Cognition 173 (2018) 34–42

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Deconstructing the human algorithms for exploration

Samuel J. Gershman*

Department of Psychology and Center for Brain Science, Harvard University, United States

ARTICLE INFO

Keywords:
Explore-exploit dilemma
Reinforcement learning
Bayesian inference

ABSTRACT

The dilemma between information gathering (exploration) and reward seeking (exploitation) is a fundamental problem for reinforcement learning agents. How humans resolve this dilemma is still an open question, because experiments have provided equivocal evidence about the underlying algorithms used by humans. We show that two families of algorithms can be distinguished in terms of how uncertainty affects exploration. Algorithms based on uncertainty bonuses predict a change in response bias as a function of uncertainty, whereas algorithms based on sampling predict a change in response slope. Two experiments provide evidence for both bias and slope changes, and computational modeling confirms that a hybrid model is the best quantitative account of the data.

1. Introduction

When rewards are uncertain, a reinforcement learning agent faces the *explore-exploit dilemma*: should she exploit the option with the highest expected reward, possibly foregoing even higher rewards from other options, or should she explore other options to gather more accurate information about their values? How humans resolve this dilemma has long puzzled psychologists and neuroscientists (Cohen, McClure, & Yu, 2007; Mehlhorn et al., 2015). Because the optimal solution is computationally intractable, humans must employ approximations or heuristics. A large menu of algorithmic possibilities has been developed in the machine learning literature, and some of these have been studied experimentally. However, these algorithms can be difficult to disentangle empirically because they seem to make similar predictions. The key contribution of this work is to show how different algorithms can in fact make quite different predictions when viewed through the appropriate analytical lens, providing new insights into how humans resolve the explore-exploit dilemma.

Research on exploration has coalesced around two big ideas. The first is that humans engage in "directed" exploration, seeking out options that are highly informative about the underlying reward distribution. This is commonly implemented by adding an "uncertainty" or "information" bonus to the estimates of expected reward (Auer, Cesa-Bianchi, & Fischer, 2002). This scheme has the virtue that exploration will decrease with uncertainty, so that eventually choices will be purely exploitative once enough information has been gathered. A number of studies have found evidence for uncertainty bonuses in human decision making (Frank, Doll, Oas-Terpstra, & Moreno, 2009; Krueger, Wilson, & Cohen, 2017; Lee, Zhang, Munro, & Steyvers, 2011; Meyer & Shi, 1995; Wilson, Geana, White, Ludvig, & Cohen, 2014; Zhang & Yu, 2013), but some have not (Daw, O'doherty, Dayan, Seymour, & Dolan, 2006; Payzan-LeNestour & Bossaerts, 2011). Wilson et al. (2014) suggested one reason why evidence for uncertainty bonuses has been equivocal: uncertainty and reward are confounded, because people tend to choose more rewarding options and hence have less uncertainty about those options. Wilson et al. (2014) demonstrated that decisive evidence for uncertainty bonuses can be obtained when this confound is controlled.

The second big idea is that humans engage in "random" exploration, produced by injecting stochasticity into their choices (Daw et al., 2006). The most widely adopted techniques use a fixed source of stochasticity (see next section), but some evidence suggests that humans use a more sophisticated form of stochasticity, which adapts to their uncertainty level (Schulz, Konstantinidis, & Speekenbrink, 2015; Speekenbrink & Konstantinidis, 2015). This is in fact one of the oldest exploration strategies in reinforcement learning, dating back to the pioneering work of Thompson (1933). However, relatively few studies have attempted to tease apart the effects of uncertainty on directed and random exploration.

The purpose of this paper is to characterize some qualitative properties of particular directed and random exploration strategies, which make them empirically distinguishable. We first provide a formal description of these strategies, and then present the results of two experiments that suggest the use of both strategies. A hybrid of directed and random exploration strategies is anticipated by recent reinforcement learning algorithms (Chapelle & Li, 2011; May, Korda, Lee, & Leslie, 2012), which have been shown to have attractive empirical and theoretical properties.

* Address: Department of Psychology, Harvard University, 52 Oxford St., Room 295.05, Cambridge, MA 02138, United States.
E-mail address: gershman@fas.harvard.edu.

https://doi.org/10.1016/j.cognition.2017.12.014
Received 17 August 2017; Received in revised form 20 December 2017; Accepted 21 December 2017
Available online 29 December 2017

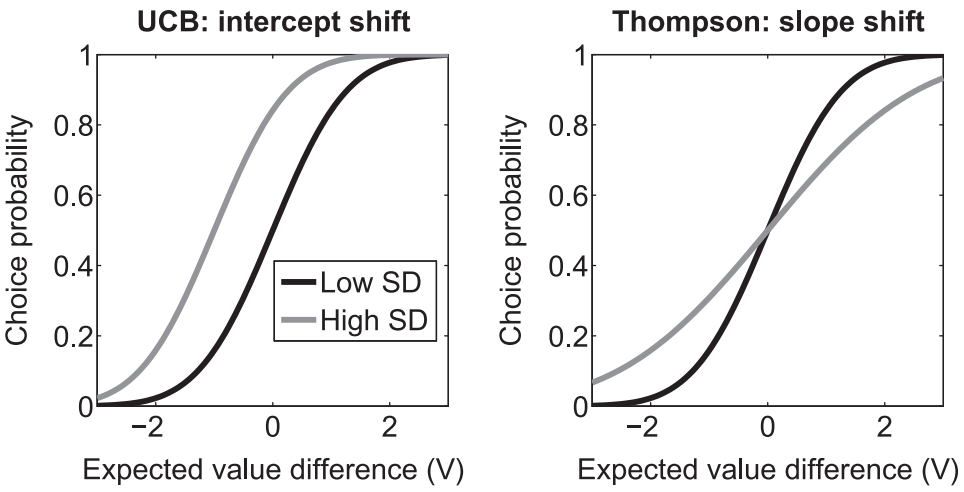

Fig. 1. Effects of uncertainty on choice probability for upper confidence bound (Left) and Thompson sampling (Right) algorithms. Each curve shows the probability of choice as a function of the difference in estimated value between the two arms, plotted separately for low/high posterior standard deviation (SD).

2. Algorithms for exploration

We focus on the two-armed bandit, in which an agent on trial t selects an action $a_t \in \{1,2\}$ and observes a reward r_t. Most models of human exploration have assumed some form of fixed random exploration (e.g., Daw et al., 2006), typically choosing action k on trial t with probability given by a softmax distribution:

$$P(a_t = k) = \frac{\exp[\beta Q_t(k)]}{\sum_{k'} \exp[\beta Q_t(k)]}, \tag{1}$$

where $Q_t(k)$ is an estimate of the expected reward (value) $\mathbb{E}[r_t|a_t = k]$ and β is an inverse temperature parameter that controls the stochasticity of action selection: lower values of β produce more stochasticity. Other forms of random exploration have also been considered: for example, Gonzalez and colleagues (Gonzalez & Dutt, 2011; Lejarraga, Dutt, & Gonzalez, 2012) add noise via an instance-based retrieval mechanism, and Barron and Erev (2003) use an ϵ-greedy strategy with a time-dependent exploration parameter. One problem with all of these strategies is that they do not take the agent's uncertainty into account. If an agent samples an action 10 times, she should have more uncertainty about its value than if she samples it 100 times, but the softmax policy will produce the same action probabilities as long as the value estimates are the same. Intuitively, the agent should be more exploratory under high uncertainty, as indeed the optimal exploration policy dictates (Wilson et al., 2014).

An alternative strategy is to adopt the principle of *optimism in the face of uncertainty*: prefer arms with greater uncertainty. The most famous operationalization of this principle is the *Upper Confidence Bound* (UCB) algorithm (Auer et al., 2002), which selects actions deterministically according to:

$$a_t = \underset{k}{\text{argmax}}\ Q_t(k) + U_t(k), \tag{2}$$

where $U_t(k)$ is the upper confidence bound that plays the role of an uncertainty bonus. The classic version of UCB (known as UCB1) uses

$$U_t(k) = \sqrt{\frac{2\log t}{N_t(k)}}, \tag{3}$$

where $N_t(k)$ is the number of times action k was chosen. While UCB1 is based on a frequentist confidence interval, Bayesian variants have been developed in which $Q_t(k)$ corresponds to the posterior mean and the uncertainty bonus is proportional to the posterior standard deviation $\sigma_t(k)$ (Srinivas, Krause, Seeger, & Kakade, 2010). Details of this computation can be found in Appendix A.

UCB can be understood as a form of directed exploration without any random component, whereas softmax is a form of random exploration without any directed component. A more sophisticated form of random exploration is *Thompson sampling* (Thompson, 1933), which draws random values from the posterior and then chooses greedily with respect to these random values. The key property of Thompson sampling that distinguishes it from softmax exploration is the uncertainty-based determination of choice stochasticity: the agent will explore more when she is more uncertain.

Viewed as hypotheses about human behavior, both UCB and Thompson sampling predict that exploration will increase with uncertainty, but they make qualitatively different predictions about the nature of this increase. This can be seen clearly by examining choice probability as a function of the estimated difference in value between the two arms (Fig. 1), where for simplicity we have assumed that arm 1 has an unknown mean and arm 2 has a known mean of 0. If we add a fixed random component to UCB (see Appendix A), then both algorithms produce sigmoidal choice probability curves. When uncertainty increases, UCB predicts a response bias (intercept) shift. This arises because the uncertainty combines additively with the value estimate, reducing the size of the value difference necessary to be indifferent between the two arms. Thompson sampling, by contrast, predicts a slope shift, because the (inverse) uncertainty combines multiplicatively with estimated value difference. Our experiments, described in the next section, capitalize on these qualitative differences to deconstruct the algorithms underlying human exploration.

3. Methods

Because the two experiments use the same methods, differing only in their reward distributions, we describe them here together. On each trial, participants chose between two arms and received reward feedback, drawn from an arm-specific Gaussian distribution, which did not change within a block (similar procedures have been used in the "decisions from experience" literature; e.g., Barron & Erev, 2003; Gonzalez & Dutt, 2011; Lejarraga et al., 2012). In Experiment 1, one arm yielded stochastic rewards and the other arm yielded a fixed reward of 0. In Experiment 2, both arms yielded stochastic rewards. We used probit regression to quantify bias and slope shifts, thereby allowing us to test the predictions of UCB and Thompson sampling algorithms.

3.1. Participants

Forty-four participants in Experiment 1 and forty-five participants in Experiment 2 were recruited via the Amazon Mechanical Turk web service and paid $1.50. The experiments were approved by the Harvard Institutional Review Board.

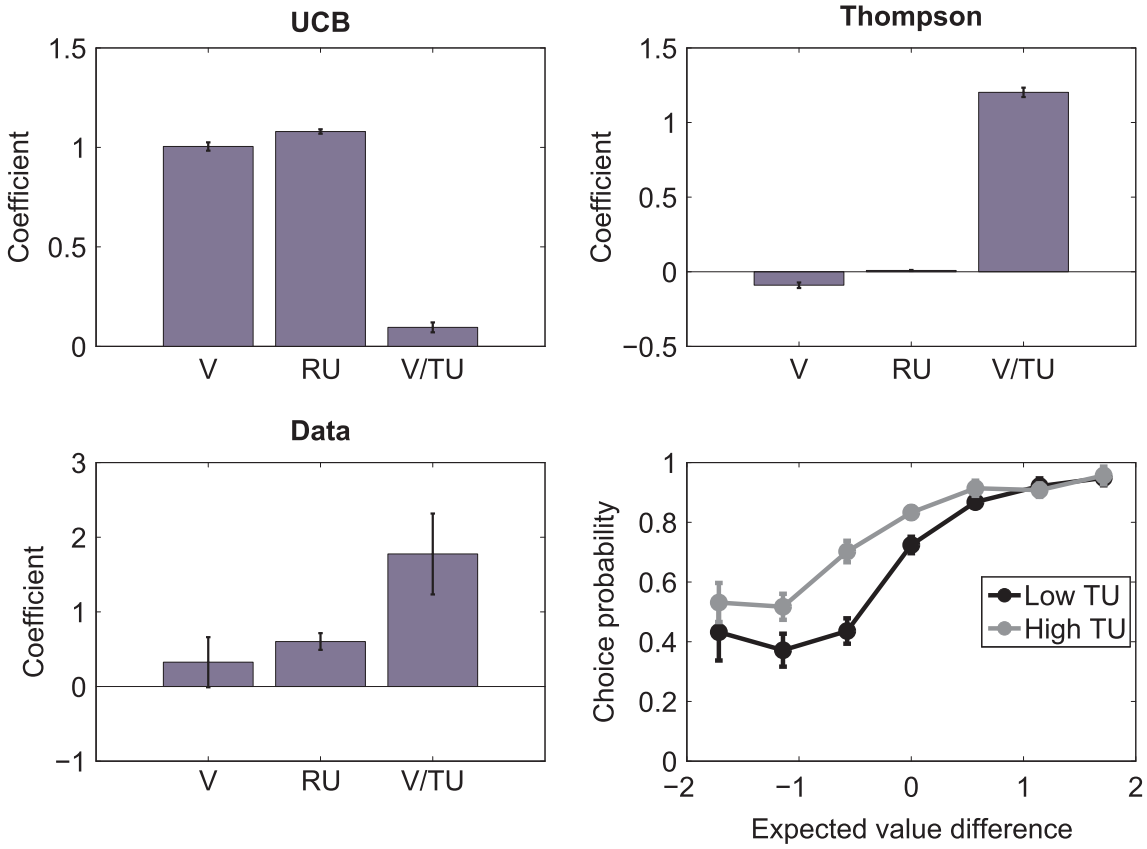

Fig. 2. Experiment 1: regression results. Choice probability was modeled using probit regression with 3 regressors: value difference (V), relative uncertainty (RU), and value difference normalized by total uncertainty (V/TU). The first 3 panels show the regression coefficients plotted separately for simulated data (UCB and Thompson sampling, top panels) and the empirical data from Experiment 1 (bottom left). The bottom right panel shows the empirical choice probability functions separately for low and high TU, based on a median split. Error bars represent standard error of the mean.

3.2. Stimuli and procedure

Participants played 20 two-armed bandits in blocks of 10 trials. On each trial, participants chose one of the arms and received reward feedback (points). They were instructed to choose the arm that maximizes their total points. In Experiment 1, the mean reward $\mu(1)$ for arm 1 on each block was drawn from a Gaussian with mean 0 and variance $\tau_0^2(1) = 10$ (thus each block had a different mean reward for arm 1), and the reward for arm 2 was fixed at 0. When participants chose arm 1, they received stochastic rewards drawn from a Gaussian with mean $\mu(1)$ and variance $\tau^2(1) = 10$. When they chose arm 2, they always received a reward of 0. The structure of Experiment 2 was identical, except that mean rewards for both arms were drawn from the same distribution, with mean 0 and variance $\tau_0^2(1) = \tau_0^2(2) = 100$, and likewise the reward feedback on each trial was drawn from a Gaussian with mean $\mu(k)$ when arm k was selected, with variance $\tau_0^2(1) = \tau_0^2(2) = 10$.

The exact instructions for participants in Experiment 1 were as follows:

> *In this task, you have a choice between two slot machines, represented by colored buttons. When you choose the left (variable) machine, you will win or lose points. The left machine will not always give you the same points, but it will tend to give points around its average value. When you choose the right (fixed) machine, you will always get 0 points. Your goal is to choose the slot machine that will give you the most points. After making your choice, you will receive feedback about the outcome. You will play 20 games, each with a different left (variable) slot machine (the right, fixed machine will always stay the same). Each game will consist of 10 trials.*

The exact instructions for participants in Experiment 2 were as follows:

> *In this task, you have a choice between two slot machines, represented by colored buttons. When you click one of the buttons, you will win or lose points. Choosing the same slot machine will not always give you the same points, but one slot machine is always better than the other. Your goal is to choose the slot machine that will give you the most points. After making your choice, you will receive feedback about the outcome. You will play 20 games, each with a different pair of slot machines. Each game will consist of 10 trials.*

3.3. Analysis

To estimate the bias and slope of the choice probability functions, we used probit regression. As shown in Appendix A, UCB and Thompson sampling can (under appropriate assumptions) be exactly formalized as probit regression models. In particular, we entered the following regressors into the probit model:

- Estimated value difference (V): $Q_t(1) - Q_t(2)$.
- Relative uncertainty (RU): $\sigma_t(1) - \sigma_t(2)$.
- Total uncertainty (TU): $\sqrt{\sigma_t^2(1) + \sigma_t^2(2)}$.

The reason total uncertainty is defined as the square root of the summed variances rather than the sum of the standard deviations (which would make it symmetric with RU) is that this definition of TU allows us to directly relate it to Thompson sampling. In particular, if V is the estimated value difference between the two arms, then choice probability is a sigmoidal function of V/TU (see Appendix A for details). Thus, Thompson sampling predicts a significant positive effect of V/TU on choice probability, but not of RU or V (Fig. 2, top right). According

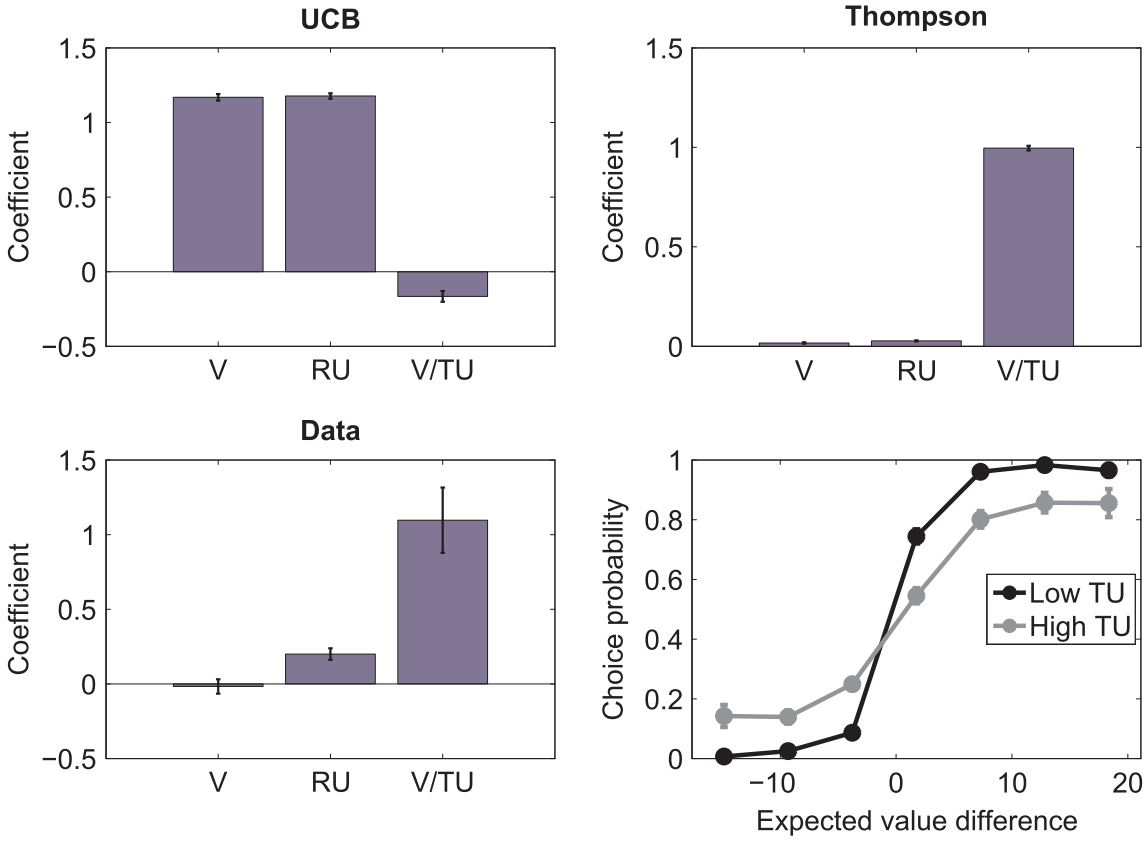

Fig. 3. Experiment 2: regression results. Choice probability was modeled using probit regression with 3 regressors: value difference (V), relative uncertainty (RU), and value difference normalized by total uncertainty (V/TU). The first 3 panels show the regression coefficients plotted separately for simulated data (UCB and Thompson sampling, top panels) and the empirical data from Experiment 2 (bottom left). The bottom right panel shows the empirical choice probability functions separately for low and high TU, based on a median split. Error bars represent standard error of the mean.

to UCB, choice probability is a sigmoidal function of V + RU. Thus, UCB predicts a significant positive effect of both V and RU, but not of V/TU (Fig. 2, top left).

For each model, we used maximum likelihood estimation to fit the coefficients ($\mathbf{w}$) in the probit regression model. The choice probability on trial t was modeled as:

$$P(a_t = 1|\mathbf{w}) = \Phi(w_1\,\mathrm{V} + w_2\mathrm{RU} + w_3\mathrm{V/TU}), \tag{4}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian distribution (mean 0 and variance 1). We compared this full regression model (which we refer to as the "hybrid" model) to two nested models: UCB (w_3 fixed to 0) and Thompson sampling (w_2 fixed to 0). For each regression model, we computed the Bayesian information criterion approximation of the log marginal likelihood (model evidence; Bishop, 2006):

$$\begin{aligned}\log P(\mathbf{a}|\mathbf{w}) &= \log \int_{\mathbf{w}} P(\mathbf{a}|\mathbf{w})P(\mathbf{w})d\mathbf{w} \\ &\approx \log P(\mathbf{a}|\mathbf{w}^*) - \tfrac{D}{2}\log T,\end{aligned} \tag{5}$$

where $\mathbf{a}$ denotes the set of all actions, D is the number of parameters, T is the number of trials, and $\mathbf{w}^* = \mathrm{argmax}_{\mathbf{w}}\, P(\mathbf{a}|\mathbf{w})$ is the maximum likelihood estimate of the coefficients. The model evidence was then entered into a random-effects model selection procedure that estimated the frequency of each model in the population (Rigoux, Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009). This procedure assumes that each participant's behavior may have been generated by a different model, drawn from some unknown population distribution. We report the protected exceedance probability (PXP) for each model, defined as the posterior probability that the model has a higher frequency than all the other models under consideration (accounting for the possibility that the differences between models arose from chance).

3.4. Supplementary material

Code and data for reproducing all analyses reported in this paper are available at https://github.com/sjgershm/exploration.

4. Results

4.1. Modeling choice probability

The probit regression analysis of the Experiment 1 data (Fig. 2, bottom) revealed effects of both RU [$t(44) = 5.41, p < .001$] and V/TU [$t(44) = 3.28, p < .005$], but not of V ($p = .34$). These results demonstrate that choices were not consistent with either a pure UCB or a pure Thompson sampling algorithm, but instead exhibited features of both. Specifically, the RU effect is consistent with the uncertainty bonus predicted by UCB, while the TU effect is consistent with the uncertainty-dependent stochasticity predicted by Thompson sampling. Furthermore, the absence of an effect of V is consistent with Thompson sampling, according to which the effect of value on choice is mediated by TU, such that V by itself does not explain any additional variance.

The probit regression analysis of the Experiment 2 data (Fig. 3, bottom) was largely consistent with the results of Experiment 1, revealing effects of both RU [$t(43) = 5.16, p < .001$] and TU [$t(43) = 5.02, p < .001$], but not of V ($p = .72$). Qualitatively, Experiment 2 appeared to produce a more pronounced slope shift compared to Experiment 1 (compare Figs. 2 and 3, bottom right).

In addition to measuring the qualitative effects on bias and slope, we used Bayesian random-effects model comparison (Rigoux et al.,

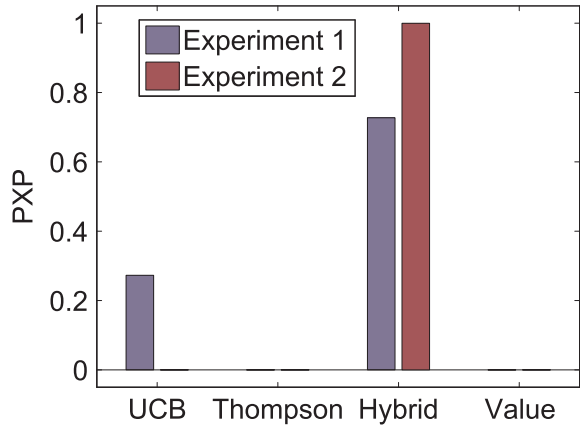

Fig. 4. Bayesian model comparison. Protected exceedance probability (PXP) for each model, estimated separately for Experiments 1 and 2. "Value" refers to the value-directed exploration model, in which choice probability is a function of the difference in value between the two arms.

2014; Stephan et al., 2009) to compare UCB and Thompson sampling quantitatively with a hybrid model which included both directed and random exploration components (see Methods for a summary of the model comparison procedure, and Appendix A for details of the Hybrid model). In addition, we included a "value-directed exploration" model in which the choice probability was only a function of the value difference (i.e., no dependence on uncertainty). For both studies, the PXPs favored the hybrid model (Fig. 4), indicating that both directed and random exploration are needed to adequately describe choice behavior in these tasks.

4.2. Deconfounding uncertainty and reward

Because participants tend to select rewarding arms, they will have less uncertainty about these arms, thus creating an uncertainty-reward confound (Wilson et al., 2014). To address this, we measured the probability of choosing arm 1 on the first trial of each block, when expected reward is equated across the two arms (i.e., the posterior mean for both arms is 0 on the first trial, because we assume a zero-mean prior). In Experiment 1, uncertainty is higher for arm 1 (which produces variable rewards) compared to arm 2 (which produces fixed rewards), but their expected reward (0) is equated at the beginning of each block. Thus, a preference for arm 1 on the first trial of each block is consistent with an uncertainty bonus, as in UCB. Consistent with this hypothesis, we found that participants preferred arm 1 on the first trial [$t(44) = 9.68, p < .001$; Fig. 5]. In Experiment 2, no such differential uncertainty exists, because both arms are equally variable. Accordingly, we found no preference for arm 1 in Experiment 2 ($p = .95$; Fig. 5).

One potential concern is that participants may have adopted a heuristic strategy in Experiment 1, whereby they preferred the stochastic arm on the first trial because there was no way they could "maximize rewards" by choosing the fixed arm, which always delivered 0 rewards. If participants are using this heuristic, then their exploratory tendency on the first trial should be uncorrelated with their uncertainty bonus fit to all the other trials. In fact, there is a significant correlation in Experiment 1 ($r = 0.6, p < .0001$; Fig. 5), demonstrating that participants who show a larger uncertainty bonus also show a greater preference for the risky option on the first trial. Moreover, participants showed a declining preference for the risky choice over the course of each block (Fig. 5), consistent with the UCB strategy. Note also that this is not an averaging artifact: the median probability of choosing arm 1 (across all trials) is 0.72, indicating that participants did not adopt a heuristic of always choosing one arm or the other.

4.3. Modeling choice response times

Response times offer an additional source of data to disambiguate directed and random exploration. Models of value-based decision making, drawing an analogy to perceptual decision making, have asserted that responses are generated when an evidence accumulation process crosses a decision threshold (Milosavljevic, Malmaud, Huth, Koch, & Rangel, 2010; Ratcliff & Frank, 2012; Tajima, Drugowitsch, & Pouget, 2016; Summerfield & Tsetsos, 2012). The "evidence" in these models corresponds to noisy samples of the value difference, weighted by their reliability. More generally, many studies have found that response time increases with uncertainty (e.g., Gershman, Tenenbaum, & Jäkel, 2016; Grossman, 1953; Hick, 1952; Hyman, 1953). Thus, we might reasonably predict that response times should be *slower* when TU is high. In contrast, an uncertainty bonus should alter the evidence itself, such that higher RU should produce *faster* response times.

We tested these predictions by modeling log response times using linear regression with V, RU and TU as regressors. We found that there was a significant effect of TU in both experiments [Experiment 1: $t(44) = 12.64, p < .0001$; Experiment 2: $t(43) = 15.52, p < .0001$], but only a significant effect of RU in Experiment 2 [$t(43) = 2.39, p < .05$; Experiment 1: $p = .25$]. Nonetheless, variation in the choice coefficient for RU across participants predicted variation in the response time coefficient in both experiments (Fig. 6): relative uncertainty sped up participants to the extent that it also biased subjects towards choosing

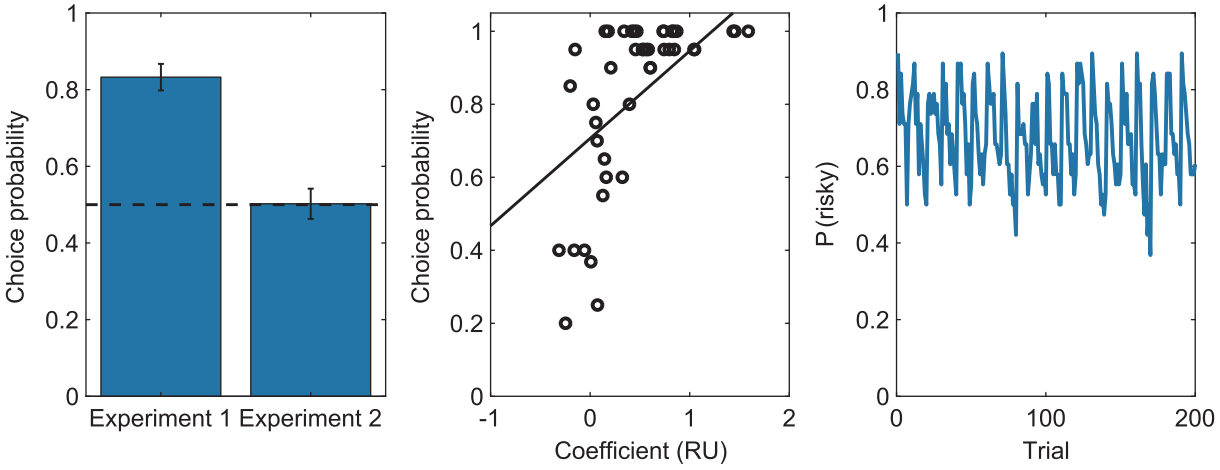

Fig. 5. Choice probabilities on the first trial of each block. (Left) Participants preferentially chose arm 1 in Experiment 1 (where it has higher uncertainty than arm 2) but not in Experiment 2 (where it has equal uncertainty). (Middle) The relative uncertainty (RU) coefficient, which captures the degree to which an individual relies on directed exploration, correlated with individual differences in arm 1 preference on the first trial. (Right) Average probability of choosing the risky option on each trial of the experiment.

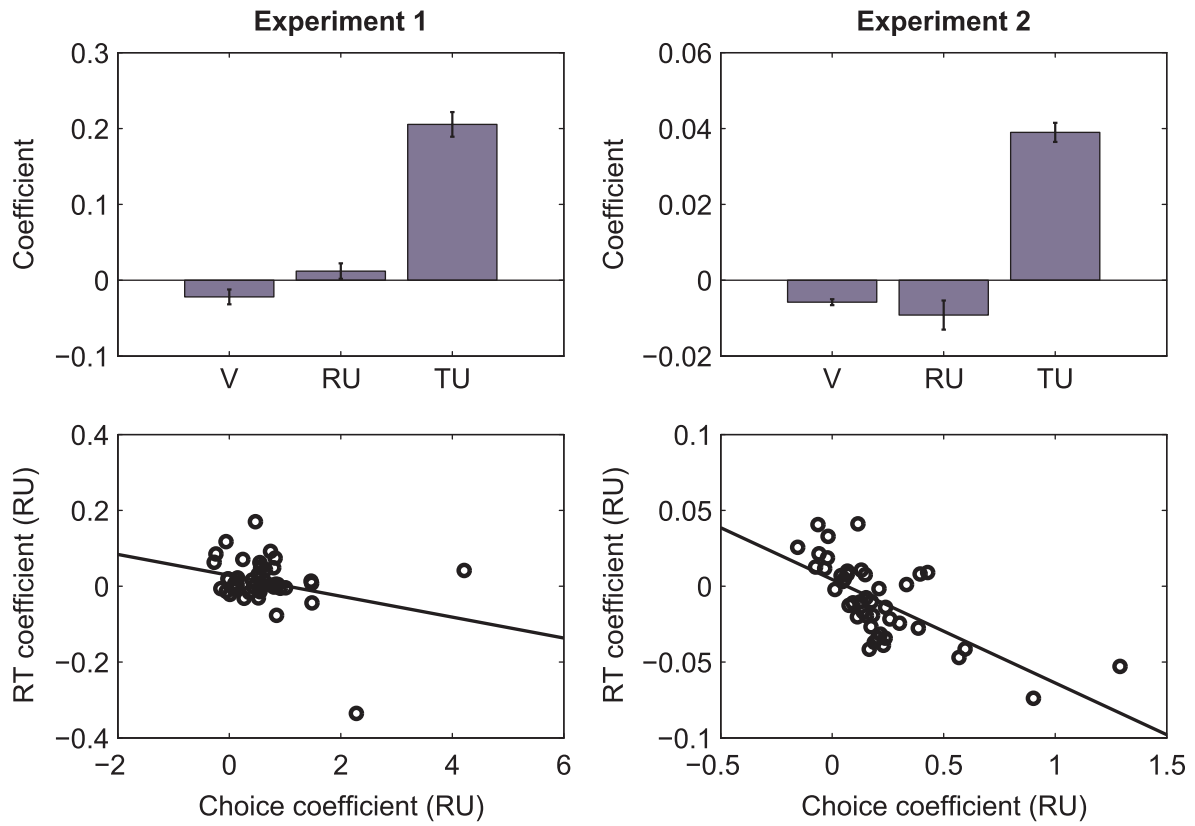

Fig. 6. Response time regression results. Log response times were modeled using linear regression with value difference (V), relative uncertainty (RU), and total uncertainty (TU) regressors. Top left: coefficients estimated from Experiment 1 data. Error bars represent standard error of the mean. Top right: coefficients estimated from Experiment 2 data. Bottom panels show the relationship between choice and response time (RT) coefficients for the relative uncertainty regressor, with superimposed least-squares line.

more uncertain arms [Experiment 1: $r = -0.3, p < .05$; Experiment 2: $r = -0.69, p < .001$].

4.4. Re-analysis of Wilson et al. (2014)

To obtain further validation of the hybrid model, we fit it to the data collected by Wilson et al. (2014). In this task, participants chose between two options with rewards drawn from a Gaussian distribution (truncated between 1 and 100, and rounded to the nearest integer). The means for the different options were different (either 40 or 60), and the standard deviation for both was set to 8. Both parameters remained fixed for each block, which lasted between 5 and 10 trials. The first 10 trials were "forced choice," a feature of the task designed to control participants' uncertainty prior to making free choices. Participants played a total of 320 blocks (see Wilson et al. (2014) for more details).

We fixed the prior variance to $\tau_0^2(1) = \tau_0^2(2) = 100$ and fit the reward variance τ^2 using grid search. The rewards were mean-centered prior to fitting, which is equivalent to setting the prior mean to 50. Only data from the free choice trials were fit, but the model used the data from the forced choice trials to update the posterior. The 4 models (UCB, Thompson, hybrid, and value-directed) were again compared using random-effects Bayesian model comparison. Recapitulating the results from Experiments 1 and 2, the data from Wilson et al. (2014) were decisively better explained by the hybrid model, with a protected exceedance probability of 0.999.

5. Discussion

The studies reported in this paper demonstrate that choice behavior on bandit problems is best explained (among the models we considered) by a combination of direct and random exploration strategies, consistent with the findings of Wilson et al. (2014) and Krueger et al. (2017). Expanding on this idea, we showed that two specific algorithmic implementations of these algorithmic strategies (UCB and Thompson sampling) produce qualitatively different effects on choice probability functions, both of which were discernible in the data. In particular, we found a downward shift in the response bias (intercept) with greater uncertainty (consistent with UCB), and an increase in choice stochasticity with greater uncertainty (consistent with Thompson sampling). Accordingly, quantitative model comparison (including a re-analysis of the data reported in Wilson et al., 2014) supported a hybrid model that combined both strategies. Finally, we found signatures of the two strategies in response times: relative uncertainty was correlated with faster response times (consistent with UCB), while total uncertainty was correlated with slower response times (consistent with Thompson sampling).

These discoveries were enabled by a detailed analysis of particular computational models, which allowed us to produce model-based uncertainty and value predictors of choice. However, we did not exhaustively survey the space of plausible models, and other models (e.g., Frank et al., 2009; Lee et al., 2011; Payzan-LeNestour & Bossaerts, 2011, 2012; Speekenbrink & Konstantinidis, 2015; Zhang & Yu, 2013) might produce empirically similar results. We conjecture that our conclusions about the signatures of directed and random exploration will hold at least qualitatively for other models which include both strategies. Conversely, our results rule out models that fail to include both strategies, such as the softmax policy with fixed inverse temperature used in the vast majority of human studies. In other words, it would be unwise at this point to make a strong claim that the uncertainty bonus and source of randomness take the particular functional forms instantiated here; rather, we are making the claim that a specific class of hybrid algorithms can explain the choice behavior reported here, and other classes cannot. So for example the knowledge gradient algorithm advocated by Zhang and Yu (2013), which implements a

different form of directed exploration, would be embraced by our theoretical framework provided that it is augmented with a form of posterior sampling.

One complexity in making these comparisons is that models based on normative considerations are typically adapted to specific environmental assumptions and task setups. For example, some models were designed specifically for non-stationary reward distributions (Payzan-LeNestour & Bossaerts, 2011, 2012; Speekenbrink & Konstantinidis, 2015; Zhang & Yu, 2013), whereas others were designed for contextual bandits, where rewards depend on contextual information provided to the agent (Schulz et al., 2015). More research is needed to understand to what extent a combination of directed and random exploration obtains in these other environments.

Wilson et al. (2014) pointed out that some studies (Daw et al., 2006; Payzan-LeNestour & Bossaerts, 2011) may have failed to support uncertainty bonuses because they were masked by a confound between uncertainty and reward: because highly rewarding arms are chosen more frequently than less rewarding arms, the highly rewarding arms will be associated with less uncertainty. Indeed, Payzan-LeNestour and Bossaerts (2012) showed that parsing uncertainty into "expected" and "unexpected" components supported a superior model that included uncertainty bonuses. We addressed this issue in our data by measuring exploration on the first trial of each block, before participants have gathered information about reward values, showing clear evidence for an uncertainty bonus.

One lingering question is *why* people would use a hybrid algorithm of the sort that best predicted choice behavior. Several papers in the machine learning literature have suggested that a hybrid approach can outperform both UCB and Thompson sampling under some conditions (Chapelle & Li, 2011; May et al., 2012), although we still lack a comprehensive understanding of what those conditions are. Indeed, it is possible that people use different exploration algorithms in different situations. To partially address the normative question, we evaluated the performance of the three algorithms on the two-armed bandit task used in Experiment 2. The results show that in general the hybrid algorithm does in fact outperform UCB and Thompson sampling (Fig. 7), thus supporting the hypothesis that people are employing a normatively justifiable exploration strategy.

Armed with fine-grained algorithmic hypotheses, it will be fruitful for future research to revisit a number of issues in contemporary research on exploration. For example, Somerville et al. (2017) found that directed exploration emerges over the course of adolescence, whereas random exploration appears to be present at the same level across developmental stages. The analyses developed in the present paper could be used to investigate whether Thompson sampling provides a good quantitative account of pre-adult exploration, with a UCB component developing later. The analyses could also be applied to relating different exploration strategies to individual differences, clinical symptoms, and neural correlates. Recent data suggest a causal role of norepinephrine in random exploration (Warren et al., 2017) and of frontopolar cortex in directed exploration (Zajkowski, Kossut, & Wilson, 2017); these neural correlates are thus good candidates for studying the modulatory role of uncertainty.

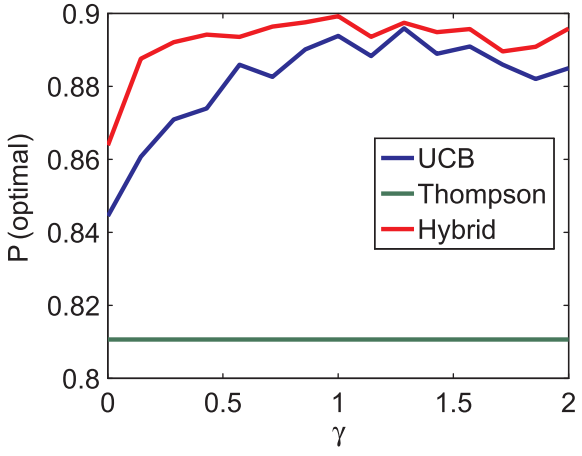

Fig. 7. Relative performance of models. Probability of choosing the optimal arm is plotted as a function of γ (undirected exploration coefficient), averaged across 500 simulated participants. The hybrid algorithm (with $\beta = 4$; see Appendix A) consistently outperforms both the UCB and Thompson sampling algorithms on the task used in Experiment 2.

While we have studied very simple (stationary, non-contextual, uncorrelated) bandit tasks, human reinforcement learning appears to make use of much more sophisticated knowledge structures. For example, humans use structured knowledge about the environment (Acuña & Schrater, 2010; Knox, Otto, Stone, & Love, 2011; Otto, Knox, Markman, & Love, 2014) and hierarchically organized priors (Gershman & Niv, 2015) to guide exploration. In principle, UCB and Thompson sampling can be applied to more structured state spaces and probabilistic models, provided that the posterior distribution over parameters can be computed. Thus, these techniques may provide a bridge between classical "model-free" reinforcement learning algorithms and structured "model-based" algorithms in the human brain (Gershman, 2017). The idea that simple, generic algorithms like UCB and Thompson sampling can be applied to a diverse range of probabilistic models provides an appealing architectural principle for the brain.

It is important, however, to recognize the limitations of algorithms like UCB and Thompson sampling. The participants in Wilson et al. (2014) modulated their exploration policy as a function of the horizon (how many trials they knew they were going to play in each block). This is at least qualitatively consistent with the optimal policy, but problematic for most implementations of UCB and Thompson sampling, which do not take the horizon into account. Thus, people may be employing model-based algorithms like tree search or dynamic programming, as has been suggested in the sequential decision making literature (Daw & Dayan, 2014; Gershman, Markman, & Otto, 2014), or they may be using hitherto unstudied approximations of these algorithms.

In summary, the key finding reported here is that reward uncertainty has two distinct effects on choice behavior: (1) It acts as a pseudo-reward, encouraging exploration of options with high uncertainty, and (2) it acts as a driver of stochasticity in choice behavior, causing choice to become more random. These dual effects suggest that humans employ a hybrid exploration algorithm, combining directed and random exploration. This hybrid algorithm can outperform pure directed and pure random exploration algorithms on the simple bandit tasks studied here, providing a normative justification for combining the two forms of exploration.

Acknowledgments

I am grateful to Eric Schulz and Felix Mauersberger for helpful discussions, and to Bob Wilson for generously sharing his data and ideas. This research was supported by the NSF Collaborative Research in Computational Neuroscience (CRCNS) Program Grant IIS-120 7833.

Appendix A. Modeling details

A.1. Belief updating

Under the Gaussian assumptions stipulated by our task, the posterior over the value of arm k is Gaussian with mean $Q_t(k)$ and variance $\sigma_t^2(k)$. Following earlier work (Daw et al., 2006; Pearson, Hayden, Raghavachari, & Platt, 2009), we used the Kalman filtering equations (Bishop, 2006) to recursively compute the posterior mean and variance for each arm:

$$Q_{t+1}(a_t) = Q_t(a_t) + \alpha_t[r_t - Q_t(a_t)] \quad (6)$$

$$\sigma_{t+1}^2(a_t) = \sigma_t^2(a_t) - \alpha_t \sigma_t^2(a_t), \tag{7}$$

where the learning rate α_t is given by:

$$\alpha_t = \frac{\sigma_t^2(a_t)}{\sigma_t^2(a_t) + \tau^2(a_t)}. \tag{8}$$

The initial values were set to the prior means, $Q_1(k) = 0$ for all k, and the initial variances were set to the prior variance, $\sigma_t^2(k) = \tau_0^2(k)$. In Experiment 1, $\tau_0^2(1) = 10$, $\tau_0^2(2) = 0$, $\tau^2(1) = 10$ and $\tau^2(2) = 0$. In Experiment 2, $\tau_0^2(1) = 100$, $\tau_0^2(2) = 100$, $\tau^2(1) = 10$ and $\tau^2(2) = 10$.

Note that although the Kalman filter is often used to model learning in dynamically changing environments, we apply it to a static environment (the reward distribution was fixed within a block). Thus, in this setting the Kalman filter is simply a recursive implementation of Bayesian inference for the mean of a Gaussian distribution.

A.2. Action policies

For the Thompson sampling policy, a random sample $\widetilde{Q}_t(k) \sim \mathcal{N}(Q_t(k), \sigma_t^2(k))$ is drawn for each arm k, and then the arm with highest $\widetilde{Q}_t(k)$ is chosen. Marginalizing over $\widetilde{V}_t = \widetilde{Q}_t(1) - \widetilde{Q}_t(2)$ gives a closed-form expression for the choice probability:

$$\begin{aligned} P(a_t = 1) &= \int_0^\infty \mathcal{N}(\widetilde{V}_t; V_t, \sigma_t^2(1) + \sigma_t^2(2)) d\widetilde{V}_t \\ &= \Phi\left(\frac{V_t}{\sqrt{\sigma_t^2(1) + \sigma_t^2(2)}}\right), \end{aligned} \tag{9}$$

where $V_t = \mathbb{E}[\widetilde{V}_t] = Q_t(1) - Q_t(2)$. This policy gives a sigmoidal "probit" function of V_t, much like softmax (which is equivalent to a logistic sigmoid for two-armed bandits), with stochasticity proportional to $\sqrt{\sigma_t^2(1) + \sigma_t^2(2)}$.

For the UCB policy, we relaxed the assumption that policies are chosen deterministically, in order to accommodate human choice stochasticity and also to facilitate the comparison with Thompson sampling. Specifically, we again used the Gaussian CDF (probit) policy:

$$P(a_t = 1) = \Phi\left(\frac{V_t + \gamma[\sigma_t(1) - \sigma_t(2)]}{\lambda}\right), \tag{10}$$

where γ is a weighting factor on the uncertainty bonus, and λ controls the choice stochasticity (analogous to temperature in the softmax policy).

For the hybrid model, we combined the directed exploration component of UCB with the random component of Thompson sampling:

$$P(a_t = 1) = \Phi\left(\beta \frac{V_t}{\sqrt{\sigma_t^2(1) + \sigma_t^2(2)}} + \gamma[\sigma_t(1) - \sigma_t(2)]\right), \tag{11}$$

where β is an additional free parameter that controls (along with γ) the balance between directed and random exploration. This is similar in spirit to the "optimistic" Thompson sampling algorithm developed by May et al. (2012).

References

Acuña, D., & Schrater, P. (2010). Structure learning in human sequential decision-making. *PLoS Computational Biology, 6*, e1001003.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning, 47*, 235–256.

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making, 16*, 215–233.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Chapelle, O., & Li, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in neural information processing systems* (pp. 2249–2257).

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*, 933–942.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society of London B, 369*, 20130478.

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

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience, 12*, 1062–1068.

Gershman, S. J. (2017). Reinforcement learning and causal models. In M. Waldmann (Ed.). *The Oxford handbook of causal reasoning*. Oxford University Press.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General, 143*, 182–194.

Gershman, S. J., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science, 7*, 391–415.

Gershman, S. J., Tenenbaum, J. B., & Jäkel, F. (2016). Discovering hierarchical motion structure. *Vision Research, 126*, 232–241.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118*, 523–551.

Grossman, E. (1953). Entropy and choice time: The effect of frequency unbalance on choice-response. *Quarterly Journal of Experimental Psychology, 5*, 41–51.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4*, 11–26.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology, 45*, 188–196.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2011). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology, 2*.

Krueger, P. M., Wilson, R. C., & Cohen, J. D. (2017). Strategies for exploration in the domain of losses. *Judgment and Decision Making, 12*, 104–117.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research, 12*, 164–174.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making, 25*, 143–153.

May, B. C., Korda, N., Lee, A., & Leslie, D. S. (2012). Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research, 13*, 2069–2106.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., et al. (2015). Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2*, 191–215.

Meyer, R. J., & Shi, Y. (1995). Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Science, 41*, 817–834.

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for value-based choice response times under high and low time pressure. *Judgment and Decision Making, 5*, 437–449.

Otto, A. R., Knox, W. B., Markman, A. B., & Love, B. C. (2014). Physiological and behavioral signatures of reflective exploratory choice. *Cognitive, Affective, & Behavioral Neuroscience, 14*, 1167–1183.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology, 7*, e1001048.

Payzan-LeNestour, E., & Bossaerts, P. (2012). Do not bet on the unknown versus try to find out more: Estimation uncertainty and "unexpected uncertainty" both modulate exploration. *Frontiers in Neuroscience, 6*.

Pearson, J. M., Hayden, B. Y., Raghavachari, S., & Platt, M. L. (2009). Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. *Current Biology, 19*, 1532–1537.

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural*

Computation, 24, 1186–1229.
Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies revisited. *NeuroImage, 84*, 971–985.
Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015). Learning and decisions in contextual multi-armed bandit tasks. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2122–2127).
Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., et al. (2017). Charting the expansion of strategic exploratory behavior during adolescence. *Journal of Experimental Psychology: General, 146*, 155–164.
Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science, 7*, 351–367.
Srinivas, N., Krause, A., Seeger, M., & Kakade, S. M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th international conference on machine learning* (pp. 1015–1022).
Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage, 46*, 1004–1017.
Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *Frontiers in Neuroscience, 6*.
Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications, 7*.
Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika, 25*, 285–294.
Warren, C. M., Wilson, R. C., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., Cohen, J. D., et al. (2017). The effect of atomoxetine on random and directed exploration in humans. *PloS One, 12*, e0176034.
Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General, 143*, 2074–2081.
Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife, 6*, e27430.
Zhang, S., & Yu, A. J. (2013). Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems* (pp. 2607–2615).