American Psychological Association

Decision

© 2018 American Psychological Association
2325-9965/19/$12.00

2019, Vol. 6, No. 3, 277–286
http://dx.doi.org/10.1037/dec0000101

BRIEF REPORT

Uncertainty and Exploration

Samuel J. Gershman
Harvard University

In order to discover the most rewarding actions, agents must collect information about their environment, potentially foregoing reward. The optimal solution to this "explore–exploit" dilemma is often computationally challenging, but principled algorithmic approximations exist. These approximations utilize uncertainty about action values in different ways. Some *random* exploration algorithms scale the level of choice stochasticity with the level of uncertainty. Other *directed* exploration algorithms add a "bonus" to action values with high uncertainty. Random exploration algorithms are sensitive to *total* uncertainty across actions, whereas directed exploration algorithms are sensitive to *relative* uncertainty. This article reports a multiarmed bandit experiment in which total and relative uncertainty were orthogonally manipulated. We found that humans employ both exploration strategies, and that these strategies are independently controlled by different uncertainty computations.

Keywords: explore–exploit dilemma, reinforcement learning, Bayesian inference

Uncertainty lies at the heart of decision making in the real world. A bee in search of nectar and a venture capitalist in search of an investment both need to explore their options in order to reduce their uncertainty, at the expense of exploiting the currently best option. In the absence of uncertainty, no explore–exploit dilemma would exist. The question addressed here is how humans use uncertainty to guide exploration.

The optimal solution to the explore–exploit dilemma is, except for some special cases (e.g., in foraging theory; Charnov, 1976; Stephens & Krebs, 1986), computationally intractable, but computer scientists have developed algorithmic approximations that can provably approach optimal behavior (Sutton & Barto, 1998). Psychologists have also studied algorithmic hypotheses about how humans balance exploration and exploitation (Cohen, McClure, & Yu, 2007; Hills et al., 2015; Mehlhorn et al., 2015), but only recently have the links between modern machine learning algorithms and psychological processes been systematically investigated (Gershman, 2018; Schulz, Konstantinidis, & Speekenbrink, 2017; Speekenbrink & Konstantinidis, 2015). Key to this synthesis is the idea that uncertainty can guide exploration in two qualitatively different ways: by adding *randomness* into choice behavior, or by *directing* choice toward uncertain options.

The pioneering work of Wilson, Geana, White, Ludvig, and Cohen (2014) demonstrated that humans use both random and directed exploration in a carefully designed two-armed bandit task. When the subject had more experience with one option (and hence less uncertainty), she favored the more uncertain option, indicating a form of directed exploration. In addition, subjects increased the randomness in their choices when they were more uncertain,

This article was published Online First October 1, 2018.

I am grateful to Eric Schulz for helpful discussions. This research was supported by the NSF Collaborative Research in Computational Neuroscience (CRCNS) Program Grant R01-1207833 and the Office of Naval Research under Grant N000141712984. This research was supported by grants from the National Institutes of Health (NIH) (R01-1207833).

Correspondence concerning this article should be addressed to Samuel J. Gershman, Department of Psychology, Harvard University, 52 Oxford St., Room 295.05, Cambridge, MA 02138. E-mail: gershman@fas.harvard.edu

spreading their choices across both high- and low-value options. Directed and random exploration strategies are dissociable, developing on different timescales across the life span (Somerville et al., 2017) and relying on different neural substrates (Zajkowski, Kossut, & Wilson, 2017).

The directed/random distinction is closely related to the distinction between two families of exploration algorithms that have been studied extensively in machine learning. Directed exploration can be realized by adding *uncertainty bonuses* to estimated values (Auer, Cesa-Bianchi, & Fischer, 2002; Brafman & Tennenholtz, 2002; Dayan & Sejnowski, 1996; Kolter & Ng, 2009; Srinivas, Krause, Seeger, & Kakade, 2010). One of the most prominent versions of this approach is known as the *upper confidence bound* (UCB) algorithm (Auer et al., 2002), which chooses action a_t on trial t according to:

$$a_t = \underset{k}{\operatorname{argmax}}[Q_t(k) + U_t(k)], \tag{1}$$

where k indexes actions and $U_t(k)$ is the upper confidence bound that plays the role of an uncertainty bonus. Under a Bayesian analysis (Srinivas et al., 2010), $Q_t(k)$ corresponds to the posterior mean and the uncertainty bonus is proportional to the posterior standard deviation $\sigma_t(k)$.

Random exploration has an even older pedigree in machine learning, dating back to Thompson's work in the 1930s (Thompson, 1933). In what is now known as *Thompson sampling*, a random value function from the posterior is drawn and then the agent chooses greedily with respect to the random draw. Like UCB, Thompson sampling uses uncertainty to promote exploration, but does so by encouraging stochasticity (i.e., a form of probability matching) rather than via a response bias.[1]

In psychophysical terms, uncertainty in Thompson sampling changes the slope of the function relating action values to choice probabilities, whereas uncertainty in UCB changes the intercept (indifference point).[2] The role of uncertainty in Thompson sampling can be understood by recognizing that sampling from a distribution over values implies that variability in this distribution directly translates to variability in choices. By contrast, the role of uncertainty in UCB can be understood by recognizing that the uncertainty bonus acts in the same way as a boost in reward, shifting choice probabilities toward the more uncertain option without altering the level of choice stochasticity.

Recognizing the dissociable (and possibly complementary) nature of directed and random exploration, computer scientists have also constructed hybrids of UCB and Thompson sampling (Chapelle & Li, 2011; May, Korda, Lee, & Leslie, 2012). A recent report (Gershman, 2018) provided the most direct evidence for such hybrids in human decision making, demonstrating that uncertainty influences both the intercept and slope of the choice probability function. Critically, the intercept and slope effects derive from different uncertainty computations: The intercept is governed by *relative* uncertainty (the difference in posterior uncertainty between the two options, defined formally below), whereas the slope is governed by *total* uncertainty (the sum of posterior uncertainty across the options). This suggests that experimental manipulations of these two factors should produce dissociable effects.

Here we pursue this line of reasoning using a two-armed bandit task, with the additional twist that we inform subjects about the riskiness of each arm (see Figure 1). On each trial, participants were given a choice between two arms, labeled either as “safe” (S) or “risky” (R). The safe arms always delivered the same reward, whereas the risky arms delivered stochastic rewards (with Gaussian noise). Denoting trial types by compound labels (e.g., “SR” denote trials in which the left arm is safe and the right arm is risky), we used the comparison between preference for Arm 1 on SR and RS trials to isolate the effects of relative uncertainty, hold-

[1] The idea that uncertainty promotes choice stochasticity is present in some theories of decision making, notably decision field theory (Busemeyer & Townsend, 1993). Thompson sampling differs formally in that stochasticity is driven by posterior uncertainty (see the formal description below), whereas in decision field theory it is driven by payoff variance. Typically, uncertainty and payoff variance are correlated, a fact that we exploit in our experimental design (see also Leuker, Pachur, Hertwig, & Pleskac, 2018).

[2] We have assumed here, following Gershman (2018), that some noise is added to the values that enter into the UCB computation. Without this assumption (or some other sources of stochasticity), we would not be able to capture variability in choices.

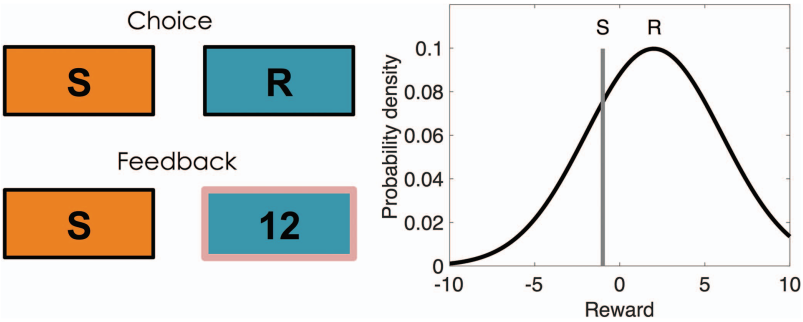

Figure 1. Task design. (Left) On each trial, subjects choose between two options and receive reward feedback in the form of points. Safe options are denoted by “S” and risky options are denoted by “R.” On each block, one or both of the options may be safe or risky (Right). The rewards for risky options are drawn from a Gaussian distribution that remains constant during each block. The rewards for safe options are deterministic. Both the mean for the risky option and the reward value of the safe option are drawn from a zero-mean Gaussian distribution that is resampled at each block transition. See the online article for the color version of this figure.

ing total uncertainty fixed, and the comparison between preference for Arm 1 on SS and RR trials to isolate the effects of total uncertainty, holding relative uncertainty fixed (see Figure 2). By independently manipulating relative uncertainty (SR vs. RS) and total uncertainty (SS vs. RR), we can go beyond the correlational findings of Gershman (2018) to causally test the predictions of UCB and Thompson sampling. In particular, we predicted that the SS condition would increase the slope of the choice probability function (affecting random but not directed exploration) relative to the RR condition, whereas the RS condition would shift the intercept of the choice probability function (affecting directed but not random exploration) relative to the SR condition.

Materials and Method

Code and data for reproducing all analyses reported in this article, as well as Javascript code for rerunning the experiments, are available at https://github.com/sjgershm/exploration_uncertainty.

Subjects

Forty-six subjects were recruited via the Amazon Mechanical Turk web service and paid \$2.00. The sample size was chosen to be comparable to previous studies using a similar experimental paradigm (Gershman, 2018). The experiments were approved by the Harvard Institutional Review Board.

Stimuli and Procedure

Participants played 30 two-armed bandits, each for one block of 10 trials. On each trial, participants chose one of the arms and received reward feedback (points). They were instructed to choose the “slot machine” (corresponding to an arm) that maximizes their total points. On each block, the mean reward $\mu(k)$ for each arm was drawn from a Gaussian with mean 0 and variance $\tau_0^2(k) = 100$. The arms were randomly designated “safe” or “risky,” indicated by an S or R, respectively, and these designations were randomly resampled after a block transition. When participants chose the risky arm, they received stochastic rewards drawn from a Gaussian with mean $\mu(R)$ and variance $\tau^2(R) = 16$. When participants chose the safe arm, they received a reward of $\mu(S)$.

The exact instructions for participants 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. Sometimes the machines are “safe” (always delivering the same feedback), and sometimes the machines are “risky” (delivering variable feedback). Before you make a choice, you will get information about each machine: “S” indicates SAFE, “R” indicates RISKY. Note that safe/risky is independent of how rewarding a

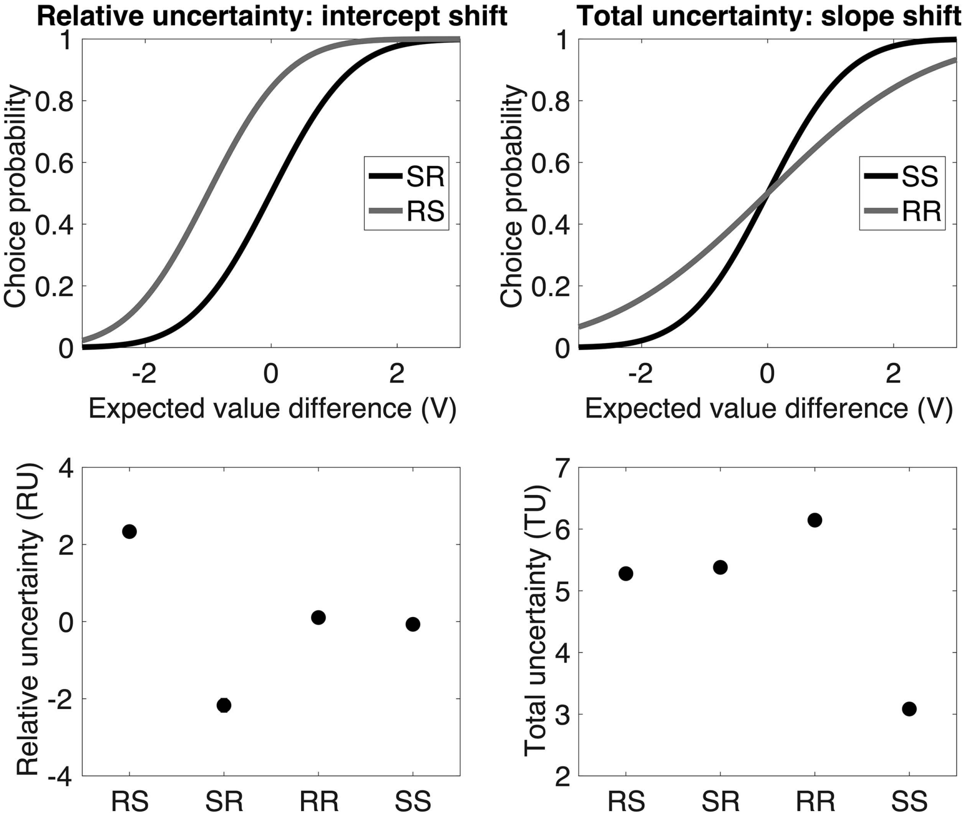

Figure 2. Relative and total uncertainty. (Top) Illustration of how the probability of choosing Option 1 changes as a function of the experimental condition and form of uncertainty. V represents the difference between the expected value of Option 1 and the expected value of Option 2. (Bottom) Average relative and total uncertainty in each condition. Conditions are denoted by safe/risky compounds; for example, "SR" denotes a trial in which Option 1 is safe and Option 2 is risky.

machine is: A risky machine may deliver more points on average than a safe machine, and vice versa. You cannot predict how good a machine is based on whether it is safe or risky. You will play 30 games, each with a different pair of slot machines. Each game will consist of 10 trials.

Belief Updating Model

To derive estimates of expected value and uncertainty, we assume that subjects approximate an ideal Bayesian learner. Given the Gaussian distributional structure underlying our task, the posterior over the value of arm k is Gaussian with mean $Q_t(k)$ and variance $\sigma_t^2(k)$. These sufficient statistics can be recursively updated using the Kalman filtering equations:

$$Q_{t+1}(a_t) = Q_t(a_t) + \alpha_t[r_t - Q_t(a_t)] \tag{2}$$

$$\sigma_{t+1}^2(a_t) = \sigma_t^2(a_t) - \alpha_t\sigma_t^2(a_t), \tag{3}$$

where a_t is the chosen arm, r_t is the received reward, and the learning rate α_t is given by:

$$\alpha_t = \frac{\sigma_t^2(a_t)}{\sigma_t^2(a_t) + \tau^2(a_t)}. \tag{4}$$

Note that only the chosen option's mean and variance are updated after each trial. 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, $\sigma_1^2(k) = \tau_0^2(k)$. The value of τ^2 was set to 16 (its true value) for risky options, and to 0.00001 for safe options (to avoid numerical issues, the variance was not set exactly to 0). Although the Kalman filter is an idealization of human learning, it has been shown to account well for human behavior in bandit tasks (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Gershman, 2018; Schulz, Konstantinidis, & Speekenbrink, 2015; Speekenbrink & Konstantinidis, 2015).

Choice Probability Analysis

Gershman (2018) showed that Thompson sampling, UCB, and a particular hybrid of the two imply a probit regression model of choice:

$$P(a_t = 1|\mathbf{w}) = \Phi(w_1 \mathrm{V}_t + w_2 \mathrm{RU}_t + w_3 \mathrm{V}_t/\mathrm{TU}_t), \tag{5}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian distribution (mean 0 and variance 1), and the regressors are defined as follows:

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

As shown in Gershman (2018), Thompson sampling predicts a significant positive effect of V/TU on choice probability, but not of RU or V, whereas UCB predicts a significant positive effect of both V and RU, but not of V/TU (Figure 2, top left). We used mixed effects estimation to fit the coefficients (**w**) in the probit regression model.

In addition to this model-based analysis, we analyzed choices as a function of experimental condition (i.e., RS, SR, RR, SS).

$$P(a_t = 1|\mathbf{w}) = \Phi\left(\sum_j w_1^j \pi_{tj} + w_2^j \pi_{tj} \mathrm{V}_t\right), \tag{6}$$

where $\pi_{tj} = 1$ if trial t is assigned to condition j, and 0 otherwise. We refer to the w_1 terms as intercepts and the w_2 terms as slopes.[3]

Response Time Analysis

We examined response times as an additional source of evidence about exploration strategies. Our hypotheses are motivated by a sequential sampling framework, according to which the value difference between two options drives a stochastic accumulator until it reaches a decision threshold (Busemeyer & Townsend, 1993; Krajbich, Armel, & Rangel, 2010; Milosavljevic, Malmaud, Huth, Koch, & Rangel, 2010; Ratcliff & Frank, 2012; Summerfield & Tsetsos, 2012). Formally, the decision variable $\mu_t(\tau)$ evolves over time (within a trial, index by τ) as follows:

$$\frac{d\mu_t(\tau)}{d\tau} = V_t + \epsilon(\tau), \tag{7}$$

where V_t (the value difference between Options 1 and 2 defined above) is a deterministic "drift" term, and $\epsilon(\tau)$ is a stochastic "diffusion" term. Without loss of generality, it is conventional to assume that $\epsilon(\tau)$ is drawn from a standard Gaussian distribution (mean of 0 and variance of 1). The decision variable evolves until it hits one of two thresholds $\pm B$ (corresponding to the two options), at which point a decision is made. If we assume, following the logic of the previous section, that both UCB and Thompson sampling are implemented by a linear transformation of the value difference (see Gershman, 2018, for more details), then the decision variable will evolve according to:

$$\frac{d\mu_t(\tau)}{d\tau} = V_t' + \epsilon(\tau), \tag{8}$$

where $V_t' = \beta \frac{V_t}{\mathrm{TU}} + \gamma \mathrm{RU}$, and β and γ are coefficients controlling the relative contribution of random and directed exploration, respectively. Under this model, the expected response time is given analytically by:

$$\mathbb{E}[\mathrm{RT}_t] = \frac{B}{V_t'} \tanh(BV_t'), \tag{9}$$

where tanh is the hyperbolic tangent function. Because the expected response time is a monotonically decreasing function of V_t', it decreases with RU and increases with TU (see also Busemeyer & Townsend, 1993).

[3] Note that although this analysis is "model-free" in the sense that it does not use the computationally derived uncertainty regressors, it is still dependent on the model-based value estimates.

Results

The hybrid random/directed exploration model hypothesizes that all three computational regressors (V, RU, V/TU) should be predictors of choice. We therefore confirmed that fixed effects estimates of the corresponding coefficients were significantly greater than 0—V: $t(13797) = 16.48$, $p < .0001$; RU: $t(13797) = 4.9$, $p < .0001$; V/TU: $t(13797) = 6.04$, $p < .0001$ (see Figure 3). Model comparison using the Bayesian and Akaike information criteria strongly favored this three-parameter model over a one-parameter model with only the V regressor, a one-parameter model with V/TU, and a two-parameter model with V and RU (see Table 1). This supports previous results showing that humans use both random and directed exploration strategies (Gershman, 2018; Somerville et al., 2017; Wilson et al., 2014).

We next addressed the central question of the article: Can random and directed exploration be independently manipulated? As predicted, changing relative uncertainty (RS vs. SR) alter the intercept of the choice probability function (see Figure 3): there was a significant difference between the intercepts for the RS and SR conditions, $F(1, 13792) = 13.92$, $p < .001$. Furthermore, the RS intercept was significantly greater than 0, $t(13792) = 2.33$, $p < .02$, and the SR intercept was significantly less than 0, $t(13792) = 4.47$, $p < .001$, indicating an uncertainty-directed choice bias, as predicted by the theory. In other words, uncertainty boosted the value of an arm, shifting choice probability toward that arm. Critically, there was no significant effect of total uncertainty (RR versus SS, $p = .40$), consistent with the hypothesis that uncertainty-directed biases only emerge when there is a difference in relative uncertainty.

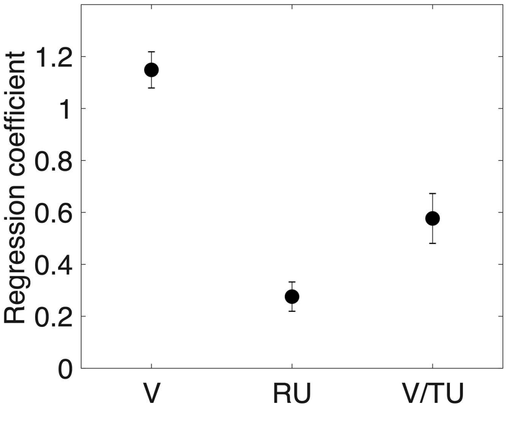

Figure 3. Probit regression results: computational variables. Fixed effects parameter estimates and standard errors for each regressor. V = estimated value difference between arms; RU = relative uncertainty; TU = total uncertainty (see Materials and Method section for details).

Table 1
Model Comparison Results

Model	BIC	AIC
V	13,493	13,478
V/TU	17,262	17,247
V + RU	12,874	12,836
V + RU + V/TU	12,689	12,621

Note. BIC = Bayesian information criterion; AIC = Akaike information criterion. Lower values indicate higher model evidence.

The pattern is flipped when we inspect the slope parameter estimates (see Figure 4): increasing total uncertainty (RR vs. SS) reduced the slope, $F(1, 13792) = 20.12$, $p < .0001$, but relative uncertainty (RS vs. SR) did not have a significant effect ($p = .72$). This finding is consistent with our hypothesis that the random component of exploration would be specifically sensitive to changes in total uncertainty. In other words, the decreased slope for RR indicates greater choice stochasticity when total uncertainty was higher. An alternative possibility is that the increased level of payoff variance in RR compared to SS causes value differences to be more strongly driven by payoff noise. In other words, even if subjects were not using random exploration, they might still show a difference in choice behavior between the RR and SS conditions. However, this hypothesis cannot account completely for the RR versus SS effect, because our analysis quantifies how choice probability differs between conditions for the same estimated value difference. Thus, even if the conditions differed in their effects on value learning, this analysis controls for such differences.

Sequential sampling models have also been shown to jointly predict choice and response time in reinforcement learning tasks, where the value is dynamically updated based on experience (Frank et al., 2015; Millner, Gershman, Nock, & den Ouden, 2018; Pedersen, Frank, & Biele, 2017). Within the sequential sampling

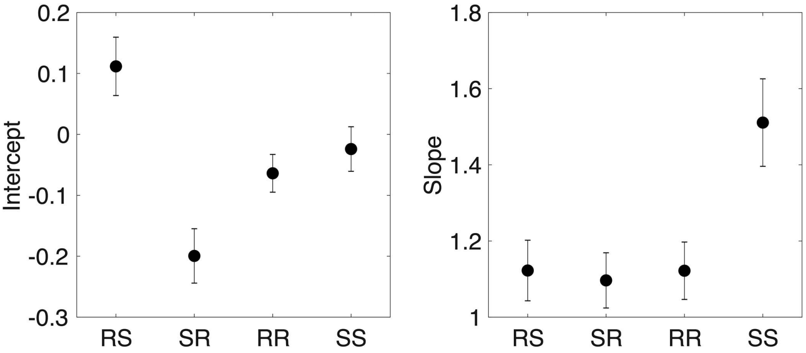

Figure 4. Probit regression results: experimental variables. Fixed effects parameter estimates and standard errors for each regressor (Left). Intercept coefficients (Right). Slope coefficients.

framework, a directed exploration strategy predicts that response times should be faster for risky choices than for safe choices on RS and SR trials, because uncertainty acts as a bonus added to the values (see Materials and Method section). This prediction was confirmed in our data, $t(45) = 2.01$, $p < .05$ (see Figure 5).

When uncertainty is equated across the arms, directed exploration does not predict any difference in response time as a function of total uncertainty. Random exploration strategies, in contrast, correctly predict that total uncertainty will act divisively on values, thereby slowing response times when both arms are risky compared to when both arms are safe, $t(45) = 2.28$, $p < .05$ (see Figure 5).

Discussion

By separately manipulating relative and total uncertainty, we were able to independently influence directed and random exploration, lending support to the contention that these strategies coexist and jointly determine exploratory behavior (Gershman, 2018; Wilson et al., 2014). In particular, increasing relative uncertainty by making one option riskier than the other caused participants to shift their preference toward the risky option in a value-independent manner, consistent with a change in the intercept (indifference point) of the choice probability function. This manipulation had no effect on the slope of the choice probability function. In contrast, increasing total uncertainty by making both options risky decreased the slope relative to when both options were safe, with no effect on the intercept. In other words, increasing total uncertainty caused choices to become more stochastic.

This dissociation between strategies guiding choice behavior was mirrored by a dissociation in response times. When one option was riskier than the other, subjects were faster in choosing the risky option. However, the same subjects were *slower* when both options were risky compared to when both options were safe. Taken together, these findings demonstrate that risk can have qualitatively different effects on both

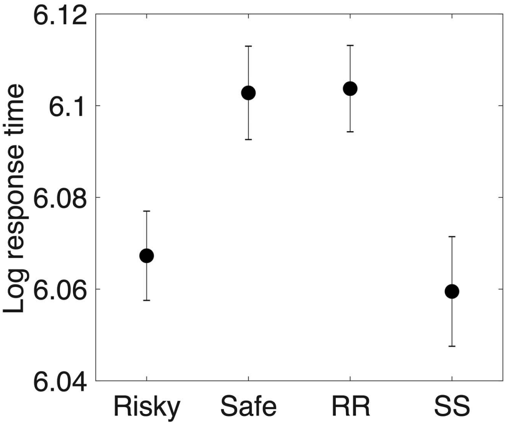

Figure 5. Response time analysis. Log response times (mean ± within-subject standard error). "Risky" denotes risky choices on SR and RS trials; "Safe" denotes safe choices on SR and RS trials. SS and RR response times are collapsed across arms.

choice and response time, depending on the underlying uncertainty computation (i.e., total vs. relative). Consistent with our findings, previous work has shown that payoff variance increases exploration (Lejarraga, Hertwig, & Gonzalez, 2012; Wulff, Mergenthaler-Canseco, & Hertwig, 2018). However, this work did not distinguish between random and directed exploration.

We note here that preference for the risky option appears to be in opposition to what would be predicted by theories of risk aversion, as pointed out by previous authors (Gershman, 2018; Payzan-LeNestour & Bossaerts, 2012; Wilson et al., 2014). This does not mean that our subjects were not risk averse; rather, we suspect that the imperative to explore in a multiarmed bandit task overwhelmed the tendency to avoid risks seen in typical gambling paradigms.

The role of uncertainty-guided exploration has come to occupy an increasingly important place in theories of reinforcement learning (Gershman & Niv, 2015; Knox, Otto, Stone, & Love, 2011; Navarro, Newell, & Schulze, 2016; Payzan-LeNestour & Bossaerts, 2011; Pearson, Hayden, Raghavachari, & Platt, 2009; Schulz et al., 2015; Speekenbrink & Konstantinidis, 2015; Zhang & Yu, 2013), superseding earlier models of exploratory choice based on a fixed source of decision noise, as in ϵ-greedy and softmax policies (e.g., Daw et al., 2006). This shift has been accompanied by a deeper understanding of how reinforcement learning circuits in the basal ganglia compute, represent, and transmit uncertainty to downstream decision-making circuits (Gershman, 2017; Lak, Nomoto, Keramati, Sakagami, & Kepecs, 2017; Starkweather, Babayan, Uchida, & Gershman, 2017). Despite this progress, we are only beginning to understand how these circuits implement dissociable channels regulating directed and random exploration (Warren et al., 2017; Zajkowski et al., 2017). Furthermore, a number of inconsistencies exist in the literature; for example, some studies do not find evidence for uncertainty bonuses in exploration (Daw et al., 2006; Riefer, Prior, Blair, Pavey, & Love, 2017). Resolving these inconsistencies will require a more complete picture of the different factors governing directed and random exploration, including individual differences, internal states (e.g., stress, cognitive load), and task structure.

References

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

Brafman, R. I., & Tennenholtz, M. (2002). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research, 3,* 213–231.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459.

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

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology, 9,* 129–136.

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

Dayan, P., & Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning, 25,* 5–22.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience, 35,* 485–494.

Gershman, S. J. (2017). Dopamine, inference, and uncertainty. *Neural Computation, 29,* 3311–3326.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition, 173,* 34–42.

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

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & Cognitive Search Research Group. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences, 19,* 46–54.

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,* 398. http://dx.doi.org/10.3389/fpsyg.2011.00398

Kolter, J. Z., & Ng, A. Y. (2009). Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning* (pp. 513–520).

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13,* 1292.

Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology, 27,* 821–832.

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition, 124,* 334–342.

Leuker, C., Pachur, T., Hertwig, R., & Pleskac, T. J. (2018). Exploiting risk–reward structures in decision making under uncertainty. *Cognition, 175,* 186–200.

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., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2,* 191–215.

Millner, A. J., Gershman, S. J., Nock, M. K., & den Ouden, H. E. (2018). Pavlovian control of escape and avoidance. *Journal of Cognitive Neuroscience, 30,* 1379–1390.

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.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. *Cognitive Psychology, 85,* 43–77.

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,* 150.

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.

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review, 24,* 1234–1251.

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.

Riefer, P. S., Prior, R., Blair, N., Pavey, G., & Love, B. C. (2017). Coherency-maximizing exploration in the supermarket. *Nature Human Behaviour, 1,* 0017.

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).

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2017). Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 927–943.

Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., & Wilson, R. C. (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).

Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience, 20,* 581.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *Frontiers in Neuroscience, 6,* 70.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

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., & Nieuwenhuis, S. (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.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin, 144,* 140–176.

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).

Received February 12, 2018
Revision received July 3, 2018
Accepted July 12, 2018 ■