Neuroscience

On the computational principles underlying human exploration

Lior Fox, **Ohad Dan**, **Yonatan Loewenstein**

The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem • Yale School of Medicine • Department of Cognitive Sciences, The Alexander Silberman Institute of Life Sciences, and The Federmann Center for the Study of Rationality

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted
October 10, 2023 (this version)

Sent for peer review
July 16, 2023

Posted to PsyArXiv
May 24, 2023

Abstract

Adapting to new environments is a hallmark of animal and human cognition, and Reinforcement Learning (RL) models provide a powerful and general framework for studying such adaptation. A fundamental learning component identified by RL models is that in the absence of direct supervision, when learning is driven by trial-and-error, *exploration* is essential. The necessary ingredients of effective exploration have been studied extensively in machine learning. However, the relevance of some of these principles to humans' exploration is still unknown. An important reason for this gap is the dominance of the Multi-Armed Bandit tasks in human exploration studies. In these tasks, the exploration component *per se* is simple, because local measures of uncertainty, most notably visit-counters, are sufficient to effectively direct exploration. By contrast, in more complex environments, actions have long-term exploratory consequences that should be accounted for when measuring their associated uncertainties. Here, we use a novel experimental task that goes beyond the bandit task to study human exploration. We show that when local measures of uncertainty are insufficient, humans use exploration strategies that propagate uncertainties over states and actions. Moreover, we show that the long-term exploration consequences are temporally-discounted, similar to the temporal discounting of rewards in standard RL tasks. Additionally, we show that human exploration is largely uncertainty-driven. Finally, we find that humans exhibit signatures of temporally-extended learning, rather than local, 1-step update rules which are commonly assumed in RL models. All these aspects of human exploration are well-captured by a computational model in which agents learn an exploration "value-function", analogous to the standard (reward-based) value-function in RL.

eLife assessment

This well-written and well-reasoned manuscript describes a behavioral and computational modeling study designed to understand pure exploration, where action selection is driven by pure information seeking and not rewards. Using a novel task, the authors find that a subset of people use information value to drive their selection behavior, consistent with a simple information maximization model of reinforcement learning. The rest of the participants did not exhibit this behavior. This **valuable** work provides intriguing, yet somewhat **incomplete**, insights into understanding directed exploration and its computational form.

Introduction

When encountered with a novel setting, animals and humans explore their environment. Such exploration is essential for learning which actions are beneficial for the organism and which should be avoided. The speed of learning, and even the learning outcome, crucially depends on the "quality" of that exploration: for example, if as a result of poor exploration some actions are never chosen, their effects are never observed, and hence cannot be learned. More generally, a fundamental difference between learning by trial and error and Supervised Learning scenarios is that in the latter, the distribution of examples is controlled by the "teacher", whereas in the former, the distribution of examples that the agent gets to observe depends on the agent's own behavioral policy. Therefore, in order to successfully learn a good policy by trial and error, agents need to take into account *uncertainty* when choosing actions, reflecting the fact that the observations collected so far might mis-represent the actual quality of the different actions.

Learning by trial and error is often abstracted in the framework of the computational problem of Reinforcement Learning (RL) (Sutton and Barto, 2018): An agent makes sequential decisions in an unknown environment; at each time-step, it observes the current state of the environment, and chooses an action from a set of possible actions. In response to this action, the environment transfers the agent to the next state, and provides a reward signal (which can also be zero or negative). The ultimate goal of the agent is to learn how to choose actions – i.e, learn a *policy* – such as to maximize some performance metric, typically the expected cumulative reward.

Exploration algorithms in RL differ in the particular way they address uncertainties. *Random exploration*, in which a random component is added to the policy (e.g., a policy otherwise maximizing based on current estimates) is, arguably, the simplest way of incorporating exploration. By adding randomness, the agent is bound to eventually accumulate information about all states and actions. More sophisticated exploration methods, referred to as *directed exploration* (Thrun, 1992), attempt to identify and actively choose the specific actions that will be more effective in reducing uncertainty. To do that, the agent needs to track and update some estimate or measures of uncertainty associated with different actions. For example, the agent can use visit-counters: keep track of the number of times each action was chosen in each state, and prioritize those actions that have previously been neglected (Auer et al., 2002; Bellemare et al., 2016; Tang et al., 2017; Ostrovski et al., 2017).

The intuition behind counter-based methods can be made precise in the important case of Multi-Armed Bandit problems (or bandit problems, for short). In a k-armed bandit, the environment is characterized by a single state and k actions ("arms"), each associated with a reward distribution. Because these distributions are unknown, and feedback (i.e., a sample from the distribution) is given only for the chosen arm at each trial, exploration is needed to guarantee that the best arm (i.e., the one associated with the highest expected reward) is identified. Bandit problems are theoretically well-understood, with various algorithms having optimality guarantees, under some statistical assumptions (for a comprehensive review see Lattimore and Szepesvári, 2020). Particularly, counter-based methods (e.g., UCB, Auer et al., 2002) can be shown to explore optimally in bandit tasks, in the online-learning sense of minimizing regret.

Human exploration has been studied extensively in bandit and bandit-like problems (Shteingart et al., 2013; Wilson et al., 2014; Mehlhorn et al., 2015; Gershman, 2018; Schulz et al., 2020). Because these are arguably the simplest form of RL problems, they offer a clean and potentially well-controlled framework for experiments (Fox et al., 2020). The strong theoretical foundations are another appeal for experimental work, because behavior can be compared with well-defined algorithms, and, potentially, also with an optimal solution.

However, generalizing conclusions about human exploration from behavior in bandit tasks to behavior in more complex environments is not trivial. In a bandit task, an action that was chosen less times is, everything else being equal, exploratory more valuable compared to one that was chosen more often. By contrast, visit-counters alone might be a poor measure of uncertainty in complex environments, because they completely ignore future consequences of the actions (**Figure 1a**). Indeed, the limitations of naive counter-based exploration in structured and complex environments have been discussed in the machine learning literature, and different exploration schemes that take into account the long-term exploratory consequences of actions have been proposed (**Storck et al., 1995**; **Meuleau and Bourgine, 1999**; **Osband et al., 2016a**,**b**; **Chen et al., 2017**; **Fox et al., 2018**).

Our goal here is to study the extent to which human exploration is sensitive to long-term consequences of actions, as opposed to counter-based exploration. Crucially, this question cannot be addressed in the common bandit problems paradigm, because general exploration algorithms are reduced to counter-based methods when they are faced with a bandit problem. Thus, even if humans do (approximately) use some general, beyond visit-counters, directed exploration strategies, they will likely manifest as counter-based strategies in bandit tasks. Therefore, we set out to study exploration in a novel task that addresses these issues. First, we show that humans take into account the long-term exploratory consequences of their actions when exploring complex environments (Experimental results). Next, we model this exploration using an RL-like algorithm, in which agents learn exploratory "action-values" and use these values to guide their exploration (Computational modeling).

Results

Experimental results

Sensitivity to future consequences of actions

To test the hypothesis that human exploration is sensitive to the long-term consequences of actions, we conducted an experiment that formalizes the intuition presented in the Introduction (see **Figure 1a**). In the experiment (denoted as "Experiment 1"), participants were instructed to explore a novel environment, a maze of rooms, by navigating through the doors connecting those rooms (**Figure 1b**). Each room was identified by a unique background, a title, and the number of doors in that room. No reward was given in this task, but participants were instructed to "understand how the rooms are connected" (see Methods). Testing participants in a task devoid of clear goal and rewards is somewhat unorthodox. We go back to this point in the Discussion section.

Three groups of participants were tested, each in a different maze as is described in **Figure 1c** (top): In all mazes, there was a start room (S) with two doors, each leading to a different room. One of these rooms, a multi-action room (M_R) was endowed with n_R doors, while the other, denoted as M_L, was endowed with n_L doors. All three mazes were unbalanced, in the sense that $n_R > n_L$. Between the different mazes, we varied $n_R - n_L$, while keeping $n_R + n_L = 7$ constant. The locations of the doors leading to M_R and M_L were counterbalanced across participants. For clarity of notation, we refer to them as "right" and "left", respectively. All other remaining rooms were endowed with only a single door. After going through these single-door rooms, a participant would reach a common terminal room (T). There, they were informed that they reached the end of the maze and then they were transported back to S. Overall, each participant visited S (of the one particular environment they were assigned to) 20 times.

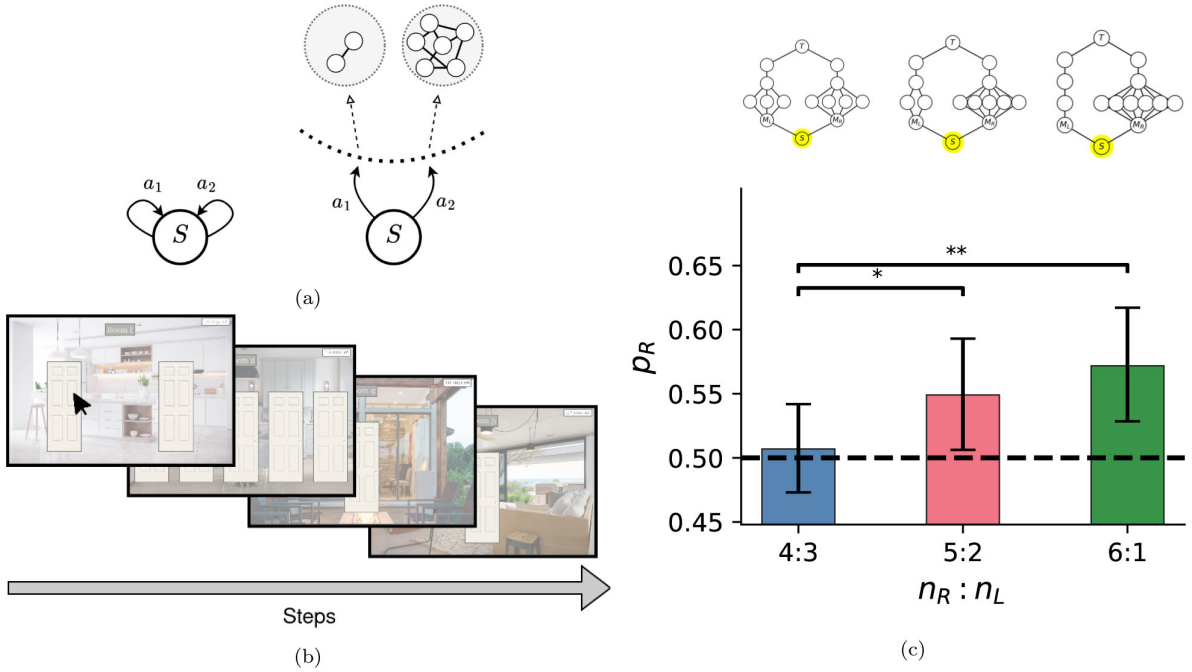

Figure 1

Directed exploration in complex environments.

(**a**) In a bandit problem (left), actions have no long-term consequences. In complex environments (right), actions have long-term consequences as particular actions might lead, in the future, to different parts of the state-space. In this example, these parts (shaded areas) are of different size. As a result, the local visit-counters are no longer a good measure of uncertainty. In this example, a_2 should be, in general, chosen more often compared to a_1 in order to exhaust the larger uncertainty associated with it. **(b)** Participants were instructed to navigate through a maze of rooms. Each room was identified by a unique background image and a title. To move to the next room, participants chose between the available doors by mouse-clicking. Background images and room titles (Armenian letters) were randomized between participants, and were devoid of any clear semantic or spatial structure. **(c)** The three maze structures in Experiment 1 (Top) have a root state *S* (highlighted in yellow) with two doors. They differ in the imbalance between the number of doors available in future rooms M_R and M_L (n_R : n_L – 4:3, 5:2, 6:1). Consistent with models of directed exploration that take into account long-term consequences of actions, and unlike counter-based models, participants exhibited bias towards room M_R, deviating from a uniform policy (Bottom, bars and error-bars denote mean and 95% confidence interval of p_R; number of participants: n = 161; 120; 137. Statistical significance, here and in following figures: * : p < 0.05, ** : p < 0.01; *** : p < 0.001).

Since there was no reward, all choices in this task are exploratory. If participant's exploration is driven by visit-counters, then we expect that the *frequencies* in which they choose each of the doors in S, denoted p_R and p_L, would be equal. By contrast, if they take into consideration the long-term consequences of their actions, then we would expect them to choose the right door more often (resulting in $p_R > p_L$). In line with the hypothesis that participants are sensitive to the long-term consequences of their actions, we found that averaged over all participants in the three conditions, $p_R > p_L$ (p_R = 0.54, 95% confidence interval: $p_R \in [0.518, 0.563]$). Considering each group of participants separately, significant bias in favor of p_R was observed in the 6:1 (p_R = 0.572, n = 137, 95% CI: [0.528, 0.617]) and the 5:2 groups (p_R = 0.549, n = 120, 95% CI: [0.506, 0.592]), but not in the 4:3 group (p_R = 0.507, n = 161, 95% CI: [0.472, 0.541]).

We hypothesized that the larger the imbalance ($n_R - n_L$), the stronger will be the bias towards M_R (larger p_R). To test this hypothesis, we compared the biases of participants in the different groups (**Figure 1c**). As expected, the average p_R in the 5:2 and 6:1 groups was significantly larger than that of the 4:3 group ($p < 0.05$ and $p < 0.01$ respectively, permutation test, see Methods). The average p_R in the 6:1 group was larger than that of the 5:2 group. However, this difference was not statistically significant ($p = 0.17$).

The results depicted in **Figure 1c** indicate that on average, human participants are sensitive to the exploratory long-term consequences of their actions. Considering individual participants, however, there was substantial heterogeneity in the biases exhibited by the different participants. While some chose the right door almost exclusively, others favored the left door. We next asked whether some of this heterogeneity across participants reflects more general individual-differences in exploratory strategies, which would also manifest in their exploration in other states. To test this hypothesis, we focused on state M_R. In this state, exploration is also required because there are n_R different alternatives to choose from. However, unlike in state S, these alternatives do not, effectively, have long-term consequences. As such, choosing an action in M_R is a bandit-like task. Thereofre, directed exploration in M_R is expected to be driven by visit-counters, such that participants would equalize the number of times each door in M_R is selected. Note that this is not a strong prediction, because random exploration will, on average, also equalize the number of choices of each door. Yet, directed and random exploration have diverging predictions with respect to the temporal pattern of choices in M_R. Specifically, with pure directed exploration (that is driven by visit-counters), participants are expected to avoid choosing the same door that they chose the last time that they visited M_R. Consequently, the probability of repeating the same choice in consecutive visits of M_R, which we denote by p_{repeat}, is expected to vanish. By contrast, random exploration predicts that $p_{\text{repeat}} = 1/n_R$. **Figure 2** (Top) depicts the histograms (over participants) of p_{repeat} in the three experimental conditions, demonstrating that participants exhibited substantial variability in p_{repeat}. While for some participants p_{repeat} was close to 0, as predicted by pure directed exploration, for others it was similar to $1/n_R$, as predicted by random exploration. Many other participants exhibited p_{repeat} that was even larger than $1/n_R$, indicating that, potentially, choice bias and / or momentum also influenced choices in the task. Based on the predictions of directed and random exploration, we divided participants into two groups, depending on the quality of exploration in M_R: "good" directed explorers, in which $p_{\text{repeat}} < 1/n_R$, and "poor" directed explorers, in which $p_{\text{repeat}} \geq 1/n_R$ (**Figure 2** Top, dots and diagonal stripes, respectively).

Is the quality of directed exploration in the bandit-like task of state M_R informative about directed exploration in S? To address this question, we computed the histograms of p_R separately for the "good" and "poor" directed explorers (**Figure 2** Bottom). Averaging within each group we found that indeed, p_R among the "poor" explorers was not significantly different from chance in any of the three conditions (**Figure 3a**), consistent with the predictions of random exploration. By contrast, among "good" explorers, there was a significant bias in the 5:2 (p_R = 0.597, n = 53, 95% CI: [0.537, 0.652]) and the 6:1 (p_R = 0.612, n = 71, 95% CI: [0.544, 0.678]) groups (**Figure 3b**). These findings show that participants that avoid repetition in the bandit task are also more sensitive to

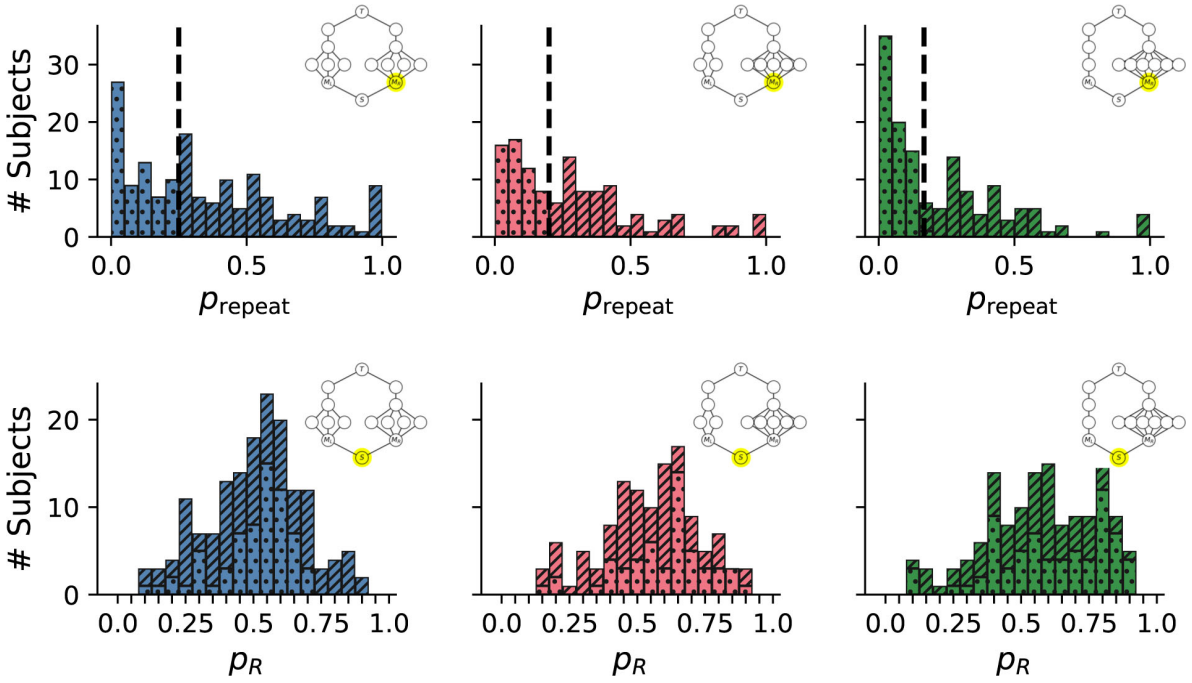

Figure 2

Heterogeneity in exploration strategies.

Top: Histograms of p_{repeat} at state M_R (highlighted in yellow) for participants in the three conditions of Experiment 1 (left to right: n_R = 4, 5, 6). Dashed vertical line represents the value expected by chance, $1/n_R$. Based on their p_{repeat} values, we divided participants into "good" and "poor" directed explorers (dotted and striped patterns, respectively; "good" explorers proportion: 40%, 44%, 51%). **Bottom**: Histograms of p_R at state S (highlighted in yellow), for the "good" and "poor" directed explorers groups.

the long-term exploratory consequences of their actions. We conclude that those participants who tend to perform good directed exploration in M_R also perform good directed exploration in S. Crucially, the *implementation* of directed exploration in the two states is rather different. In M_R, where different actions have no long-term consequences, "good" explorers rely on visit-counters that are the relevant measure of uncertainty, resulting in an overall uniform choice. By contrast in S, actions do have long-term consequences, and "good" explorers go beyond the visit-counters, biasing their choices in favor of the action associated with more future uncertainty.

Temporal discounting

In the previous section we showed that if the future exploratory consequences of the actions are one trial ahead, humans are sensitive to these consequences. It is well known that in humans and animals, the value of a reward is discounted with its delay (**Vanderveldt et al., 2016**). We hypothesized that similar temporal discounting will manifest in evaluating the exploratory "usefulness" of actions. To test this prediction, we conducted Experiment 2 on a new set of participants. Similar to Experiment 1, Experiment 2 consisted of 3 different maze structures. The imbalance between the number of possible outcomes was kept fixed across 3 mazes, at $n_R = 5$ and $n_L = 2$. However, the *depth* at which these outcomes occur, relative to the root state S, varied between 1 (as in Experiment 1) to 3 (**Figure 4**, Top). The depth of M_R determines the delay between the choice made at S and its exploratory benefit. In the presence of temporal discounting of exploration, we therefore expect p_R to decrease with the depth of M_R.

To test this prediction, we divided participants to "good" and "poor" directed explorers, as in Experiment 1, based on the degree of p_{repeat} in M_R. As depicted in **Figure 4**, both the "poor" and "good" explorers exhibited a bias in favor of "right" in S. For the "good" explorers, a larger delay was also associated with a smaller bias.

The dynamics of exploration

Insofar, we demonstrated that human participants exhibit directed exploration in which they take into their considerations the future exploratory consequences of their action. To better understand the computational principles underlying this directed exploration, we revisit the question of why explore in the first place. One possible answer to this question is that exploration is required for learning. According to this view, actions are favorable from an exploratory point of view when they are associated with, or lead to other actions associated with, high uncertainty, missing knowledge, and other related quantities (**Schmidhuber, 1991**; **Still and Precup, 2012**; **Little and Sommer, 2014**; **Houthooft et al., 2016**; **Pathak et al., 2017**; **Burda et al., 2019**). An alternative, that has received some attention in the machine learning literature, is that exploration could be driven by its own normative objective (**Machado and Bowling, 2016**; **Hazan et al., 2019**; **Zhang et al., 2020**; **Zahavy et al., 2021**). For example, such objective could be to maximize the entropy of the discounted distribution of visited states and chosen actions (**Hazan et al., 2019**). Experimentally, the difference between the two approaches will be particularly pronounced towards the end of a long experiment. When all states and actions had been visited sufficiently many times, everything that can be learned has already been learned. Thus, if the goal of exploration is to facilitate learning, then exploratory behavior is expected to fade over time. By contrast, if exploration is driven by a normative objective, then we generally expect behavior to converge to a one that (approximately) maximizing this objective, and hence maintaining asymptotic exploratory behavior.

Specifically considering Experiment 1 and 2, we do not expect any bias in S ($p_R = 0.5$) in the beginning of the task, because participants are naive and are unaware of the different long-term consequences of the two actions. With time and learning, we expect participants to favor M_R over M_L ($p_R > 0.5$). This prediction holds either if participants are driven by the goal of reducing the (long-term) uncertainty associated with M_R, or by the goal of optimizing some exploration objective, such as to match the choices per door in M_R and M_L. In other words, both approaches

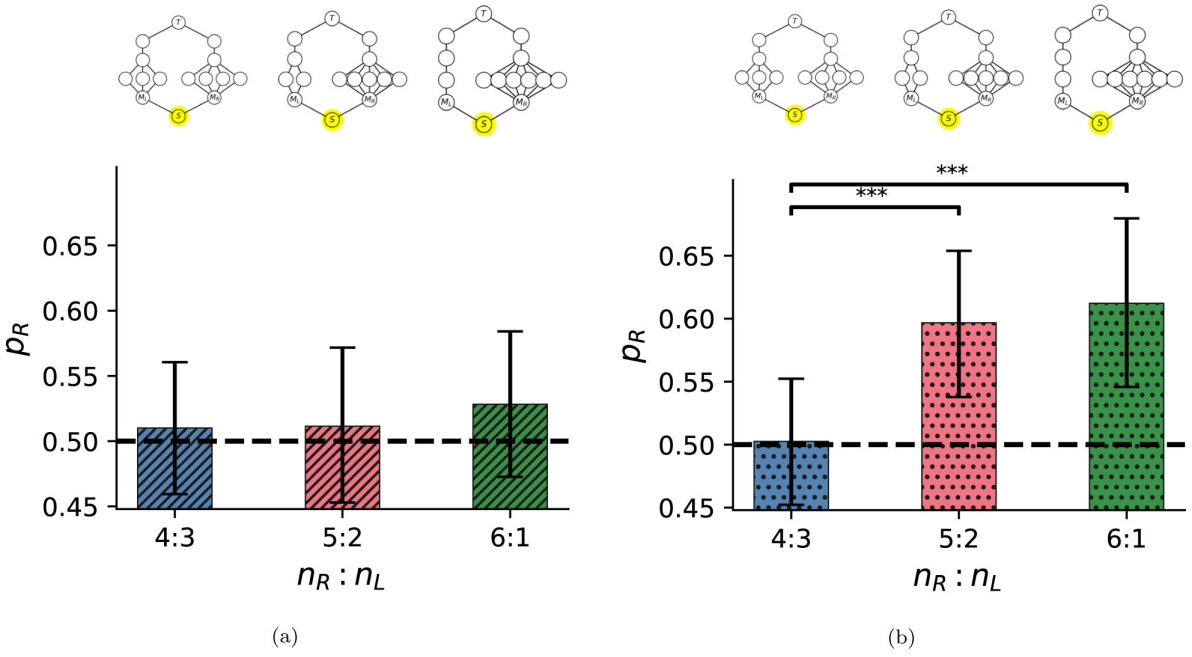

Figure 3

"Poor" and "good" directed explorers.

Choice biases at state *S* (p_R) analyzed separately for "poor" and "good" explorers (striped and dotted patterns; divided based on their exploration in M_R, see **Figure 2**) in the 3 conditions of Experiment 1. While behavior of the "poor" explorers was not significantly different from chance (consistent with the prediction of random exploration), "good" explorers in the $n_R = 5$, 6 conditions exhibited significant bias towards "right". Bars and error bars denote mean and 95% confidence interval of p_R; number of participants n = 95; 66 67; 53, 66; 71 ("poor"; "good").

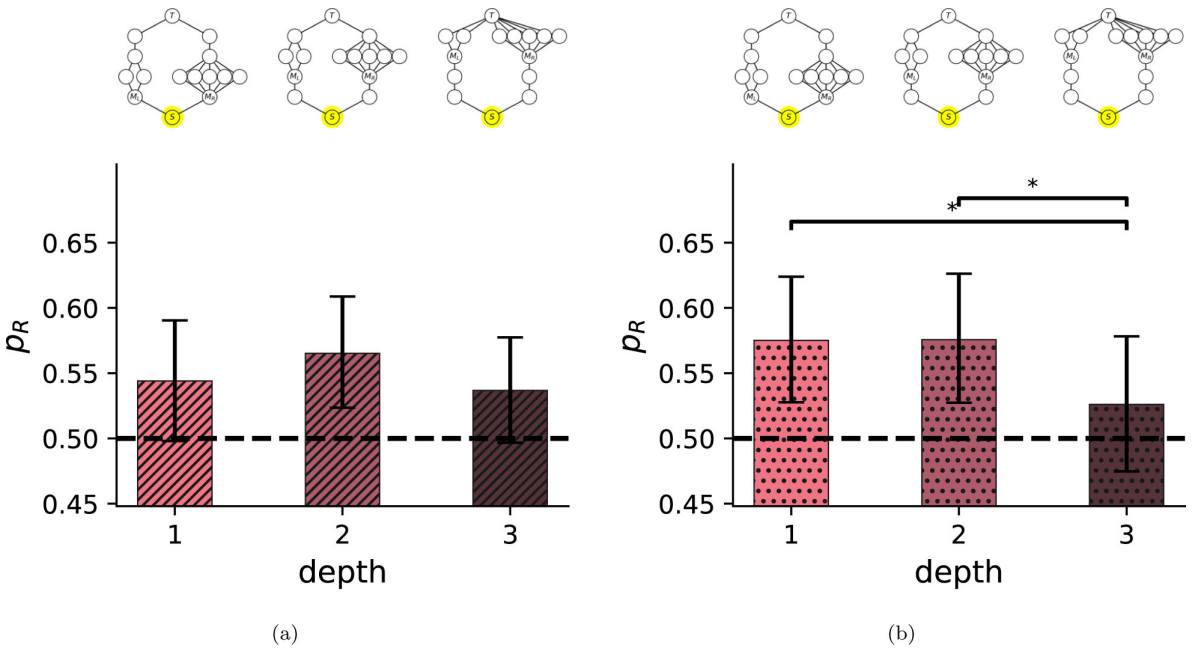

Figure 4

Temporal discounting of exploratory consequences.

The three mazes in Experiment 2 (Top) had the same imbalance ($n_R = 5$, $n_L = 2$), however we varied the depth of M_R (and M_L) relative to the root state S (left to right: depth = 1, 2, 3). "Poor" and "good" directed explorers (striped and dotted patterns, respectively) were divided by their p_{repeat} value at M_R (same as in Experiment 1, see **Figure 2**). Bars and error-bars denote mean and 95% confidence interval of p_R. Number of participants n = 99; 92, 121; 84, 153; 85 ("poor"; "good").

predict that with time, p_R will increase. With more time elapsing, however, the predictions of the two approaches diverge. As uncertainty decreases, uncertainty-driven exploration predicts a decay of p_R to its baseline value $\left(p_R^* = 0.5\right)$. By contrast, the normative approach predicts that p_R will converge to a $p_R^* > 0.5$ steady-state.

Figure 5 depicts the temporal dynamics of p_R (t), as a function of the number of times t that S was visited (defined as "episodes"). The learning curves are shown separately for the "poor" (**Figure 5a**) and "good" (**Figure 5b**) explorers, averaged over all 6 conditions of Experiments 1 and 2. As expected, there was no preference in the first episodes. However, with time, the participants developed a bias in favor of M_R, which was more pronounced in the "good" directed explorers group. In this group, participants exhibited a significant bias, p_R (t) > 0.5 from the 3rd episode. Notably, this increased bias was followed by a decrease to a steady state bias value (episodes 10 – 20). This steady state value was lower than its peak transient value (consistent with uncertainty-driven exploration), but was higher than baseline level before learning (consistent with a normative exploration objective).

Computational modeling

The model

Together, the two experiments of the previous sections provide us with the following insights: (1) Humans exploration is affected by long-term consequences of actions (**Figure 1c**); (2) Both the number of future states and their depth affect this exploration (**Figure 3** and **Figure 4**); and finally, (3) Exploration dynamics peaks transiently and then decays, consistent with an uncertainty-driven exploration (**Figure 5**).

In theorizing about effective exploration we have alluded to concepts such as "exploratory value" or "usefulness" of particular actions, but did not provide a precise working definition for it. In this section we consider a specific computational model for directed exploration, and test this model in view of these experimental findings. The model is a general-purpose algorithm for directed exploration, which formalizes the intuition that the challenge of exploration in complex environments is analogous to the standard credit-assignment problem in RL (in the reward-maximization sense).

According to the model, the agent observes the current state of the environment s at each time-step and chooses an action a from the set of possible actions. In response to this action, the environment transfers the agent to the next state s', at which the agent chooses action a'. Each state-action pair (s, a) is associated with an exploration value, denoted E (s, a) (**Fox et al., 2018**). These exploration values represent a current estimate of "missing knowledge", such that a high value indicates that further exploration of that action is beneficial. At the beginning of the process, E-values are initialized to a positive constant (specifically E = 1), representing the largest possible missing knowledge. Each transition from s, a to s', a' triggers an update to E (s, a) according to the following update rule:

$$E\left(s,a\right) \leftarrow E\left(s,a\right) + \eta\left(-E\left(s,a\right) + \gamma E\left(s',a'\right)\right)$$

In words, the change in E (s, a) is a sum of two contributions. The first, $-E$ (s, a), is the immediate reduction in the uncertainty regarding state s and action a due to the current visit of that state-action. The second, γE (s', a') represents future uncertainty propagating back to (s, a). This second part is weighted by a discount-factor parameter, $0 \le \gamma \le 1$. The overall update magnitude is controlled by a learning-rate parameter $0 < \eta < 1$. In the particular case that s' is a terminal state, its exploration value is always defined as 0.

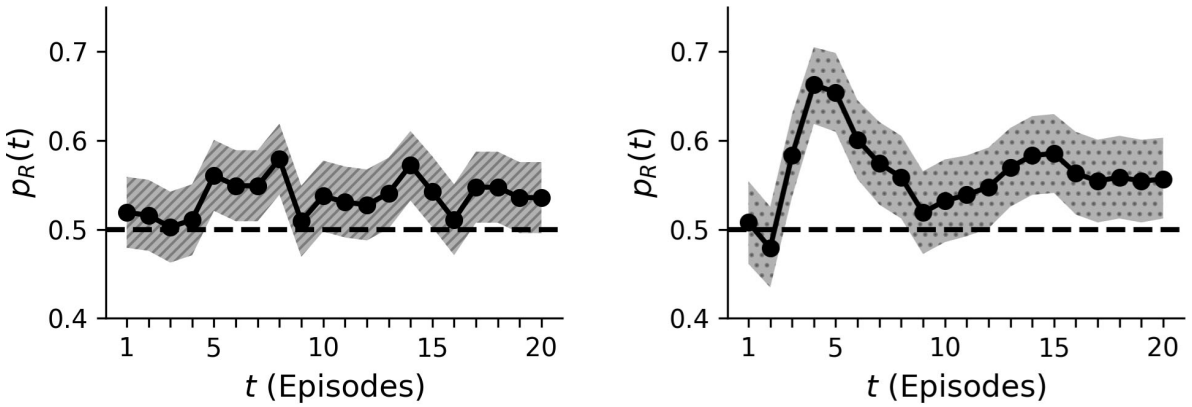

Figure 5

Learning dynamics.

Bias towards M_R as a function of training episode ($p_R(t)$), averaged over participants in all 6 conditions (Experiments 1 & 2), shown for the "poor" (a) and "good" (b) groups. The "good" explorers exhibited a transient peak in $p_R(t)$, consistent with models of uncertainty-driven exploration. However, the steady-state value p_R^* was still slightly larger than chance, consistent with an objective-driven exploration component. Dots and shaded areas denote mean and 95% confidence interval of $p_R(t)$.

To complete the model specification, we define the *policy* as derived directly from these exploration values. We use a standard softmax policy, in which the probability of choosing an action a in state s is given by:

$$\pi\left(a|s\right)=\frac{e^{\beta E(s,a)}}{\sum_{a'}e^{\beta E(s,a')}}$$

where $\beta \geq 0$ is a gain parameter. A gain value of $\beta = 0$ corresponds to random exploration, with all actions chosen at equal probability, while a positive gain corresponds to (stochastically) preferring actions associated with a larger *E*-value (and hence higher uncertainty).

Conceptually, this model is similar to standard RL algorithms (specifically the SARSA algroithm, **Rummery and Niranjan, 1994**) that are used to account for operant learning in animals and humans. There, a similar update rule is used to learn the expected discounted sum of future rewards (and a similar rule is assumed for action-selection). Therefore, similar cognitive mechanisms that account for operant learning, can account for this type of directed exploration (at least to the extent that standard RL models are indeed a good descriptions of operant learning; see **Mongillo et al., 2014**; **Fox et al., 2020**).

To gain insight into the properties of the *E*-values, we consider first the case of "infinite" discounting, namely $\gamma = 0$. In that case, the update rule of **Equation 1** becomes:

$$E\left(s,a\right)\leftarrow\left(1-\eta\right)E\left(s,a\right)$$

and hence, after n visits of (s, a), the associated *E*-value is $E(s, a) = (1 - \eta)^n$, such that $-\log E \propto n$.[1] In other words, when $\gamma = 0$, and long-term consequences are completely ignored, the *E*-value is effectively a visit-counter.

When $\gamma > 0$, the change in the value of $E(s, a)$ following a visit of (s, a) is more complex. In addition to the decay term, a term that is proportional to $E(s', a')$ is added to $E(s, a)$. Notably, $E(s', a')$ depends on the number of past visits of (s', a'), (as well its *own* future states (s'', a'') and so on). Consequently, the number of *actual* visits that is required to reduce the *E*-values by a given amount is larger in state-actions leading to many future states than in state-actions leading to fewer future states. In that sense, the *E*-values are a *generalization* of visit-counters.

Finally (and regardless of the value of γ), the softmax policy of **Equation 2** favors actions associated with larger *E*-values. Because choosing these actions will generally lead to a reduction in their associated *E*-values, the result will be a policy that effectively attempts to equalize the *E*-values of all available actions (within a given state). In the case of $\gamma = 0$, this will result in a preference toward those actions that were chosen less often. In the case of $\gamma > 0$, it will result in a preference that is also sensitive to (the number of) future potential states reachable through the different actions.

To conclude, the model therefore encapsulates the three principles identified in human behavior – it propagates information to track long-term uncertainties associated with individual state-actions, it temporally discounts future exploratory consequences, and it uses estimated uncertainties to derive a behavioral policy.

Directed-exploration in the maze task

We now return to the maze task and study the behavior of the model there. In state M_R, where the *E*-values correspond to visit-counters, the attempt to equalize the *E*-values will result in a bias against repeating the same action, yielding a low p_{repeat} value and on average, a uniform policy. To demonstrate this, we simulated behavior of the model in the 3 conditions of Experiments 1.

Indeed, as depicted in **Figure 6a**, the values of p_{repeat} in the simulations were smaller than chance-level. Unlike the population of human participants, simulated agents are more homogeneous, as reflected in the narrower histograms of p_{repeat}. This is due to the fact that the model is designed to perform directed exploration, that is, to model the behavior of the "good" directed explorers. Nevertheless, the model can also produce random exploration if the gain parameter is set to $\beta = 0$ (see also Discussion).

More interesting is the behavior of the model in state S. The larger n_R, the smaller will be the decay of E ($s = S$, a = right) per a single visit of ($s = S$, a = right). Therefore, the model will tend to choose "right" more often ($p_R > 0.5$), a bias that is expected to increase with n_R. Indeed, similar to the behavior of the "good" human explorers, the simulated agents exhibited a preference towards "right" in S, a preference that increased with $n_R - n_L$ (**Figure 6b**).

The model is sensitive to long-term consequences because it propagates future uncertainty, from the next visited state-action back to the current state-action. This future uncertainty, however, is weighted by $\gamma < 1$, such that the effect of further away states on E (s, a) is expected to decrease with distance. In the environments of experiment 2, where we manipulated the depth of M_R (relative to S), this will result in a decrease of the bias (p_R) at S, as demonstrated in **Figure 6c**.

Because the policy in the model is derived from the E-values, the temporal pattern of exploration is expected to be transient. In the first episodes, when E ($s = S$, a = right) = E ($s = S$, a = left), the result is $p_R = 0.5$. With sufficient learning, exploration values of all visited state-actions decay to 0 and in this limit, $p_R = 0.5$ as well. Therefore, we expect the learning dynamics to exhibit a transient increase in bias, followed by a decay back to chance level. This is demonstrated in **Figure 6d** where we plot p_R (t), averaged over the simulations of the model in all six conditions of Experiments 1 and 2.

Qualitatively, the transient dynamics resemble the experimental results (**Figure 5b**). However, there are two important differences. First, while the human participants exhibited what seems like a steady-state bias even at the end of the experiment, p_R in the model decays to chance level. As discussed above, the decay to chance in the simulations is expected because exploration in the model is uncertainty-driven. In the framework of this model, steady-state exploration can be achieved if we assume that β is not stationary, but rather increases over episodes. However, we hypothesize that to capture this aspect of humans' exploration, we may need to go beyond this class of uncertainty-driven models. Second, the transient dynamics of the model are longer than that of the human participants. While the learning speed in the model is largely controlled by the learning-rate parameter η, the value of η cannot by itself explain this gap. This is because in the model $\eta < 1$, and the dynamics cannot be arbitrarily fast. Particularly, in the simulations of **Figure 6d** we have used a large learning-rate of $\eta = 0.9$, but learning was still considerably slower compared to human participants. We further discuss the issue of learning speed in the next section.

Learning dynamics: 1-step updates and trajectory-based updates

To learn to prefer "right" in S, the agent needs to learn that this action leads, in the future, to M_R, which from an exploratory point of view is superior to M_L. This kind of learning of delayed outcomes is typical of RL problems, in which the agent needs to learn that the value of a particular action stems from its consequences, which can be delayed. For example, an action may valuable because it leads to a large reward, even if this reward is delayed. In the RL literature this is known as the *credit assignment* problem, because during learning, upon observing a desired outcome (in "standard" RL, getting a large reward; here, arriving at M_R), the agent needs to properly assign credit for *past* actions that have led to this outcome.

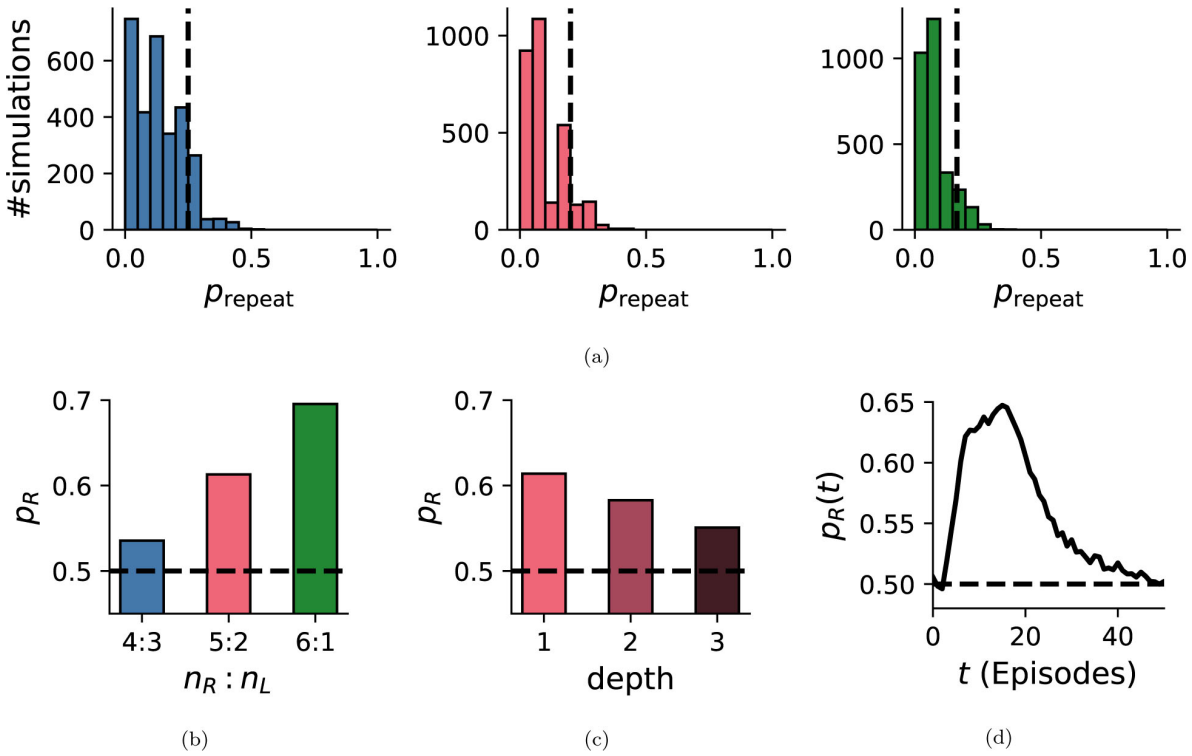

Figure 6

Simulations results.

Simulating behavior of the *E*-values model (**Equation 1**–**2**) reproduces the main findings of directed exploration in the maze task. **(a)** In M_R, the model exhibits directed exploration which manifests in low values of p_{repeat} (shown for the 3 conditions of Experiment 1; dashed line denote chance-level expected for random exploration, $1/n_R$) **(b)** In the environments of Experiment 1, agents exhibited bias towards M_R that increased with imbalance of $n_R : n_L$, reflecting the propagation of long-term uncertainties over states. **(c)** In the environments of Experiment 2, the bias decreased with depth, reflecting temporal discounting. **(d)** Bias towards M_R peaks transiently, followed by a decay to baseline at steady-state, as expected from uncertainty-driven exploration (average results over all 6 environments). Results are based on 3,000 simulations in each environment. Bars and histograms in (a)-(c) are shown for the first 20 episodes for comparison with the behavioral experiments. Error bars are negligible and therefore are not shown. Model parameters: $\eta = 0.9$, $\beta = 5$, $\gamma = 0.6$.

RL algorithms typically address the credit assignment problem by propagating information about the reward backwards through sequences of visited states and actions (**Sutton, 1988**; **Watkins and Dayan, 1992**; **Dayan, 1992**). According to some RL algorithms, the information about the reward propagates backwards one state at a time. By contrast, in other algorithms, a trace of the entire trajectory is maintained, allowing the information to "jump" backwards over a large number of states and actions. We refer to these alternatives as 1-step and trajectory-based updates, respectively.

The *E*-values model can be understood as an RL algorithm that propagates visitations information (rather than reward information). Specifically, it uses 1-step updates (**Equation 1**) such that with each observation (a transition of the form *s, a, s', a'*) only immediate information, from (*s', a'*), is used to update the exploration value of (*s, a*). With 1-step updates it takes time (episodes) for information from M_R to reach back to *S*. We hypothesized that this reliance on 1-step updates might be an important source for the difference in learning speed between the model and humans, who might use more temporally-extended learning rules. To test this, we considered an extension to the exploration model in which *E*-values are learned using a trajectory-based update rule. Technically, this corresponds to changing the TD algorithm of **Equation 1** to a TD (λ) algorithm (see Methods, **Algorithm 1**). Simulating this extended model we found that, similar to the original model, it reproduces the main experimental findings (**Figure S1**, compare with **Figure 6**). Moreover, as predicted, learning is faster than that the learning in the original model (**Figure S1d**, compare with **Figure 6d**). Nevertheless, even this faster learning is still slower than the rapid learning observed in human participants, suggesting further components of human learning that are not captured by either of the models (we get back to this point in the Discussion).

Another way of distinguishing between 1-step and trajectory-based updates is to consider the predictions they make in Experiment 2. Recall that the three conditions in Experiment 2 differ in the delay (in the sense of number of states) between *S* and M_R. If information (about the exploratory "value" of M_R) propagates one step at a time, then the time it takes to learn that "right" is preferable in *S* will increase with the delay: it will be shortest in Condition 1, in which M_R and M_L are merely one step ahead of *S*, and longest in Condition 3, in which M_R and M_L are three steps away from *S* (**Figure 7**, top left). By contrast, if information about M_R and M_L can "jump" directly to *S* within each episode, as in trajectory-based updates, learning speed will be comparable in all three conditions (**Figure 7**, top right). A more thorough analysis of the model dependence on the parameters γ and λ is depicted in **Figure S2**. Finally, **Figure 7** (bottom) depicts the learning dynamics of the "good" human explorers, analyzed separately in the three conditions of Experiment 2. We did not find evidence supporting the hypothesis that learning time increases with depth. These results further support the hypothesis that human learning relies on more global, temporally-extended update rules in which information can "jump" backwards over several states and actions.

Discussion

Exploration is a wide phenomenon that has been linked to different aspects of behavior, including foraging (**Mobbs et al., 2018**; **Kolling and Akam, 2017**), curiosity (**Gottlieb and Oudeyer, 2018**), and creativity (**Hart et al., 2018**). In this study, we focused on exploration as part of learning. For that, we use the framework of RL, in which exploration is an essential component. Particularly, we study the computational principles underlying human exploration in complex environments – sufficiently complex such that exploration *per se* requires learning, due to delayed and long-term consequences of actions. Our approach builds on the analogy between the challenges of learning to explore, and the challenges of learning to maximize reward – the latter being the standard RL scenario. In both cases, the agent needs to represent information, propagate it, and use it to choose actions. In the former case it is information about uncertainty and in the latter it is information about expected reward.

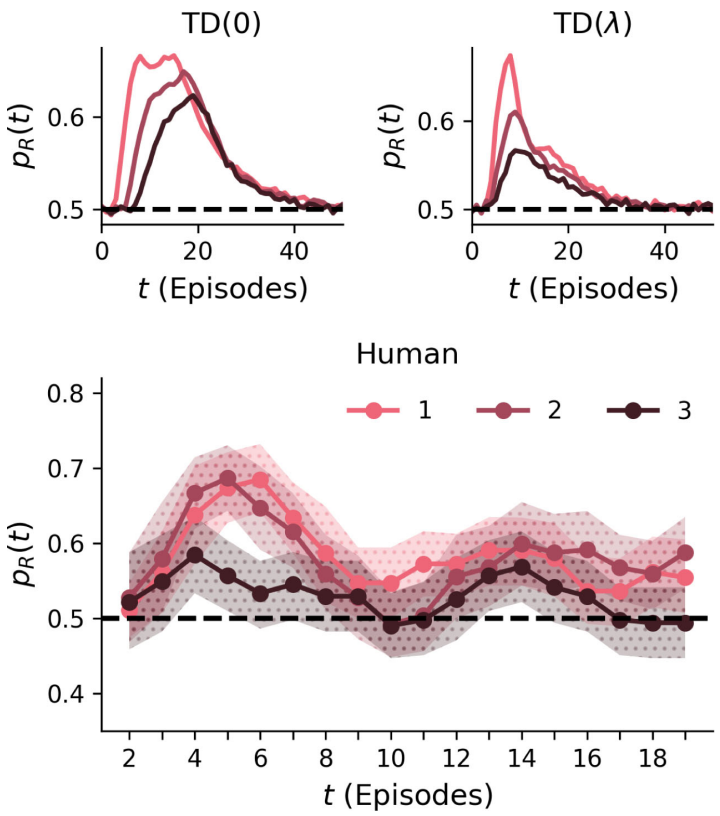

Figure 7

1-step backups and trajectory-based updates.

Learning dynamics simulated by the *E*-values model using the 1-step backup learning rule of TD (0) (**Equation 1**–**2**; **top left**) and the trajectory-based learning rule TD (λ) (Methods, **Algorithm 1**; **top right**) in the 3 environments of Experiment 2. With TD (0), the depth of M_R relative to S (depth = 1, 2, 3) affects both the peak value of $p_R(t)$ (due to temporal discounting) and the time it takes the model to learn (due to the longer sequence of states over which the information has to be propagated). By contrast, with TD (λ), different depths result in a different maximum bias (due to temporal discounting), but the learning time is comparable (because information is propagated over multiple steps in each update). For the same reason, learning is overall faster with TD (λ). In humans (**bottom**), peak bias decreased with depth (consistent with temporal-discounting), but there was no noticeable difference in learning speed (consistent with trajectory-based updates). Learning curves of human participants are shown with a moving-average of 3 episodes. Dots and shaded areas denote means and 70% confidence intervals of $p_R(t)$. Model results are average over 30, 000 simulations; model parameters: $\eta = 0.9$, $\beta = 5$, $\gamma = 0.6$, and $\lambda = 0.6$ (for the TD (λ) model).

We found that while exploring in complex environments, humans are sensitive to long-term consequences of actions and not only to local measures of uncertainty. Moreover, such longterm exploratory consequences are temporally-discounted, similar to the discounting of future rewards. Finally, the dynamics of exploration is consistent with the predictions of uncertainty-driven exploration, in which directed exploratory behavior peaks transiently, and then decay to a more random exploration (supposedly when most of the uncertainty have been resolved). To account for these experimental results, we introduce a computational model that uses a RL-like learning rule implementing the aforementioned principles. In the model, information about state-action visits, rather than about reward as in standard RL algorithms, is being propagated (and discounted) over sequences visited state-actions. This results in a set of "exploration values" (analogous to reward-based values) which are then used to choose actions.

Directed exploration beyond bandit tasks Previous studies have identified some components of directed exploration in human behavior using bandit tasks (**Wilson et al., 2014**; **Gershman, 2018**, **2019**), particularly, the use of counter-based methods such as Upper Confidence Bounds (UCB, **Auer et al., 2002**). Going beyond the bandit, we were able to show that these counter-based strategies might be a special case implementation (appropriate for bandit tasks) of more general principles. To study and identify these principles, it is therefore necessary to test human exploration in environments that are more complex than the bandit task. Indeed a more recent study have shown that more general principles might underlie human exploration, both random and directed, in sequential tasks (**Wilson et al., 2020**). However, unlike our experiments, in that study actions did not have long-term consequences in the sense of state transitions. Finally, the necessity of going beyond simple bandit tasks is not unique to the study of exploration alone. It is present also when studying other components of RL algorithms underlying operant learning. For example, it is impossible to distinguish in a bandit task between *model-based* and *model-free* RL, because there is no "model" to be learned in those tasks (**Daw et al., 2011**).

Non-stationary aspects of exploration While the analogy between learning to explore and learning to maximize rewards is a useful one, there are some important differences. One difference is that while in RL, rewards (more precisely, the distribution thereof) are typically assumed to be Markovian and stationary, exploration has a fundamental non-stationary nature. This is due to the fact that if exploration is interpreted as part of the learning process, or is uncertainty driven, then the exploratory "reward" from a given state-action will *decrease* over time, because uncertainty will reduce with visits of that state-action. This non-stationarity poses a challenge for exploration algorithms. The *E*-values model circumvents that by assuming a stationary (and constant) zero fictitious "reward", combined with an optimism bias at initialization (**Fox et al., 2018**).

A different solution to the challenge of non-stationarity is to posit an exploration objective function which is by itself independent of learning. The predictions of the two classes of models differ with respect to the expected steady-state behavior. In the former, exploration will diminish over time while in the latter, it will be sustained. The observation that human participants maintain a preference (albeit relatively small) for "right" even at the end of the experiment suggests that human exploration is driven, at least in part, by more than just uncertainty. A more complete characterization of these two components will be an interesting topic for future work.

Pure-exploration and the role of reward It has been long argued that at least part of human and animal behavior is driven by intrinsic motivation, which is largely independent of external rewards (**Oudeyer and Kaplan, 2009**; **Barto, 2013**). Pure exploration tasks can be used to characterize aspects of such intrinsic motivation. In this study, the "desire" to visit less-visited states is one such intrinsic motivation factor. Additional factors that are based on information-theoretic quantities (**Still and Precup, 2012**; **Little and Sommer, 2014**; **Houthooft et al., 2016**) or prediction errors of non-reward signals (**Pathak et al., 2017**; **Burda et al., 2019**) have also been proposed in the literature. While many of these will, in general, be correlated, and

hence difficult to identify experimentally, we believe that future studies of pure-exploration in complex environments will allow to better relate these concepts, mostly discussed in the theoretical and computational literature, to the learning and behavior of humans and animals.

To dissect the exploratory component of behavior, we focused on a pure-exploration, reward-free task. This allowed us to neutralize the exploration-exploitation dilemma, focusing on the unique challenges for exploration itself. More generally, we expect the identified exploration principles to be relevant also in the reward maximization scenario. Indeed, it has been shown theoretically and empirically that the naive use of counter-based methods (or other "local" exploration techniques) can be highly sub-optimal for learning an optimal policy (in the reward maximization sense) in complex environments (**Osband et al., 2016a**,**b**; **Chen et al., 2017**; **Fox et al., 2018**; **Oh and Iyengar, 2018**). How humans deal with the exploration-exploitation dilemma in complex environments is an important open question.

Implications for neuroscience Algorithms such as TD-learning hold considerable sway in neuroscience. For example, it is generally believed that dopaminergic neurons encode reward prediction errors, which are used for learning the "values" of states and actions (**Schultz et al., 1997**; **Glimcher, 2011**, but see also **Elber-Dorozko and Loewenstein, 2018**). More recent studies suggest that in fact, the brain maintains a separate representation of different reward dimensions (**Smith et al., 2011**; **Grove et al., 2022**). Given that our formalism of uncertainty (*E*-values) is identical to that of other types of value, it would be interesting to test whether the representation of uncertainty in the brain is similar to that of other reward types. For example, whether dopaminergic neurons also represent the equivalent of *E*-values TD-error. Along the same lines, it would be interesting to check whether the finding that dopaminergic neurons encode what seems to be reward-independent features of the task (**Engelhard et al., 2019**) can be better understood assuming that uncertainty is a reward-like measure.

Heterogeneity There was a substantial heterogeneity among participants in both Experiments 1 and 2. We used this heterogeneity to divide participants into "good" and "poor" explorers in terms of the "directedness" of their exploration. However, this division is somewhat crude. For example, while bias in favor of M_R was smaller in the "poor" explorers, it was still larger than the baseline level of 0.5 predicted by a true random exploration behavior (**Figure 5a**). This separation can be understood as a first approximation, highlighting the more prominent source of exploratory behavior at the individual subject basis. Moreover, even within the "good" explorers, there was considerable variability. Heterogeneity in the parameters of the computational model can, perhaps, explain some of the heterogeneity, but parameters variability alone (within the *E*-values model) certainly cannot explain all of the heterogeneity in participants' behavior. For example, consider again the division to "poor" and "good" directed explorers. In principle, such a division could be modeled through the gain parameter β, with random explorers having a value of $\beta = 0$ (and directed explorers a value of $\beta > 0$). Even with random exploration, the model prediction for p_{repeat} is $1/n_R$. By contrast, many participants exhibited values of p_{repeat} larger than this chance-level, all the way up to $p_{\text{repeat}} = 1$. Similarly, considering behavior at S as measured by p_R, no combination of model parameters predict p_R values which are smaller than 0.5. This is because even random exploration will result in $p_R = 0.5$. Values of p_R that are close to 1 are also impossible in the model, because they imply under-exploration of the left-hand-side of the maze. Yet some human participants exhibited extreme (close to 0 or 1) values of p_R. Other factors, such as (task-independent) choice bias (**Baum, 1974**; **Laquitaine et al., 2013**; **Lebovich et al., 2019**) and tendency to repeat actions (**Urai et al., 2019**) are likely to contribute to participants' choices.

Learning speed Another limitation of the model is the gap between the learning speed of human participants and the learning speed of the model. Overall, humans learned considerably faster than the model, even with a large learning-rate. On average participants exhibited a bias as soon as the 3$^{\text{rd}}$ episode, which is faster than the theoretical limit possible for the TD(0) model in

this task. While some of this discrepancy can be attributed to the model's reliance on 1-step backups, it is noteworthy that even in comparison with TD(λ), humans' learning is faster than the that of the model. The rapid learning in humans suggest mechanisms that go beyond simple *model-free* learning as implemented in our models. In our model, the fact that "right" is favorable can only be learned implicitly, by actually visiting more unique states following M_R (compared to M_L). This is because the only information that is available to the agent is the identity of states and actions. By contrast, a single visit of both M_R an M_L is likely sufficient for humans to learn that the number of doors in M_R is larger than in M_L, a fact which can by itself bias their following choices in favor of "right". Indeed by using this (possibly salient) feature, of the number of doors, as an explicit part of the state representation, one could infer that M_R is more favorable over M_L already after 2 episodes even with model-free learning. While such strategy is not as *general* as the computational principles encapsulated by our models, in the *specific* task at hand it will be rather effective. The ability of humans to rapidly form and utilize such heuristics and generalizations is likely an important part of their ability to rapidly adapt and learn in novel situations. The interplay between basic, more general-purpose, computational principles, and heuristic, more ad-hoc, principles remains an important challenge for computational modeling in the cognitive sciences.

Generalization, priors, and "natural" exploration The goal of this study was to identify computational principles underlying exploration in a "general" setting. To that goal, we used a task in which the semantic content attached to states was minimal, with no a-priori indication of any structure (temporal, geometric, spatial, etc.) of the state-space. The motivation behind this design was to de-emphasize, as much as possible, behavior components stemming from participants' prior knowledge and generalization abilities, and focus on core exploratory strategies. This also justified the models that we used: general-purpose, simplistic, learning models that operate on an abstract notion of states and actions. On the other hand, the abstract design of the task limits its applicability to more realistic tasks and natural behavior. Indeed in complex environments, it has been demonstrated that humans rely largely on both priors and generalizations to achieve efficient learning and exploration (**Dubey et al., 2018**; **Schulz et al., 2020**). How such priors, semantic knowledge, and generalization interact with more abstract and general principles of exploration and decision-making is an important open question. Notably, we have found that humans are capable of performing directed exploration of complex environments even in the absence of a readily-available semantic structure to guide their exploration. This is in contrast to the recent work of **Brändle et al. (2022)**, that demonstrated directed exploration (interpreted as driven by the information-theoretic quantity of *empowerment*) in complex environments with available semantic structure, that was not observed in a structurally identical task where the semantic structure has been masked.

Methods

Online experiments and data collection

The study was approved by the Hebrew University Committee for the Use of Human Subjects in Research. Participants were recruited using the Amazon MechanicalTurk online platform, and were randomly assigned to one of the conditions in each experiment. Participants were instructed to "understand how the rooms are connected", and were informed regarding the test phase: "At the end of the task, a test will check how quickly can you get from one specific room to a different one.". The training phase of the experiment consisted of 120 trials, corresponding to 20 episodes. Between 20% to 30% of participants (depending on the experiment and condition) performed a longer experiment of 250 trials corresponding to 42 episodes, but for these participants only the first 20 episodes were analyzed. The end of each episode (reaching the terminal state T) was signaled by a message screen ("Youv'e reached a dead-end room, and will be moved back to the first room"). After the training episodes, there was a test phase in which participants were asked

to navigate to a target room in the minimal number of steps possible, starting from a particular start room (which was not the initial state *S*). An online working copy of the experiment can be accessed at: *https://decision-making-lab.com/lf/eee_rep/Instructions.php*.

For each participant, we recorded the sequence of visited rooms (states) and chosen doors (actions), in the train and test phases. No other details (including demographics details, questionnaire, or comments about the experiment) were collected from participants. Test performance was used as a criterion for filtering. Out of the total participants who finished the experiment (i.e., finished both training and test phases), we rejected those who did not finish the test phase in a number of steps smaller than expected by chance (e.g., the expected number of steps it would take to reach the target by random walk). We also rejected participants who, during training, did not choose both "right" and "left" at least twice. The test start and target rooms were identical for all participants, and were chosen as to maximize the difference between performance (i.e., number of steps) expected by chance to that of the optimal (shortest path) policy. The number of participants in each experiment is given in *Table 1*, and their division into "Good" and "Poor" explorers is given in *Table 2*.

Estimating policy from behavior

For the average results, we computed for each participant their p_R value as the number of "right" choices divided by the total (and fixed) number of visits to *S*. Similarly, p_{repeat} was calculated for individual participants as the number of visits to M_R in which the chosen action was identical to the one chosen in their previous visit of M_R, divided by the total visits of M_R minus one. Note that the total number of visits to M_R was different for different participants, as it depended on their policy at *S*. We have used the same measurements for the results of the model simulations for consistency. Note that, in principle, the model allows to measure the policy of individual agents (at individual time-points) directly, without the need to estimate it from behavior (i.e., the generated stochastic choices). To estimate learning dynamics, we can no longer estimate $p_R(t)$ on an individual level, because each participant only made one binary choice at a given episode. Therefore, we computed $p_R(t)$ at the population level, as the number of participants who chose "right" in the t^{th} episode divided by the total number of participants (possibly within a particular group, for example only "good" explorers). Alternatively, when considering specific experimental conditions, we have estimated $p_R(t)$ for individual participants using a moving-average over a window of 3 consecutive episodes.

Statistical analysis

Confidence Intervals (CI) for p_R were computed using bootstrapping, by resampling participants and choices. Comparisons between different conditions were computed using a permutation test, by shuffling all participants of the two groups being compared, and resampling under the null hypothesis of no group difference. With this resampling we computed the distribution of p_R (A) – p_R (B) for two random shuffled groups of participants A and B. Reported p-value is the CDF of this distribution evaluated at the real (unshuffled) groups.

TD (λ) learning for *E*-values

We start by proving a short, non-technical description of the TD and TD (λ) value-learning algorithms. The *value* of a state-action (denoted $Q(s, a)$), is defined as the expected sum of (discounted) rewards achieved following that state-action. The goal of the algorithms is to learn these values. To that end, the agent maintains and updates estimates $\hat{Q}(s, a)$ of the true state-action values $Q(s, a)$. In TD-learning, Upon observing a transition (s, a, r, s', a'), the estimated value $(\hat{Q}(s, a))$ is updated towards $r + \gamma\hat{Q}(s', a')$. Crucially, $\hat{Q}(s', a')$ is also, on its own, an estimated value. This usage of (a part of) the current estimator to form the target for updating the

Exp.	Env.	Completed	Included
1	$3:4$	191	161
	$2:5$	174	120
	$1:6$	176	137
2	$d=1$	244	191
	$d=2$	269	205
	$d=3$	282	238

Table 1

Number of participants in Experiments 1 and 2.

Exp.	Env.	"good" explorers	"poor" explorers
1	$3:4$	66	95
	$2:5$	53	67
	$1:6$	71	66
2	$d=1$	92	99
	$d=2$	84	121
	$d=3$	85	153

Table 2

Participant groups in Experiments 1 and 2

same estimator is known as *bootstrapping*. TD learning therefore breaks the estimation of value – the sum of rewards – into two parts: the first reward, which is taken from the environment, and the rest of the sum, which is bootstrapped.

It is possible, however, to estimate the values while breaking the sum of rewards in other ways. For example one could sum the first two rewards based on observations, and bootstrap the rest, that is, from time-step 3 on-wards. Importantly, this would result in information (about the rewards) propagating backwards 2-steps in a single update, rather than 1-step. More generally, breaking the sum after n steps will result in an n-step backup learning rule. It is also possible to average multiple n-step backups in a single update. The TD (λ) algorithm is a particular popular scheme to do that: it can be understood as combining *all* possible n-step backups, with a weighting function that decays exponentially with n (i.e., the weight given to the n-step backup is λ^{n-1}, where λ is a parameter). With $\lambda = 0$ the algorithm recovers the standard 1-step backup algorithm, or in other words, TD (0) is simply TD. A value of $\lambda = 1$ corresponds to no bootstrapping at all, relying instead on *Monte Carlo* estimates of the action value by collecting direct samples (sum of rewards over complete trajectories).[2]

Equation 1 can be understood as a TD algorithm (specifically, using the sarsa algorithm (*Rummery and Niranjan, 1994*; *Sutton and Barto, 2018*)) in the particular case that all the rewards signals are assumed to be $r = 0$, and estimates are initialized at 1. The extended model (*Algorithm 1*) is a direct generalization of that correspondence to the TD (λ) case.

Algorithm 1 TD (λ) learning for E-values

Require: Parameters η, λ, γ
  Initialize $E(s,a) = 1$ for all s, a
  for all episodes **do**
    set $\varepsilon(s,a) = 0$ for all s, a ▷ eligibility-traces
    set $\tau = \{\}$ ▷ trajectory in this episode
    set s to the initial state and choose action a
    while s is not a terminal state **do**
      sample the next state and action s', a'
      increment $\varepsilon(s,a) \leftarrow \varepsilon(s,a) + 1$, and concatenate (s,a) to τ
      for all (s_t, a_t) in τ **do**
        $E(s_t,a_t) \leftarrow E(s_t,a_t) + \eta\varepsilon(s_t,a_t)(\gamma E(s',a') - E(s,a))$ ▷ update E-value
        $\varepsilon(s_t,a_t) \leftarrow \gamma\lambda\varepsilon(s_t,a_t)$ ▷ decay eligibility-trace
      end for
      $s \leftarrow s'$, $a \leftarrow a'$
    end while
    $E(s,a) \leftarrow (1-\eta)E(s,a)$ ▷ update in terminal-state
  end for

Supplementary material

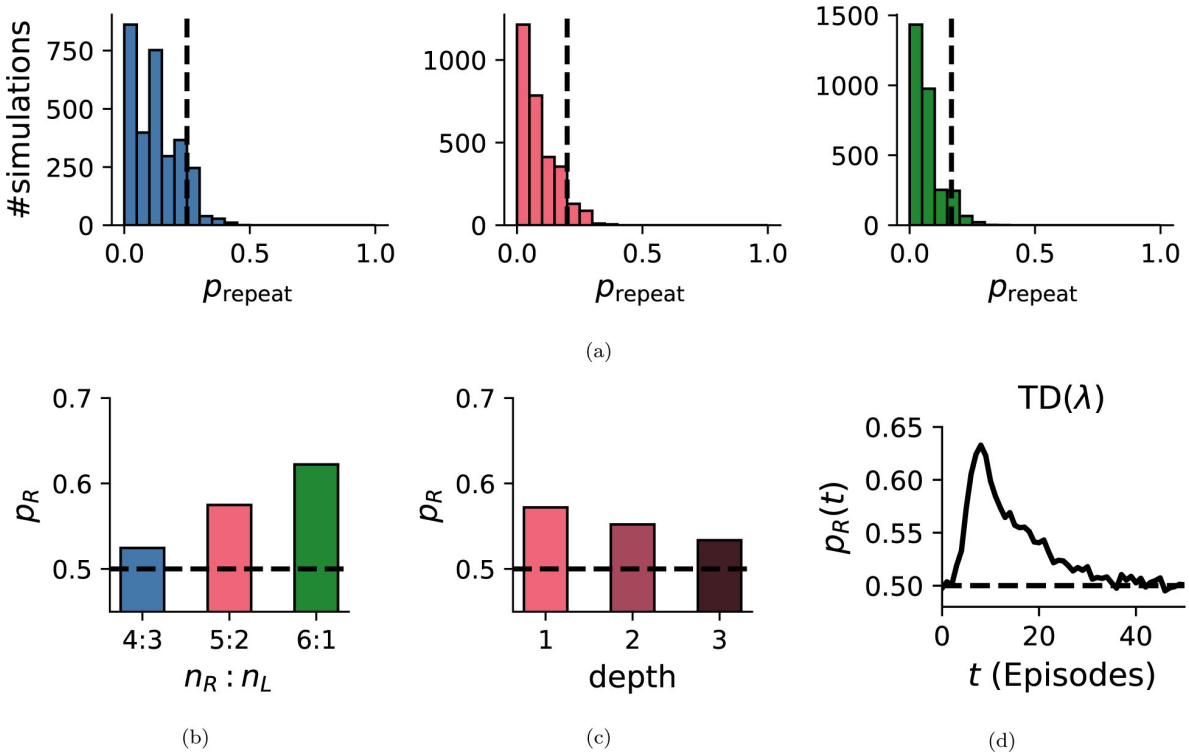

Figure S1

Simulations results of TD (λ).

Simulating behavior of the *E*-values model with the TD (λ) learning rule (Methods, *Algorithm 1*) reproduces the main findings of directed exploration in the maze task. **(a)** In M_R, the model exhibits directed exploration which manifests in low values of p_{repeat} (shown for the 3 conditions of Experiment 1; dashed line denote chance-level expected for random exploration, $1/n_R$) **(b)** In the environments of Experiment 1, agents exhibited bias towards M_R that increased with imbalance of $n_R : n_L$, reflecting the propagation of long-term uncertainties over states. **(c)** In the environments of Experiment 2, the bias decreased with depth, reflecting temporal discounting. **(d)** Bias towards M_R peaks transiently, followed by a decay to baseline at steady-state, as expected from uncertainty-driven exploration (average results over all 6 environments). The learning dynamics is faster than that of the 1-step update model. Results are based on 3,000 simulations in each environment. Bars and histograms in (a)-(c) are shown for the first 20 episodes to match the behavioral experiments. Model parameters: $\eta = 0.9$, $\beta = 5$, $\gamma = 0.6$, $\lambda = 0.6$.

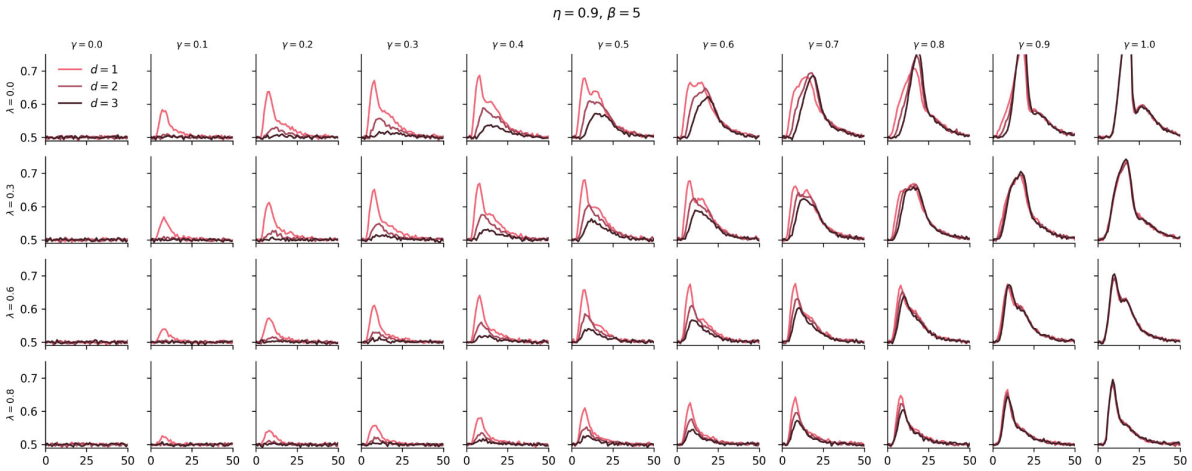

Figure S2

Model parameters.

Learning curves of the TD (λ) model in the 3 environments of Experiment 2 for different values of γ,λ (with fixed $\eta = 0.9$, $\beta = 5$). With infinite discounting ($\gamma = 0$), future consequences are neglected, resulting in a uniform (counter-based like) policy with no bias. With no discounting ($\gamma = 1$), information from the terminal state T dominates, resulting in a bias towards "right" (since there are more routes to the terminal states via the "right" branch) that is not dependent of the depth of M_R. For intermediate values of γ, transient exploration opportunities (i.e., in M_R) becomes important, resulting in a bias towards M_R that decreases with depth, reflecting temporal-discounting. In this regime, one-step backup learning rule ($\lambda = 0$) results in difference learning speed for different depths, while for trajectory-based learning rules ($\lambda > 0$) learning speed is comparable for the different depths. Each learning curve is the average of 30, 000 simulations.

References

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

Barto A. G., Baldassarre G., Mirolli M. (2013) **Intrinsic motivation and reinforcement learning** *Intrinsically motivated learning in natural and artificial systems* :17–47

Baum W. M. (1974) **On two types of deviation from the matching law: Bias and undermatching** *Journal of the Experimental Analysis of Behavior* **22**:231–242

Bellemare M., Srinivasan S., Ostrovski G., Schaul T., Saxton D., Munos R., Lee D. D., Sugiyama M., Luxburg U. V., Guyon I., Garnett R. (2016) **Unifying count-based exploration and intrinsic motivation** *Advances in Neural Information Processing Systems 29* :1471–1479

Brändle F., Stocks L. J., Tenenbaum J., Gershman S. J., Schulz E. (2022) **Intrinsically motivated exploration as empowerment** *PsyArXiv Preprints*

Burda Y., Edwards H., Storkey A., Klimov O. (2019) **Exploration by random network distillation**

Chen R. Y., Sidor S., Abbeel P., Schulman J. (2017) **Ucb exploration via q-ensembles** *arXiv preprint* https://doi.org/10.48550/arXiv.1706.01502

Daw N. D., Gershman S. J., Seymour B., Dayan P., Dolan R. J. (2011) **Model-based influences on humans' choices and striatal prediction errors** *Neuron* **69**:1204–1215

Dayan P. (1992) **The convergence of td (λ) for general λ** *Machine learning* **8**:341–362

Dubey R., Agrawal P., Pathak D., Griffiths T., Efros A., Dy J., Krause A. (2018) **Investigating human priors for playing video games** :1349–1357

Elber-Dorozko L., Loewenstein Y. (2018) **Striatal action-value neurons reconsidered** *eLife* **7**

Engelhard B., Finkelstein J., Cox J., Fleming W., Jang H. J., Ornelas S., Koay S. A., Thiberge S. Y., Daw N. D., Tank D. W. (2019) **Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons** *Nature* **570**:509–513

Fox L., Choshen L., Loewenstein Y. (2018) **DORA the explorer: Directed outreaching reinforcement action-selection**

Fox L., Dan O., Elber-Dorozko L., Loewenstein Y. (2020) **Exploration: from machines to humans** *Current Opinion in Behavioral Sciences* :104–111

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

Gershman S. J. (2019) **Uncertainty and exploration** *Decision* **6**

Glimcher P. W. (2011) **Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis** *Proceedings of the National Academy of Sciences* **108**:15647–15654

Gottlieb J., Oudeyer P.-Y. (2018) **Towards a neuroscience of active sampling and curiosity** *Nature Reviews Neuroscience* **19**:758–770

Grove J. C., Gray L. A., La Santa Medina N., Sivakumar N., Ahn J. S., Corpuz T. V., Berke J. D., Kreitzer A. C., Knight Z. A. (2022) **Dopamine subsystems that track internal states** *Nature* **608**:374–380

Hart Y., Goldberg H., Striem-Amit E., Mayo A. E., Noy L., Alon U. (2018) **Creative exploration as a scale-invariant search on a meaning landscape** *Nature communications* **9**:1–11

Hazan E., Kakade S., Singh K., Van Soest A., Chaudhuri K., Salakhutdinov R. (2019) **Provably efficient maximum entropy exploration** :2681–2691

Houthooft R., Chen X., Duan Y., Schulman J., De Turck F., Abbeel P. (2016) **Vime: Variational information maximizing exploration** *Advances in Neural Information Processing Systems* :1109–1117

Kolling N., Akam T. (2017) **(reinforcement?) learning to forage optimally** *Current Opinion in Neurobiology* :162–169

Laquitaine S., Piron C., Abellanas D., Loewenstein Y., Boraud T. (2013) **Complex population response of dorsal putamen neurons predicts the ability to learn** *PLOS ONE* **8**

Lattimore T., Szepesvári C. (2020) **Bandit Algorithms**

Lebovich L., Darshan R., Lavi Y., Hansel D., Loewenstein Y. (2019) **Idiosyncratic choice bias naturally emerges from intrinsic stochasticity in neuronal dynamics** *Nature Human Behaviour* **3**:1190–1202

Little D. Y., Sommer F. T. (2014) **Learning and exploration in action-perception loops** *Closing the Loop Around Neural Systems*

Machado M. C., Bowling M. (2016) **Learning purposeful behaviour in the absence of rewards** *arXiv preprint* https://doi.org/10.48550/arXiv.1605.07700

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**

Meuleau N., Bourgine P. (1999) **Exploration of multi-state environments: Local measures and back-propagation of uncertainty** *Machine Learning* **35**:117–154

Mobbs D., Trimmer P. C., Blumstein D. T., Dayan P. (2018) **Foraging for foundations in decision neuroscience: insights from ethology** *Nature Reviews Neuroscience* **19**:419–427

Mongillo G., Shteingart H., Loewenstein Y. (2014) **The misbehavior of reinforcement learning** *Proceedings of the IEEE* **102**:528–541

Oh M.-h., Iyengar G. (2018) **Directed exploration in pac model-free reinforcement learning** *arXiv preprint* https://doi.org/10.48550/arXiv.1808.10552

Osband I., Blundell C., Pritzel A., Van Roy B. (2016) **Deep exploration via bootstrapped dqn** *Advances in neural information processing systems* :4026–4034

Osband I., Roy B. V., Wen Z., Balcan M. F., Weinberger K. Q. (2016) **Generalization and exploration via randomized value functions** :2377–2386

Ostrovski G., Bellemare M. G., Oord A. v. d., Munos R. (2017) **Count-based exploration with neural density models** *arXiv preprint* https://doi.org/10.48550/arXiv.1703.01310

Oudeyer P.-Y., Kaplan F. (2009) **What is intrinsic motivation? a typology of computational approaches** *Frontiers in neurorobotics* **1**

Pathak D., Agrawal P., Efros A. A., Darrell T. (2017) **Curiosity-driven exploration by self-supervised prediction**

Rummery G. A., Niranjan M. (1994) **On-line Q-learning using connectionist systems**

Schmidhuber J. (1991) **Curious model-building control systems** :1458–1463

Schultz W., Dayan P., Montague P. R. (1997) **A neural substrate of prediction and reward** *Science* **275**:1593–1599

Schulz E., Franklin N. T., Gershman S. J. (2020) **Finding structure in multi-armed bandits** *Cognitive Psychology* **119**

Shteingart H., Neiman T., Loewenstein Y. (2013) **The role of first impression in operant learning** *Journal of Experimental Psychology: General* **142**

Smith K. S., Berridge K. C., Aldridge J. W. (2011) **Disentangling pleasure from incentive salience and learning signals in brain reward circuitry** *Proceedings of the National Academy of Sciences* **108**:E255–E264

Still S., Precup D. (2012) **An information-theoretic approach to curiosity-driven reinforcement learning** *Theory in Biosciences* **131**:139–148

Storck J., Hochreiter S., Schmidhuber J. (1995) **Reinforcement driven information acquisition in non-deterministic environments** :159–164

Sutton R. S. (1988) **Learning to predict by the methods of temporal differences** *Machine learning* **3**:9–44

Sutton R. S., Barto A. G. (2018) **Reinforcement learning : an introduction**

Tang H., Houthooft R., Foote D., Stooke A., Chen O. X., Duan Y., Schulman J., DeTurck F., Abbeel P. (2017) **# exploration: A study of count-based exploration for deep reinforcement learning** *Advances in Neural Information Processing Systems* :2753–2762

Thrun S. B. (1992) **Efficient exploration in reinforcement learning**

Urai A. E., de Gee J. W., Tsetsos K., Donner T. H. (2019) **Choice history biases subsequent evidence accumulation** *eLife* **8**

Vanderveldt A., Oliveira L., Green L. (2016) **Delay discounting: pigeon, rat, human—does it matter?** *Journal of Experimental Psychology: Animal learning and cognition* **42**

Watkins C. J., Dayan P. (1992) **Q-learning** *Machine learning* **8**:279–292

Wilson R., Wang S., Sadeghiyeh H., Cohen J. D. (2020) **Deep exploration as a unifying account of explore-exploit behavior** *PsyArXiv preprint*

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**

Zahavy T., O' Donoghue B., Desjardins G., Singh S., Ranzato M., Beygelzimer A., Dauphin Y., Liang P., Vaughan J. W. (2021) **Reward is enough for convex mdps** *Advances in Neural Information Processing Systems* :25746–25759

Zhang J., Koppel A., Bedi A. S., Szepesvari C., Wang M., Larochelle H., Ranzato M., Hadsell R., Balcan M. F., Lin H. (2020) **Variational policy gradient method for reinforcement learning with general utilities** *Advances in Neural Information Processing Systems* :4572–4583

Article and author information

Lior Fox
The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem
For correspondence: lior.fox@mail.huji.ac.il

Ohad Dan
Yale School of Medicine
For correspondence: ohad.dan@gmail.com

Yonatan Loewenstein
The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Department of Cognitive Sciences, The Alexander Silberman Institute of Life Sciences, and The Federmann Center for the Study of Rationality
For correspondence: yonatan@huji.ac.il

Copyright

Editors

Reviewing Editor
Timothy Verstynen
Carnegie Mellon University, United States of America

Senior Editor
Timothy Behrens
University of Oxford, United Kingdom

Reviewer #1 (Public Review):

Summary:

Fox, Dan, and Loewenstein investigated how people explored six maze-like environments. They show that roughly one-third of their participants make choices now that increase the potential for future information gain and also temporally discount potential information gain based on how far in the future potential gains might be. The authors argue that rather than valuing exploration in its own right, participant behavior is most consistent with using exploration as a way to reduce uncertainty. They then propose a reinforcement learning (RL) model in which agents estimate an "exploration value" (the expected cumulative information gained by taking a given action in a given state) using standard RL techniques for estimating value (expected cumulative reward). They find that this model exhibits several qualitative similarities with human behavior and that it best captures the temporal dynamics of human exploration when propagating information through the entire history of a behavioral episode (as opposed to merely propagating it in a single step as some of the simplest RL models do).

While the core insight and basic method of the paper are compelling, the way in which both the behavioral experiment and computational modeling were conducted raise concerns that mean that, in their present form, the results do not fully justify the conclusions. After resolving these issues, the work would demonstrate how human exploration is sensitive to long-range dependencies in information gain, as well as valuable insights about how best to characterize this behavior computationally. I am not particularly well-versed in the literature on exploration so cannot comment on novelty here.

Strengths:
The entire paper is logically well-motivated. It builds on a valuable basic insight, namely that while bandit tasks are an ideally minimal platform for testing certain questions about decision-making and exploration, richer paradigms are needed to capture the long-range informational dependencies distinguishing between various approaches to exploration.

Even so, the maze navigation paradigm explored here remains simple. Participants navigate a maze with two main branches which are identical save for minimal, theoretically motivated differences. Moreover, the tested differences are designed to clearly and explicitly test well-identified questions. The task, and really the entire paper, is clearly organized, and each component is logically connected to a larger argument.

The proposed model is also simple, clearly presented, and a clever way of applying ideas typically used to reason about reward-motivated behavior to reason here about information-motivated behavior.

One other strength of this work is that it combines behavioral experiments with computational modeling. This approach pairs a detailed and objectively specified theory (i.e. the model) with novel data specifically designed to test that theory and thus in principle presents a particularly strong test of the authors' hypotheses.

Weaknesses:
Despite many strengths in the underlying logic of the paper, the presented evidence does not provide compelling support for the conclusions. In particular:

- The main claims are based on the behavior of 452 participants classed as good explorers, out of 1,052 participants included in the analyses and 1,336 participants who completed the study. That is, the authors' broad claims about human exploration are based on a third of their total sample; the other two-thirds displayed very different behavior, including 20% who

performed at or below chance levels. That is, while a significant sub-population may demonstrate the claimed abilities, it is far from clear that they are universal.

- While the experimental manipulations are elegant, the behavioral study seems underpowered. In each of the primary manipulations, key theoretical predictions are not statistically validated. For example, in Experiment 1, the preference for the right door increases from the 4:3 condition to the 5:2 condition, but not when moving from the 5:2 condition to the 6:1 condition, as predicted (Figure 1c). Similar results can be seen for other analyses in Figures 3b and 4b. Relatedly, the experiments comprised just 20 episodes, and it is unclear whether that was sufficiently long for participants to demonstrate asymptotic behavior (e.g. Figure 5b). Either more participants or greater differences between conditions (e.g. testing 9:8, 12:5, and 15:2 conditions in a revised Experiment 1), as well as running a greater number of total episodes, would be needed to resolve this concern.

- The model is presented after the behavioral results, giving the impression that it was perhaps constructed to fit the data. No attempt is made to fit the model to a subset of the data and then validate the rest or give any clear indication as to how the model parameters were set. Moreover, as noted, even where the model is successful, it only explains the behavior of a minority of the total participants. No modeling work is done to explain the behavior of the other two-thirds of the participants.

- The authors helpfully discuss several meaningful alternative models of exploration, such as visit-counting and incorporating an objective function sensitive to information gain. They do not, however, compare their model against these or any other meaningful baselines. Moreover, the comparison between model and human participants is qualitative rather than quantitative. These issues could be resolved by introducing a more rigorous analysis quantitatively comparing a variety of theoretically relevant models as quantitative explanations of the human data.

Reviewer #2 (Public Review):

Summary:
In this article, the authors develop an algorithm for exploration inspired by the classic, state-action-reward-state-action (SARSA) reinforcement learning algorithm. Designed to account for exploration in multi-state environments, this algorithm computes the expected discounted return from selecting an action in a state and uses that value to update the cached value of taking a given action in a given state. The value represents the uncertainty in a given state, and the backed-up value is computed from the discounted future return plus the immediate reduction in uncertainty regarding the state.

Strengths:
The article is ambitious and seeks to characterize human exploration in a novel task using zero rewards. That characterization is useful.

Weaknesses:
The paper suffers from many problems. Here, I will mention three. First, the algorithm is very poorly motivated-exploration is central to many behaviors, but the algorithm computes the value of exploration independent of any long-run considerations of exploitation. Second, the article attempts to recover the observed exploratory behavior in two different multi-state choice tasks. But the algorithm does not explain that behavior, and there is no performance metric on the model, nor a comparison to other models. Third, the article frames the algorithm in terms of uncertainty, but there is no measure of uncertainty.

In short, in many ways this manuscript is 'half an article', and the authors have much work to do. They could decide to dive into the convergence proofs and other theoretical properties of the model. However, as far as I understand the model, it is literally an optimistic SARSA,

whose characteristics are well-understood. Or, they could compare the model's performance to a number of other exploration models (UCB, Thompson sampling, infomax, infotaxis-there are so many!). However the authors need to choose one or the other. I urge the authors to properly compare their model to other models.

1. Motivation
The algorithm is poorly motivated. Exploration is valuable for a time but quickly becomes less valued as more is learned about the environment. The algorithm attempts to account for this by the ad hoc nature of the backup: the immediate outcome is -E(s,a), which represents a reduction in uncertainty. So in the long run, the exploratory value will decrease to zero. But this is ad hoc; why not add E(s,a)? In addition, exploration values are set to 1. But this is also ad hoc; why should E(s,a) start at 1? They have cherry-picked their starting values and the nature of the back-up to yield exploratory behavior.

2. Performance
The authors wish to compare the model's performance to observed exploration behavior. However, their model does a poor job of explaining the behavior. What's confusing is that the authors note the ways the model deviates. There are two principal deviations. First, the model exhibits an exploratory transient, but it is too wide to match the humans. Second, the model fails to exhibit the low-level persistent exploration characteristic of humans in their task.

The next natural step would be to augment the model in different ways to attempt to describe the behavior. The authors do attempt to import td-λ aspects into their exploration model. They determine that importation fails to capture the observed behavior. But why stop there? Why not continue? Why not follow through and change the model in a way that can capture the dynamics of exploration?

In addition, a natural complement would be to compare the model's ability to describe human performance to other models. This would require model fitting, recovery, and validation. However the authors don't engage in that model fitting exercise.

They note that a model-based learning strategy could account for the speed of learning in humans. However they don't comment generally on how model-based strategies could explain their findings nor how they relate to their model. They should comment on this. In particular, the participants are likely learning a model of their environment, and this can be done using non-parametric Bayesian inference (along the lines of Gershman or Collins's work). The authors should model their task using these models and compare this to their algorithm.

The authors state that there was no reward. Were subjects paid for their time? Also, the lack of a reward is unusual, and even if unconsciously, participants may have been engaged in reward-seeking. The authors should try to model the behavior with a pseudo-reward to see how that accounts for their findings. This is especially true from the perspective of computational RL. On that theory, the only object 'in' the agent is the policy; everything else is considered 'in' the environment. This means that rewards in RL need not be from environmental returns but could also be from inside the organism (even if modeled as 'outside' the agent in the RL framework). So they need to model the behavior using 'pseudorewards' to see if that can account for their findings. Finally, though trivially, a reward of 0 is technically a reward, and the model's exploratory drive comes from settling on the true values of the states (i.e., 0).

3. Uncertainty
The authors frame their model in terms of uncertainty, but their model does not measure uncertainty at all. The model makes choices on the basis of optimistic initial Q-values and then searches on that basis, backing up the 0 rewards until the true values are more or less

hit upon. But that is not a measure of uncertainty in any sense; rather, it is an optimistic Q-value bias that drives exploration. However, I may simply fail to understand their model.

Reviewer #3 (Public Review):

Summary:
In this article, Fox and colleagues describe the results of a novel and innovative task, coupled with a modified computational model, to explore pure directed exploration (not quite a pun, but intended nonetheless). In their task, participants make a series of discrete choices, importantly with no reward feedback, to navigate a nested series of rooms in a virtual environment. The initial 2-door choice is used as the primary probe and the complexity of the series of rooms behind each choice is used as the critical independent variable. The authors find that, as the number of follow-up options behind a door increases, "good" participants are more likely to choose the door that leads to the more complex choices. As the depth of the search increased (i.e. the room with the most doors was presented "farther" down the search), these same participants were less likely to choose the door leading to the more complex route. Finally, these same "good" participants showed an initial boost in preference towards the more complex exploration option after a few learning episodes that settled down after about 10 episodes, with a modest reliable preference towards the more complex route. This reflected the fact that information value decays over time in stable situations. Using an adaptation of standard Q-learning, with a proxy of information value being substituted for reward value, the authors show how their model can qualitatively capture most of the observed experimental effects, although with some critical differences in the temporal dynamics of learning, suggesting that the memory horizon for humans is longer than in the adapted Q-learning model.

Strengths:
1. Clever experimental design
The novel task is really clever and gets around many of the limitations for understanding directed exploration that have plagued prior research (which typically involve some use of reward feedback). Finding a way to provide direct information that can be experimentally manipulated, without needing to provide any explicit reward feedback, makes this one of the few pure exploration tasks that I am aware of.

2. Compelling results
The effect of manipulating choice complexity and depth on initial choice probability for "good" directed learners seems fairly strong, as do the learning dynamics. The heterogeneity in exploration style across participants is also interesting and brings up more questions that are useful for follow-up research.

3. Simple model
The computational model used is a simple adaptation of standard reinforcement learning models, specifically Q-learning models. This is elegant as it doesn't require major changes in the dynamics of learning, simply a revision of the variables going into the update. The simplicity of this change, coupled with the ability to capture the results of the "good" directed explorers makes a strong case that information seeking and reward-seeking may share common underlying mechanisms (as shown previously by Kobayashi, K., & Hsu, M. (2019). Common neural code for reward and information value. Proceedings of the National Academy of Sciences, 116(26), 13061-13066.).

Weaknesses:

1. "Good" vs. "poor"
There is an odd circularity, and implicit value judgment, in the classification of participants into "good" and "poor" directed explorers. The logic, based on the visit-counter model of directed exploration, is that the probability of repeating a choice (at the initial decision trial)

should be low for directed explorers vs. random explorers. Doing the median split on repetition probability seems intuitively fine here, but it does bring up two issues. First, the labels "good" vs. "poor" seem arbitrarily judgemental, after all random exploration is a viable exploration strategy in many contexts. Would "directed" vs. "random" be more appropriate labels based on how the decision was made to categorize participants? Second, how much of the "good" participant performance is driven by the extreme non-repeaters? For example, if a tertiary split was performed instead of a binary median split, would the middle group show a weaker version of the effects seen in the "good" group or appear more like the "poor" group?

2. Characterization of information value
The authors discuss primarily methods that can be summarized by visit counters as a description for all directed exploration models. However, that doesn't seem to be a good summary of the overall literature in this space. There are also entropy-based approaches, that quantify information value based on the statistics of the feedback. For example, in machine learning methods like the KL divergence are often used to represent the information value of a channel. A few such papers are highlighted below. Now it is entirely possible that these approaches can be extrapolated to simple visit-count approaches, but I am unaware of anything showing this. I think it would be good to broaden the discussion on directed exploration models beyond visit-counter methods like UCB, highlighting the other methods used to promote directed exploration.

Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., & Abbeel, P. (2016). Vime: Variational information maximizing exploration. Advances in neural information processing systems, 29.

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. arXiv preprint arXiv:1802.06070.

Hazan, E., Kakade, S., Singh, K., & Van Soest, A. (2019, May). Provably efficient maximum entropy exploration. In International Conference on Machine Learning (pp. 2681-2691). PMLR.

3. Model vetting
The model used to simulate the behavioral results is interesting and intuitive. However, there seem to be some things left on the table and unresolved. First, the definition of information value (E) that is maximized is assumed to satisfy the same constraints as typical reward does in the Bellman solution for reinforcement learning. This is the only way it can be substituted into the typical Q-learning method. Is that true here?

Second, the advantage of these simpler computational-level models is that they can be effectively fit to behavior. The model outlined in the paper has only a few free parameters (some of which can be fixed for convenience purposes). Was there an attempt to fit each participant's data into the model? This would be a powerful way of highlighting where exactly the differences between the "good" and "bad" participants arise.