, by definition the action value is just the expected value of the reward from each option; i.e.,

$$Q(a_H, S_H) = \frac{R_H^{a_H}}{n_H^{a_H}} \tag{8}$$

Finally the optimal action is to choose the option for which has the highest value on the first free trial, i.e.

$$c_1 = \underset{a}{\text{argmax}}\ Q(a, S_1) \tag{9}$$

This analysis allows us to compute the optimal behavior on the task. To compute the optimal ambiguity bonus in Figures 2D, we simulated choices from this optimal model on the same set of problems faced by the participants. We then fit this simulated data using the same logistic regression model used to fit participants behavior. Finally, we note that, because this algorithm is deterministic, the optimal decision noise is zero for all horizon conditions.

References

Akaishi, R., Umeda, K., Nagase, A., & Sakai, K. (2014). Autonomous Mechanism of Internal Choice Estimate Underlies Decision Inertia. *Neuron 81*, 195–206

Bellman, R. E. (1957). *Dynamic programming.* Princeton, NJ: Princeton University Press.

Blanchard, T. C., Wolfe, L. S., Vlaev, I., Winston, J. S., & Hayden, B. Y. (2014). Biases in preferences for sequences of outcomes in monkeys. *Cognition 130*, 289–299

Do, A. M., Rupert, A. V., & Wolford, A. G. (2008). Evaluations of pleasurable experiences: the peak– end rule. *Psychonomic Bulletin & Review*, *15*(1), 96-98

Duff, M. (2002). *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*, Ph.D. thesis, University of Massachusetts Amherst.

Ito, M. & Doya, K. (2009). Validation of Decision-Making Models and Analysis of Decision Variables in the Rat Basal Ganglia. *The Journal of Neuroscience, 29*(31):9861–9874

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*(6), 401–405.

Lau, B. & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior, 84*, 555-579

Loewenstein, G. F., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review, 100*(1), 91–108.

Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making, 27*, 146-156.

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*(1), 3–8.