Published in final edited form as:

Curr Opin Behav Sci. 2021 April ; 38: 49–56. doi:10.1016/j.cobeha.2020.10.001.

Balancing exploration and exploitation with information and randomization

Robert C. Wilson[1,2,3,*], **Elizabeth Bonawitz**[4], **Vincent D. Costa**[5], **R. Becket Ebitz**[6]

[1]Department of Psychology, University of Arizona, Tucson AZ USA

[2]Cognitive Science Program, University of Arizona, Tucson AZ USA

[3]Evelyn F. McKnight Brain Institute, University of Arizona, Tucson AZ USA

[4]Department of Psychology, Rutgers University - Newark, Newark NJ USA

[5]Department of Behavioral Neuroscience, Oregon Health and Science University, Portland OR USA

[6]Department of Neuroscience, University of Montréal, Montréal, Québec, Canada

Abstract

Explore-exploit decisions require us to trade off the benefits of exploring unknown options to learn more about them, with exploiting known options, for immediate reward. Such decisions are ubiquitous in nature, but from a computational perspective, they are notoriously hard. There is therefore much interest in how humans and animals make these decisions and recently there has been an explosion of research in this area. Here we provide a biased and incomplete snapshot of this field focusing on the major finding that many organisms use two distinct strategies to solve the explore-exploit dilemma: a bias for information ('directed exploration') and the randomization of choice ('random exploration'). We review evidence for the existence of these strategies, their computational properties, their neural implementations, as well as how directed and random exploration vary over the lifespan. We conclude by highlighting open questions in this field that are ripe to both explore and exploit.

Introduction

When you go to your favorite restaurant, do you always get the same thing (the pizza you know and love), or are you ever tempted by something else (the ravioli you've never tried)? Exploiting an old favorite guarantees the reward of a good meal, but is uninformative (you already know the pizza is good). On the other hand, exploring something new brings no guarantees about the quality of the meal, but does guarantee that you learn something (in this case, whether you like the ravioli or not). This is an example of the explore-exploit dilemma,

[*]Correspondence to: bob@arizona.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

a behavioral dilemma that occurs any time our desire for information conflicts with our need for reward [1, 2, 3, 4, 5, 6].

Outside of the contrived examples at the start of explore-exploit papers, the explore-exploit dilemma is ubiquitous. It is faced by birds in the sky as they forage for food [7] and fish in the sea as they decide where to hunt [8]. Monkeys face it when navigating changing environments [9] and rats face it all the time in the lab [10]. Even slime molds, without a brain but with a rudimentary ability to process information and a drive to find food, confront the explore-exploit dilemma [11]. As humans we face it when we travel to work [12], decide where to fish [13], or pick out food to order online [14]. As scientists we face it when choosing the next experiment to run. And as experts on explore-exploit decisions, the authors of this paper face it every time we introduce the dilemma to a new audience in our papers and talks: should we exploit the tried-and-trusted restaurant examples we have used before [15], or branch out and explore something new?

Given the ubiquity of the explore-exploit dilemma, an obvious question is: How do we solve it? Based on recent literature in psychology and neuroscience, we argue that humans and animals combine two major (but not mutually exclusive) strategies for exploration: information-seeking (directed exploration) and behavioral variability (random exploration). These strategies are associated with different neural correlates and the natural balance between them changes over the lifespan. Ultimately, we argue that a holistic model of human exploration must account for this duality and we review recent progress in developing such models. We end by highlighting some of the major open questions in the field.

Directed and random strategies for exploration

Optimal solutions to the explore-exploit dilemma are intractable in all but the simplest cases [16, 17, 18, 19]. The reason for this difficulty is that optimal solutions require us to perform massive simulations of the future — considering how choices impact future outcomes and how those outcomes will impact future choices. Because of this computational complexity, most theoretical work has turned to approximate methods. While there are literally hundreds of such approximate algorithms [20], across all of these approaches, two general strategies have proven to be especially effective: an explicit bias for information, ‘directed exploration’ [16, 17], and the randomization of choice, ‘random exploration’ [21, 22, 2] (Figure 1).

In directed exploration, exploration is ‘directed’ towards more informative options by a deterministic ‘information bonus,’ which increases the value of informative options. Mathematically, one way to implement this strategy is by increasing the value of informative options such that the value of taking action a

$$Q(a) = r(a) + IB(a) \tag{1}$$

where $r(a)$ is how good we expect option a to be and $IB(a)$ is the information bonus for each action. In this simple model, the choice is made by picking the option with the highest $Q(a)$. Different algorithms have different forms for the information bonus. One popular approach, epitomized by the ‘Upper Confidence Bound’ (or UCB) algorithm (Figure 1c), is to set the information bonus proportional to the *uncertainty* about the expected payoff from each option [23]. In some situations the performance of this strategy can be close to optimal despite having a fraction of the computational cost.

In random exploration, random decision noise drives exploration by chance. Mathematically, one way to implement random exploration is by adding noise to the computation of value

$$Q(a) = r(a) + \eta(a) \tag{2}$$

where $\eta(a)$ is zero-mean random noise sampled from some probability distribution and again the choice is the action with largest $Q(a)$. The exact flavor of random exploration is determined by the form of the noise distribution. For example, adding logisticially-distributed random noise to values will yield the famous softmax choice function (Figure 1d). Random forms of exploration can perform quite well in many environments and other noise distributions can solve other problems. For example, annealing algorithms reduce the noise over time, while an algorithm called Thompson Sampling scales the noise with the agent's uncertainty. Both approaches lead to high levels of random exploration when the environment is unknown, but reduce random exploration later on allowing the agent to exploit what it has learned and leading to near optimal performance in some cases (Figure 1e) [21, 24].

Of course, directed and random exploration are not mutually exclusive and it is possible to use both strategies at once. In fact, there is strong evidence that humans and animals use both strategies; how exactly this holistic strategy could be implemented is a question we address towards the end.

The existence of directed and random exploration in behavior

While the success of directed and random exploration in artificial systems suggests that they may be good strategies for humans and animals, identifying these strategies in behavior proved difficult because of a number of psychological confounds. For example, directed exploration is often opposed by risk and ambiguity aversion [25, 26], while random exploration is difficult to disentangle from boredom or disengagement from the task. As a result, early work looking for directed exploration led to mixed results (e.g. [27, 28]), and the interpretation of choice variability as random exploration has been controversial [29, 30, 31].

A key advance in proving the existence of directed and random exploration was the design of tasks that manipulate how valuable it is to explore, independent of confounding factors. This approach builds on normative models, which identify the environmental features that *should* influence exploration because they alter the utility of new information. By manipulating these environmental features, we create conditions where exploration has value and conditions where it does not. We then identify directed and random exploration by looking for changes in information seeking and behavioral variability between these conditions.

Manipulations that increase the value of exploration include: increasing the time horizon [32, 15], increasing the uncertainty participants have about different options [33], and adding completely novel options that participants have not seen before [34, 35, 36, 37, 38]. Conversely, manipulations that increase the value of exploitation include raising the stakes of a one-off decision and decoupling information from choice [30]. In all of these cases, information seeking and behavioral variability increase when it is valuable to explore and decrease when it is valuable to exploit, providing strong evidence for the existence of both directed and random exploration in behavior.

The existence of directed and random exploration in the brain

Findings from neuroscience support the existence of directed and random exploration and further suggest that the neural processes underlying them may be dissociable.

Directed exploration, but not random exploration, has been associated with several prefrontal structures and systems, including activity in frontal pole [39, 40], mesocorticolimbic regions [36, 37], frontal theta oscillations [41], and prefrontal dopamine [28, 35, 42, 43]. A causal role for frontal regions in directed exploration is suggested by [44], who showed that continuous theta-burst transcranial magnetic stimulation to frontal pole abolishes horizon-dependent directed exploration, but leaves random exploration untouched. Building on a long literature in navigation and foraging [45], Johnson and colleagues [46] proposed that hippocampus should also play a key role in directed exploration.

The unique neural correlates of random exploration are less circumscribed. A number of studies have suggested that increased neural variability as a mechanism for generating exploratory noise. For example, random exploration is associated with a sudden increase in neural variability and a loss of choice tuning in decision-making circuits in monkeys [47]. Likewise, increased neural variability in motor related circuits in both humans [40], rodents [48], and birds [49, 50] is also related to the use of random exploration. However, choice tuning and changes in neural variability have also been associated with directed exploration [51] and could reflect other computations [31].

There is also some evidence that random exploration may be modulated by catecholamines norepinephrine and dopamine. Norepinephrine levels and related measures, like pupil size [52, 53], predict increased noise in behavior [54, 55, 56, 57] and the brain [58, 31]. However, results from systemic pharmacological studies tend to be more mixed [59, 60], and one study even found that blocking norepinephrine only affected exploration of completely novel options, a form of directed exploration [38], that is also modulated by dopamine [35]. Dopamine has been also been associated with random exploration, with decreased tonic dopamine associated increased behavioral variability [61] in rats, and dopaminergic neurons being found to modulate song variability in song birds [62].

Directed and random exploration develop differently over the lifespan

Multiple lines of evidence show that directed and random exploration develop differently over the lifespan. Even infants and preschoolers recognize uncertainty in their environment and make exploratory choices to deconfound variables [63], resolve belief violations [64, 65], and reduce ambiguity in intuitive theories [66]. Traditional bandits tasks reveal that the trade-off between exploration and exploitation shifts over development. Preschoolers are more likely to engage in directed exploration than 7–9 year-olds who are more directed than adults [67, 68].

In naturalistic tasks, preschoolers are sensitive to reward expectation and balance this against factors like whether information is decoupled from reward [69, 70]. However, in more cognitively-demanding tasks, factors like horizon appear to more weakly influence directed exploration for young teens (age 10–12), and grow in influence through adolescence and adulthood (early 20s) [71], remaining stable into old age [72].

Conversely, random exploration as behavioral variability, (e.g. noise over choice behavior) is prevalent in preschoolers and decreases through adulthood [67, 68, 73]. However, horizon-dependent random exploration appears more constant over adolescence and young adulthood [71], though is also reduced in healthy older adults [72]. Taken together, these results suggest that some forms of directed and random exploration may be most prevalent in early childhood and diminish with age, but that sensitivity to factors like horizon may follow different trajectories, highlighting the dissociability of the different types of exploration.

Integrating directed and random exploration

Although directed and random exploration rely on dissociable cognitive and neural systems, it is clear that both are used in behavior. In this section we discuss holistic models for how directed and random exploration could be combined.

Perhaps the simplest holistic strategy is to combine the information bonus and decision noise into one equation to estimate the value of choosing a particular option [15, 6]

$$Q(a) = r(a) + IB(a) + n(a)$$

In this case, directed and random exploration have additive effects on value, with exploration occurring whenever the information bonus or decision noise tips the balance away from the exploitative choice.

Another approach would be to mix the distinct strategies for directed exploration, random exploration, and exploitation over time. In this case, a higher-order process decides *when* to explore (and perhaps also whether to use a directed or random strategy), while the decision of *what* to choose is driven by component processes of exploitation, and directed and random exploration. Consistent with this separation of what and when, several studies argue for dissociable exploratory and exploitative states, on the basis of qualitative differences in behavior and neural activity [74, 75, 53, 36]. Further, directed exploration can occur at random times, an observation that is difficult to reconcile with a unified mechanism for deciding when and what to explore [76].

A third approach to combining directed and random exploration, contends that the two strategies emerge as different behavioral facets of a unified algorithm known in the machine learning literature as Deep Exploration [77]. In this model [78], the explore-exploit choice is made by mental simulation of a small number of plausible futures that are 'deep,' in that they extend multiple time steps into the future, but narrow, in that the number of simulations used is small. Random exploration arises from this model because the simulations are stochastic. More subtly, directed exploration also arises from Deep Exploration if the mental simulations extend multiple time steps into he future and include lose-shift behavior, i.e. mentally simulating a switch away from an explore option if the simulated outcome is bad. Such lose-shift simulations limit the simulated downside of exploring, which boosts the simulated value of exploring and hence biases the exploration towards informative options.

Beyond the existence of directed and random exploration, Deep Exploration also accounts for the horizon dependence, uncertainty dependence and feedback dependence of directed and random exploration. Moreover, it predicts that there should be a tradeoff between directed and random exploration. As people use more simulations to make their decision they should exhibit more directed exploration and less random exploration, a prediction that holds both across the population and within subject [78].

Open questions

What should we continue to exploit?

There are outstanding issues in almost every section above and much future work will involve exploiting these ideas. Open questions include: What are the neural mechanisms of directed and random exploration? And how do they relate to other cognitive processes such as working memory [79] and motor control [47, 40, 48, 49, 50]? How does explore-exploit behavior in general (and directed and random exploration in particular) vary across species and cultures? Are individual differences in directed and random exploration stable traits? Do directed and random exploration change in mental illness [80, 81, 82, 83, 84, 85]? And is modifying explore-exploit, with neural stimulation or behavioral therapy, a potential therapeutic target [86, 44]?

In regard to the computations underlying explore-exploit decisions: Are they really made by mental simulation as implied by Deep Exploration? And if Deep Exploration is really at play, how can this unifying theory explain the dissociation between directed and random exploration? Are there separate decision processes for 'when' versus 'what' to explore? And, more generally, are exploration and exploitation distinct behavioral states or are they drives that work together in a more continuous fashion?

What is left to explore?

There are a number of more open-ended questions we should explore, although this is by no means an exhaustive list.

How does explore-exploit behavior relate to other constructs?—Because we still lack precise operational definitions, there is perennial confusion about the relationship between explore-exploit behavior and a number of other related constructs. These include 'classic' measures in behavioral economics such as risk taking, ambiguity attitude, and loss aversion, as well as measures from ecology such as foraging, stopping problems, and search

[19]. The relationship with foraging and search has been questioned empirically [87] and whether these are distinct behaviors or share some variance is a key question.

Outside of the physical world, explore-exploit problems may also occur in the mental domain. Indeed, mental search and theory change has framed learning as an explore-exploit problem in which we choose between exploiting our current beliefs about the world or mentally exploring for new hypotheses (see, e.g. [88, 89, 90, 91]). The degree to which the processes of mental search align with those in the physical world remain a topic for future work; additionally behavioral methods to tease apart directed versus random search in mental space remain elusive.

Finally, it is an open question how explore-exploit behavior relates to emotions such as curiosity, interest, and affect [92, 93, 94, 95, 96, 97]. While these concepts are thought to inspire or inhibit exploration or exploitation, the relationship between them and explore-exploit behavior has not been sufficiently explored. There is a conceptual bias that exploration is an intrinsically appetitive act, intended to maximize future rewards and minimize future losses. However, depending on the decision context and choice outcomes exploring could be construed as aversive and avoided. By examining directed and random exploration in both appetitive and aversive environments we can understand how explore-exploit decisions differ from simpler approach-avoidance behaviors.

How do we explore complex environments?—Many explore-exploit papers begin in the real-world, with a motivating example like the restaurant example in this paper, but they almost always end up in the lab, with participants performing highly controlled tasks with few options and simple probability distributions over rewards. The real world does not look like this, and the obvious question is whether that matters for explore-exploit behavior or not?

That is, are directed and random exploration enough for the complexity of real-world explore-exploit problems? Or are we missing something fundamental from the theory? A new type of exploration perhaps [38]? Or exploration based on stereotyped behaviors, that are not based on information or randomization, but are nevertheless effective [98]?

Relatedly, does the complexity of the real world, make implementing the more sophisticated algorithms for directed and random exploration too difficult? For example, Thompson Sampling and Deep Exploration require the ability to simulate future outcomes, a computation that gets more difficult as the complexity of those outcomes increases. Conversely, simpler strategies such as a fixed bias for novel stimuli [99] or adding decision noise that is not tuned to simulated outcomes may perform well enough in more complex settings that a more difficult computation is not justified.

Towards this end of understanding how directed and random exploration are implemented in more complex settings, is recent work by [100] who used a task with hundreds of options and a complex reward structure. While this increased complexity required a new model of *learning* to explain behavior, the decision process, based on an information bonus and decision noise, did not. Similarly, work on exploration using online food orders found evidence for uncertainty driven exploration in this real-world case [14]. As impressive as this work is, we have only just begun to scratch the surface of the complexity of the real world and there is, quite literally, a whole world out there left to explore!

Acknowledgements

This work was supported by grants from the National Institute on Aging (R01AG061888 to RCW), the National Science Foundation (SES #1627971 to EB), the McDonnell Foundation (to EB and to RCW), the Jacobs Foundation (to EB), a Brain and Behavior Foundation Young Investigator Award (#27298 to RBE) and an Unfettered Research Grant from the Momental Foundation (to RBE).

References

[1]. Kaelbling Leslie Pack, Littman Michael L, and Moore Andrew W. Reinforcement learning: A survey. Journal of artificial intelligence research, 4:237–285, 1996.

[2]. Sutton Richard Sand Barto Andrew G. Reinforcement learning: An introduction. MIT press, 2018.

[3]. Cohen Jonathan D, McClure Samuel M, and Yu Angela J. 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, 2007.

[4]. Hills Thomas T, Todd Peter M, Lazer David, Redish A David, Couzin Iain D, Cognitive Search Research Group, et al. Exploration versus exploitation in space, mind, and society. Trends in cognitive sciences, 19(1):46–54, 2015. [PubMed: 25487706]

[5]. Mehlhorn Katja, Newell Ben R, Todd Peter M, Lee Michael D, Morgan Kate, Braithwaite Victoria A, Hausmann Daniel, Fiedler Klaus, and Gonzalez Cleotilde. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. Decision, 2(3):191, 2015.

[6]. Schulz Eric and Gershman Samuel J. The algorithmic architecture of exploration in the human brain. Current opinion in neurobiology, 55:7–14, 2019. [PubMed: 30529148]

[7]. Krebs John R, Kacelnik Alejandro, and Taylor Peter. Test of optimal sampling by foraging great tits. Nature, 275(5675):27–31, 1978.

[8]. Sims David W, Southall Emily J, Humphries Nicolas E, Hays Graeme C, Bradshaw Corey JA, Pitchford Jonathan W, James Alex, Ahmed Mohammed Z, Brierley Andrew S, Hindell Mark A, et al. Scaling laws of marine predator search behaviour. Nature, 451(7182):1098–1102, 2008. [PubMed: 18305542]

[9]. Thatcher Harriet R, Downs Colleen T, and Koyama Nicola F. Anthropogenic influences on the time budgets of urban vervet monkeys. Landscape and Urban Planning, 181:38–44, 2019.

[10]. Jackson Brian J, Fatima Gusti Lulu, Oh Sujean, and Gire David H. Many paths to the same goal: balancing exploration and exploitation during probabilistic route planning. Eneuro, 7(3), 2020.

[11]. Reid Chris R, Latty Tanya, Dussutour Audrey, and Beekman Madeleine. Slime mold uses an externalized spatial "memory" to navigate in complex environments. Proceedings of the National Academy of Sciences, 109(43):17490–17494, 2012.

[12]. Larcom Shaun, Rauch Ferdinand, and Willems Tim. The benefits of forced experimentation: striking evidence from the london underground network. The Quarterly Journal of Economics, 132(4):2019–2055, 2017.

[13]. O'Farrell Shay, Sanchirico James N, Spiegel Orr, Depalle Maxime, Haynie Alan C, Murawski Steven A, Perruso Larry, and Strelcheck Andrew. Disturbance modifies payoffs in the explore-exploit trade-off. Nature communications, 10(1):1–9, 2019.

[14]. Schulz Eric, Bhui Rahul, Love Bradley C, Brier Bastien, Todd Michael T, and Gershman Samuel J. Structured, uncertainty-driven exploration in real-world consumer choice. Proceedings of the National Academy of Sciences, 116(28):13903–13908, 2019.** Leverages a large data set of food orders to show that real-world explore-exploit decisions are modulated by uncertainty

[15]. Wilson Robert C, Geana Andra, White John M, Ludvig Elliot A, and Cohen Jonathan D. Humans use directed and random exploration to solve the explore–exploit dilemma. Journal of Experimental Psychology: General, 143(6):2074, 2014. [PubMed: 25347535]

[16]. Bellman Richard. A problem in the sequential design of experiments. Sankhyā: The Indian Journal of Statistics (1933–1960), 16(3/4):221–229, 1956.

[17]. Gittins John C. Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society: Series B (Methodological), 41(2):148–164, 1979.

[18]. Zhang Shunan and Angela J Yu. Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In Advances in neural information processing systems, pages 2607–2615, 2013.

[19]. Averbeck Bruno B. Theory of choice in bandit, information sampling and foraging tasks. PLoS computational biology, 11(3), 2015.

[20]. Bubeck Sébastien and Cesa-Bianchi Nicolo. Regret analysis of stochastic and non-stochastic multi-armed bandit problems. arXiv preprint arXiv:1204.5721, 2012.

[21]. Thompson William R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3/4):285–294, 1933.

[22]. Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. PhD thesis, King's College, Cambridge, 1989.

[23]. Auer Peter, Cesa-Bianchi Nicolo, and Fischer Paul. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2–3):235–256, 2002.

[24]. Agrawal Shipra and Goyal Navin. Analysis of thompson sampling for the multi-armed bandit problem. In Conference on learning theory, pages 39–1, 2012.

[25]. Ellsberg Daniel. Risk, ambiguity, and the savage axioms. The quarterly journal of economics, pages 643–669, 1961.

[26]. Camerer Colin and Weber Martin. Recent developments in modeling preferences: Uncertainty and ambiguity. Journal of risk and uncertainty, 5(4):325–370, 1992.

[27]. Daw Nathaniel D, O'doherty John P, Dayan Peter, Seymour Ben, and Dolan Raymond J. Cortical substrates for exploratory decisions in humans. Nature, 441(7095):876–879, 2006. [PubMed: 16778890]

[28]. Frank Michael J, Doll Bradley B, Oas-Terpstra Jen, and Moreno Francisco. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature neuroscience, 12(8):1062, 2009. [PubMed: 19620978]

[29]. Nassar Matthew R and Frank Michael J. Taming the beast: extracting generalizable knowledge from computational models of cognition. Current opinion in behavioral sciences, 11:49–54, 2016. [PubMed: 27574699]

[30]. Findling Charles, Skvortsova Vasilisa, Dromnelle Remi, Palminteri Stefano, and Wyart Valentin. Computational noise in reward-guided learning drives behavioral variability in volatile environments. Nature neuroscience, pages 1–12, 2019.** Demonstrates that most noise in reinforcement learning tasks is learning noise that has autocorrelation over time. However, the noise used for random exploration is different and appears to have no autocorrelation.

[31]. Muller Timothy H, Mars Rogier B, Behrens Timothy E, and O'Reilly Jill X. Control of entropy in neural models of environmental state. Elife, 8:e39404, 2019. [PubMed: 30816090]

[32]. Kacelnik Alejandro. Studies of foraging behaviour and time budgeting in great tits (Parus major). PhD thesis, University of Oxford, 1979.

[33]. Gershman Samuel J. Deconstructing the human algorithms for exploration. Cognition, 173:34–42, 2018. [PubMed: 29289795]

[34]. Wittmann Bianca C, Daw Nathaniel D, Seymour Ben, and Dolan Raymond J. Striatal activity underlies novelty-based choice in humans. Neuron, 58(6):967–973, 2008. [PubMed: 18579085]

[35]. Costa Vincent D, Tran Valery L, Turchi Janita, and Averbeck Bruno B. Dopamine modulates novelty seeking behavior during decision making. Behavioral neuroscience, 128(5):556, 2014. [PubMed: 24911320]

[36]. Costa Vincent D, Mitz Andrew R, and Averbeck Bruno B. Subcortical substrates of explore-exploit decisions in primates. Neuron, 103(3):533–545, 2019. [PubMed: 31196672] ** Monkeys use directed exploration to manage explore-exploit tradeoffs and these signals are coded in motivational brain regions (amygdala and ventral striatum).

[37]. Costa Vincent Dand Averbeck Bruno B. Primate orbitofrontal cortex codes information relevant for managing explore–exploit tradeoffs. Journal of Neuroscience, 40(12):2553–2561, 2020. [PubMed: 32060169]

[38]. Dubois Magda, Habicht Johanna, Michely Jochen, Moran Rani, Dolan Ray, and Hauser Tobias. Noradrenaline modulates tabula-rasa exploration. bioRxiv, 2020.

[39]. Badre David,Doll Bradley B, Long Nicole M, and Frank Michael J. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron, 73(3):595–607, 2012. [PubMed: 22325209]

[40]. Tomov Momchil S, Truong Van Q, Hundia Rohan A, and Gershman Samuel J. Dissociable neural correlates of uncertainty underlie different exploration strategies. Nature communications, 11(1):1–12, 2020.** Uses neuroimaging to identify different areas of the brain associated with directed and random exploration.

[41]. Cavanagh James F, Figueroa Christina M, Cohen Michael X, and Frank Michael J. Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. Cerebral cortex, 22(11):2575–2586, 2012. [PubMed: 22120491]

[42]. Gershman Samuel J and Tzovaras Bastian Greshake. Dopaminergic genes are associated with both directed and random exploration. Neuropsychologia, 120:97–104, 2018. [PubMed: 30347192]

[43]. Chakroun Karima, Mathar David, Wiehler Antonius, Ganzer Florian, and Peters Jan. Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. Elife, 9:e51260, 2020. [PubMed: 32484779]

[44]. Zajkowski Wojciech K, Kossut Malgorzata, and Wilson Robert C. A causal role for right frontopolar cortex in directed, but not random, exploration. Elife, 6:e27430, 2017. [PubMed: 28914605]

[45]. O’keefe John and Nadel Lynn. The hippocampus as a cognitive map. Oxford: Clarendon Press, 1978.

[46]. Johnson Adam, Varberg Zachary, Benhardus James, Maahs Anthony, and Schrater Paul. The hippocampus and exploration: dynamically evolving behavior and neural representations. Frontiers in human neuroscience, 6:216, 2012. [PubMed: 22848196]

[47]. R Becket Ebitz Eddy Albarran, and Moore Tirin. Exploration disrupts choicepredictive signals and alters dynamics in prefrontal cortex. Neuron, 97(2):450–461, 2018. [PubMed: 29290550] ** Suggests a possible mechanism for generating random exploration with random neural activity

[48]. Murakami Masayoshi, Shteingart Hanan, Loewenstein Yonatan, and Mainen Zachary F. Distinct sources of deterministic and stochastic components of action timing decisions in rodent frontal cortex. Neuron, 94(4):908–919, 2017. [PubMed: 28521140]

[49]. Dhawale Ashesh K, Smith Maurice A, and Ölveczky Bence P. The role of variability in motor learning. Annual review of neuroscience, 40:479–498, 2017.

[50]. Kojima Satoshi, Kao Mimi H, Doupe Allison J, and Brainard Michael S. The avian basal ganglia are a source of rapid behavioral variation that enables vocal motor exploration. Journal of Neuroscience, 38(45):9635–9647, 2018. [PubMed: 30249800]

[51]. Ebitz R Becket, Tu Jiaxin Cindy, and Hayden Benjamin Y. Rule adherence warps decision-making. BioRxiv, 2019.

[52]. Joshi Siddhartha and Gold Joshua I. Pupil size as a window on neural substrates of cognition. Trends in Cognitive Sciences, 2020.

[53]. Ebitz R Becket and Moore Tirin. Both a gauge and a filter: Cognitive modulations of pupil size. Frontiers in neurology, 9:1190, 2019. [PubMed: 30723454]

[54]. Jepma Marieke and Nieuwenhuis Sander. Pupil diameter predicts changes in the exploration–exploitation trade-off: Evidence for the adaptive gain theory. Journal of cognitive neuroscience, 23(7):1587–1596, 2011. [PubMed: 20666595]

[55]. Tervo DG, Proskurin M, Manakov M, Kabra M, Vollmer A, Branson K, and Karpova AY. Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. Cell, 159(1):21–32, 2014. [PubMed: 25259917]

[56]. Ohira Hideki, Ichikawa Naho, Kimura Kenta, Fukuyama Seisuke, Shinoda Jun, and Yamada Jitsuhiro. Neural and sympathetic activity associated with exploration in decision-making: further

evidence for involvement of insula. Frontiers in behavioral neuroscience, 8:381, 2014. [PubMed: 25426038]

[57]. Ebitz R Becket, Pearson John M, and Platt Michael L. Pupil size and social vigilance in rhesus macaques. Frontiers in neuroscience, 8:100, 2014. [PubMed: 24834026]

[58]. O Martins Ana Raquel and Froemke Robert C. Coordinated forms of noradrenergic plasticity in the locus coeruleus and primary auditory cortex. Nature neuroscience, 18(10):1483–1492, 2015. [PubMed: 26301326]

[59]. Warren Christopher M, Wilson Robert C, Van Der Wee Nic J, Giltay Eric J, Van Noorden Martijn S, Cohen Jonathan D, and Nieuwenhuis Sander. The effect of atomoxetine on random and directed exploration in humans. PloS one, 12(4):e0176034, 2017. [PubMed: 28445519]

[60]. Jepma Marieke, Te Beek Erik T, Wagenmakers Eric-Jan, Van Gerven Joop, and Nieuwenhuis Sander. The role of the noradrenergic system in the exploration-exploitation trade-off: a pharmacological study. Frontiers in human neuroscience, 4:170, 2010. [PubMed: 21206527]

[61]. Cinotti François, Fresno Virginie, Aklil Nassim, Coutureau Etienne, Girard Benoît, Marchand Alain R, and Khamassi Mehdi. Dopamine blockade impairs the exploration-exploitation trade-off in rats. Scientific reports, 9(1):1–14, 2019. [PubMed: 30626917]

[62]. Budzillo Agata, Duffy Alison, Kimberly E Miller, Adrienne L Fairhall, and David J Perkel. Dopaminergic modulation of basal ganglia output through coupled excitation–inhibition. Proceedings of the National Academy of Sciences, 114(22):5713–5718, 2017.

[63]. Schulz Laura E and Bonawitz Elizabeth Baraff. Serious fun: preschoolers engage in more exploratory play when evidence is confounded. Developmental psychology, 43(4):1045, 2007. [PubMed: 17605535]

[64]. Bonawitz Elizabeth Baraff, van Schijndel Tessa JP, Friel Daniel, and Schulz Laura. Children balance theories and evidence in exploration, explanation, and learning. Cognitive psychology, 64(4):215–234, 2012. [PubMed: 22365179]

[65]. Stahl Aimee Eand Feigenson Lisa. Observing the unexpected enhances infants' learning and exploration. Science, 348(6230):91–94, 2015. [PubMed: 25838378]

[66]. Wang J, Yang Y, Macias C, and Bonawitz E. Children with more immature intuitive theories seek domain-relevant information. in revision.

[67]. Schulz Eric, Wu Charley M, Ruggeri Azzurra, and Meder Björn. Searching for rewards like a child means less generalization and more directed exploration. Psychological science, 30(11):1561–1572, 2019. [PubMed: 31652093] ** Investigates the developmental profile of directed and random exploration in a bandits task with a large number of options.

[68]. Meder Björn, Wu Charley M, Schulz Eric, and Ruggeri Azzurra. Development of directed and random exploration in children. 2020.

[69]. Bonawitz Elizabeth, Bass Ilona, and Lapidow Elizabeth. Choosing to learn: Evidence evaluation for active learning and teaching in early childhood In Active Learning from Infancy to Childhood, pages 213–231. Springer, 2018.

[70]. Lapidow E and Bonawitz E. Explore-exploit: Ambiguity, expectation, and information gain influence preschooler's choices in exploration. in review.

[71]. Somerville Leah H, Sasse Stephanie F, Garrad Megan C, Drysdale Andrew T, Akar Nadine Abi, Insel Catherine, and Wilson Robert C. Charting the expansion of strategic exploratory behavior during adolescence. Journal of experimental psychology: general, 146(2):155, 2017. [PubMed: 27977227]

[72]. Mizell Jack-Morgan, Wang Siyu, Frisvold Alec, Alvarado Lily, Alex Farrell-Skupny Waitsang Keung, Sundman Mark H., Franchetti Mary-Kathryn, Chou Ying-hui, Alexander Gene E., and Wilson Robert C.. Differential impacts of healthy cognitive aging on directed and random exploration.

[73]. Plate Rista C, Fulvio Jacqueline M, Shutts Kristin, Green C Shawn, and Pollak Seth D. Probability learning: Changes in behavior across time and development. Child development, 89(1):205–218, 2018. [PubMed: 28121026]

[74]. Steyvers Mark, Lee Michael D, and Wagenmakers Eric-Jan. A bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology, 53(3):168–179, 2009.

[75]. D Lee Michael, Zhang Shunan, Munro Miles, and Steyvers Mark. Psychological models of human and optimal performance in bandit problems. Cognitive Systems Research, 12(2):164–174, 2011.

[76]. Ebitz R Becket, Sleezer Brianna J, Jedema Hank P, Bradberry Charles W, and Hayden Benjamin Y. Tonic exploration governs both flexibility and lapses. PLoS computational biology, 15(11):e1007475, 2019. [PubMed: 31703063] * Directed exploration occurs even when exploration has no value, suggesting a random mechanism for deciding when to explore.

[77]. Osband Ian, Blundell Charles, Pritzel Alexander, and Van Roy Benjamin. Deep exploration via bootstrapped dqn. In Advances in neural information processing systems, pages 4026–4034, 2016.

[78]. Wilson Robert, Wang Siyu, Sadeghiyeh Hashem, and Cohen Jonathan D. Deep exploration as a unifying account of explore-exploit behavior. 2020.** Proposes Deep Exploration as a unifying account of explore-exploit behavior.

[79]. Dezza Irene Cogliati, Cleeremans Axel, and Alexander William. Should we control? the interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. Journal of Experimental Psychology: General, 148(6):977, 2019. [PubMed: 30667262]

[80]. Averbeck Bruno B, Djamshidian Atbin, O'Sullivan Sean S, Housden Charlotte R, Roiser Jonathan P, and Lees Andrew J. Uncertainty about mapping future actions into rewards may underlie performance on multiple measures of impulsivity in behavioral addiction: Evidence from parkinson's disease. Behavioral neuroscience, 127(2):245, 2013. [PubMed: 23565936]

[81]. Strauss Gregory P, Frank Michael J, Waltz James A, Kasanova Zuzana, Herbener Ellen S, and Gold James M. Deficits in positive reinforcement learning and uncertainty-driven exploration are associated with distinct aspects of negative symptoms in schizophrenia. Biological psychiatry, 69(5):424–431, 2011. [PubMed: 21168124]

[82]. Wiehler Antonius, Chakroun Karima, and Peters Jan. Attenuated directed exploration during reinforcement learning in gambling disorder. BioRxiv, page 823583, 2019.

[83]. Dezza Irene Cogliati, Noel Xavier, Cleeremans Axel, and Yu Angela. Novelty-seeking impairment in addiction. BioRxiv, 2020.

[84]. Waltz JA, Wilson RC, Albrecht MA, Frank MJ, and Gold JM. Differential effects of psychotic illness on directed and random exploration. Computational Psychiatry, 2020.* Schizophrenia is associated with changes in directed exploration, but not random exploration

[85]. Cathomas Flurin, Klaus Federica, Guetter Karoline, Chung Hui-Kuan, Beharelle Anjali Raja, Spiller Tobias, Schlegel Rebecca, Seifritz Erich, Hartmann-Riemer Matthias, Tobler Philippe N, et al. Increased random exploration in schizophrenia is associated with inflammation. bioRxiv, 2020.

[86]. Beharelle Anjali Raja, Polanía Rafael, Hare Todd A, and Ruff Christian C. Transcranial stimulation over frontopolar cortex elucidates the choice attributes and neural mechanisms used to resolve exploration–exploitation trade-offs. Journal of Neuroscience, 35(43):14544–14556, 2015. [PubMed: 26511245]

[87]. von Helversen Bettina, Mata Rui, Samanez-Larkin Gregory R, and Wilke Andreas. Foraging, exploration, or search? on the (lack of) convergent validity between three behavioral paradigms. Evolutionary Behavioral Sciences, 12(3):152, 2018.

[88]. Bonawitz Elizabeth, Denison Stephanie, Griffiths Thomas L, and Gopnik Alison. Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. Trends in cognitive sciences, 18(10):497–500, 2014. [PubMed: 25001609]

[89]. Bonawitz Elizabeth, Denison Stephanie, Gopnik Alison, and Griffiths Thomas L. Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. Cognitive psychology, 74:35–65, 2014. [PubMed: 25086501]

[90]. Bonawitz Elizabeth, Ullman Tomer D, Bridgers Sophie, Gopnik Alison, and Tenenbaum Joshua B. Sticking to the evidence? a behavioral and computational case study of micro-theory change in the domain of magnetism. Cognitive science, 43(8):e12765, 2019. [PubMed: 31446650]

[91]. Ullman Tomer D, Goodman Noah D, and Tenenbaum Joshua B. Theory learning as stochastic search in the language of thought. Cognitive Development, 27(4):455–480, 2012.

[92]. Berlyne Daniel E. Curiosity and exploration. Science, 153(3731):25–33, 1966. [PubMed: 5328120]

[93]. Kidd Celeste and Hayden Benjamin Y. The psychology and neuroscience of curiosity. Neuron, 88(3):449–460, 2015. [PubMed: 26539887]

[94]. Gottlieb Jacqueline and Oudeyer Pierre-Yves. Towards a neuroscience of active sampling and curiosity. Nature Reviews Neuroscience, 19(12):758–770, 2018. [PubMed: 30397322]

[95]. Geana Andra, Wilson Robert, Daw Nathaniel D, and Cohen Jonathan D. Boredom, information-seeking and exploration. In CogSci, 2016.

[96]. Hidi Suzanne and Renninger K Ann. The four-phase model of interest development. Educational psychologist, 41(2):111–127, 2006.

[97]. Colantonio Joseph and Bonawitz Elizabeth. Awesome play: Awe increases preschooler's exploration and discovery. 2018.

[98]. Towal R Blythe and Hartmann Mitra JZ. Variability in velocity profiles during freeair whisking behavior of unrestrained rats. Journal of neurophysiology, 100(2):740–752, 2008. [PubMed: 18436634]

[99]. Redgrave Peter and Gurney Kevin. The short-latency dopamine signal: a role in discovering novel actions? Nature reviews neuroscience, 7(12):967–975, 2006. [PubMed: 17115078]

[100]. Wu CM, Schulz E, Speekenbrink M, Nelson JD, and Meder B. Generalization guides human exploration in vast decision spaces. Nature human behaviour, 2(12):915–924, 2018.* Explore-exploit behavior probed using a bandits task with a large number of options. Also notable for the sophisticated computational model needed to analyze the data.

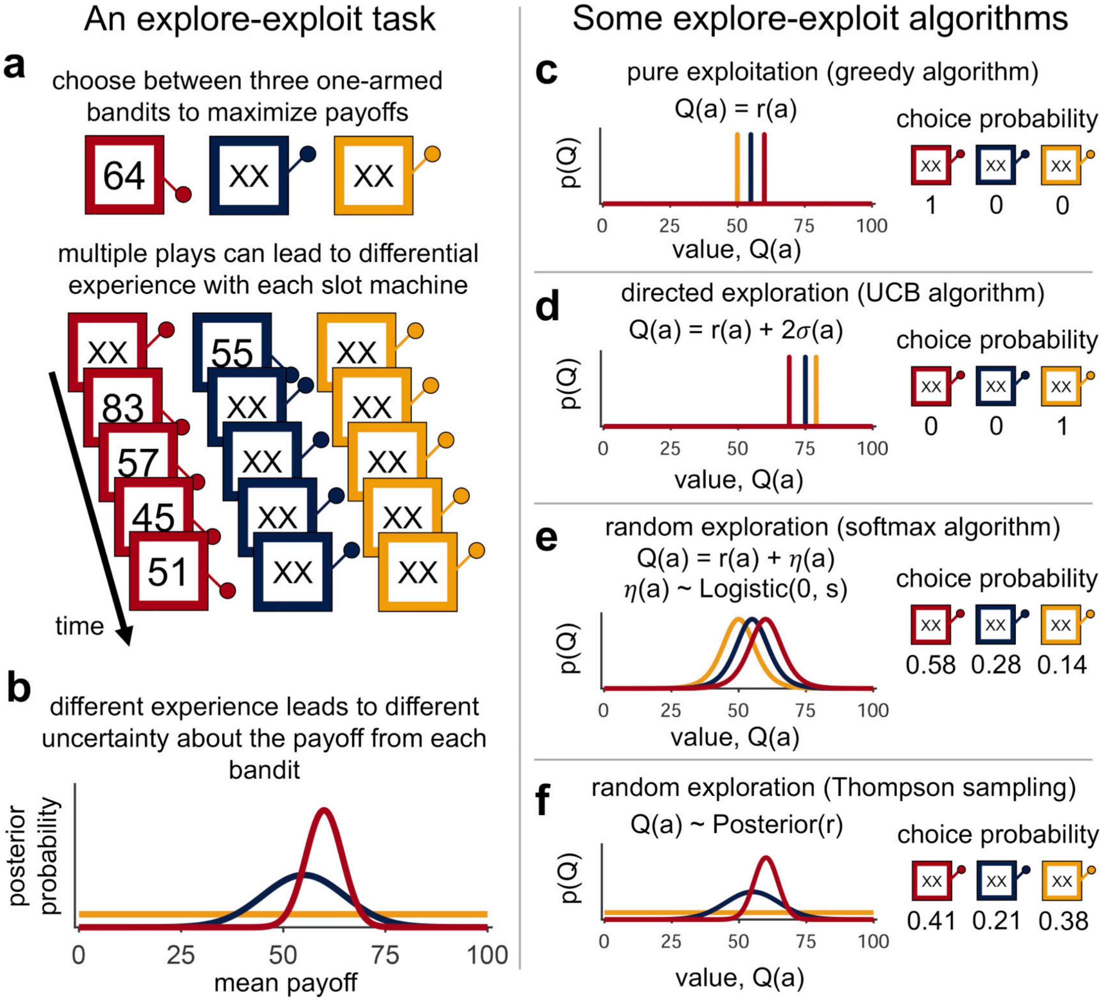

Figure 1:

Illustration of a simple explore-exploit task and algorithms that attempt to solve it. (**a**) In this ‘bandits task,’ participants choose between three slot machines, or ‘one-armed bandits,’ loosely based on those in a casino. In this case the bandits pay out rewards sampled from Gaussian distributions whose mean is different for each machine. Participants are not told these mean payoffs and so must explore in order to learn which bandit is best. However, because they receive the reward from the bandit they choose, they also face pressure to exploit the bandit they currently believe to be best. After several plays, participants can have quite different information about each bandit, which can be summarized as a posterior distribution over the mean payoff from each option, (**b**). In this case, the participant has the most information about the red option, which they also believe to be best (narrow distribution with highest mean), some information about the blue distribution (wide distribution, lower mean), and no information about the yellow option (uniform distribution). (**c** - **f**) Different algorithms make use of the information in different ways. (**c**) A purely exploitative algorithm sets the value of each option, $Q(a)$, deterministically to the mean of the posterior, $r(a)$. This algorithm always exploits the option with highest $r(a)$, in this case choosing the red option with 100% probability. (**d**) A purely directed algorithm (in this case the Upper Confidence Bound algorithm, [23]) adds a deterministic information bonus, set to twice the standard deviation of the posterior distribution for each bandit, to the value of each option. This information bonus is largest for the yellow option and the algorithm chooses this exploratory option with 100% probability. (**e**) A simple random strategy for exploration is to add random noise to the value of all options [22]. This algorithm is not deterministic, and any option can be chosen. However, because the values still include the expected payoff, this option is most likely to choose the option it believes to be best, the red option. (**f**) A more sophisticated random exploration strategy is Thompson sampling [21]. In this algorithm the value of each option is sampled randomly from the posterior distribution over the mean payoff from each option. The mean of this sampled value is still equal to the expected reward, but its variability - i.e. the decision noise - scales with the uncertainty in each option. Thus random exploration in Thompson sampling is highest when the algorithm is uncertain and reduces to zero when the payoff structure is fully known.