Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in
Neurobiology

The algorithmic architecture of exploration in the human brain

Eric Schulz and Samuel J Gershman

Balancing exploration and exploitation is one of the central problems in reinforcement learning. We review recent studies that have identified multiple algorithmic strategies underlying exploration. In particular, humans use a combination of random and uncertainty-directed exploration strategies, which rely on different brain systems, have different developmental trajectories, and are sensitive to different task manipulations. Humans are also able to exploit sophisticated structural knowledge to aid their exploration, such as information about correlations between options. New computational models, drawing inspiration from machine learning, have begun to formalize these ideas and offer new ways to understand the neural basis of reinforcement learning.

Address
Department of Psychology and Center for Brain Science, Harvard University, 52 Oxford Street, Cambridge, MA 02138, USA

Corresponding author:
Gershman, Samuel J. (gershman@fas.harvard.edu)

Current Opinion in Neurobiology 2018, **55**:7–14

This review comes from a themed issue on **Machine Learning, Big Data, and Neuroscience**

Edited by **Maneesh Sahani** and **Jonathan Pillow**

https://doi.org/10.1016/j.conb.2018.11.003

Introduction

In order to maximize its long-term rewards, an agent must collect information about the environment, possibly at the expense of temporarily choosing less rewarding actions. This dilemma between exploration and exploitation lies at the heart of reinforcement learning. Computer scientists and engineers have developed a broad array of tractable algorithms for balancing exploration and exploitation, but only recently have these ideas begun to penetrate computational neuroscience [1,2] and psychology [3].

We review recent progress in research on the algorithmic architecture of exploration in the human brain. This review mostly focuses on *multi-armed bandit tasks*, which are frequently used to study the trade-off between exploration and exploitation experimentally [1]. The term multi-armed bandit comes from a casino metaphor where there is a row of slot machines and each slot machine has an independent payoff distribution. It is then an agents goal to maximize rewards by repeatedly selecting an arm and observing and collecting the resulting reward.

We first summarize evidence that humans use two distinct exploration strategies [4,5]: *random exploration*, which increases choice stochasticity to the agent's uncertainty about the values of available actions, and *directed exploration* which adds a bonus to each action in proportion to the agent's uncertainty about each action's value. These two algorithms offer heuristic yet efficient solutions to the exploration–exploitation dilemma. Signatures of directed and random exploration can be observed in human choice behavior, develop differently across the lifespan, and recruit distinct neural mechanisms.

In addition to using uncertainty to guide exploration, recent evidence suggests that humans use structured knowledge about the environment. For example, knowing that two options yield correlated rewards can enable more sophisticated exploration policies. Recent work also suggests that humans can adopt non-myopic policies, evaluating the benefit of future information gain. Finally, we review progress towards revealing the neural architecture underlying these exploration strategies.

From optimality to heuristics

Optimal exploration involves combining the immediate reward and the value of information for each action. This is accomplished by thinking through future actions and calculating how much future rewards could increase if more knowledge about actions is collected. Except for some special cases, optimal exploration is computationally intractable. Intuitively, this is because the value of information depends on how the information affects an agent's later choices, but these later choices may also result in new information; thus, an optimal agent would need to consider the full "policy tree" that describes all possible future trajectories. Because the size of this tree is an exponential function of the planning horizon, it cannot be computed efficiently.

One approach to bypassing this problem is to start from the observation that the optimal exploration policy will almost always deviate from the greedy policy (which only takes the action with the highest average payoff) on some proportion of choices. Thus, one heuristic strategy is to dispense entirely with computing the value of information, and instead simply select a random action on some proportion (ε) of trials. Using this ε-greedy strategy, and gradually decreasing ε, an agent will eventually learn the correct values (expected payoffs) for each action [6].

The ε-greedy heuristic, while computationally efficient, can be wasteful, because an exploratory choice is just as likely to sample the action with the worst average payoff as it is to sample the action with the second-best average payoff. An alternative heuristic, known as softmax exploration, implements a more graceful form of random sampling in which options with greater average payoff are chosen with higher probability:

$$P(a=1)=\frac{\exp\quad[\beta\mu_1]}{\sum_k\quad\exp\quad[\beta\mu_k]}, \qquad (1)$$

where μ_k denotes the average payoff of option k, and the “inverse temperature” parameter β maps out a spectrum of policies ranging from a uniform distribution over actions ($\beta\rightarrow 0$) to deterministically choosing the action with the highest experienced payoff ($\beta\rightarrow\infty$). Softmax exploration is the standard assumption in most studies of reinforcement learning, due to its simplicity, biological plausibility [7], and empirical support [8,9]. It is also closely related to a number of other ideas about choice behavior in psychology, such as probability matching [10 ,11], Luce’s choice axiom [12], and the drift diffusion model [13].

Uncertainty-based exploration

Decision making is normally affected by two types of variances, risk and posterior uncertainty. Risk can be defined as irreducible and expected payoff stochasticity, posterior uncertainty is a form of uncertainty that can be reduced through information gathering. It is this second kind of reducible uncertainty that is sought out by uncertainty-based exploration strategies.

The exploration heuristics reviewed in the previous section only depend on the average payoffs for each action. However, accumulating evidence suggests that people are also sensitive to the variability of the payoffs, in two distinct ways.

First, payoff variability increases the stochasticity of choice, a phenomenon known as the *payoff variability effect* [14–16]. Second, payoff variability can sometimes systematically attract choices, resulting in a form of risk-seeking [17,18]. Reinforcement learning theory offers an algorithmic rationalization of these effects.

The payoff variability effect can be understood as a consequence of *random* exploration strategies, which increase choice stochasticity when an agent is more uncertain. The classic example of such a strategy is *Thompson sampling*, which draws a random sample from the posterior distribution over action values and then chooses greedily with respect to this random sample [19]. When there are only two options and the posterior distribution is Gaussian, Thompson sampling is equivalent to a “probit” policy [20]:

$$P(a=1)=\Phi\left(\frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right), \qquad (2)$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function and σ_k^2 is the posterior variance for option k. Since posterior variance will generally increase with payoff variability, Thompson sampling will increase choice stochasticity when payoff variability is high. Despite its simplicity, Thompson sampling is known to produce competitive performance [21], and has recently gained in popularity due to its empirical robustness [22].

Risk-seeking can be understood as a consequence of *directed* exploration strategies, which attach an “uncertainty bonus” to each action value [23,24•,25]. The most well-known directed exploration strategy is the Upper Confidence Bound (UCB) algorithm [26]. If we assume that the action values are corrupted by a fixed amount of Gaussian noise with variance τ^2, then the UCB algorithm can also be expressed as a probit policy [20]:

$$P(a=1)=\Phi\left(\frac{\mu_1-\mu_2+\gamma[\sigma_1-\sigma_2]}{\tau}\right), \qquad (3)$$

where γ is a parameter governing the strength of the exploration bonus. UCB sampling tries to approximate optimal exploration by adding a proxy for the value of information (based on current uncertainty) to each action. Like Thompson sampling, UCB has strong theoretical properties [26,27], and is widely used in machine learning applications.

We can understand the difference between these two effects by visualizing the psychometric function relating choice probability to the difference in average payoff between two options (see Figure 1). For Thompson sampling, increasing the total uncertainty across options has the effect of reducing the slope of the choice probability function. For UCB, increasing the relative uncertainty between options shifts the intercept of the choice probability function.

A recent study [28] capitalized on these distinct psychometric signatures by orthogonally manipulating total and relative uncertainty in a bandit task with two independent options. Specifically, relative uncertainty was manipulated by making one option “risky” (paying off stochastically) and the other option “safe” (paying off deterministically). Total uncertainty was manipulated by making

Figure 1

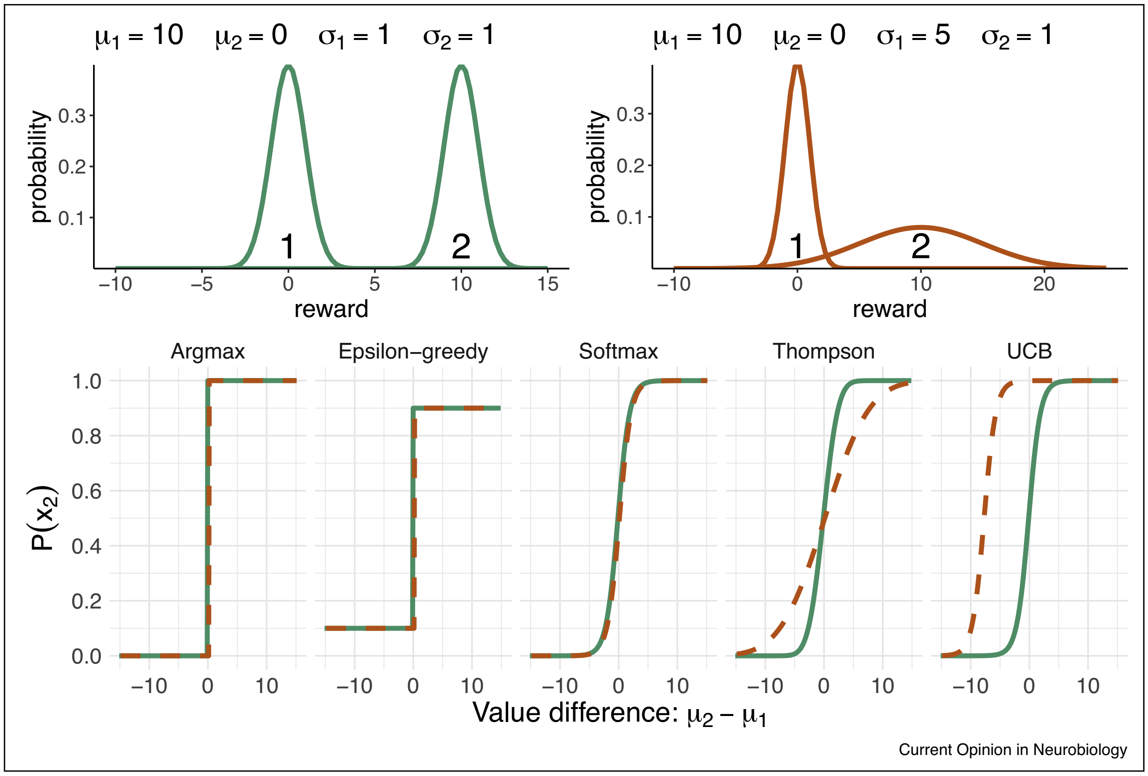

Illustration of the effect of uncertainty on choosing an option x_2 over choosing an option x_1. Upper panel shows two distributions with the same level of uncertainty on the left, and different levels of uncertainty on the right. Lower panel shows the probability of choosing x_2 over x_1 as predicted by different exploration strategies in dependency of the value difference $\mu_1 - \mu_2$.

both options risky or both options safe. Consistent with earlier results [20], subjects exhibited sensitivity to both manipulations: the slope of the choice probability function changed with the total uncertainty manipulation, and the intercept changed with the relative uncertainty manipulation.

In a large-scale comparison of models applied to a bandit task with drifting rewards [24•], Speekenbrink and Konstantinidis found that Thompson sampling accounted for human decision making better than other standard forms of random exploration, including softmax and ε-greedy sampling.

While both Thompson sampling and UCB assume that the exploration policy depends only on summary statistics of past experience (the posterior mean and variance), another possibility is that people are able to modify their policies based on beliefs about the future. To pursue this idea, Wilson et al. [4] developed a “horizon task” in which subjects played a bandit task in two contexts that differed in the number of trials (the planning horizon). Participants were allowed to make either a single choice in each game, or six sequential choices, giving them more opportunity to explore. Wilson and colleagues found that subjects increased both directed and random exploration with the longer horizon, broadly consistent with the optimal (non-myopic) exploration policy. Intuitively, an agent should explore more when there are more opportunities to compensate for early exploratory choices. Random (but not directed) exploration also decreased over the course of a game in the horizon 6 condition, consistent with the intuition that an agent should increasingly focus on the option with highest payoff as the horizon nears. Note, however, that random exploration in the horizon task does not necessarily imply Thompson sampling, but could alternatively be formalized using other forms of explorations such as softmax exploration, as long as the temperature parameter is horizon-sensitive.

In addition to experimental dissociations, directed and random exploration have been developmentally dissociated. Somerville and colleagues [29] found that directed exploration emerges during adolescence and maintains its level through early adulthood, whereas random exploration happens at comparable levels over the whole age range (though see [30] for contrasting findings).

The horizon task has also been used to probe the neural correlates of uncertainty-based exploration. Zajkowski and colleagues [31••] found that inhibition of the right frontopolar cortex (using transcranial magnetic stimulation) reduced directed exploration while leaving random exploration intact. This suggests a causal role for right frontopolar cortex in directed exploration and also that the two strategies could rely on dissociable neural systems.

In another study, atomoxetine (a norepinephrine transporter blocker that increases extracellular levels of norepinephrine) was administered to subjects and their behavior in the horizon task compared to a placebo control group [32]. Whereas transient increases in norepinephrine (NE) can be advantageous for task-relevant behavior when applied at the right time, high NE can also propagate the influence of noise and induce more variable behavior. Accordingly, NE levels may govern the balance between exploitative choices and random exploration [33]. Specifically, the authors assessed if tonic increases in NE levels from intermediate to high levels promote disengagement from current behaviors thereby increasing exploratory decision noise. Unexpectedly, atomoxetine reduced rather than increased random exploration and had no effect on directed exploration as was predicted by the authors. The authors speculated that this effect could potentially be explained by the non-linear relationship between tonic norepinephrine and exploration, as well as an unforeseen interaction between atomoxetine and other neuromodulators.

Other more belief-directed strategies of exploration have also been observed. For example, in a sequential decision making task involving approach and avoidance decisions, subjects increased their exploration when the expected number of future encounters with an option was known to be large [34••]. Furthermore, subjects were sensitive to the relative frequency of future encounters when this frequency was unknown and had to be inferred. This suggests that people can adaptively use information about the future when deciding to explore.

Exploration in structured spaces

Although standard multi-armed bandit problems capture something quintessential about learning and decision making, many realistic problems contain more structure than choosing from a set of independent options. As long as choices only concern deciding between independent options as in standard bandit tasks or the horizon task, it can be hard to distinguish between strategic random choice behavior and randomness caused by an increased task difficulty and thus slower learning rates. Therefore, tasks with additional structure have been proposed in which simple mean tracking models without exploration cannot reproduce human-like performance.

Additional structure has been investigated in several extensions of multi-armed bandit tasks, where the reward probabilities for pairs of options were correlated across trials. For example, in a version of the "acquired equivalence" paradigm [35,36], subjects played a 4-armed bandit in which (unbeknownst to them) the arms were organized into two pairs with yoked reward probabilities that changed gradually over time. Thus, the "true" number of arms was two, but subjects had to discover this fact through trial and error. After this training phase, subjects were exposed to one arm from each pair with differential reinforcement. In a subsequent test phase, subjects generalized their learned preference to the other arm of each pair, demonstrating that they had developed an associative structure mentally linking the yoked arms. This generalization was accompanied by functional connectivity between hippocampus and striatum [36], consistent with the hypothesis that the hippocampus encodes the underlying structure of the state space [37].

Stojic et al. [38] designed a feature-based multi-armed bandit task (also known as a "contextual bandit") where multiple alternatives were characterized by two features (the lengths of a horizontal and a vertical line) that mapped onto an option's expected reward by an underlying linear function. Their results showed that participants used the feature information to direct their exploration towards promising alternatives.

Using another contextual multi-armed bandit task, in which global features (the conditions of fictitious galaxies) related to different options' expected rewards (the number of emeralds mined on a chosen planet) by different functions, Schulz and colleagues [39] showed that human generalization could be well-captured with a Gaussian process regression framework, consistent with other results in the human function learning literature [40,41]. This framework combines powerful non-linear function approximation with analytically tractable computations. Model comparison indicated that combining Gaussian process generalization with a directed exploration strategy (UCB) produced the best account of human choice behavior.

Human reinforcement learning has also been investigated in spatially correlated multi-armed bandits [42], where rewards are distributed on a grid and each tile of the grid is the arm of a bandit; crucially, different arms' rewards are spatially correlated such that proximal arms produce similar rewards, enabling participants to generalize over many options. Human exploration in the spatially correlated multi-armed bandit is best predicted by a Gaussian process regression model paired with a sampling strategy that combines structured generalization, directed (UCB) and random (softmax) exploration [43]. In all of these experiments, a combination of UCB and random exploration performed better than either softmax-exploration or ε-greedy sampling.

The studies reviewed above ask whether humans can take advantage of correlations between options to explore more efficiently. Others studies examine whether humans can take advantage of dynamical structure. For example, Knox et al. [44] developed a "leapfrog" task, a two-armed bandit in which the payoff for one arm jumped at random intervals, surpassing the payoff of the other arm. Thus, the optimal arm switches according to a change process, and the question is whether subjects could learn and use this structure ("reflective" exploration) or if they treated the task as a standard two-armed bandit ("reflexive" exploration). Using a model-based analysis, Knox and colleagues found that human choice behavior followed the predictions of a reflective exploration strategy: the probability of exploration increased with number of trials since the last experienced jump. However, subjects did not show consideration of future states, but rather only planned myopically.

The leapfrog task has also been used to probe the role of dopamine in exploration [45•]. The catechol-O-methyltransferase (COMT) gene modulates dopamine levels in prefrontal cortex, such that Met allele carriers have lower COMT enzyme activity and thus higher dopamine levels compared to Val allele carriers. Met carriers showed a greater tendency to explore reflectively compared to Val/Val homozygotes when put under cognitive load.

Non-myopic policies

The exploration algorithms discussed in previous section are for the most part myopic: they do not consider explicitly the value of future information, but rather utilize heuristics based on summary statistics such as reward mean and variance for each option. Whether people engage non-myopic planning during exploration remains controversial. Some of the studies reviewed above find evidence for long-range planning [4], whereas some do not [44]. However, given that there are two repeated conditions of different sampling lengths in the horizon task, the degree of exploration by condition could still be tuned by a myopic and model-free learning system.

Some recent evidence in favor of non-myopic exploration comes from tasks that assess how humans explore their environment in adaptive planning tasks with sequential state dependence. These tasks are unlike traditional bandit tasks in that past choices affect future states. This means that parts of the state space might never be revisited again. Exploration is particularly important in such sequentially structured tasks.

For example, people plan ahead in a complex foraging task not by updating all possible states but rather by initializing beliefs first and then thinking ahead, a strategy similar to random exploration as characterized by Thompson sampling [46]. In another study [47], participants performed an adaptive control task in which they had to steer a boat through perilous sea in a simple video game. The results of this study showed that participants explored strategically, executing "test trials" of later trajectories during times of free exploration.

A similar strategy of exploration, which takes into account the future value of information, has also been formalized and assessed behaviorally [48]. In a bandit task with binary outcomes, human choices can be well-captured by a model which combined exact Bayesian learning with a decision policy that maximized a combination of immediate rewards and long-term information gain. This idea was then developed further to investigate behavior in a task in which the informational value and the potential rewards were directly manipulated on each trial [25]. As before, the best-fitting computational model augmented standard myopic algorithms by additionally incorporating a value of information.

Disentangling the neural correlates of exploration and exploitation

Although the distinction between exploration and exploitation is well-defined computationally, it is much more difficult to distinguish them empirically. Many theories posit that choice is fundamentally stochastic (e.g., [13 ,15]), which complicates the interpretation of apparently exploratory choices. If a person chooses an option that has a lower average payoff than another option, is that because they were exploring, or because they made a random mistake? This ambiguity means that one must be careful when interpreting parameter estimates from reinforcement learning models; if the estimated inverse temperature is low, this could mean that an individual is very exploratory, but it could also mean that she is more "noisy." The same ambiguity applies to neuroimaging correlates of exploration [8,49].

One solution to this problem is to use a task that explicitly separates exploration and exploitation. Tversky and Edwards [50] devised such a task, in which subjects choose on repeated and independent trials to either observe a reward (without collecting the payoff) or to bet (collecting the payoff without observing it). Observing corresponds unambiguously to exploration, and betting corresponds unambiguously to exploitation. Interest in this task has recently revived [51••], using new computational methods to analyze how subjects choose to observe or bet. Although people often employed suboptimal strategies at the beginning, most of them were able to approximate the correct strategy after only minimal experience. Moreover, people deviated from the optimal strategy by repeatedly switching between observation and betting at the start—a strategy that is well-suited for dynamic problems but ill-suited for static ones.

Using the observe-or-bet task in the fMRI scanner, a recent study found that insula and dorsal anterior cingulate cortex showed greater activity on observe trials compared to bet trials [52•]. This suggests that these regions play a role in driving pure exploration, consistent with some earlier studies [53–55]. The activity of these areas during exploratory choices cannot be explained by simple value effects (i.e., that participants received surprising rewards, as these choices were purely observatory without actually gaining rewards).

Surprisingly, this study did not find a signature of exploration in rostral prefrontal or frontopolar cortex, in contrast with the results of several earlier studies [8,49,56,57], possibly indicating that these earlier results were not entirely pure signatures of exploration.

Conclusions

Problems requiring a trade-off between gathering information and collecting rewards are ubiquitous in human learning and decision making. We have reviewed recent developments in research on the algorithmic architecture of exploration in the human brain. These developments have begun to coalesce around a few key computational ideas. In particular, humans seem to employ a combination of both random and uncertainty-directed exploration strategies, which rely on different brain systems, have distinct developmental trajectories, and are sensitive to different task manipulations. These two strategies, when implemented mechanistically, correspond to two algorithms found in the machine learning literature (Thompson and Upper Confidence Bound sampling). Humans also appear to take advantage of latent structure to explore more efficiently, and in some cases explore non-myopically.

Other explanations of participants' sensitivity to the variability of rewards also exist. For example, risk aversion and risk seeking can also arise from nonlinear marginal utility functions, although some of these interpretations can be ruled out by computational modeling [20,24•]. Additionally, an increase in participants' choice variability with outcome variance can arise from learning alone, even in simple mean tracking models that fully ignore uncertainty. However, mean tracking models alone cannot capture all of the results presented here (e.g., [42,43]). Future work should still try to further disentangle these competing interpretations.

Another promising future avenue for research on human exploration could be to further assess hybrid algorithms of random and directed exploration strategies that have been postulated in machine learning. For example, May et al. [58] proposed a sampling strategy called optimistic Bayesian sampling which, like Thompson sampling, optimizes based on sampled beliefs, but additionally inflates the probability of choosing an action based on the uncertainty in the estimate of the action value. Another possible hybrid sampling strategy finds inspiration from work on sampling strategies that switch between globally directed and locally random sampling [59]. We believe that further formalizing and testing ways in which people combine random and directed exploration will provide useful insights into both human and artificial reinforcement learning.

Conflicts of interest statement

Nothing declared.

Acknowledgments

This research was supported by the Office of Naval Research (Award N000141712984) and the Harvard Data Science Initiative.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Cohen JD, McClure SM, Yu AJ: **Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration**. *Philos Trans R Soc Lond B: Biol Sci* 2007, **362**:933-942.

2. Laureiro-Martínez D, Brusoni S, Zollo M: **The neuroscientific foundations of the exploration–exploitation dilemma**. *J Neurosci Psychol Econ* 2010, **3**:95-115.

3. Mehlhorn K, Newell BR, Todd PM, Lee MD, Morgan K, Braithwaite VA, Hausmann D, Fiedler K, Gonzalez C: **Unpacking the exploration–exploitation tradeoff: a synthesis of human and animal literatures**. *Decision* 2015, **2**:191-215.

4. Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD: **Humans use directed and random exploration to solve the explore–exploit dilemma**. *J Exp Psychol Gen* 2014, **143**:2074-2081.

5. Gershman SJ: **Reinforcement learning and causal models**. In *The Oxford Handbook of Causal Reasoning.* Edited by Waldmann M. Oxford University Press; 2017.

6. Sutton RS, Barto AG: *Reinforcement Learning: An Introduction*. MIT Press; 1998.

7. Collins AG, Frank MJ: **Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive**. *Psychol Rev* 2014, **121**:337-366.

8. Daw ND, O'doherty JP, Dayan P, Seymour B, Dolan RJ: **Cortical substrates for exploratory decisions in humans**. *Nature* 2006, **441**:876-879.

9. Yechiam E, Busemeyer JR: **Comparison of basic assumptions embedded in learning models for experience-based decision making**. *Psychon Bull Rev* 2005, **12**:387-402.

10. Neimark ED, Shuford E: **Comparison of predictions and estimates in a probability learning situation**. *J Exp Psychol* 1959, **57**:294-298.

11. Vulkan N: **An economist's perspective on probability matching**. *J Econ Surv* 2000, **14**:101-118.

12. Pleskac TJ: **Decision and choice: Luce's choice axiom.**. *International Encyclopedia of the Social & Behavioral Sciences* 2015:895-900.

13. Pedersen ML, Frank MJ, Biele G: **The drift diffusion model as the choice rule in reinforcement learning**. *Psychon Bull Rev* 2017, **24**:1234-1251.

14. Myers JL, Sadler E: **Effects of range of payoffs as a variable in risk taking**. *J Exp Psychol* 1960, **60**:306-309.

15. Busemeyer JR, Townsend JT: **Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment**. *Psychol Rev* 1993, **100**:432-459.

16. Erev I, Barron G: **On adaptation, maximization, and reinforcement learning among cognitive strategies**. *Psychol Rev* 2005, **112**:912-931.

17. Hertwig R, Barron G, Weber EU, Erev I: **Decisions from experience and the effect of rare events in risky choice**. *Psychol Sci* 2004, **15**:534-539.

18. Weber EU, Shafir S, Blais AR: **Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation**. *Psychol Rev* 2004, **111**:430-445.

19. Thompson WR: **On the likelihood that one unknown probability exceeds another in view of the evidence of two samples**. *Biometrika* 1933, **25**:285-294.

20. Gershman SJ: **Deconstructing the human algorithms for exploration**. *Cognition* 2018, **173**:34-42.

21. Agrawal S, Goyal N: **Analysis of Thompson sampling for the multi-armed bandit problem**. *Conference on Learning Theory* 2012:39-41.

22. Chapelle O, Li L: **An empirical evaluation of Thompson sampling**. *Advances in Neural Information Processing Systems* 2011:2249-2257.

23. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F: **Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation**. *Nat Neurosci* 2009, **12**:1062-1068.

24. Speekenbrink M, Konstantinidis E: **Uncertainty and exploration**
• **in a restless bandit problem**. *Top Cogn Sci* 2015, **7**:351-367.
Behavioral evidence for uncertainty bonuses

25. Dezza IC, Angela JY, Cleeremans A, Alexander W: **Learning the value of information and reward over time when solving exploration–exploitation problems**. *Sci Rep* 2017, **7**:16919.

26. Auer P, Cesa-Bianchi N, Fischer P: **Finite-time analysis of the multiarmed Bandit problem**. *Mach Learn* 2002, **47**:235-256.

27. Srinivas N, Krause A, Seeger M, Kakade SM: **Gaussian process optimization in the Bandit setting: no regret and experimental design**. *Proceedings of the 27th International Conference on Machine Learning* 2010:1015-1022.

28. Gershman SJ: **Uncertainty and exploration**. *bioRxiv* 2018:265504.

29. Somerville LH, Sasse SF, Garrad MC, Drysdale AT, Abi Akar N, Insel C, Wilson RC: **Charting the expansion of strategic exploratory behavior during adolescence**. *J Exp Psychol Gen* 2017, **146**:155-164.

30. Schulz E, Wu CM, Ruggeri A, Meder B: **Searching for rewards like a child means less generalization and more directed exploration**. *bioRxiv* 2018:327593.

31. Zajkowski WK, Kossut M, Wilson RC: *eLife* 2017, **6**:e27430.
•• Transcranial magnetic stimulation study dissociating different forms of exploration

32. Warren CM, Wilson RC, van der Wee NJ, Giltay EJ, van Noorden MS, Cohen JD, Nieuwenhuis S: **The effect of atomoxetine on random and directed exploration in humans**. *PLoS One* 2017, **12**:e0176034.

33. Aston-Jones G, Cohen JD: **An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance**. *Annu Rev Neurosci* 2005, **28**:403-450.

34. Rich, A.S., Gureckis, T.M. Exploratory Choice Reflects the Future
•• Value of Information.
Evidence for non-myopic exploration.

35. Daw ND, Shohamy D: **The cognitive neuroscience of motivation and learning**. *Soc Cogn* 2008, **26**:593-620.

36. Wimmer GE, Daw ND, Shohamy D: **Generalization of value in reinforcement learning by humans**. *Eur J Neurosci* 2012, **35**:1092-1104.

37. Stachenfeld KL, Botvinick MM, Gershman SJ: **The hippocampus as a predictive map**. *Nat Neurosci* 2017, **20**:1643-1653.

38. Stojic H, Analytis PP, Speekenbrink M: **Human behavior in contextual multi-armed bandit problems**. *In Proceedings of the 37th Annual Meeting of the Cognitive Science Society* 2015:2290-2295.

39. Schulz E, Konstantinidis E, Speekenbrink M: **Putting bandits into context: how function learning supports decision making**. *J Exp Psychol Learn Memory Cogn* 2017.

40. Lucas CG, Griffiths TL, Williams JJ, Kalish ML: **A rational model of function learning**. *Psychon Bull Rev* 2015, **22**:1193-1215.

41. Schulz E, Tenenbaum JB, Duvenaud D, Speekenbrink M, Gershman SJ: **Compositional inductive biases in function learning**. *Cognit Psychol* 2017, **99**:44-79.

42. Wu CM, Schulz E, Speekenbrink M, Nelson JD, Meder B: **Mapping the unknown: the spatially correlated multi-armed bandit**. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* 2017:1357-1362.

43. Wu C, Schulz E, Speekenbrink M, Nelson JD, Meder B: **Exploration and generalization in vast spaces**. *bioRxiv* 2017:171371.

44. Knox WB, Otto AR, Stone P, Love BC: **The nature of belief-directed exploratory choice in human decision-making**. *Front Psychol* 2012:2.

45. Blanco NJ, Love BC, Cooper JA, McGeary JE, Knopik VS,
• Maddox WT: **A frontal dopamine system for reflective exploratory behavior**. *Neurobiol Learn Mem* 2015, **123**:84-91.
Genetic analysis of belief-directed exploration

46. Krusche MJF, Schulz E, Guez A, Speekenbrink M: **Adaptive planning in human search**. *bioRxiv* 2018.

47. Schulz E, Klenske E, Bramley N, Speekenbrink M: **Strategic exploration in human adaptive control**. *bioRxiv* 2017:110486.

48. Zhang S, Yu AJ: **Forgetful Bayes and myopic planning: human learning and decision-making in a bandit setting.**. *Advances in Neural Information Processing Systems* 2013:2607-2615.

49. Boorman ED, Behrens TE, Woolrich MW, Rushworth MF: **How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action**. *Neuron* 2009, **62**:733-743.

50. Tversky A, Edwards W: **Information versus reward in binary choices**. *J Exp Psychol* 1966, **71**:680-683.

51. Navarro DJ, Newell BR, Schulze C: **Learning and choosing in an**
•• **uncertain world: an investigation of the explore–exploit dilemma in static and dynamic environments**. *Cognit Psychol* 2016, **85**:43-77.
Revival of Tversky and Edward's observe-or-bet task, which cleanly dissociates exploration and exploitation

52. Blanchard TC, Gershman SJ: **Pure correlates of exploration and**
• **exploitation in the human brain**. *Cogn Affect Behav Neurosci* 2018, **18**:117-126.
Neural analysis of the observe-or-bet task

53. Kolling N, Behrens TE, Mars RB, Rushworth MF: **Neural mechanisms of foraging**. *Science* 2012, **336**:95-98.

54. Boorman ED, Rushworth MF, Behrens TE: **Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice**. *J Neurosci* 2013, **33**:2242-2253.

55. Li J, McClure SM, King-Casas B, Montague PR: **Policy adjustment in a dynamic economic game**. *PLoS One* 2006, **1**:e103.

56. Badre D, Doll BB, Long NM, Frank MJ: **Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration**. *Neuron* 2012, **73**:595-607.

57. Beharelle AR, Polanía R, Hare TA, Ruff CC: **Transcranial stimulation over frontopolar cortex elucidates the choice attributes and neural mechanisms used to resolve exploration–exploitation trade-offs**. *J Neurosci* 2015, **35**:14544-14556.

58. May BC, Korda N, Lee A, Leslie DS: **Optimistic Bayesian sampling in contextual-bandit problems**. *J Mach Learn Res* 2012, **13**:2069-2106.

59. McLeod M, Osborne MA, Roberts SJ: **Optimization, fast and slow: optimally switching between local and Bayesian optimization**. *ArXiv* 2018. [e-prints].