Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Exploration: from machines to humans

Lior Fox[1,5], Ohad Dan[2,3,5], Lotem Elber-Dorozko[1,5] and Yonatan Loewenstein[1,2,3,4]

Consider a wildlife photographer that has just entered a rainforest that she has never visited. Looking for a good spot for animal photos, she can spend all her time in the first hideout that she found, slowly learning which animals visit that spot. Alternatively, she can consider other locations, which are potentially better but might also be worse. To identify these better locations she needs to leave her hideout and walk further into the forest, thus missing the opportunity to learn more about the qualities of her first hideout. How should she explore the forest? How does she explore it? Here we describe the computational principles and algorithms underlying exploration in the field of Machine Learning and discuss their relevance to human behavior.

Addresses
[1] The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Israel
[2] Dept. of Cognitive Sciences, The Hebrew University, Jerusalem, Israel
[3] The Federmann Center for the Study of Rationality, The Hebrew University, Jerusalem, Israel
[4] The Alexander Silberman Institute of Life Sciences, The Hebrew University, Jerusalem, Israel

Corresponding author: Loewenstein, Yonatan (yonatan@huji.ac.il)
[5] These authors contributed equally.

Current Opinion in Behavioral Sciences 2020, **35**:104–111

This review comes from a themed issue on **Curiosity (Explore versus exploit)**

Edited by **Daphna Shohamy** and **Ran Hassin**

https://doi.org/10.1016/j.cobeha.2020.08.004

2352-1546/© 2020 Published by Elsevier Ltd.

> “ . . . As she gazed, she sniffed and sighed. ‘The sea is deep and the world is wide! How I long to sail!’ Said the tiny snail.”
> — Julia Donaldson, The Snail and The Whale [1]

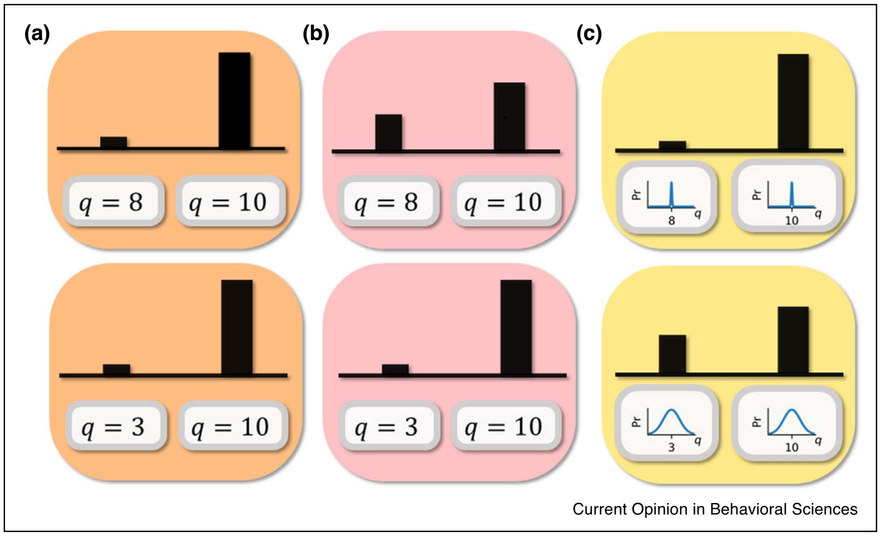
Introduction

Whether it is a wildlife photographer in a forest, looking for a good spot for animal photos or a rat in a subway station looking for food and shelter, exploring one’s environment is an essential component of Reinforcement Learning (RL). In Machine Learning (ML), exploration is typically studied in the framework of Markov Decision Processes (MDPs) [2,3]. MDPs are characterized by states and actions. Taking an action in a specific state can result in a transition of the agent to a different state and the delivery of a reward, with fixed probabilities. The Markov property dictates that given its current state and action, the transition to the next state and the delivery of reward are independent of the agent’s history of states, actions and rewards. Considering the photographer example, she is rewarded for taking good pictures of animals, whose probabilities depend on her current location (state). At every time point, she must decide which action to take, whether to take a photo, or to execute a different action, for example, to walk or to climb a tree, which will result in a change of her location. The goal of the agent in an MDP is to maximize the expected cumulative rewards (often with some discounting of future rewards). If the MDP is fully known, there exist efficient algorithms that can guide the agent to select, in each state, the optimal action with respect to its goal [2]. However, when the MDP is unknown, the agent must learn the optimal mapping from states to actions (‘policy’) by interacting with the environment. This learning requires exploration, and how to explore well is an active topic of research in ML.

Random exploration

To learn about the consequences of the different actions in the different states, all of the actions in all of the states must be taken. If the MDP is stochastic, they must be taken many times (in fact, infinitely many times). This can be achieved by choosing actions at random. However, this approach will not only perform poorly with respect to reward accumulation, learning this way will, in practice, be highly inefficient. This is because such exploration does not utilize the knowledge that has already been gained about the environment. Specifically, a photographer that has already identified several potentially good photo locations should give more attention to those spots, rather than explore spots that have previously proven to be lean. A standard solution to this problem is to utilize an estimate of the cumulative rewards following each action in each state, a quantity known as ‘*action-value*’, and to select with a higher probability actions which are associated with a higher action-value. This results in exploration that is still random, but is no longer uniform. Rather, it is biased in favor of actions which are deemed better. The most standard application of this approach in ML is known as ‘*ε-greedy*’ (Figure 1a): with high probability $(1-\varepsilon)$, the agent selects the alternative deemed best

Figure 1



Current Opinion in Behavioral Sciences

Random exploration strategies in the 2-armed bandit task.
(a) ε-*greedy*: the alternative associated with the larger action-value (q) is chosen with probability $(1-\varepsilon)$ and that associated with the smaller action-value is chosen with probability ε, independently of its action-value (compare top and bottom). **(b)** *Softmax*: the probability of choice is proportional to the (scaled) exponentiated action-value. As a result, the probability of choice depends on the specific action-values, and not only their ranking (compare top and bottom). **(c)** Thompson sampling: rather than using point estimates of the action-values, the agent estimates the action-values using probability distributions. In each trial, the agent samples an action-value from each distribution and greedily chooses the action associated with the largest sampled action-value. As a result, the probability of choice depends not only on the mean of the distribution but also on its higher moments. Specifically, in this example, an action associated with a smaller *mean* action-value may be chosen more often than one with a larger action-value if the *variance* over its distribution is larger (compare action 'Left' in Top and Bottom). Estimated actions-values q ((a) and (b)) and their distributions are presented in the rounded rectangles. Black bars' length depict the probabilities of choice of the corresponding actions. Top and Bottom panels portray different action-values.

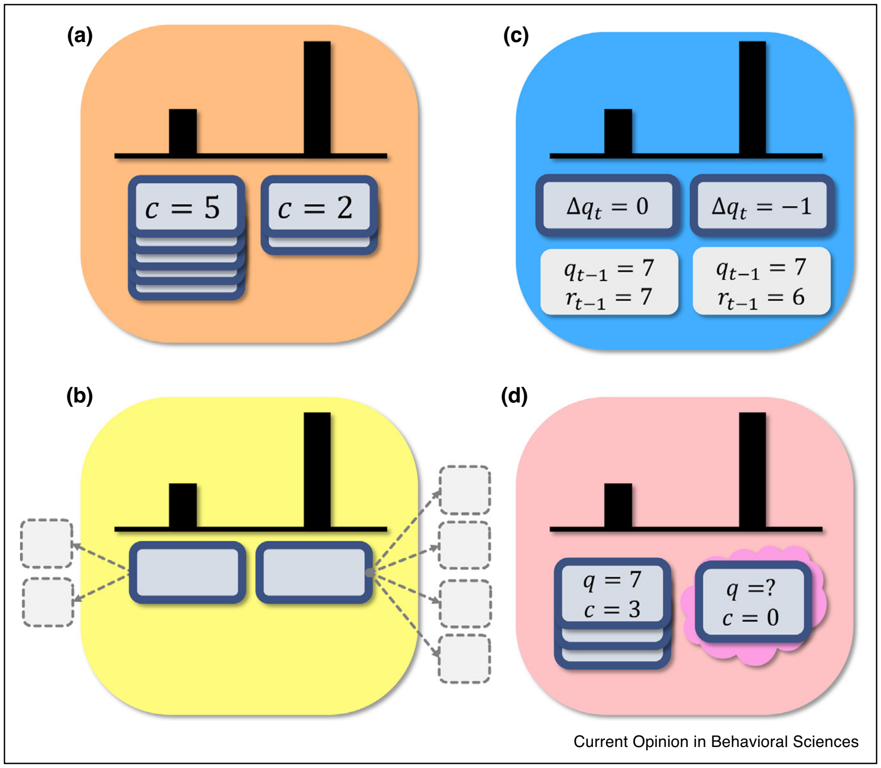
with respect to rewards (greedy). With a low probability (ε), the agent explores by randomly selecting another action. Exploration this way, however, does not distinguish between the non-greedy alternative actions. Therefore, a more graded approach, in which alternative actions that are deemed better are chosen with a higher probability is often used. Typically, this is achieved using a '*softmax*' function (Figure 1b), which can be justified as resulting from constraints on the entropy of the policy [4]. Finally, in *Thompson sampling* (Figure 1c) the posterior distributions over action-values are estimated, and actions are chosen by randomly sampling from these distributions and greedily choosing with respect to these random samples [5]. This allows for stochasticity, whose magnitude decreases with the certainty in the estimation of the action-values.

Directed exploration

The goal of exploration is to gain new knowledge. Therefore, exploration should ideally be directed towards actions that are more useful in that respect [6,7]. Choosing an action randomly, or according to its action-value is not efficient in that perspective. Rather, an agent can more efficiently explore if it tracks its own past behavior and chooses actions according to their predicted exploratory value. Methods that preferentially choose more uncertain options are termed '*directed exploration*'. A simple way of keeping track of how 'well-explored' a particular action is, is to use *visit counters* (Figure 2a). For each action and state, count how many times this action has been selected (in the given state) and prioritize actions that were previously selected less often [8–10]. In recent years, the concept of visit counters has been extended in several ways. Most notably, there are (a) techniques to apply counter-based methods in large or continuous problems (in which it is unfeasible, or not helpful, to actually 'count' visits of individual states) [11•,12,13,14•]; and (b) the introduction of *generalized counters* (Figure 2b), used to evaluate the long-term exploratory consequences of actions, beyond the immediate, one-step-ahead information represented by standard visit counters [14•,15].

Tracking its own learning process can also inform the agent about gaps in its knowledge about the world. A surprising outcome of an action in a particular state (relative to what the agent has predicted based on its past experience) is an indication of missing knowledge that should drive exploratory choices in that direction. For example, in many algorithms, the *reward prediction error*, a measure of the surprise (with respect to reward)

Figure 2



Directed exploration.

In directed exploration, actions associated with more uncertainty are chosen more often. Here we describe a few methods for directed exploration. **(a)** *Counters*: choices are biased in favor of actions that were previously chosen less often. **(b)** *Generalized counters*: choices are also biased in favor of actions that are likely to lead to *other* actions that were previously chosen less often. **(c)** *Surprise*: choices are also biased in favor of actions that yielded surprising results. The magnitude of the reward prediction error is one way of measuring ’surprise’. In this example, choice is biased in favor of the action associated with the larger magnitude reward prediction error, despite it being negative. **(d)** *Optimism*: action-values’ estimates are initialized using a large number. As a result, a greedy choice would initially favor those actions that were previously chosen less often. Estimated actions-values q, visit counters c and prediction errors Δq_t are presented in the rounded rectangles. Black bars’ length depict the probabilities of choice of the corresponding actions.

associated with the outcome of an action, is used to update the estimated value of the chosen action. This prediction error can also serve as a signal for guiding exploration (Figure 2c). This is because actions associated with high prediction error (in absolute value) are ones for which learning has probably not converged yet and thus requires further exploration [16,17]. The same logic can be applied to prediction errors arising in learning of quantities other than the expected reward, such as the prediction error for the next state given the current state and action [18]. Surprisingly, it turns out that even prediction errors arising from learning a fixed, random function, can be sufficient for successfully guiding effective exploration [19]. Other methods to quantify and utilize surprise use information-theoretic quantities such as *information gain* to guide exploration [20–22]. Finally, a popular method for exploration is known as *optimism in the face of uncertainty* [23,24]. The idea is to optimistically initialize the estimated action-values in the learning process (Figure 2d). If exploration is directed in favor of actions that seem more valuable then by construction, those actions less visited will be favored.

These different methods for directed exploration can be incorporated in the process of learning in various ways. An *exploration bonus* that is based on one of the principles outlined above can be added to the reward, such that reward-seeking will result also in exploration [9,11•,19]. Alternatively, action-selection can directly incorporate a term that favors exploration [8,14•]. Finally, these different principles can be combined. For example, optimism in the face of uncertainty can be combined with measures of uncertainty or missing knowledge such as counters. An agent can adopt an optimistic belief for actions which have not been explored enough yet, and trust its unbiased estimate for actions which have been explored sufficiently many times. This approach underlies several algorithms that are theoretically guaranteed to efficiently explore [25,26].

Studying exploration in humans using the bandit task

A most popular paradigm used to uncover the computational principles underlying exploration in humans is the bandit task (see for example Refs. [27•,28,29]). A participant is instructed to repeatedly choose between k alternatives (often, $k = 2$), that are characterized by different reward-distributions. To uncover exploration in this task, it is assumed that the participant has estimated the action-values associated with the different actions and that her overall objective is to maximize cumulative rewards. An action that is associated with the largest action-value is interpreted as reflecting the exploitation of the already-obtained information, while any deviation from such greedy behavior is interpreted as reflecting exploration, whose goal is to add information about the other action-values. The mapping from action-values to choices has been measured non-parametrically, revealing that humans utilize an action-selection function that combines ε-greedy and softmax functions [30•]. Later studies have revealed that the magnitude of exploration depends on its usefulness. Specifically, in a ‘horizon task’, in which the number of remaining trials is large, participants tend to explore more compared to tasks in which a single trial remains [31,32].

Several studies have shown that in addition to random exploration, uncertainty also directs human exploration [27•,29,33–35]. Developmental [36], genetic [37,38], imaging [39], pharmacological [40] and transcranial magnetic stimulation [41] studies suggest that anatomically distinct cognitive modules underlie random and directed explorations. Indeed, directed, but not random exploration is correlated with the extent to which participants care about future rewards (their temporal discounting function [32]). Similarly, frequent gamblers exhibit a specific reduction in directed exploration, but not in random exploration [42].

By construction, the bandit task cannot address a fundamental aspect of exploration — the long-term exploratory consequences of actions. For example, the photographer may choose to climb down a tree not because she is interested in photos associated with the climb, but because she is interested in moving to a different location in the forest. Studying this kind of exploration requires more complex experimental designs (see also below) [43].

Challenges in identifying human exploration in the bandit task

To relate participants choices to exploration, researchers typically estimate the action-values utilized by the participants (Figure 3a). This procedure implicitly postulates that participants indeed compute and utilize action-values in their learning behavior. However, there is no guarantee that this is indeed the case [44]. In fact, several operant learning algorithms that are devoid of any explicit or even implicit representation of action-values (e.g. based on policy gradient) (Figure 3b) explain behavior well in bandit-like tasks [45–47]. It is not even clear how to define exploratory behavior in the absence of value representation, as it can no longer be related to choosing lower-valued options. One may be tempted to identify stochastic choice with exploration. However, while the existence of an optimal deterministic policy is guaranteed in fully observable MDPs, this is not the case when considering reactive policies in the more realistic partially observable MDPs (POMDPs) [48,49]. On the other hand, some exploration algorithms are fully deterministic [14•].

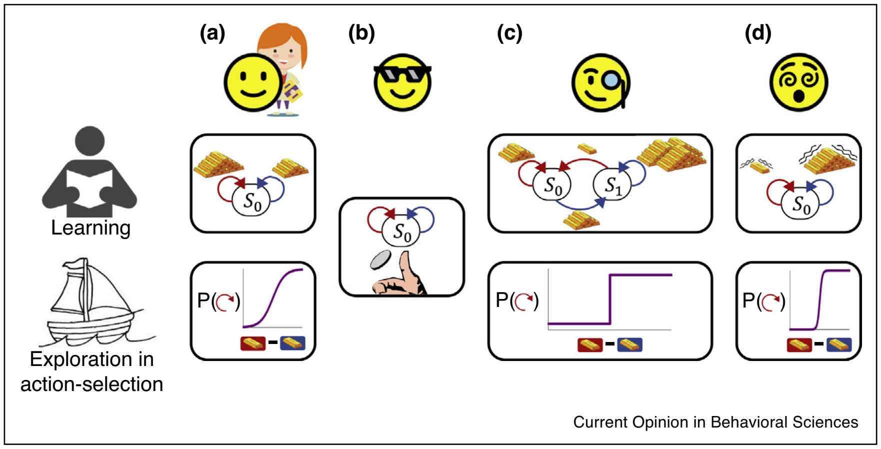
Moreover, in the framework of action-value estimation in the bandit task, it is typically assumed that the participants estimate action-values as if they are in a one-state MDP. However, it is well known that humans ’detect’ temporal structures even in random sequences [50,51]. This result suggests that participants are likely to utilize a more sophisticated model than a one-state MDP when tested in the bandit task (Figure 3c) [45,46]. Indeed, given the same sequence of outcomes, participants' behavior critically depends on whether they fully understand the stochastic mechanism that maps actions to rewards [52•]. The one-state MDP assumption is further challenged by the fact that in some tasks participants' behavior is consistent with the belief that they operate in a non-observable MDP (a POMDP with just one observation) [49]. Indeed, in many bandit experiments, the task is not a one-state MDP and the (unknown) reward probabilities change throughout the task. A recent study has demonstrated the difficulty in identifying exploratory behavior in the framework of action-value learning. Studying choices in a bandit task, it was shown that the majority of non-greedy decisions is due to limited computational precision rather than reflecting human exploration [53•] (Figure 3d).

Finally, the challenge of identifying the model underlying behavior is not unique to exploration. In general, the internal models participants employ are underdetermined by their behavior [54]. To deal with this issue, models are compared and their parameters are estimated using methods such as maximum-likelihood. However, despite substantial progress, a comprehensive understanding of human behavior in the bandit task is still lacking [55].

Ecological exploration

The k-armed bandit task is relatively easy to model and to relate to the general-purpose ML algorithms described above. However, it does not take into account an essential aspect of human learning and exploration — prior

Figure 3


(a)
(b)
(c)
(d)
Learning
Exploration in action-selection
S_0
S_1
P(↻)
Current Opinion in Behavioral Sciences

Repertoire of possible mental models.

It is difficult to identify exploration because the participant may utilize (unknown) different models when learning in the two-armed bandit task. **(a)** The participant (smiley) may assume that the world is a one-state (S_0) MDP (Top), learn the two action-values (gold bars) and choose between the two actions (arrows) using a softmax function (Bottom). This is the model researchers typically use to quantify behavior. **(b)** However, the participant may utilize a very different learning model. For example, the participant may learn the policy directly, without estimating action-values. In this case, it is not even clear how to define exploration. **(c)** The participant may assume an MDP that is more complex than the true one. She may also use a different action-selection function. **(d)** Finally, noise in the action-values' estimation may be erroneously interpreted as 'exploration'. This could lead to an underestimation of the slope of the action-selection function. Each model is described in one column, where the top panel depicts the learning and the bottom panel the action-selection. Gold bars, ship, reading person and scientist are adapted from Ref. [75].

knowledge about the structure of the MDP. Let us reconsider the photographer example. The photographer enters the forest, which she has never visited with extensive knowledge about it. For example, she knows that if she moves left — she will find herself to the left of her previous state. She knows that if she climbs up a tree, she will need to climb it down in order to move to a different location in the forest (unless she is Tarzan). These trivial facts, which will dominate the photographer's exploratory behavior, are typically lacking from the standard ML algorithms, which were constructed to learn general MDPs. The dependence of human learning (but not of machine learning) on such priors has been demonstrated in an experiment that compared computer-game learning of humans and machines. Humans learned the game much faster than machines. However, their learning ability substantially deteriorated when objects (ladders for climbing, demons as game-ending enemies) were masked by re-rendering their pixels. By contrast, the ML algorithm was insensitive to this manipulation [56]. Another study demonstrated that participants utilize spatial cues when learning in a bandit task with a large number of possible actions [57]. Even infants, the ultimate candidates to be considered as tabula-rasa agents, have expectations of their environment and insights on its structure [58,59]. It has been argued that the artificial environments that are utilized in lab experiments are too different from ecological-relevant exploration. As a result, the relevance of the resultant conclusions to natural behavior is questionable [60]. This lacuna can be addressed by utilizing more ecologically valid experimental paradigms [33,43].

Exploration has also been studied in the context of foraging, which is perhaps ecologically more relevant than the bandit task [61,62]. The foraging decision is whether to exploit a current option or explore, looking for a better one. The experimental design can be similar to that of the bandit task, but the magnitude or probability of reward diminishes with the number of times that the alternative was chosen. Foraging is typically analyzed in the framework of the Marginal Value Theorem [63], which describes the strategy that maximizes the cumulative rewards when returns decrease with time spent exploiting an option. This is because a general MDP that does not take into account prior knowledge about the diminishing nature of returns does not seem relevant to human behavior. This poses a challenge when attempting to relate contemporary machine-learning exploration algorithms to behavior in these foraging tasks [61].

Finally, people tend to overestimate the probability of positive outcomes, and underestimate that of negative outcomes, a phenomenon known as 'optimism bias' [64]. This could reflect a biased prior knowledge about the world. To the best of our knowledge this bias has not been directly linked to human exploration. It would be interesting to test whether it contributes to human exploration in a similar way that 'optimism in the face of uncertainty' contributes to exploration in ML.

Exploration and curiosity

Broadly speaking, curiosity is often defined as the desire for information [65–67]. In the framework of RL, curiosity has been traditionally related to exploration, either by using exploration as a measurement for curiosity [35,68,69], or by considering a (model of) curiosity as a form of an exploratory drive [7,18,20]. While curiosity in general, as well as other 'intrinsic' drives, might be broader than the notion of exploration in RL contexts [70,71], some hypotheses about curiosity can be directly formulated in the language of RL, and particularly exploration strategies [72]. For example, one theory states that novel objects create more curiosity [69] while another theory states that people are more curious about information gaps - specific cases of high uncertainty [73,74]. The first theory is in line with 'visit counters' exploration (Figure 2a), while the second is in line with exploration that is motivated by prediction-error or information-gain (Figure 2c).

Concluding remarks

Substantial progress has been made in recent years in the development of algorithms for efficient exploration, and in understanding the computational principles underlying human exploration. While bandit tasks have been pivotal for understanding many aspects of the computational principles underlying exploratory behavior, they failed to capture what we view as the major difference between human and machine exploration — the extensive use of prior knowledge in human learning. In machine learning, this prior knowledge is implicitly embedded in the specific hypothesis classes used for function approximation. This prior knowledge, however, is very different from that utilized by humans, as described above. One exception may be the weight sharing and local connectivity in convolutional neural networks, where prior knowledge about the homogeneity of low-level statistical dependencies in natural images is implemented in the structure and learning of the network. The difference between humans and machines may be easy to miss in bandit tasks, but it is easily seen in more ecological tasks that have a complex structure [43,61]. Such tasks will not only allow us to more fully understand human behavior, their focus on prior knowledge can aid us in creating ML algorithms that better solve real-life problems.

Conflict of interest statement

Nothing declared.

Funding

This work was supported by the Israel Science Foundation (Grants 757/16 and 3213/19), and by the Gatsby Charitable Foundation. Lotem Elber-Dorozko is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship and Ohad Dan would like to acknowledge the support of the The Hoffman Leadership and Responsibility Fellowship Program.

Acknowledgement

We thank Gianluigi Mongillo for carefully reading the manuscript and for his helpful comments.

References and recommended reading

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

- of special interest

1. Donaldson J: *The Snail and the Whale*. Puffin Books; 2006.
2. Sutton RS, Barto AG: *Reinforcement Learning: An Introduction*. MIT Press; 1998.
3. Kaelbling LP, Littman ML, Moore AW: **Reinforcement learning: a survey**. *J Artif Intell Res* 1996, **4**:237-285.
4. Achbany Y, Fouss F, Yen L, Pirotte A, Saerens M: **Tuning continual exploration in reinforcement learning: an optimality property of the Boltzmann strategy**. *Neurocomputing* 2008, **71**:2507-2520.
5. Russo DJ, Van Roy B, Kazerouni A, Osband I, Wen Z: **A tutorial on Thompson sampling**. *Found Trends Mach Learn* 2018, **11**:1-96.
6. Thrun SB: *Efficient Exploration in Reinforcement Learning*. 1992.
7. Schmidhuber J: **Curious model-building control systems**. *Proceedings of the IEEE International Joint Conference on Neural Networks* 1991:1458-1463.
8. Auer P, Cesa-Bianchi N, Fischer P: **Finite-time analysis of the multiarmed bandit problem**. *Mach Learn* 2002, **47**:235-256.
9. Strehl AL, Littman ML: **An analysis of model-based interval estimation for Markov decision processes**. *J Comput Syst Sci* 2008, **74**:1309-1331.
10. Kolter JZ, Ng AY: **Near-Bayesian exploration in polynomial time**. *Proceedings of the 26th Annual International Conference on Machine Learning* 2009:513-520.
11. • Bellemare M, Srinivasan S, Ostrovski G, Schaul T, Saxton D, Munos R: **Unifying count-based exploration and intrinsic motivation**. In *Advances in Neural Information Processing Systems 29*. Edited by Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R. Curran Associates, Inc.; 2016:1471-1479.

This paper studied how counter-based exploration can be applied in environments characterized by a large state-space (potentially continuous). The basic idea is to learn a density model over the states and use it to extract counter-like variables that can be used to drive exploration.

12. Ostrovski G, Bellemare MG, van den Oord A, Munos R: **Count-based exploration with neural density models**. *Proceedings of the 34th International Conference on Machine Learning* 2017:2721-2730. 70.
13. Tang H, Houthooft R, Foote D, Stooke A, Chen OX, Duan Y, Schulman J, DeTurck F, Abbeel P: **#Exploration: a study of count-based exploration for deep reinforcement learning**. *Advances in Neural Information Processing Systems* 2017:2753-2762.
14. • Fox L, Choshen L, Loewenstein Y: **DORA the explorer: directed outreaching reinforcement action-selection**. *International Conference on Learning Representations* 2018.

Standard RL algorithms are designed to maximize not only the immediate reward but also to take into consideration the long-term consequences of actions. This paper presents a novel algorithm that is based on a similar principle for exploration. It introduced a generalization of visit-counters, such that in states that can lead, in the future, to the exploration of less-visited states, the generalized counters grow more slowly than in 'less-promising' states. This approach can also be applied for large (or continuous) problems using function-approximation methods.

15. Oh M, Iyengar G: **Directed exploration in PAC model-free reinforcement learning**. *arXiv Prepr* 2018. arXiv180810552.

16. Tokic M, Palm G: **Value-difference based exploration: adaptive control between epsilon-greedy and softmax**. *KI 2011: Advances in Artificial Intelligence*. Springer; 2011:335-346.

17. Simmons-Edler R, Eisner B, Yang D, Bisulco A, Mitchell E, Seung S, Lee D: *QXplore: Q-learning Exploration by Maximizing Temporal Difference Error*. 2019.

18. Pathak D, Agrawal P, Efros AA, Darrell T: **Curiosity-driven exploration by self-supervised prediction**. *Proceedings of the 34th International Conference on Machine Learning* 2017:2778-2787.

19. Burda Y, Edwards H, Storkey A, Klimov O: **Exploration by random network distillation**. *International Conference on Learning Representations* 2019.

20. Still S, Precup D: **An information-theoretic approach to curiosity-driven reinforcement learning**. *Theory Biosci* 2012, **131**:139-148.

21. Little DY, Sommer FT: **Learning and exploration in action-perception loops**. *Closing Loop Around Neural Syst* 2014, **7**:37.

22. Houthooft R, Chen X, Duan Y, Schulman J, De Turck F, Abbeel P: **VIME: variational information maximizing exploration**. *Advances in Neural Information Processing Systems*. 2016:1109-1117.

23. Even-Dar E, Mansourt Y: **Convergence of optimistic and incremental Q-learning**. *Advances in Neural Information Processing Systems*. 2002:1499-1506.

24. Tosatto S, D'Eramo C, Pajarinen J, Restelli M, Peters J: **Exploration driven by an optimistic bellman equation**. *2019 International Joint Conference on Neural Networks (IJCNN)* 2019:1-8.

25. Kearns M, Singh S: **Near-optimal reinforcement learning in polynomial time**. *Mach Learn* 2002, **49**:209-232.

26. Brafman RI, Tennenholtz M: **R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning**. *J Mach Learn Res* 2003, **3**:213-231.

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

Do participants utilize directed exploration in two-armed bandit tasks? To address this question, the effects of uncertainties in the estimated action-values on participants' choice behavior were studied. The paper reports that uncertainty in an action-value affects both the slope of the action-selection function– an indication for sampling based random exploration, as well as the bias of the action-selection function – an indication for directed exploration. The conclusion is that participants' utilize both directed-exploration and random-exploration in their learning behavior.

28. 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.

29. Schulz E, Franklin NT, Gershman SJ: **Finding structure in multi-armed bandits**. *Cogn Psychol* 2020, **119**:101261.

30. Shteingart H, Neiman T, Loewenstein Y: **The role of first**
• **impression in operant learning**. *J Exp Psychol Gen* 2013, **142**:476-488.

In this paper, the behavior of human participants in a two-armed bandit task is analyzed. It characterizes non-parametrically the action-selection function in humans, underlying random exploration. Specifically, while humans are sensitive to the difference in action-values (as in softmax), they exhibit substantial exploration even when the difference between these values is large.

31. 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.

32. Sadeghiyeh H, Wang S, Alberhasky MR, Kyllo HM, Shenhav A, Wilson RC: **Temporal discounting correlates with directed exploration but not with random exploration**. *Sci Rep* 2020, **10**:4020.

33. Schulz E, Bhui R, Love BC, Brier B, Todd MT, Gershman SJ: **Structured, uncertainty-driven exploration in real-world consumer choice**. *Proc Natl Acad Sci U S A* 2019, **116**:13903-13908.

34. Gershman SJ: **Uncertainty and exploration**. *Decision* 2019, **6**:277-286.

35. Dubey R, Griffiths TL: **Reconciling novelty and complexity through a rational analysis of curiosity**. *Psychol Rev* 2020, **127**:455-476.

36. 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.

37. 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.

38. Gershman SJ, Tzovaras BG: **Dopaminergic genes are associated with both directed and random exploration**. *Neuropsychologia* 2018, **120**:97-104.

39. Tomov MS, Truong VQ, Hundia RA, Gershman SJ: **Dissociable neural correlates of uncertainty underlie different exploration strategies**. *Nat Commun* 2020, **11**:2371.

40. Warren CM, Wilson RC, 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.

41. Zajkowski WK, Kossut M, Wilson RC: **A causal role for right frontopolar cortex in directed, but not random, exploration**. *eLife* 2017, **6**:e27430.

42. Wiehler A, Chakroun K, Peters J: **Attenuated directed exploration during reinforcement learning in gambling disorder**. *bioRxiv* 2019 http://dx.doi.org/10.1101/823583.

43. Javadi A-H, Patai EZ, Margois A, Tan H-RM, Kumaran D, Nardini M, Penny W, Duzel E, Dayan P, Spiers HJ: **Spotting the path that leads nowhere: modulation of human theta and alpha oscillations induced by trajectory changes during navigation**. *bioRxiv* 2018 http://dx.doi.org/10.1101/301697.

44. Elber-Dorozko L, Loewenstein Y: **Striatal action-value neurons reconsidered**. *eLife* 2018, **7**:e34248.

45. Shteingart H, Loewenstein Y: **Reinforcement learning and human behavior**. *Curr Opin Neurobiol* 2014, **25**:93-98.

46. Mongillo G, Shteingart H, Loewenstein Y: **The misbehavior of reinforcement learning**. *Proc IEEE* 2014, **102**:528-541.

47. Loewenstein Y, Seung HS: **Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity**. *PNAS* 2006, **103**:15224-15229.

48. ICML: **Learning without state-estimation in partially observable Markovian decision processes**. *Proceedings of the Eleventh International Conference on International Conference on Machine Learning* 1994:284-292.

49. Loewenstein Y, Prelec D, Seung HS: **Operant matching as a nash equilibrium of an intertemporal game**. *Neural Comput* 2009, **21**:2755-2773.

50. Oskarsson AT, Van Boven L, McClelland GH, Hastie R: **What's next? Judging sequences of binary events**. *Psychol Bull* 2009, **135**:262-285.

51. Neiman T, Loewenstein Y: **Reinforcement learning in professional basketball players**. *Nat Commun* 2011, **2**:569.

52. Morse EB, Runquist WN: **Probability-matching with an**
• **unscheduled random sequence**. *Am J Psychol* 1960, **73**:603-607.

This paper describes participants' repeated choice behavior in two, very similar, two-alternative choice tasks. In the first, participants were instructed to predict whether a rod that is dropped would cross lines drawn on the floor. In the second, they had to predict which of two bulbs would turn on in the trial. Despite the fact that both groups of participants observed the exact same sequence of binary events, their behaviors differed. They tended to maximize in the first task and to probabilitymatch in the second. These results highlight the importance of a world model in learning.

53. Findling C, Skvortsova V, Dromnelle R, Palminteri S, Wyart V:
• **Computational noise in reward-guided learning drives behavioral variability in volatile environments**. *Nat Neurosci* 2019, **22**:2066-2077.

Non-greedy choices in a two-armed bandit task experiment are typically interpreted as reflecting exploration. By comparing variability in a partial-feedback task to that in a full-feedback task (in which no exploration is expected), it is shown that the majority of non-greedy decisions stem from learning noise, rather than reflecting exploration.

54. Ng AY, Russell SJ: **Algorithms for inverse reinforcement learning**. In *Proceedings of the Seventeenth International Conference on Machine Learning*; *Morgan Kaufmann Publishers Inc.: 2000:663-670*.

55. Dan O, Loewenstein Y: **From choice architecture to choice engineering**. *Nat Commun* 2019, **10**:2808.

56. Dubey R, Agrawal P, Pathak D, Griffiths TL, Efros AA: **Investigating human priors for playing video games**. *Proceedings of the 35th International Conference on Machine Learning* 2018:1349-1357. PMLR 80.

The authors systematically modified video-games' environment in order to mask visual information that could be used as priors. It turns out the human participants' learning in the game heavily relies on such priors. Specifically, they exhibit different patterns of learning and exploration in the 'masked' conditions. By contrast, artificial agents are largely unaffected by the masking of almost all visual priors.

57. Wu CM, Schulz E, Speekenbrink M, Nelson JD, Meder B: **Generalization guides human exploration in vast decision spaces**. *Nat Hum Behav* 2018, **2**:915-924.

58. Arterberry ME, Bornstein MH: **Three-month-old infants' categorization of animals and vehicles based on static and dynamic attributes**. *J Exp Child Psychol* 2001, **80**:333-346.

59. Setoh P, Wu D, Baillargeon R, Gelman R: **Young infants have biological expectations about animals**. *Proc Natl Acad Sci U S A* 2013, **110**:15937-15942.

60. Mobbs D, Trimmer PC, Blumstein DT, Dayan P: **Foraging for foundations in decision neuroscience: insights from ethology**. *Nat Rev Neurosci* 2018, **19**:419-427.

61. Kolling N, Akam T: **(Reinforcement?) Learning to forage optimally**. *Curr Opin Neurobiol* 2017, **46**:162-169.

62. Trapanese C, Meunier H, Masi S: **What, where and when: spatial foraging decisions in primates**. *Biol Rev* 2019, **94**:483-502.

63. Charnov EL: **Optimal foraging, the marginal value theorem**. *Theor Popul Biol* 1976, **9**:129-136.

64. Sharot T, Riccardi AM, Raio CM, Phelps EA: **Neural mechanisms mediating optimism bias**. *Nature* 2007, **450**:102-105.

65. Berlyne DE: **Curiosity and exploration**. *Science* 1966, **153**:25-33.

66. Voss H-G, Keller H: *Curiosity and Exploration Theories and Results*. Elsevier Inc.; 1983.

67. Kashdan TB, Stiksma MC, Disabato DJ, McKnight PE, Bekier J, Kaji J, Lazarus R: **The five-dimensional curiosity scale: capturing the bandwidth of curiosity and identifying four unique subgroups of curious people**. *J Res Pers* 2018, **73**:130-149.

68. Berlyne DE: **A theory of human curiosity**. *Br J Psychol Gen Sect* 1954, **45**:180-191.

69. Smock CD, Holt BG: **Children's reactions to novelty: an experimental study of "curiosity motivation"**. *Child Dev* 1962, **33**:631-642.

70. Gottlieb J, Oudeyer P-Y: **Towards a neuroscience of active sampling and curiosity**. *Nat Rev Neurosci* 2018, **19**:758-770.

71. Oudeyer P-Y, Kaplan F: **What is intrinsic motivation? A typology of computational approaches**. *Front Neurorobot* 2009, **1**:6.

72. Barto AG: **Intrinsic motivation and reinforcement learning**. In *Intrinsically Motivated Learning in Natural and Artificial Systems.* Edited by Baldassarre G, Mirolli M. Berlin Heidelberg: Springer; 2013:17-47.

73. Loewenstein G: **The psychology of curiosity: a review and reinterpretation**. *Psychol Bull* 1994, **116**:75-98.

74. Kang MJ, Hsu M, Krajbich IM, Loewenstein G, McClure SM, Wang JT, Camerer CF: **The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory**. *Psychol Sci* 2009, **20**:963-973.

75. Http://clipart-library.com/gold-cliparts.html, Https://www.pinterest.cl/pin/240450067594092613/, Https://www.clipart.email/download/11007237.html, Https://www.pngegg.com/en/png-bzpdh: Clipart websites. 2020.