aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
1–9

DOI: 10.1177/09637214221105051
www.psychologicalscience.org/CDPS


Chasing Unknown Bandits: Uncertainty Guidance in Learning and Decision Making

Maarten Speekenbrink
Department of Experimental Psychology, University College London, and The Alan Turing Institute, London, England

Abstract

In repeated decision problems for which it is possible to learn from experience, people should actively seek out uncertain options, rather than avoid ambiguity or uncertainty, in order to learn and improve future decisions. Research on human behavior in a variety of multiarmed-bandit tasks supports this prediction. Multiarmed-bandit tasks involve repeated decisions between options with initially unknown reward distributions and require a careful balance between learning about relatively unknown options (exploration) and obtaining high immediate rewards (exploitation). Resolving this exploration-exploitation dilemma optimally requires considering not only the estimated value of each option, but also the uncertainty in these estimations. Bayesian learning naturally quantifies uncertainty and hence provides a principled framework to study how humans resolve this dilemma. On the basis of computational modeling and behavioral results in bandit tasks, I argue that human learning, attention, and exploration are guided by uncertainty. These results support Bayesian theories of cognition and underpin the fundamental role of subjective uncertainty in both learning and decision making.

Keywords

experience-based decisions, exploration-exploitation dilemma, Bayesian learning

> I used to be uncertain, but now I'm not so sure.
> (inspired by Tommy Cooper)

Many crossroads in life require deciding between actions with unknown consequences. These decisions can be mundane, such as whether to take the bus or cycle to work, or more profound, such as whether to accept a new job offer or to continue in a current job. The more mundane decisions afford more opportunity for learning, as we face them and their consequences many times. Nevertheless, the consequences of mundane decisions can be profound and life changing: Arriving late for a crucial meeting might result in being fired and subsequently looking for a new job.

How do we decide to act under uncertainty? And how do we learn to improve our decisions in the future? Generally, there is an inherent tie between our actions and our experience of the world. We do not know what would have happened if we had taken a different course of action. The resulting conundrum, known as the *exploration-exploitation dilemma* (Cohen et al., 2007), is easy to describe, yet difficult to resolve: Should we choose options that we know we like (exploit our knowledge), or should we choose more uncertain options so that we might learn about them and improve our future decisions (explore to acquire knowledge)?

Classic results suggest that people generally choose known rather than uncertain alternatives. In the Ellsberg paradox (Ellsberg, 1961), people are presented with two urns: a *known urn* with exactly 50 red and 50 black balls and an *ambiguous urn* with 100 red and black balls in an unknown proportion. When betting on whether a randomly drawn ball will be black or red, people prefer to draw from the known urn. Such ambiguity, or uncertainty, aversion has been observed many times (Camerer & Weber, 1992). Yet there are many situations in which people actively seek out uncertain alternatives. If you are allowed to play the Ellsberg game

Corresponding Author:
Maarten Speekenbrink, Department of Experimental Psychology, University College London
Email: m.speekenbrink@ucl.ac.uk

repeatedly, choosing the ambiguous urn would prove advantageous. Over time, you could learn whether there are more red or black balls in the urn and bet accordingly, winning more often than possible with the known urn. When (a) learning is possible and (b) future decisions can be improved by learning, uncertainty should—and indeed does—guide learning and decision making.

Evidence for a guiding role of uncertainty in learning and decision making is interesting for a variety of reasons. First, it indicates that uncertainty is cognitively accessible and affects behavior. Evaluating the uncertainty of knowledge is perhaps one of the primary metacognitive abilities. Second, it provides additional support for Bayesian theories of cognition. Theories of learning and decision making that concern only expectancies and not their associated uncertainties should be deemed incomplete.

Multiarmed Bandits

Multiarmed-bandit tasks provide a useful experimental paradigm to study how people navigate the exploration-exploitation dilemma. In a multiarmed-bandit task, participants are repeatedly presented with a set of options, each with an initially unknown distribution of rewards. The goal is to accumulate as much reward as possible. After participants choose an option, a reward is randomly drawn from the chosen option's reward distribution. Crucially, participants do not see the rewards they would have obtained had they chosen differently. These tasks are analogous, for example, to ordering at a restaurant. After you order a dish, you do not know how much you would have enjoyed a different dish. This is the informational bottleneck that leads to the exploration-exploitation dilemma: By sticking to one option, you forgo the opportunity to learn about other options. But exploring other options comes with a potential cost, as you may not enjoy them as much as your current favorite. Exploration should therefore focus on *promising* options, options that have a reasonable chance of being better than the current favorite. This probability is inherently tied to both value (expected reward) and uncertainty about that value. Hence, good strategies for exploration need to take both value and uncertainty into consideration.

Multiarmed-bandit tasks are a simple type of *reinforcement-learning task*, a term that broadly refers to "learning by doing." In *standard bandit* tasks, all options are independent, and the reward distributions are static (they do not change over time). An example would be choosing what to order in a restaurant where dishes are always prepared by the same chef who is consistently excellent. Although there might be variations in your experience, due to slight variations in the quality of ingredients, quantity of spices, and so forth, these are random, and your average enjoyment will not fluctuate. In such tasks, exploration should be front loaded; that is, it should occur only during the initial stages. Once the value of each option is estimated with sufficient precision, you can safely exploit the option deemed best forever after.

In *restless-bandit* tasks (Daw et al., 2006; Knox et al., 2012; Speekenbrink & Konstantinidis, 2015), the reward distributions vary over time. An example would be choosing what to order at a restaurant where the dishes are prepared by different chefs who can learn to perfect their skills and who also have periods of underperformance. At such a restaurant, a once favorite dish may become relatively poor for a prolonged period in time. The rational approach to such tasks would be to continue exploration throughout, because during the time spent exploiting one option, other options might have surpassed it.

In *contextual-bandit* tasks (Schulz, Konstantinidis, & Speekenbrink, 2018; Stojić, Schulz et al., 2020; Wu et al., 2018), options and the environment come with reward-predictive features. An example would be choosing what to order at a restaurant where, before making your choice, you can observe the chefs at work and other patrons consuming their dishes. Learning the feature-reward relations allows generalizing experience across options. The rational approach would again be for exploration to be front loaded, but now with respect to the options' features. Moreover, it is no longer necessary to explore all options. Once the feature-reward relations are learned with sufficient precision, unpromising options can be avoided altogether without ever trying them. Figure 1 illustrates the differences between the three types of bandit tasks.

The goal of exploration is to improve future decisions. When a bandit task approaches its end, the risk of exploration will likely start to outweigh its potential benefits, simply because there are fewer future decisions to improve. Therefore, exploration should decrease toward the end of all types of bandit tasks.

Types of uncertainty

Multiarmed bandits involve two types of uncertainty. The first, called *aleatoric uncertainty* or *risk*, concerns inherent variability in rewards. Even if we know the reward distributions exactly, variability in rewards means we cannot know exactly what reward will follow a given choice. Aleatoric uncertainty is therefore also called *irreducible uncertainty*, as no amount of learning can reduce

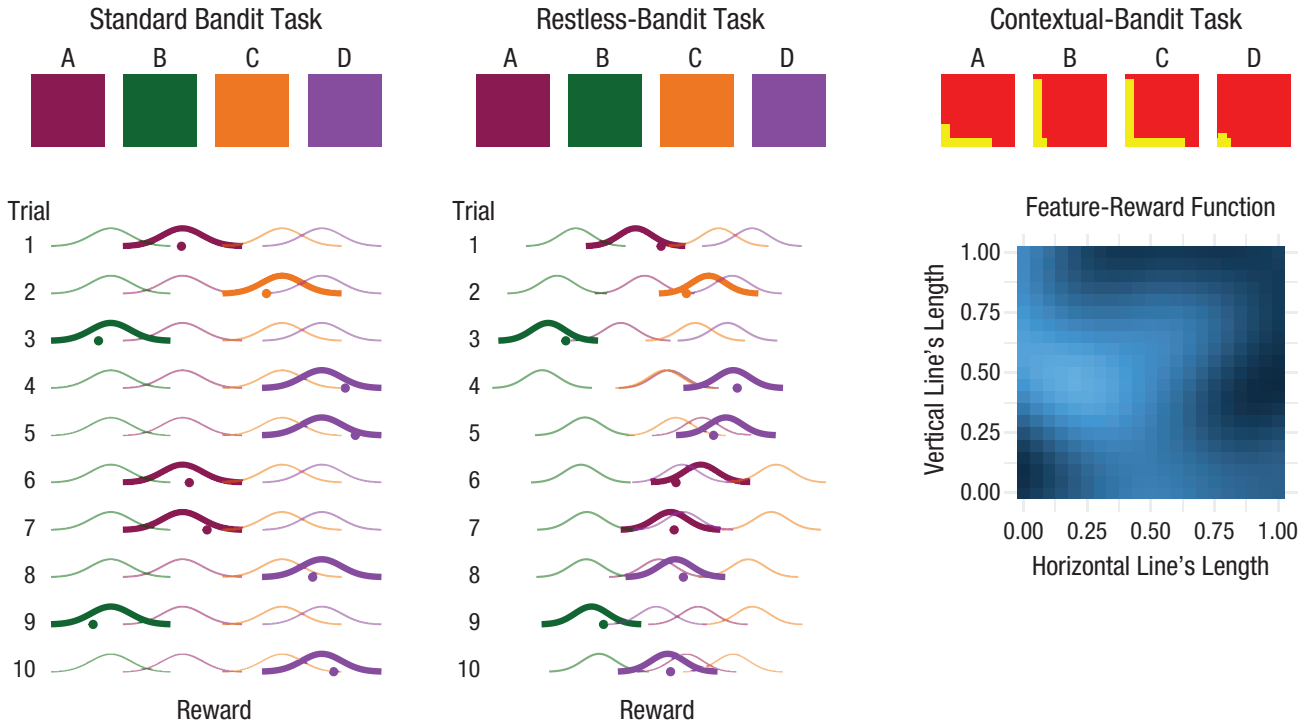

Fig. 1. Three types of multiarmed-bandit tasks. The squares at the top show how options might be presented to participants. In a standard bandit task, participants are provided with options with unknown but static reward distributions. The options may be distinguished by a label or color (as shown here), but these features are not related to rewards. In a restless-bandit task, participants are also provided with options with unknown reward distributions, but these distributions vary over time, such that an option that once had the highest value may be surpassed by other options. As in the standard bandit task, the features of the options are not related to rewards. The lower plots for these two tasks show the reward distributions of chosen (thick lines) and nonchosen (thin lines) options over 10 trials. The horizontal location of each reward distribution reflects that option's value. Thus, in the standard bandit task, the distributions have the same locations from trial to trial, whereas in the restless-bandit task, the distributions' locations vary over trials. On each trial, after an option is chosen, a random reward (shown as a dot) is drawn from that option's reward distribution. In a contextual-bandit task, options come with reward-predictive features, such as horizontal and vertical lines that differ in length across options. These features are related to the value of an option, as illustrated in the plot on the lower right; lighter colors indicate higher average reward.

it. The second form of uncertainty, called *epistemic uncertainty*, concerns the accuracy of our beliefs about the world. Within this type of uncertainty, a further distinction can be made between *estimation uncertainty* and *structural uncertainty*. Although we may assume that we have a structurally correct model of the world, some aspects of this model (the model parameters) may be unknown. This type of uncertainty is called estimation uncertainty. For example, in the repeated Ellsberg game, the only uncertainty is about the proportion of red balls. This uncertainty is reducible, as we can estimate the proportion with increasing precision by drawing more balls from the ambiguous urn. Structural uncertainty concerns uncertainty about whether our structural model of the world is accurate. Perhaps the experimenter lied, and the ambiguous urn also contains green and blue balls. Or perhaps after some time one ambiguous urn is replaced by another one with a different proportion of red balls. This uncertainty is also reducible, as observations can be used to select among a set of competing models of the world.

The different types of uncertainty can be formalized precisely within a Bayesian learning framework (see Fig. 2). In Bayesian learning, a main driver of learning is the level of epistemic uncertainty relative to aleatoric uncertainty. When epistemic uncertainty is relatively high, many states of affairs (e.g., possible values of an option) are initially plausible. When aleatoric uncertainty is relatively low, a reward provides relatively precise information about an option's value, such that many of the initially plausible states of affairs become implausible. When aleatoric uncertainty increases, an observed reward provides less information about value, and so less can be learned. When epistemic uncertainty decreases, more is known, and fewer states of affairs become plausible. This leaves less room to shift beliefs, and hence less can be learned from an observed reward. So, all else being equal, higher epistemic uncertainty

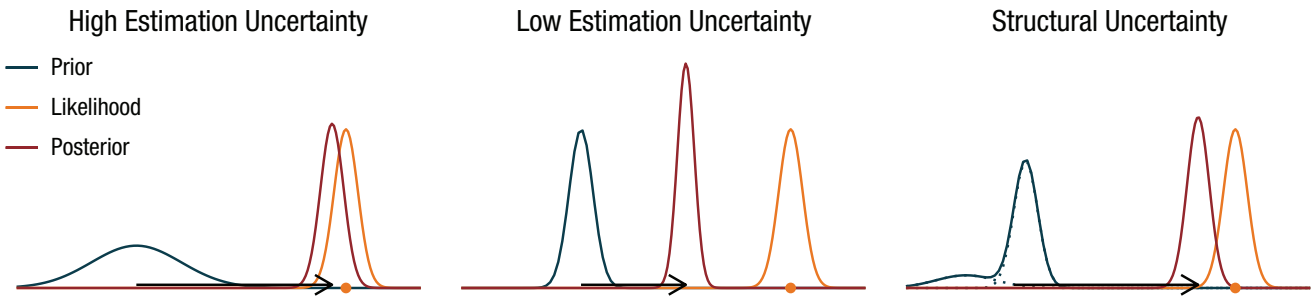

Fig. 2. Types of uncertainty in Bayesian learning. In Bayesian reinforcement learning, the prior distribution reflects beliefs about the value (e.g., average reward) of an option. The wider the prior distribution, the higher the estimation uncertainty. The likelihood represents variability in rewards (aleatoric uncertainty). An observed reward (shown as dots) provides new information, allowing belief to be updated from the prior to the posterior distribution. The *learning rate* (distance between the means of the prior and posterior distributions, shown as arrows; larger distance implies higher learning rate) depends on the relative magnitude of epistemic uncertainty (estimation or structural uncertainty) compared with aleatoric uncertainty. When estimation uncertainty is high relative to aleatoric uncertainty (left plot), the learning rate is relatively high. Reducing estimation uncertainty (middle plot) reduces the learning rate. Structural uncertainty refers to uncertainty about the underlying structural model. This can be represented through a prior distribution that is a weighted sum over models, each with associated estimation uncertainty. An observed reward provides information about the identity of the underlying model as well as the parameters of the possible models. Structural uncertainty (right plot) increases the learning rate compared with no structural uncertainty (middle plot), for reasons similar to those that account for the increase in learning rate when estimation uncertainty increases.

implies a higher learning rate, but higher aleatoric uncertainty implies a lower learning rate.

Exploration strategies

Determining the optimal trade-off between the costs and benefits of exploration is generally not possible (May et al., 2012). It involves planning ahead by considering each potential outcome of each potential decision, in terms of immediate rewards as well as how these would change one's beliefs and subsequent decisions (and their potential outcomes) in the future. In all but restricted cases, the sheer number of possible futures makes optimal planning impossible.

A heuristic solution, known as the upper-confidence-bound rule, is to approximate the informational value of exploration by adding an *uncertainty bonus* to the estimated value of each option. As the term suggests, the resulting sum of the expectation and uncertainty bonus equals an upper confidence bound on the option's value. The rule states that the option with the highest upper bound should be selected. An alternative form of uncertainty-guided exploration, called Thompson sampling, is to add to the estimated values momentary random noise that reflects the uncertainty of these estimates. This can be implemented by randomly sampling a momentary expected value from the current prior distribution for each option and choosing the option with the highest sampled value. An uncertainty-ignorant heuristic is the *softmax* strategy, in which the same level of momentary random noise—which does not depend on uncertainty—is added to the estimated values of all options, and the option with the highest sum is chosen. Another uncertainty-ignorant heuristic is the *epsilon-greedy* strategy, which introduces randomness in choice by sometimes (with a probability ε) choosing an option from all available options completely at random, irrespective of the options' estimated values or uncertainty.

Uncertainty-guided heuristics generally outperform uncertainty-ignorant heuristics in maximizing the accumulated rewards (May et al., 2012). For example, if the epsilon-greedy and softmax strategies are used, a given option would be explored equally often regardless of whether the subjective probability that it is the best option is substantial or vanishingly small (see Fig. 3). Uncertainty-guided heuristics lead to exploration of an option only in the former case, which is sensible because there is little to be gained from exploring an option when it is almost certainly worse than the other options.

Uncertainty Guides Learning

There is a wealth of evidence that, in accordance with Bayesian principles, epistemic uncertainty increases and aleatoric uncertainty decreases how much is learned from experience (e.g., Behrens et al., 2007; Dayan et al., 2000; Nassar et al., 2010; Payzan-LeNestour & Bossaerts, 2011; Speekenbrink & Shanks, 2010; Stojić, Orquin, et al., 2020). This evidence often relies on computational modeling in which models with a constant learning rate are compared with models in which the learning rate is modulated by epistemic and aleatoric uncertainty. Such work has shown that the latter models describe behavior better than the former.

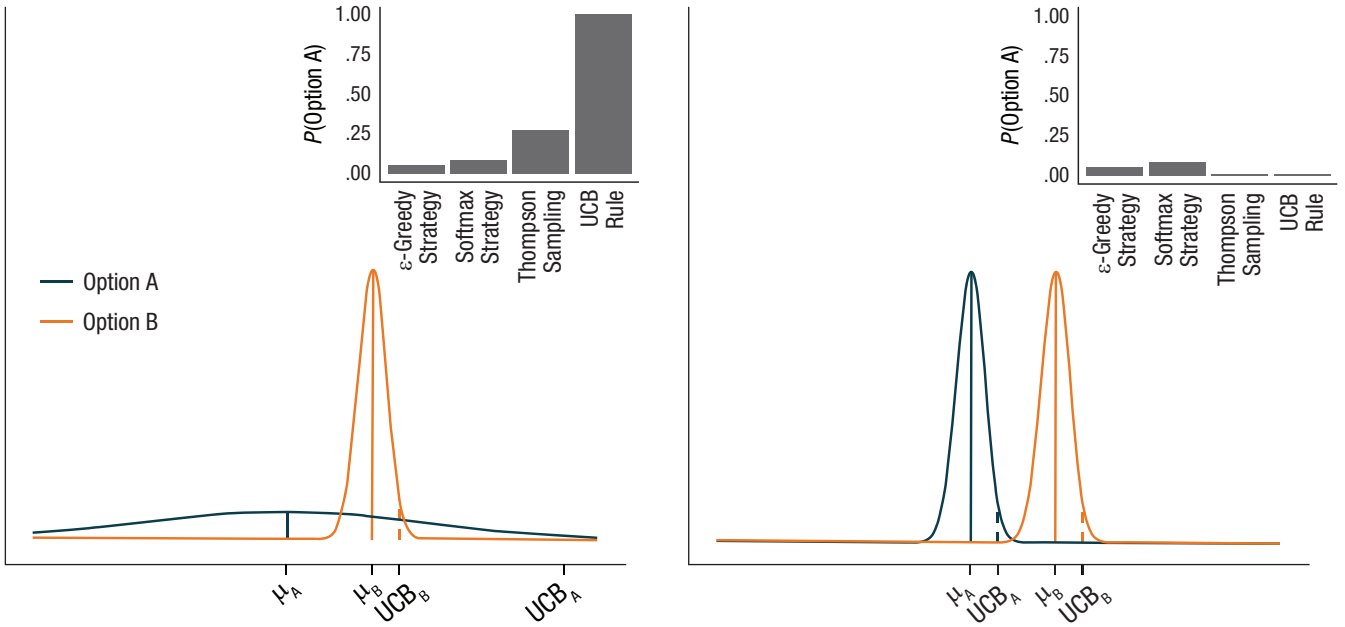

Fig. 3. Illustration of the performance of exploration heuristics in two situations with different estimation uncertainty. The plots show the current prior expectations about the value of two options, and the insets show the probability of exploring Option A according to each heuristic. In both situations, Option A is expected to provide a lower average reward than Option B (the prior mean of Option A, μ_A, is lower than that of Option B, μ_B). In the left plot, there is much more uncertainty about Option A than Option B, which is also evident in the 95% upper confidence bounds (UCBs): UCB_A is substantially higher than UCB_B. As a result, the uncertainty-guided heuristics, the UCB rule and Thompson sampling, predict a relatively high probability of exploring Option A. The UCB rule always chooses the option with the highest UCB, whereas Thompson sampling chooses the option with the highest sampled value from the prior distribution. When the estimation uncertainty of Option A is reduced (right plot), these uncertainty-guided heuristics predict that Option A will not be explored. In contrast, the uncertainty-ignorant heuristics, the epsilon-greedy and softmax strategies, predict equal rates of exploration of Option A in these two situations.

In behavioral studies, structural uncertainty is often introduced by abruptly changing the value of options (e.g., Behrens et al., 2007; Payzan-LeNestour & Bossaerts, 2011). Increased structural uncertainty should speed learning (Fig. 2, right), as it indicates that formerly held beliefs, and the data that supported these, may no longer be relevant. This prediction has generally been supported (Gallistel, 2012). A complicating factor is that the speed of learning can be assessed only through changes in choice probabilities, and choices reflect not only learned value (i.e., exploitation) but also exploration. Prediction tasks (Nassar et al., 2010; Speekenbrink & Shanks, 2010), in which people are asked to predict the next value of a variable, allow for more direct measures of learning rate. Results for such tasks also indicate that learning rate increases with epistemic uncertainty and decreases with aleatoric uncertainty, as expected from Bayesian principles.

Uncertainty Guides Attention and Information Gathering

Before one decides on a course of action, information can be obtained about the current state of the world. This information can be perceptual as well as memory based (Shadlen & Shohamy, 2016). In contextual-bandit tasks and studies on associative learning, participants are presented with reward-predictive cues. Determining the identity of these cues helps reduce uncertainty about the consequences of possible actions (e.g., choosing an option). Eye-tracking studies indicate that people preferentially focus their attention on more predictive cues (Leong et al., 2017; Walker et al., 2019). This is sensible, as these cues reduce uncertainty most.

Even in noncontextual-bandit tasks, people allocate their attention to options in an uncertainty-guided manner. My colleagues and I (Stojić, Orquin, et al., 2020) found that, before making their choice, people looked more at options with a higher value (expected reward) as well as those with higher estimation uncertainty. Focusing on particular options in the absence of further cues may help retrieval of prior experiences from memory, and thus help in determining the value of options. But it may also bias choice directly toward attended options (Krajbich et al., 2010). Directing attention toward options according to both value and uncertainty is then a way to employ this bias to balance exploitation and uncertainty-guided exploration. We (Stojić,

Orquin, et al.) found that time spend looking at an option had a strong influence on subsequent choice, which could be only partially explained by the option's estimated value and estimation uncertainty. But value and estimation uncertainty also had additional unique effects beyond those mediated by looking time. Thus, there appear to be multiple routes by which uncertainty guides decisions: a direct route via an internal valuation process and an indirect route via visual attention.

Uncertainty Guides Exploration

Although some studies have found no evidence for uncertainty-guided exploration (Daw et al., 2006; Payzan-LeNestour & Bossaerts, 2011), many studies have found such evidence (Frank et al., 2009; Gershman, 2018; Knox et al., 2012; Speekenbrink & Konstantinidis, 2015; Stojić, Orquin, et al., 2020; Stojić, Schulz, et al., 2020). Whether uncertainty adds a bonus or noise to options' estimated value is not clear, and it likely does both (Gershman, 2018; Wilson et al., 2014). Most of these studies have relied on computational modeling and shown that models with an uncertainty-driven exploration component describe people's behavior better than those with uncertainty-ignorant exploration strategies. Wilson et al. (2014) showed that after a forced-choice stage, people prefer to choose options they experienced less and therefore are more uncertain about. Gershman (2019) generated different bandit tasks with all possible combinations of *safe* options, which always give the same reward, and *risky* options, which give variable rewards, and found behavioral evidence for uncertainty-guided exploration when the options differed in uncertainty (one option was safe and the other risky), as well as evidence for random exploration.

Deriving unambiguous behavioral signatures of exploration strategies is complicated by the fact that uncertainty is inherently tied to a statistical model of the world, and people may entertain different models of a given task. If someone assumes that a task is a static-bandit task, in which the value of options remains unchanged over time, uncertainty-driven exploration should be high initially and switch to pure exploitation when sufficient knowledge is acquired (front-loaded exploration). By contrast, uncertainty-ignorant exploration should persist throughout the task. If someone assumes that a task is a restless-bandit task, in which the value of options changes over time, both uncertainty-guided and uncertainty-ignorant exploration should persist throughout the task. However, only uncertainty-guided exploration would predict exploration of an option to increase with the time since it was last chosen. When people are explicitly instructed whether the task is a static- or restless-bandit task, these predictions for uncertainty-guided exploration are often confirmed. Navarro et al. (2016) used an observe-or-bet task to cleanly separate exploration and exploitation. Participants could either chose to observe the reward of an option without reaping it or bet on an option by choosing it without observing the reaped reward. In this task, participants front-loaded exploration of static bandits but not restless bandits, although front loading required prior experience with static-bandit tasks. Knox et al. (2012) instructed participants about the changing value of the options in their task and found, as predicted, that the probability of exploration increased with the time since an option was last explored. Using another restless-bandit task, Konstantinidis and I (Speekenbrink & Konstantinidis, 2015) additionally found that people switched more between options (a rough measure of exploration) in periods of rapid change with high epistemic uncertainty.

In another study, my colleagues and I (Stojić, Schulz, et al., 2020) derived qualitative predictions for uncertainty guidance in a contextual-bandit task in which, after some time, a novel option was introduced. We found evidence that people explore a novel option much more when the learned relationship between features and rewards indicates that the new option will yield high rewards. In addition, we found evidence that people explore a novel option when its features are relatively dissimilar to those of previously encountered options. As exploring options with novel features should reduce uncertainty about the relationship between features and rewards substantially, this behavioral pattern points to *functional* uncertainty guidance in exploration. Using a different contextual-bandit task in which the features were the spatial locations of options, my colleagues and I (Wu et al., 2018) found related evidence for functional uncertainty guidance.

Open Questions

I have reviewed a range of results supporting the idea that uncertainty plays a guiding role in learning and experience-based decision making. In this section, I describe important open questions that can be addressed in future research.

Balancing the costs and benefits of exploration requires consideration of immediate risk and potential future benefits. As the end of a task approaches and there are only a limited number of decisions left, the potential benefits of newly acquired knowledge tend to be outweighed by the costs of acquiring more knowledge. The heuristic exploration strategies considered here ignore this planning aspect of exploration. Whether humans do so as well is not entirely clear. Although people explore less in shorter compared with longer

tasks (Rich & Gureckis, 2018; Wilson et al., 2014), and when they expect to encounter an unknown option less often (Rich & Gureckis, 2018; Wulff et al., 2015), this behavior is likely based on heuristics rather than a planning process that optimally weights the risks and benefits of acquiring new information (Knox et al., 2012).

Krueger et al. (2017) compared situations in which all options provide rewards and those in which all options provide losses and found evidence for uncertainty-guided exploration in both. Although the uncertainty bonus was larger when the options yielded losses, participants relied on qualitatively similar exploration strategies in both situations. However, as in most studies discussed in this review, the cost of exploration was relatively benign (e.g., a small loss in points or financial reward). In real life, the cost is often more profound. For example, a foraging animal exploring a new area may die if it encounters a predator or no food. More research on how people safely explore in high-stakes environments is needed. Initial results (e.g., Schulz, Wu, et al., 2018) indicate that people adapt their exploration strategies when disastrous outcomes have to be avoided. Purposeful exploration has been found to be greater in periods when it is relatively safe than when the stakes are high (Schulz et al., 2017). Linking such findings to those of research on exploratory play, curiosity, risk sensitivity, and changing patterns of exploration over the life span is an important avenue for future research.

Reinforcement-learning models (RLMs) focus on how sequences of decisions over trials are shaped by obtained rewards (and sometimes uncertainty). Evidence-accumulation models (EAMs) focus on the intratrial processes that affect the timing and nature of isolated decisions. How these evidence-accumulation processes link with longer-term learning and afford uncertainty-guided exploration is an open question. Although recent advances have integrated RLMs and EAMs (e.g., Pedersen et al., 2017), these efforts have generally ignored the role of uncertainty. There are different ways to relate the parameters of EAMs to uncertainty and value estimates. Gershman (2018) found evidence that uncertainty bonuses decrease response times, whereas uncertainty-modulated noise added to options' values increases response times. This indicates that multiple parameters of EAMs may be related to uncertainty (e.g., the initial level and variability of evidence). Attention to options may also directly influence the evidence-accumulation process (e.g., Krajbich et al., 2010), although work in this area tends to treat attention as given, rather than investigating why and when attention is directed to options. Our recent work shows that attention is guided by both value and uncertainty (Stojić, Orquin, et al., 2020), which may inspire new ways of integrating RLMs and EAMs.

Conclusion

Research with a variety of multiarmed-bandit tasks shows that both learning and decisions are guided by uncertainty. This indicates that theories of behavior that ignore uncertainty are incomplete. Rather than avoiding uncertainty, we should embrace it and let uncertainty guide our learning and decisions in scientific endeavors as well as daily life.

Recommended Reading

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & the Cognitive Search Research Group. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences, 19*(1), 46–54. https://doi.org/10.1016/j.tics.2014.10.004. Discusses how the exploration-exploitation dilemma permeates many aspects of cognition, from searching in memory to determining social structure.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior, 36*(3), 210–226. https://doi.org/10.3758/LB.36.3.210. Reviews Bayesian and traditional theories of associative learning and extensions to active learning (exploration without an exploitation trade-off).

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2*(3), 191–215. https://doi.org/10.1037/dec0000033. Reviews a broad range of psychological research on how humans and other animals solve the exploration-exploitation dilemma.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology, 55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003. Provides a review of computational approaches to solving the exploration-exploitation dilemma, and the behavioral and neuroscientific evidence for use of these approaches, with a particular focus on the difference between random and uncertainty-guided exploration.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026. Introduces the distinction between expected (estimation) uncertainty and unexpected (structural) uncertainty and discusses how these two kinds of uncertainty might be encoded through different neuromodulators.

Transparency

Action Editor: Robert L. Goldstone
Editor: Robert L. Goldstone
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Maarten Speekenbrink https://orcid.org/0000-0003-3221-1091

Acknowledgments

The author is grateful to Michele Nathan for her thorough and detailed editorial efforts to make this article more readable and understandable.

References

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*(4), 325–370. https://doi.org/10.1007/BF00122575

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 B: Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*(11), 1218–1223. https://doi.org/10.1038/81504

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669. https://doi.org/10.2307/1884324

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*(8), 1062–1068. https://doi.org/10.1038/nn.2342

Gallistel, C. R. (2012). Extinction from a rationalist perspective. *Behavioural Processes*, *90*(1), 66–80. https://doi.org/10.1016/j.beproc.2012.02.008

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, *6*(3), 277–286. https://doi.org/10.1037/dec0000101

Knox, W., Otto, A., Stone, P., & Love, B. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology*, *2*, Article 398. https://doi.org/10.3389/fpsyg.2011.00398

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Krueger, P. M., Wilson, R. C., & Cohen, J. D. (2017). Strategies for exploration in the domain of losses. *Judgment and Decision Making*, *12*(2), 104–117. https://journal.sjdm.org/14/141223a/jdm141223a.pdf

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*(2), 451–463. https://doi.org/10.1016/j.neuron.2016.12.040

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. https://www.jmlr.org/papers/volume13/may12a/may12a.pdf

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, *30*(37), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010

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. https://doi.org/10.1016/j.cogpsych.2016.01.001

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLOS Computational Biology*, *7*(1), Article e1001048. https://doi.org/10.1371/journal.pcbi.1001048

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, *24*(4), 1234–1251. https://doi.org/10.3758/s13423-016-1199-y

Rich, A. S., & Gureckis, T. M. (2018). Exploratory choice reflects the future value of information. *Decision*, *5*(3), 177–192. https://doi.org/10.1037/dec0000074

Schulz, E., Klenske, E. D., Bramley, N. R., & Speekenbrink, M. (2017). Strategic exploration in human adaptive control. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *CogSci 2017: Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1047–1052). Cognitive Science Society. https://cogsci.mindmodeling.org/2017/papers/0204/paper0204.pdf

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(6), 927–943. https://doi.org/10.1037/xlm0000463

Schulz, E., Wu, C. M., Huys, Q. J. M., Krause, A., & Speekenbrink, M. (2018). Generalization and search in risky environments. *Cognitive Science*, *42*(8), 2592–2620. https://doi.org/10.1111/cogs.12695

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939. https://doi.org/10.1016/j.neuron.2016.04.036

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*(2), 351–367. https://doi.org/10.1111/tops.12145

Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, *139*(2), 266–298. https://doi.org/10.1037/a0018620

Stojić, H., Orquin, J. L., Dayan, P., Dolan, R. J., & Speekenbrink, M. (2020). Uncertainty in learning, choice, and visual fixation. *Proceedings of the National Academy of Sciences, USA*, *117*(6), 3291–3300. https://doi.org/10.1073/pnas.1911348117

Stojić, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2020). It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *Journal of Experimental Psychology: General*, *149*(10), 1878–1907. https://doi.org/10.1037/xge0000749

Walker, A. R., Luque, D., Le Pelley, M. E., & Beesley, T. (2019). The role of uncertainty in attentional and choice exploration. *Psychonomic Bulletin & Review*, *26*(6), 1911–1916. https://doi.org/10.3758/s13423-019-01653-2

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*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924. https://doi.org/10.1038/s41562-018-0467-4

Wulff, D. U., Hills, T. T., & Hertwig, R. (2015). How short- and long-run aspirations impact search and choice in decisions from experience. *Cognition*, *144*, 29–37. https://doi.org/10.1016/j.cognition.2015.07.006