Exploration in Complex Environments: Computational Modeling and Human Behavior

Thesis for the degree of
"Doctor of Philosophy"

by
Lior Fox

Submitted to the Senate of the Hebrew University of Jerusalem
August 2023

This work was carried out under the supervision of

Professor Yonatan Loewnestein

To Ophir

In loving memory of Michael (Mike) Fox

Acknowledgments

This thesis would not have been written if not for the guidance, help, and support of many. Mentioning them here is but an insufficient way to express my deep gratitude:

To Yonatan, my advisor, whose absolute commitment for truth, clarity, and understanding is my inspiration, and my role-model, of what a scientist ought to be. Our time together has been all I could ever have asked for in this journey.

For making our room in the lab a second home, thanks to my friends Lotem and Ohad.

To David, for his presence, and for his consulting.

And to my friends outside the lab, the gang-of-four, to Leshem, Guy, Gal, and Yoav, for endless invaluable discussions in our *shallow conversation on deep learning* discussion group and in real life, and – above all – for true friendship.

I owe a special thanks to Nili, for the support, the attention, and the confident steering of what are sometimes troubled waters of university administrations.

To my former students in the *Computation and Cognition* course, for teaching me how to teach, and for their patience.

Finally, to my family. To my parents, for helping me find my way, and for their love, that has been guiding me through it.

To Ophir, my daughter, the light of my life — watching you grow have elucidated to me how far we are from analytically understanding what learning, intelligence, and behavior truly are, and how far will we forever remain.

To Yaara, for believing in me.

Abstract

Exploration is a fundamental aspect of learning from trial-and-error for two main reasons. First, in this type of learning, the feedback provided for agents on their decisions is often sparse, delayed, and partial; and second, the very distribution of observations available for agents depends on their own behavior. These two important factors differentiate learning by trial-and-error from learning with direct supervision, where in the latter a "teacher" is responsible both for providing the target responses (full feedback), as well as for sampling the learning examples or observations. Thus, agents learning by trial-and-error need to guarantee that their behavior provide a diverse set of examples, representative of the true underlying task or environment; in other words – they need to explore their environment. Exploration, however, is challenging. It is particularly challenging in complex environments, where decisions typically have long-term (exploratory) consequences, which should be taken into account. Thus, learning of the environment is required in order to effectively explore it, but such learning requires, by itself, exploration of the environment.

The first part of this thesis describes two algorithmic approaches that deal with that challenge. One is an uncertainty-driven exploration mechanism, relying on the intuition that this challenge of "learning to explore" is analogous to the challenge of learning a value-function in Reinforcement Learning problems. The method constructs an "exploration value function" (E-value), that can be learned online from observations

and behavior, and serves as a measure for missing knowledge, or uncertainty, of individual state-actions. It is further shown that E-values can be applied for large (or continuous) state-space problems, using function approximation techniques. The second approach is a normative one, in which *optimal* exploration is defined as the one maximizing a particular objective – the entropy of the visitation distribution over the states and actions, induced by the behavioral policy. Solving this optimization problem generates non-trivial policies even in the complete absence of external rewards, as well as in the absence of learning – hence solving the problem of "planning to explore", if a complete knowledge of the environment is given. In the more realistic regime where learning *is* required, it is shown how the (approximately) optimal exploration policy can be learned from observations.

The second part of this thesis studies human exploratory behavior in light of the computational principles identified by the models. For that end, human exploration was evaluated in a set of complex environments, in which long-term consequences of actions exist in the first place. It is shown that human exploration is indeed sensitive to such consequences, suggesting exploration strategies that propagate uncertainties over states and actions, going beyond local measures of uncertainty. Several aspects of human behavior in this task, including some related to learning dynamics, can be well accounted for using the E-values model.

Letter of Contribution

This thesis includes three main chapters and an appendix, detailed as follows:

1. **Lior Fox***, Leshem Choshen*, and Yonatan Loewenstein. DORA the explorer: Directed outreaching reinforcement action-selection. *International Conference on Learning Representations* (ICLR), 2018.

2. **Lior Fox** and Yonatan Loewenstein. Learning Optimal Exploration: a Maximum Entropy Approach. Unpublished, 2019.

3. **Lior Fox**, Ohad Dan, and Yonatan Loewenstein. On the computational principles underlying human exploration. *PsyArxiv preprint*, 2023.
(currently under review).

A. **Lior Fox***, Ohad Dan*, Lotem Elber-Dorozko*, and Loewenstein, Y. Exploration: from machines to humans. *Current Opinion in Behavioral Sciences, 35*, 2020

* denotes equal contribution

These chapters represent my own scientific work. In all of the above I have made significant contributions on all levels, including theoretical contribution, the development and analysis of algorithmic aspects, designing and performing experiments and/or simulations, analysis, and writing. All of these works were done under the supervision of, and in collaboration with, Yonatan Loewenstein. My specific contribution to the aforementioned chapters is as follows:

1. Together with Leshem Choshen, I have developed the algorithm, performed and analyzed simulations, prepared figures, and wrote the paper.

2. I led the research project, developed the algorithms, performed and analyzed simulations, prepared figures, and wrote the paper.

3. I led the research project, designed and programmed the experiment together with Ohad, analyzed the data, performed the computational modeling and simulations, prepared figures, and wrote the paper.

A. The appendix chapter is the result of a joint work with my co-authors Ohad and Lotem. Together, we have constructed the arguments, designed the figures, and wrote the paper.

Contents

Chapter 1

Introduction

עלמא בחרובא אשכחתיה,
כי היכי דשתלי לי אבהתי שתלי נמי לבראי

תענית כג,א

1.1 Reinforcement Learning

The ability to learn rich purposeful behaviors is a hallmark of both animal and human cognition. Understanding how do humans and animals learn – by trial-and-error, in unknown environments, and without direct supervision – has been a fundamental question in psychology, cognitive sciences, and neuroscience for over a century. At the same time, designing systems that can learn similarly, based on their own "experience" of trial-and-error, has long been a motivating goal in the fields of Artificial Intelligence and Machine Learning.

The computational framework of Reinforcement Learning (RL) provides one possible formal point of contact between these two efforts (Sutton and Barto, 2018; Kaelbling et al., 1996). RL offers a set of abstractions, techniques, and algorithms, to analyze, and construct, different forms of trial-and-error learning. It combines concepts from control theory and optimization with concepts from machine learning, resulting in a

flexible framework that can be used to model behavior in a broad range of tasks, from game-playing (Tesauro, 1992; Mnih et al., 2015; Silver et al., 2017), to robotics (Gu et al., 2017), to the control of nuclear reactors (Degrave et al., 2022). Concepts and ideas from RL theory have also been influential in neuroscience and cognitive sciences, for studying both behavior and its underlying neural basis (Schultz et al., 1997; Glimcher, 2011). One reason for that success is that RL models offer a formalization of fundamental psychological principles (Dayan and Niv, 2008; Niv, 2009; Mongillo et al., 2014), hence supporting the generation of precise and quantitative predictions concerning behavior and learning in diverse general environments or conditions. Moreover, the theory can provide a *normative* explanation for behavior, allowing to compare it with optimal solutions.

The modern "canonical" formulation of the RL problem is that of a Markov Decision Process, or MDP (Puterman, 1994). An MDP models the iterative interaction between an *agent* and an *environment*. At each time-step, the agent observes the current state of the environment, and chooses an action. In response to that action, the environment transfers to the next state, and provides a reward signal. The *Markovian* assumption in MDPs is that both the transition to the next state and the reward (both of which can be probabilistic) are *conditionally independent* of past states and actions, given the current state and action. The objective for the agent is to find an optimal *policy* – a mapping from states to actions – such as to maximize the expected cumulative (and often temporally-discounted) reward.

Compared to the Supervised Learning (SL) scenario, RL presents a set of unique challenges, stemming from the more complicated properties of the agent-environment interaction. First, unlike the typical SL scenario, RL tasks are often temporally (and spatially) structured, such that observations made by the agent at different time-steps are not independent and identically distributed (i.i.d). Second, in RL the agent receives only partial feedback from environment, unlike the SL case in which the "teacher" provides the correct response for each observation (Sutton, 1992). Finally, in RL the distribution of examples available for the agent crucially depends on the agent's own behavioral policy, while in SL the examples are sampled from some (fixed, unknown) underlying distribution, independently of the agent's performance.

1.2 Exploration: an Overview

These challenges, and in particular the last one, make *exploration* a crucial component of RL. Put simply, in order to learn an optimal policy, the agent must take into account the fact that observations collected so far might be mis-representative of the true task.

The requirement for exploration is often discussed in the context of the *exploration-exploitation dilemma*. The dilemma presents itself in the *online learning* case, where the agent's goal is to simultaneously maximize reward while learning. In that case, there is, on the one hand, an incentive to repeat actions that have already been proven to be beneficial, i.e., exploiting current knowledge. On the other hand, choosing solely based on current knowledge might be suboptimal, because there could be other actions which are better, but are not yet recognized as such. Therefore, exploration is also necessary, even at the expense of temporary collecting more reward.

Despite much attention given to it, the exploration-exploitation dilemma is not the only exploration-related challenge in RL. Indeed, for the reasons mentioned before, exploration is required even in an "offline learning" settings in which the agent performance is not evaluated concurrently with its learning. In that case, even if there is no need to answer the question of *when* to explore (i.e., balancing exploration with exploitation), there is still the question of *how* to explore. This question, in turn, yields its own "dilemma", which we here term the ***exploration conundrum***: In order to effectively explore, the agent needs knowledge about the environment; but to gain such knowledge the agent needs to effectively explore it in the first place.

One way to overcome this problem is to give up on the "effective" part of exploration. A straightforward implementation is to add a random component to the behavior, a technique known as *random exploration*. By adding randomness, the agent can typically guarantee that all reachable states and actions will, eventually, be visited. However, random exploration can be prohibitively slow and inefficient. Therefore, a second way to overcome the said problem is to have the agent "learning to explore". To do so, the agent has to track, estimate, and update any quantity that can be used to determine what actions are valuable from an exploratory point of view. By prioritizing such actions the agent can actively direct its exploration towards useful states or actions, resulting in what is often termed *directed exploration* (Thrun, 1992).

Despite the contrast in names, directed exploration can be manifested as a random policy, and does not have to be deterministic. Rather, the main difference is whether the exploratory "component" of behavior is adaptive, and relies on some information already collected by the agent, or is it a simple random-walk like policy. The natural question to follow is, therefore: What kind of information can be useful to direct exploration? Different answers have been proposed for that question, spanning a wide range of ideas. These include measures of novelty or surprisal (Pathak et al., 2017), visit-counters (Auer et al., 2002; Kolter and Ng, 2009; Bellemare et al., 2016; Ostrovski et al., 2017; Tang et al., 2017), prediction-errors (Tokic and Palm, 2011; Burda et al., 2019), and information-theoretic quantities such as information gain (Still and Precup, 2012; Little and Sommer, 2014; Houthooft et al., 2016). Broadly speaking, the motivation behind such measures is to represent some form of *uncertainty* the agent has about the environment.[1]

A key observation for the works presented in this thesis is that useful measures for exploration must take into account the *long-term consequences* of actions, and not only their immediate outcomes. This is due to the fact that in complex environments, the uncertainty structure can be complex: in order to fully learn about a given state-action, the agent has to sufficiently explore the other reachable state-actions that follow it. This general principle is relevant regardless of the particular exploration measure employed by the agent. Indeed, it will be shown how this general idea can be incorporated both in the context of counter-based exploration (Chapter 2) and in the context of information-theoretic objectives for exploration (Chapter 3).

Accounting for the long-term exploratory consequences of actions is challenging. On the algorithmic side, it requires models that propagate uncertainty along states and actions, and represent "global" quantities which are sensitive to the environment structure. On the experimental side, studying whether and how humans implement such strategies requires appropriate experimental paradigms, in which long-term exploratory consequences are present in the task to begin with. This thesis consists of two parts,

[1] As opposed to uncertainty that is due to a stochastic nature of the environment itself, which in general cannot be resolved by exploration. Uncertainty due to missing knowledge on the agent part is known as *epistemic uncertainty*, while that due to stochasticity is known as *aleatoric uncertainty*. However, in what follows we will not make much use of these terms.

aimed at addressing both type of challenges. The first part, which includes Chapters 2 and 3, deals with the algorithmic and computational aspects, and form the theoretical contribution of this work. The second part, which includes Appendix A and Chapter 4, deals with the application of the models to study exploratory behavior in humans.

1.3 Computational Modeling

The challenge of learning about the long-term consequences of *exploration* resembles the standard challenge in RL of learning about delayed *rewards*. Indeed, as the goal in RL is typically to maximize the expected cumulative reward, learning about the immediate rewards alone is insufficient, and instead RL algorithms integrate reward information over trajectories in order to learn an optimal policy. This analogy suggests that standard RL techniques, used for learning about reward maximization, could potentially also be useful for learning about exploration.

In Chapter 2, we build on this intuition and present a method for learning an "exploration value-function". The resulting values, termed E-values, serves as a generalization of the familiar visit-counters. While visit-counters are local, measuring only the immediate outcomes of each state-action (i.e., the number of times it has been visisted), our "generalized counters" are sensitive to long-term outcomes such that in actions leading to many future potential states, each (actual) visit contributes less to the generalized visit-counter, compared to each visit of an action that only leads to fewer future states. This property makes E-values a useful measure for directed exploration in complex environments, in which standard visit-counters can be a poor measure of uncertainty.

Another limitation of standard visit-counters that is alleviated by E-values is applicability for problems with large or continuous state spaces. In such problems, standard visit-counters, being a "tabular" method, are impractical, as some generalization over states is essential. Because learning E-values is mathematically equivalent to learning a (particular) Value function in an MDP, it can be readily applied to large or continuous state-space problems by using function approximation techniques.

The E-values method can be thought of as a particular form of *intrinsic motivation*,

a concept which is deeply connected to that of directed exploration. Broadly speaking, the intrinsic motivation idea is that besides the external reward, there are additional factors driving the agent learning and behavior (Schmidhuber, 1991; Storck et al., 1995; Oudeyer and Kaplan, 2009; Barto, 2013). Intrinsically-motivated agents can generate non-trivial exploratory policies even in the complete absence of external rewards. However, exploration in these methods will generally depend on the agent "epistemic state"; that is, the learning process and the information collected by the agent so far.

In Chapter 3 we take a step further and ask whether exploration can be guided by a well-defined objective, which is independent both of external rewards *and* of a particular state in a learning process. Such normative approach for pure exploration can be applied even in the somewhat-extreme case of complete environmental knowledge. The problem then becomes *planning* to explore rather than *learning* to explore – again analogous to the optimal-control problems underlying RL. In the learning regime, an objective function for exploration can guide the agent by providing a well-defined, stationary, target for behavior, which does not itself change as learning progresses.

We propose an objective function for optimal exploration based on the information-theoretical concept of a maximum-entropy distribution. Crucially, the distribution whose entropy we seek to maximize is the distribution over visited state-actions induced by the policy, and not the entropy of the policy itself. While the latter is a local objective, which in the absence of reward will simply result in a uniform random exploration, the former is a global measure which is highly sensitive to the environment structure. Indeed, the resulting optimal exploration policies are structured and highly non-trivial even in the absence of any external reward.

1.4 Human Behavior

Appendix A bridges between the two parts of this thesis. We discuss the applicability of computational models of exploration from RL to the study of human behavior. Under experimental perspective, the role of the computational models changes – we require of them not only to be able to solve the task at hand, but also to provide a good or useful description of the way humans solve it. What makes a description "good" is

a complicated question, and we focus on the ability to identify, based on model predictions, different *principles* that are manifested in human exploration. One reason this might be challenging is that in the theory, the assumptions (either implicit or explicit) that are built into the agents and algorithms are congruent with the actual nature of the task. Humans, on the other hand, have their own internal and mental models of the world, including the world of laboratory experiments, and these might be incongruent with the task design as intended by the researcher. We point out some particular examples and implications of this general issue for the study of exploratory behavior.

Another issue we discuss is that sometimes the task itself might not be rich enough to enable the detection of some principles in the first place. As such, a major shortcoming of prior studies of human exploration is their almost exclusive use of the multi-armed bandit problem as the experimental paradigm. Specifically, in a bandit problem actions do not have long-term consequences, since the environment is characterized by a single state. Therefore, the extent to which humans rely on more global measures and strategies of exploration (of the form discussed previously) cannot be determined by studying their behavior in bandit tasks alone.

In Chapter 4 we take a step towards solving this limitation. We study human exploration behavior using a novel experimental task, that goes beyond the bandit. In our task actions have long-term exploratory consequences, and it is designed such that "local" exploration techniques (e.g., counter-based) and more "global" exploration techniques (e.g., E-values) have opposing predictions regarding behavior. Using this task we are able to show that human exploration is in fact sensitive to long-term exploratory consequences. These findings provide a new perspective for previous works that demonstrated directed exploration of humans in bandit tasks. The specific strategies (e.g., counter-based) identified in bandit tasks might only be a special case implementation of more general principles underlying human exploration: tracking and propagating uncertainties over states and actions, and using this uncertainty to guide exploration.

Finally, we use the E-values model (from Chapter 2) to study the patterns of learning dynamics human participants exhibit in this task. We discuss some aspects of human behavior which remains unexplained by this model alone, most notably their rapid learning in this task.

Chapter 2

DORA The Explorer: Directed Outreaching Reinforcement Action-Selection

יחוגו וינועו כשכור וכל חכמתם תתבלע

תהלים קז

Status: Published

Citation: Fox, L.*, Choshen, L.*, and Loewenstein, Y. (2018). DORA the explorer: Directed outreaching reinforcement action-selection. In *International Conference on Learning Representations* (ICLR).

https://openreview.net/forum?id=ry1arUgCW

DORA THE EXPLORER: DIRECTED OUTREACHING REINFORCEMENT ACTION-SELECTION

Leshem Choshen*
School of Computer Science and Engineering
and Department of Cognitive Sciences
The Hebrew University of Jerusalem
leshem.choshen@mail.huji.ac.il

Lior Fox*
The Edmond and Lily Safra Center for Brain Sciences
The Hebrew University of Jerusalem
lior.fox@mail.huji.ac.il

Yonatan Loewenstein
The Edmond and Lily Safra Center for Brain Sciences,
Departments of Neurobiology and Cognitive Sciences
and the Federmann Center for the Study of Rationality
The Hebrew University of Jerusalem
yonatan@huji.ac.il

ABSTRACT

Exploration is a fundamental aspect of Reinforcement Learning, typically implemented using stochastic action-selection. Exploration, however, can be more efficient if directed toward gaining new world knowledge. Visit-counters have been proven useful both in practice and in theory for directed exploration. However, a major limitation of counters is their locality. While there are a few model-based solutions to this shortcoming, a model-free approach is still missing. We propose E-values, a generalization of counters that can be used to evaluate the propagating exploratory value over state-action trajectories. We compare our approach to commonly used RL techniques, and show that using E-values improves learning and performance over traditional counters. We also show how our method can be implemented with function approximation to efficiently learn continuous MDPs. We demonstrate this by showing that our approach surpasses state of the art performance in the Freeway Atari 2600 game.

1 INTRODUCTION

"If there's a place you gotta go - I'm the one you need to know."
(Map, Dora The Explorer)

We consider Reinforcement Learning in a Markov Decision Process (MDP). An MDP is a five-tuple $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ where $\mathcal{S}$ is a set of *states* and $\mathcal{A}$ is a set of *actions*. The dynamics of the process is given by $P(s'|s,a)$ which denotes the *transition probability* from state s to state s' following action a. Each such transition also has a distribution $R(r|s,a)$ from which the *reward* for such transitions is sampled. Given a *policy* $\pi : \mathcal{S} \to \mathcal{A}$, a function – possibly stochastic – deciding which actions to take in each of the states, the state-action value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ satisfies:

$$Q^\pi(s,a) = \mathop{\mathbb{E}}_{r,s' \sim R \times P(\cdot|s,a)} [r + \gamma Q^\pi(s', \pi(s'))]$$

where γ is the *discount factor*. The agent's goal is to find an optimal policy π^* that maximizes $Q^\pi(s, \pi(s))$. For brevity, $Q^{\pi^*} \triangleq Q^*$. There are two main approaches for learning π^*. The first is a *model-based* approach, where the agent learns an internal model of the MDP (namely P and R). Given a model, the optimal policy could be found using dynamic programming methods such as Value Iteration (Sutton & Barto, 1998). The alternative is a *model-free* approach, where the agent learns only the value function of states or state-action pairs, without learning a model (Kaelbling et al., 1996)[1].

*These authors contributed equally to this work

[1]Supplementary code for this paper can be found at https://github.com/borgr/DORA/

The ideas put forward in this paper are relevant to any model-free learning of MDPs. For concreteness, we focus on a particular example, Q-Learning (Watkins & Dayan, 1992; Sutton & Barto, 1998). Q-Learning is a common method for learning Q^*, where the agent iteratively updates its values of $Q(s,a)$ by performing actions and observing their outcomes. At each step the agent takes action a_t then it is transferred from s_t to s_{t+1} and observe reward r. Then it applies the update rule regulated by a *learning rate* α:

$$Q(s_t,a_t) \leftarrow (1-\alpha)\, Q(s_t,a_t) + \alpha\left(r + \gamma \max_a Q(s_{t+1},a)\right).$$

1.1 Exploration and Exploitation

Balancing between *Exploration* and *Exploitation* is a major challenge in Reinforcement Learning. Seemingly, the agent may want to choose the alternative associated with the highest expected reward, a behavior known as exploitation. However, in that case it may fail to learn that there are better options. Therefore exploration, namely the taking of new actions and the visit of new states, may also be beneficial. It is important to note that exploitation is also inherently relevant for learning, as we want the agent to have better estimations of the values of valuable state-actions and we care less about the exact values of actions that the agent already knows to be clearly inferior.

Formally, to guarantee convergence to Q^*, the Q-Learning algorithm must visit each state-action pair infinitely many times. A naive random walk exploration is sufficient for converging asymptotically. However, such random exploration has two major limitations when the learning process is finite. First, the agent would not utilize its current knowledge about the world to guide its exploration. For example, an action with a known disastrous outcome will be explored over and over again. Second, the agent would not be biased in favor of exploring unvisited trajectories more than the visited ones – hence "wasting" exploration resources on actions and trajectories which are already well known to it.

A widely used method for dealing with the first problem is the ϵ-greedy schema (Sutton & Barto, 1998), in which with probability $1-\epsilon$ the agent greedily chooses the best action (according to current estimation), and with probability ϵ it chooses a random action. Another popular alternative, emphasizing the preference to learn about actions associated with higher rewards, is to draw actions from a *Boltzmann Distribution* (*Softmax*) over the learned Q values, regulated by a *Temperature* parameter. While such approaches lead to more informed exploration that is based on learning experience, they still fail to address the second issue, namely they are not **directed** (Thrun, 1992) towards gaining more knowledge, not biasing actions in the direction of unexplored trajectories.

Another important approach in the study of efficient exploration is based on *Sample Complexity of Exploration* as defined in the PAC-MDP literature (Kakade et al., 2003). Relevant to our work is Delayed Q Learning (Strehl et al., 2006), a model-free algorithm that has theoretical PAC-MDP guarantees. However, to ensure these theoretical guarantees this algorithm uses a conservative exploration which might be impractical (see also (Kolter & Ng, 2009) and Appendix B).

1.2 Current Directed Exploration and its Limitations

In order to achieve directed exploration, the estimation of an exploration value of the different state-actions (often termed *exploration bonus*) is needed. The most commonly used exploration bonus is based on **counting** (Thrun, 1992) – for each pair (s,a), store a counter $C(s,a)$ that indicates how many times the agent performed action a at state s so far. Counter-based methods are widely used both in practice and in theory (Kolter & Ng, 2009; Strehl & Littman, 2008; Guez et al., 2012; Busoniu et al., 2008). Other options for evaluating exploration include **recency** and **value difference** (or **error**) measures (Thrun, 1992; Tokic & Palm, 2011). While all of these exploration measures can be used for directed exploration, their major limitation in a *model-free* settings is that the exploratory value of a state-action pair is evaluated with respect only to its immediate outcome, one step ahead. It seems desirable to determine the exploratory value of an action not only by how much new immediate knowledge the agent gains from it, but also by how much more new knowledge *could* be gained from a trajectory starting with it. The goal of this work is to develop a measure for such exploratory values of state-action pairs, in a *model-free* settings.

2 LEARNING EXPLORATION VALUES

2.1 PROPAGATING EXPLORATION VALUES

The challenge discussed in 1.2 is in fact similar to that of learning the value functions. The value of a state-action represents not only the immediate reward, but also the temporally discounted sum of expected rewards over a trajectory starting from this state and action. Similarly, the "exploration-value" of a state-action should represent not only the immediate knowledge gained but also the expected future gained knowledge. This suggests that a similar approach to that used for value-learning might be appropriate for learning the exploration values as well, using exploration bonus as the immediate reward. However, because it is reasonable to require exploration bonus to decrease over repetitions of the same trajectories, a naive implementation would violate the Markovian property.

This challenge has been addressed in a *model-based* setting: The idea is to use at every step the current estimate of the parameters of the MDP in order to compute, using dynamic programming, the future exploration bonus (Little & Sommer, 2014). However, this solution cannot be implemented in a *model-free* setting. Therefore, a satisfying approach for propagating directed exploration in *model-free* reinforcement learning is still missing. In this section, we propose such an approach.

2.2 E-VALUES

We propose a novel approach for directed exploration, based on two parallel MDPs. One MDP is the original MDP, which is used to estimate the value function. The second MDP is identical except for one important difference. We posit that there are no rewards associated with any of the state-actions. Thus, the true value of all state-action pairs is 0. We will use an RL algorithm to "learn" the "action-values" in this new MDP which we denote as E-values. We will show that these E-values represent the missing knowledge and thus can be used for propagating directed exploration. This will be done by initializing E-values to 1. These positive initial conditions will subsequently result in an optimistic bias that will lead to directed exploration, by giving high estimations only to state-action pairs from which an optimistic outcome has not yet been excluded by the agent's experience.

Formally, given an MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ we construct a new MDP $M' = (\mathcal{S}, \mathcal{A}, P, \mathbf{0}, \gamma_E)$ with $\mathbf{0}$ denoting the identically zero function, and $0 \leq \gamma_E < 1$ is a discount parameter. The agent now learns both Q and E values concurrently, while initially $E(s,a) = 1$ for all s, a. Clearly, $E^* = \mathbf{0}$. However intuitively, the value of $E(s,a)$ at a given timestep during training stands for the knowledge, or uncertainty, that the agent has regarding this state-action pair. Eventually, after enough exploration, there is no additional knowledge left to discover which corresponds to $E(s,a) \to E^*(s,a) = 0$.

For learning E, we use the SARSA algorithm (Rummery & Niranjan, 1994; Sutton & Barto, 1998) which differs from Watkin's Q-Learning by being *on-policy*, following the update rule:

$$E(s_t, a_t) \leftarrow (1 - \alpha_E) E(s_t, a_t) + \alpha_E (r + \gamma_E E(s_{t+1}, a_{t+1}))$$

Where α_E is the learning rate. For simplicity, we will assume throughout the paper that $\alpha_E = \alpha$.

Note that this learning rule updates the E-values based on $E(s_{t+1}, a_{t+1})$ rather than $\max_a E(s_{t+1}, a)$, thus not considering potentially highly informative actions which are never selected. This is important for guaranteeing that exploration values will decrease when repeating the same trajectory (as we will show below). Maintaining these additional updates doesn't affect the asymptotic space/time complexity of the learning algorithm, since it is simply performing the same updates of a standard Q-Learning process twice.

2.3 E-VALUES AS GENERALIZED COUNTERS

The logarithm of E-Values can be thought of as a generalization of visit counters, with propagation of the values along state-action pairs. To see this, let us examine the case of $\gamma_E = 0$ in which there is no propagation from future states. In this case, the update rule is given by:

$$E(s,a) \leftarrow (1 - \alpha) E(s,a) + \alpha (0 + \gamma_E E(s', a')) = (1 - \alpha) E(s,a)$$

So after being visited n times, the value of the state-action pair is $(1 - \alpha)^n$, where α is the learning rate. By taking a logarithm transformation, we can see that $\log_{1-\alpha}(E) = n$. In addition, when s is a terminal state with one action, $\log_{1-\alpha}(E) = n$ for any value of γ_E.

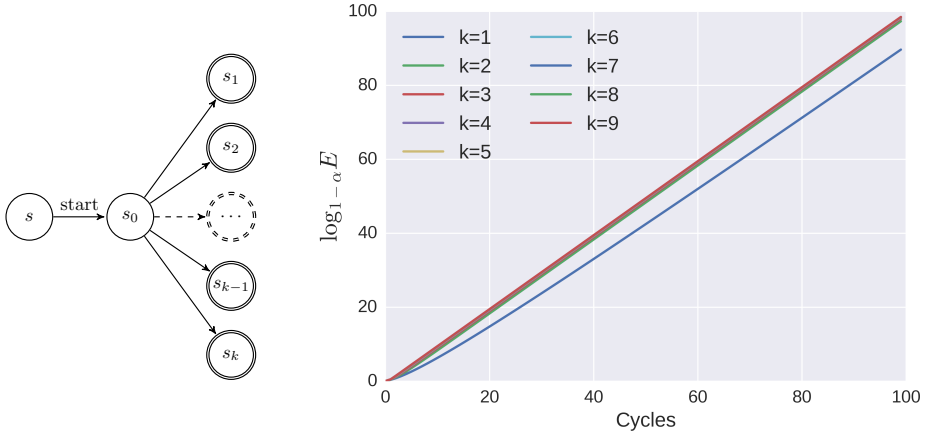


Figure 1: Left: Tree MDP, with k leaves. Tree: $\log_{1-\alpha} E\,(s, \text{start})$ as function of visit cycles, for different trees of k leaves (color coded). For each k, a cycle consists of visiting all leaves, hence k visits of the start action. $\log_{1-\alpha} E$ behaves as a generalized counter, where each cycle contributes approximately one generalized visit.

When $\gamma_E > 0$ and for non-terminal states, E will decrease more slowly and therefore $\log_{1-\alpha} E$ will increase more slowly than a counter. The exact rate will depend on the MDP, the policy and the specific value of γ_E. Crucially, for state-actions which lead to many potential states, each visit contributes less to the generalized counter, because more visits are required to exhaust the potential outcomes of the action. To gain more insight, consider the MDP depicted in Figure 1 left, a tree with the root as initial state and the leaves as terminal states. If actions are chosen sequentially, one leaf after the other, we expect that each complete round of choices (which will result with k actual visits of the (s, start) pair) will be roughly equivalent to one generalized counter. Simulation of this and other simple MDPs show that E-values behave in accordance with such intuitions (see Figure 1 right).

An important property of E-values is that they decrease over repetitions. Formally, by completing a trajectory of the form $s_0, a_0, \ldots, s_n, a_n, s_0, a_0$ in the MDP, the maximal value of $E\,(s_i, a_i)$ will decrease. To see this, assume that $E\,(s_i, a_i)$ was maximal, and consider its value after the update:

$$E\,(s_i, a_i) \leftarrow (1-\alpha)\,E\,(s_i, a_i) + \alpha\gamma_E E\,(s_{i+1}, a_{i+1})$$

Because $\gamma_E < 1$ and $E\,(s_{i+1}, a_{i+1}) \leq E\,(s_i, a_i)$, we get that after the update, the value of $E\,(s_i, a_i)$ decreased. For any non-maximal (s_j, a_j), its value after the update is a convex combination of its previous value and $\gamma_E E\,(s_k, a_k)$ which is not larger than its composing terms, which in turn are smaller than the maximal E-value.

3 Applying E-Values

The logarithm of E-values can be considered as a generalization of counters. As such, algorithms that utilize counters can be generalized to incorporate E-values. Here we consider two such generalizations.

3.1 E-Values as Reward Exploration Bonus

In model-based RL, counters have been used to create an augmented reward function. Motivated by this result, augmenting the reward with a counter-based exploration bonus has also been used in model-free RL (Storck et al., 1995; Bellemare et al., 2016). E-Values can naturally generalize this approach, by replacing the standard counter with its corresponding generalized counter ($\log_{1-\alpha} E$).

To demonstrate the advantage of using E-values over standard counters, we tested an ϵ-greedy agent with an exploration bonus of $\frac{1}{\log_{1-\alpha} E}$ added to the observed reward on the bridge MDP (Figure 2). To measure the learning progress and its convergence, we calculated the mean square error

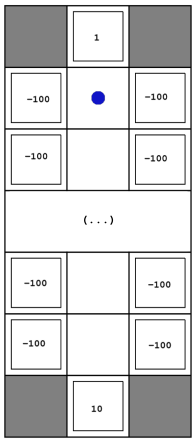


Figure 2: Bridge MDP

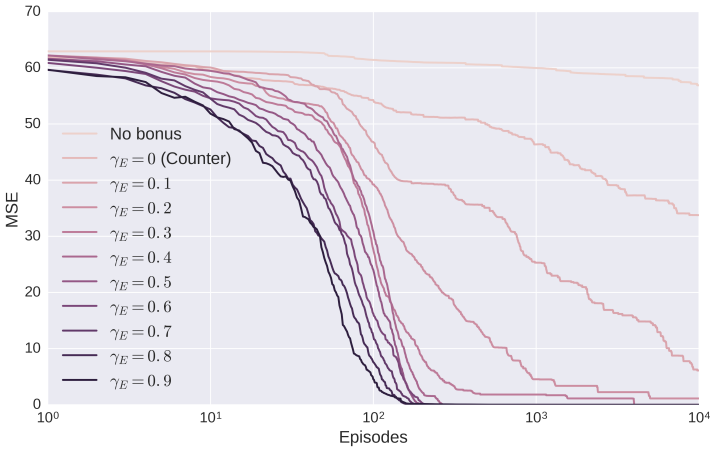


Figure 3: MSE between Q and Q^* on optimal policy per episode. Convergence of ϵ-greedy on the short bridge environment ($k = 5$) with and without exploration bonuses added to the reward. Note the logarithmic scale of the abscissa.

$\mathbb{E}_{P(s,a|\pi^*)}\left[\left(Q\left(s,a\right) - Q^*\left(s,a\right)\right)^2\right]$, where the average is over the probability of state-action pairs when following the optimal policy π^*. We varied the value of γ_E from 0 – resulting effectively in standard counters – to $\gamma_E = 0.9$. Our results (Figure 3) show that adding the exploration bonus to the reward leads to faster learning. Moreover, the larger the value of γ_E in this example the faster the learning, demonstrating that generalized counters significantly outperforming standard counters.

3.2 E-Values and Action-Selection Rules

Another way in which counters can be used to assist exploration is by adding them to the estimated Q-values. In this framework, action-selection is a function not only of the Q-values but also of the counters. Several such action-selection rules have been proposed (Thrun, 1992; Meuleau & Bourgine, 1999; Kolter & Ng, 2009). These usually take the form of a deterministic policy that maximizes some combination of the estimated Q-value with a counter-based exploration bonus. It is easy to generalize such rules using E-values – simply replace the counters C by the generalized counters $\log_{1-\alpha}\left(E\right)$.

3.2.1 Determinization of Stochastic Decision Rules

Here, we consider a special family of action-selection rules that are derived as deterministic equivalents of standard stochastic rules. Stochastic action-selection rules are commonly used in RL. In their simple form they include rules such as the ϵ-greedy or Softmax exploration described above. In this framework, exploratory behavior is achieved by stochastic action selection, independent of past choices. At first glance, it might be unclear how E-values can contribute or improve such rules. We now turn to show that, by using counters, for every stochastic rule there exist equivalent deterministic rules. Once turned to deterministic counter-based rules, it is again possible improve them using E-values.

The stochastic action-selection rules determine the frequency of choosing the different actions in the limit of a large number of repetitions, while abstracting away the specific order of choices. This fact is a key to understanding the relation between deterministic and stochastic rules. An equivalence of two such rules can only be an in-the-limit equivalence, and can be seen as choosing a specific realization of sample from the distribution. Therefore, in order to derive a deterministic equivalent of a given stochastic rule, we only have to make sure that the frequencies of actions selected under both rules are equal in the limit of infinitely many steps. As the probability for each action is likely to depend on the current Q-values, we have to consider fixed Q-values to define this equivalence.

We prove that given a stochastic action-selection rule $f(a|s)$, every deterministic policy that does not choose an action that was visited too many times until now (with respect to the expected number according to the probability distribution) is a determinization of f. Formally, lets assume that given a certain Q function and state s we wish a certain ratio between different choices of actions $a \in A$ to hold. We denote the frequency of this ratio $f_Q(a|s)$. For brevity we assume s and Q are constants and denote $f_Q(a|s) = f(a)$. We also assume a counter $C(s,a)$ is kept denoting the number of choices of a in s. For brevity we denote $C(s,a) = C(a)$ and $\sum_a C(s,a) = C$. When we look at the counters after T steps we use subscript $C_T(a)$. Following this notation, note that $C_T = T$.

Theorem 3.1. *For any sub-linear function $b(t)$ and for any deterministic policy which chooses at step T an action a such that $\frac{C_T(a)}{T} - f(a) \leq b(t)$ it holds that $\forall a \in \mathcal{A}$*

$$\lim_{T\to\infty} \frac{C_T(a)}{T} = f(a)$$

Proof. For a full proof of the theorem see Appendix A in the supplementary materials □

The result above is not a vacuous truth – we now provide two possible determinization rules that achieves it. One rule is straightforward from the theorem, using $b = \mathbf{0}$, choosing $\arg\min_a \frac{C(a)}{C} - f(a)$. Another rule follows the probability ratio between the stochastic policy and the empirical distribution: $\arg\max_a \frac{f(a)}{C(a)}$. We denote this determinization LLL, because when generalized counters are used instead of counters it becomes $\arg\max_a \mathbf{log} f(s,a) - \mathbf{log}\mathbf{log}_{1-\alpha} E(s,a)$.

Now we can replace the visit counters $C(s,a)$ with the generalized counters $\log_{1-\alpha}(E(s,a))$ to create Directed Outreaching Reinforcement Action-Selection – DORA the explorer. By this, we can transform any stochastic or counter-based action-selection rule into a deterministic rule in which exploration propagates over the states and the expected trajectories to follow.

Input: Stochastic action-selection rule f, learning rate α, Exploration discount factor γ_E
initialize $Q(s,a) = 0$, $E(s,a) = 1$;
foreach *episode* **do**
  init s;
  while *not terminated* **do**
    Choose $a = \arg\max_x \log f_Q(x|s) - \log\log_{1-\alpha} E(s,x)$;
    Observe transitions (s,a,r,s',a');
    $Q(s,a) \leftarrow (1-\alpha) Q(s,a) + \alpha(r + \gamma \max_x Q(s',x))$;
    $E(s,a) \leftarrow (1-\alpha) E(s,a) + \alpha\gamma_E E(s',a')$;
  end
end

Algorithm 1: DORA algorithm using LLL determinization for stochastic policy f

3.3 Results – Finite MDPs

To test this algorithm, the first set of experiments were done on Bridge environments of various lengths k (Figure 2). We considered the following agents: ϵ-greedy, Softmax and their respective LLL determinizations (as described in 3.2.1) using both counters and E-values. In addition, we compared a more standard counter-based agent in the form of a UCB-like algorithm (Auer et al., 2002) following an action-selection rule with exploration bonus of $\sqrt{\frac{\log t}{C}}$. We tested two variants of this algorithm, using ordinary visit counters and E-values. Each agent's hyperparameters (ϵ and temperature) were fitted separately to optimize learning. For stochastic agents, we averaged the results over 50 trials for each execution. Unless stated otherwise, $\gamma_E = 0.9$.

We also used a normalized version of the bridge environment, where all rewards are between 0 and 1, to compare DORA with the Delayed Q-Learning algorithm (Strehl et al., 2006).

Our results (Figure 4) demonstrate that E-value based agents outperform both their counter-based and their stochastic equivalents on the bridge problem. As shown in Figure 4, Stochastic and counter-based ϵ-greedy agents, as well as the standard UCB fail to converge. E-value agents are the first to reach low error values, indicating that they learn faster. Similar results were achieved

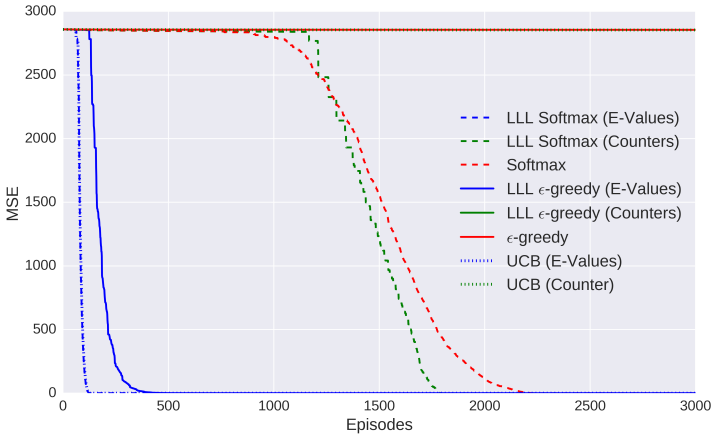


Figure 4: MSE between Q and Q^* on optimal policy per episode. Convergence measure of all agents, long bridge environment ($k = 15$). E-values agents are the first to converge, suggesting their superior learning abilities.

on other gridworld environments, such as the Cliff problem (Sutton & Barto, 1998) (not shown). We also achieved competitive results with respect to Delayed Q Learning (see supplementary B and Figure 7 there).

The success of E-values based learning relative to counter based learning implies that the use of E-values lead to more efficient exploration. If this is indeed the case, we expect E-values to better represent the agent's missing knowledge than visit counters during learning. To test this hypothesis we studied the behavior of an E-value LLL Softmax on a shorter bridge environment ($k = 5$). For a given state-action pair, a measure of the missing knowledge is the normalized distance between its estimated value (Q) and its optimal-policy value (Q^*). We recorded C, $\log_{1-\alpha}(E)$ and $\left|\frac{Q-Q^*}{Q^*}\right|$ for each s, a at the end of each episode. Generally, this measure of missing knowledge is expected to be a monotonously-decreasing function of the number of visits (C). This is indeed true, as depicted in Figure 5 (left). However, considering all state-action pairs, visit counters do not capture well the amount of missing knowledge, as the convergence level depends not only on the counter but also on the identity of the state-action it counts. By contrast, considering the convergence level as a function of the generalized counter (Figure 5, right) reveals a strikingly different pattern. Independently of the state-action identity, the convergence level is a unique function of the generalized counter. These results demonstrate that generalized counters are a useful measure of the amount of missing knowledge.

4 E-VALUES WITH FUNCTION APPROXIMATION

So far we discussed E-values in the tabular case, relying on finite (and small) state and action spaces. However, a main motivation for using model-free approach is that it can be successfully applied in large MDPs where tabular methods are intractable. In this case (in particular for continuous MDPs), achieving directed exploration is a non-trivial task. Because revisiting a state or a state-action pair is unlikely, and because it is intractable to store individual values for all state-action pairs, counter-based methods cannot be directly applied. In fact, most implementations in these cases adopt simple exploration strategies such as ϵ-greedy or softmax (Bellemare et al., 2016).

There are standard model-free techniques to estimate value function in function-approximation scenarios. Because learning E-values is simply learning another value-function, the same techniques can be applied for learning E-values in these scenarios. In this case, the concept of visit-count – or a generalized visit-count – will depend on the representation of states used by the approximating function.

To test whether E-values can serve as generalized visit-counters in the function-approximation case, we used a linear approximation architecture on the MountainCar problem (Moore, 1990) (Appendix

C). To dissociate Q and E-values, actions were chosen by an ϵ-greedy agent independently of E-values. As shown in Appendix C, E-values are an effective way for counting both visits and generalized visits in continuous MDPs. For completeness, we also compared the performance of LLL agents to stochastic agents on a sparse-reward MountainCar problem, and found that LLL agents learns substantially faster than the stochastic agents (Appendix D).

4.1 Results – Function Approximation

To show our approach scales to complex problems, we used the Freeway Atari 2600 game, which is known as a hard exploration problem (Bellemare et al., 2016). We trained a neural network with two streams to predict the Q and E-values. First, we trained the network using standard DQN technique (Mnih et al., 2015), which ignores the E-values. Second, we trained the network while adding an exploration bonus of $\frac{\beta}{\sqrt{-\log E}}$ to the reward (In all reported simulations, $\beta = 0.05$). In both cases, action-selection was performed by an ϵ-greedy rule, as in Bellemare et al. (2016).

Note that the exploration bonus requires $0 < E < 1$. To satisfy this requirement, we applied a logistic activation fucntion on the output of the last layer of the E-value stream, and initialized the weights of this layer to 0. As a result, the E-values were initialized at 0.5 and satisfied $0 < E < 1$ throughout the training. In comparison, no non-linearity was applied in the last layer of the Q-value stream and the weights were randmoly initialized.

We compared our approach to a DQN baseline, as well as to the density model counters suggested by (Bellemare et al., 2016). The baseline used here does not utilize additional enhancements (such as Double DQN and Monte-Carlo return) which were used in (Bellemare et al., 2016). Our results, depicted in Figure 6, demonstrate that the use of E-values outperform both DQN and density model counters baselines. In addition, our approach results in better performance than in (Bellemare et al., 2016) (with the mentioned enhancements), converging in approximately $2 \cdot 10^6$ steps, instead of $10 \cdot 10^6$ steps[2].

5 Related Work

The idea of using reinforcement-learning techniques to estimate exploration can be traced back to Storck et al. (1995) and Meuleau & Bourgine (1999) who also analyzed propagation of uncertainties and exploration values. These works followed a model-based approach, and did not fully deal with the problem of non-Markovity arising from using exploration bonus as the immediate reward. A related approach was used by Little & Sommer (2014), where exploration was investigated by information-theoretic measures. Such interpretation of exploration can also be found in other works (Schmidhuber (1991); Sun et al. (2011); Houthooft et al. (2016)).

Efficient exploration in model-free RL was also analyzed in PAC-MDP framework, most notably the Delayed Q Learning algorithm by Strehl et al. (2006). For further discussion and comparison of our approach with Delayed Q Learning, see 1.1 and Appendix B.

In terms of generalizing Counter-based methods, there has been some works on using counter-like notions for exploration in continuous MDPs (Nouri & Littman, 2009). A more direct attempt was recently proposed by Bellemare et al. (2016). This generalization provides a way to implement visit counters in large, continuous state and action spaces by using density models. Our generalization is different, as it aims first on generalizing the notion of visit counts themselves, from actual counters to "propagating counters". In addition, our approach does not depend on any estimated model – which might be an advantage in domains for which good density models are not available. Nevertheless, we believe that an interesting future work will be comparing between the approach suggested by Bellemare et al. (2016) and our approach, in particular for the case of $\gamma_E = 0$.

[2]We used an existing implementation for DQN and density-model counters available at https://github.com/brendanator/atari-rl. Training with density-model counters was an order of magnitude slower than training with two-streamed network for E-values

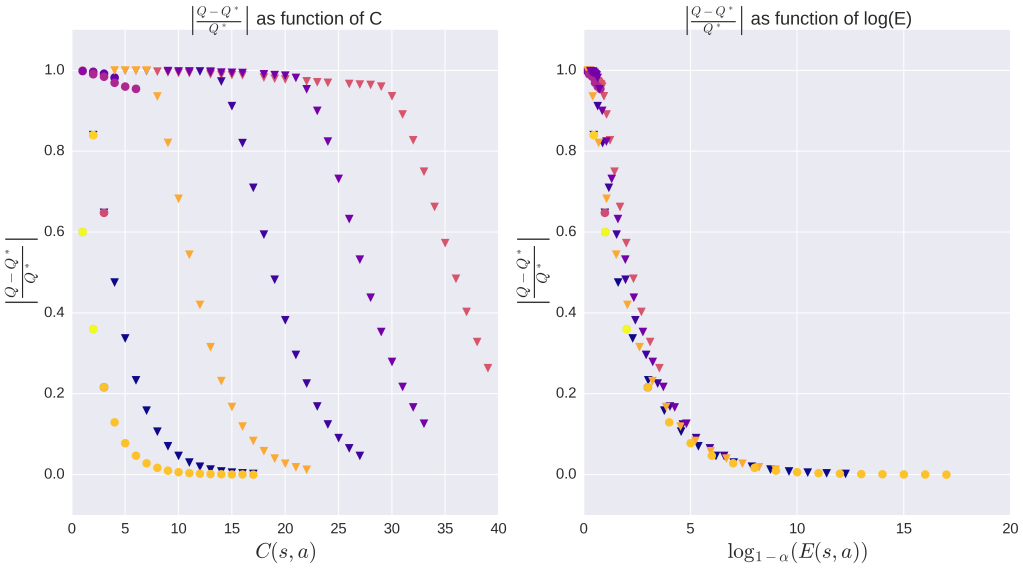


Figure 5: Convergence of Q to Q^* for individual state-action pairs (each denoted by a different color), with respect to counters (left) and generalized counters (right). Results obtained from E-Value LLL Softmax on the short bridge environment ($k = 5$). Triangle markers indicate pairs with "east" actions, which constitute the optimal policy of crossing the bridge. Circle markers indicate state-action pairs that are not part of the optimal policy. Generalized counters are a useful measure of the amount of missing knowledge.

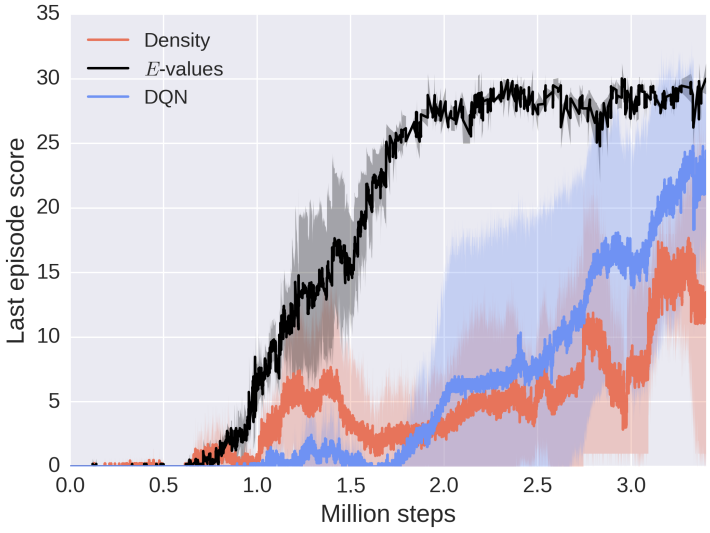


Figure 6: Results on Freeway game. All agents used ϵ-greedy action-selection rule without exploration bonus (DQN, blue), with a bonus term based on density model counters (Density, orange) added to the reward, or with bonus term based on E-values (black).

6 Acknowledgments

We thank Nadav Cohen, Leo Joskowicz, Ron Meir, Michal Moshkovitz, and Jeff Rosenschein for discussions. This work was supported by the Israel Science Foundation (Grant No. 757/16) and the Gatsby Charitable Foundation.

REFERENCES

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

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1471–1479. Curran Associates, Inc., 2016.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, And Cybernetics-Part C: Applications and Reviews, 38 (2), 2008*, 2008.

Arthur Guez, David Silver, and Peter Dayan. Efficient bayes-adaptive reinforcement learning using sample-based search. In *Advances in Neural Information Processing Systems*, pp. 1025–1033, 2012.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pp. 1109–1117, 2016.

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

Sham Machandranath Kakade et al. *On the sample complexity of reinforcement learning*. PhD thesis, 2003.

J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 513–520. ACM, 2009.

Daniel Y Little and Friedrich T Sommer. Learning and exploration in action-perception loops. *Closing the Loop Around Neural Systems*, pp. 295, 2014.

Nicolas Meuleau and Paul Bourgine. Exploration of multi-state environments: Local measures and back-propagation of uncertainty. *Machine Learning*, 35(2):117–154, 1999.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Andrew William Moore. Efficient memory-based learning for robot control. 1990.

Ali Nouri and Michael L Littman. Multi-resolution exploration in continuous spaces. In *Advances in neural information processing systems*, pp. 1209–1216, 2009.

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering, 1994.

Jürgen Schmidhuber. Curious model-building control systems. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, pp. 1458–1463. IEEE, 1991.

Jan Storck, Sepp Hochreiter, and Jürgen Schmidhuber. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the international conference on artificial neural networks, Paris*, volume 2, pp. 159–164. Citeseer, 1995.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 881–888. ACM, 2006.

Yi Sun, Faustino Gomez, and Jürgen Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *International Conference on Artificial General Intelligence*, pp. 41–51. Springer, 2011.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

Sebastian B. Thrun. Efficient exploration in reinforcement learning, 1992.

Michel Tokic and Günther Palm. Value-difference based exploration: adaptive control between epsilon-greedy and softmax. In *KI 2011: Advances in Artificial Intelligence*, pp. 335–346. Springer, 2011.

Thomas J Walsh, István Szita, Carlos Diuk, and Michael L Littman. Exploring compact reinforcement-learning representations with linear regression. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 591–598. AUAI Press, 2009.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

A PROOF OF THE DETERMINIZATION THEOREM

The proof for the determinization mentioned in the paper is achieved based on the following lemmata.

Lemma A.1. *The absolute sum of positive and negative differences between the empiric distribution (deterministic frequency) and goal distribution (non-deterministic frequency) is equal.*

$$\sum_{a:f(a)\geq\frac{C(a)}{C}} f\left(a\right)-\frac{C\left(a\right)}{C}=-\sum_{a:f(a)<\frac{C(a)}{C}} f\left(a\right)-\frac{C\left(a\right)}{C}$$

Proof. Straightforward from the observation that

$$\sum_{a} f\left(a\right)=\sum_{a}\frac{C\left(a\right)}{C}=1$$

□

Lemma A.2. *For any t*

$$\max_{a}\left\{\frac{C_{t}\left(a\right)}{t}-f\left(a\right)\right\}\leq\frac{1+b\left(t\right)}{t}$$

Proof. The proof of A.2 is done by induction. For $t=1$

$$\forall a\in A:\frac{C_{t}\left(a\right)}{t}-f\left(a\right)=\max_{a}\left\{\frac{C_{t}\left(a\right)}{t}-f\left(a\right)\right\}$$

Hence we look at $a\in A$.

$$\begin{aligned}\frac{C_{t}\left(a\right)}{t}-f\left(a\right)&\leq\frac{C_{t}\left(a\right)}{t}\\&\leq\frac{1+b\left(1\right)}{1}\end{aligned}$$

assume the claim is true for $t=T$ then for $t=T+1$ There exists a such that $C_{T}\left(a\right)/T-f\left(a\right)\leq b\left(t\right)$ which the algorithm chooses for this a. For it

$$\begin{aligned}\frac{C_{T+1}\left(a\right)}{T+1}-f\left(a\right)&=\frac{C_{T}\left(a\right)+1}{T+1}-f\left(a\right)\\&=\frac{C_{T}\left(a\right)}{T+1}-f\left(a\right)+\frac{1}{T+1}\\&=\frac{C_{T}\left(a\right)-\left(T+1\right)f\left(a\right)}{T+1}+\frac{1}{T+1}\\&\leq\frac{1+b\left(t\right)}{T+1}\end{aligned}$$

It also holds that $\forall a'\in A$ s.t. $a'\neq a$

$$\begin{aligned}\frac{C_{T+1}\left(a\right)}{T+1}-f\left(a\right)&=\frac{C_{T}\left(a\right)}{T+1}-f\left(a\right)\\&=\frac{C_{T}\left(a\right)-\left(T+1\right)f\left(a\right)}{T+1}\\&<\frac{C_{T}\left(a\right)-Tf\left(a\right)}{T+1}\\&\leq\frac{1+b\left(t\right)}{T+1}\end{aligned}$$

□

Proof of 3.1. It holds from A.2 together with A.1 that in the step t in the worst case all but one of the actions have $\frac{C_t(a)}{t}-f(a)=\frac{1}{t}$ and the last action has $f(a)-\frac{C_t(a)}{t}=-\frac{|A|-1}{t}$. So by the bound on sum of positives and negatives we get:

$$\lim_{T\to\infty}\frac{C_T(a)}{T}=f(a)$$

□

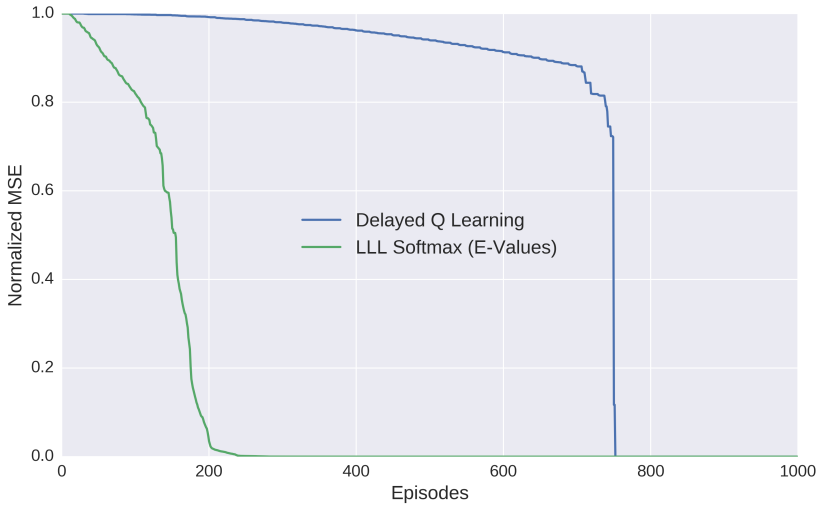


Figure 7: Normalized MSE between Q and Q^* on optimal policy per episode. Convergence of E-value LLL and Delayed Q-Learning on, normalized bridge environment ($k = 15$). MSE was noramlized for each agent to enable comparison.

B Comparison with Delayed Q-Learning

Because Delayed Q learning initializes its values optimistically, which result in a high MSE, we normalized the MSE of the two agents (separately) to enable comparison. Notably, to achieve this performance by the Delayed Q Learning, we had to manually choose a low value for m (in Figure 7, $m = 10$), the hyperparameter regulating the number of visits required before any update. This is an order of magnitude smaller than the theoretical value required for even moderate PAC-requirements in the usual notion of ϵ, δ, such m also implies learning in orders of magnitudes slower. In fact, for this limit of $m \to 1$ the algorithm is effectively quite similar to a "Vanilla" Q-Learning with an optimistic initialization, which is possible due to the assumption made by the algorithm that all rewards are between 0 and 1. In fact, several exploration schemes relying on *optimism in the face of uncertainty* were proposed (Walsh et al., 2009). However, because our approach separate reward values and exploratory values, we are able to use optimism for the latter without assuming any prior knowledge about the first – while still achieving competitive results to an optimistic initialization based on prior knowledge.

C Evaluating E-Values dynamics in function-approximation

To gain insight into the relation between E-values and number of visits, we used the linear-approximation architecture on the MountainCar problem. Note that when using E-values, they are generally correlated with visit counts both because visits result in update of the E-values through learning and because E-values affect visits through the exploration bonus (or action-selection rule). To dissociate the two, Q-values and E-values were learned in parallel in these simulation, but action-selection was independent of the E-values. Rather, actions were chosen by an ϵ-greedy agent. To estimate visit-counts, we recorded the entire set of visited states, and computed the empirical visits histogram by binning the two-dimensional state-space. For each state, its visit counter estimator $\tilde{C}(s)$ is the value of the matching bin in the histogram for this state. In addition, we recorded the learned model (weights vector for E-values) and computed the E-values map by sampling a state for each bin, and calculating its E-values using the model. For simplicity, we consider here the resolution of states alone, summing over all 3 actions for each state. That is, we compare $\tilde{C}(s)$ to $\sum_a \log_{1-\alpha} E(s,a) = C_E(s)$. Figure 8 depicts the empirical visits histogram (left) and the estimated E-values for the case of $\gamma_E = 0$ after the complete training. The results of the analysis show that, roughly speaking, those regions in the state space that were more often visited, were also associated with a higher $C_E(s)$.

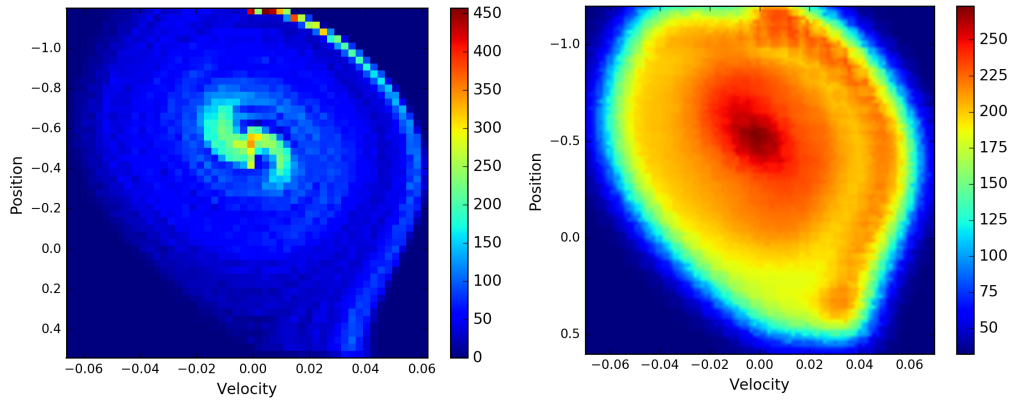


Figure 8: Empirical visits histogram (left) and learned C_E (right) after training, $\gamma_E = 0$.

To better understand these results, we considered smaller time-windows in the learning process. Specifically, Figure 9 depicts the empirical visit histogram (left), and the corresponding $C_E(s)$ (right) in the first 10 episodes, in which visits were more centrally distributed. Figure 10 depicts the *change* in the empirical visit histogram (left), and *change* in the corresponding $C_E(s)$ (right) in the last 10 episodes of the training, in which visits were distributed along a spiral (forming an near-optimal behavior). These results demonstrate high similarity between visit-counts and the E-value representation of them, indicating that E-values are good proxies of visit counters.

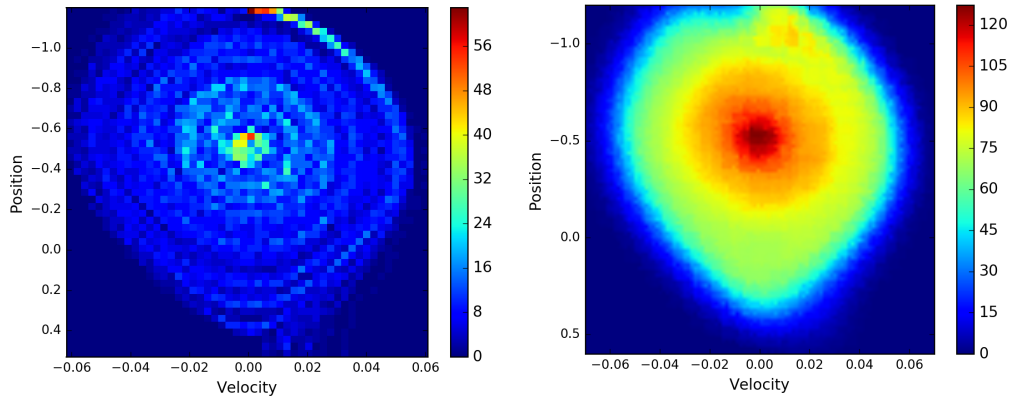


Figure 9: Empirical visits histogram (left) and learned C_E (right) in the first 10 training episodes, $\gamma_E = 0$.

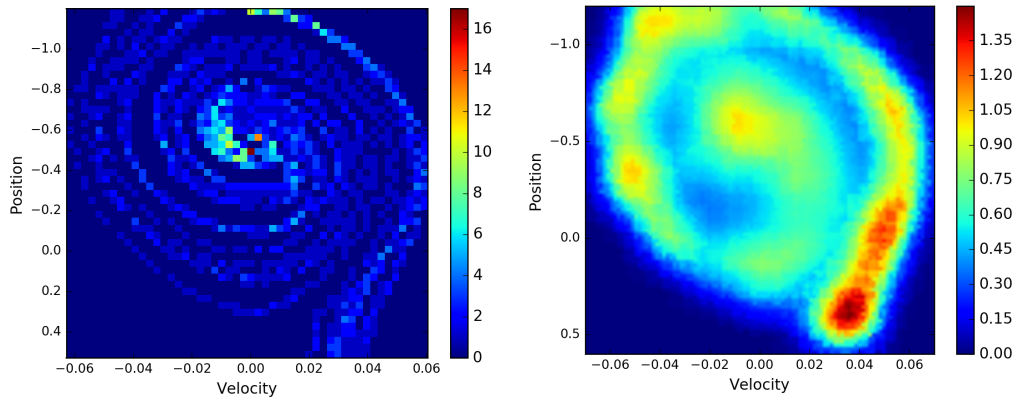


Figure 10: Difference in empirical visits histogram (left) and learned C_E (right) in the last 10 training episodes, $\gamma_E = 0$.

The results depicted in Figures 9 and 10 were achieved with $\gamma_E = 0$. For $\gamma_E > 0$, we expect the generalized counters (represented by E-values) to account not for standard visits but for "generalized visits", weighting the trajectories starting in each state. We repeated the analysis of Figure 10 for the case of $\gamma_E = 0.99$. Results, depicted in Figure 11, shows that indeed for terminal or near-terminal states (where position> 0.5) generalized visits, measured by difference in their generalized counters, are higher – comparing to far-from terminal states – than the empirical visits of these states (comparing to far-from terminal states).

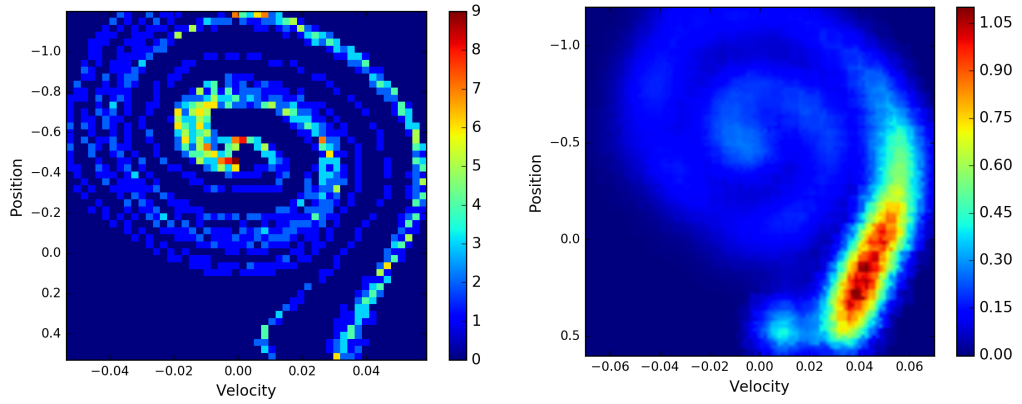


Figure 11: Difference in empirical visits histogram (left) and learned C_E (right) in the last 10 training episodes, $\gamma_E = 0.99$. Note that the results are based on a different simulation than those in Figure 10.

To quantify the relation between visits and E-values, we densely sampled the (achievable) state-space to generate many examples of states. For each sampled state, we computed the correlation coefficient between $C_E(s)$ and $\tilde{C}(s)$ throughout the learning process (snapshots taken each 10 episodes). The values $\tilde{C}(s)$ were estimated by the empirical visits histogram (value of the bin corresponding to the sampled state) calculated based on visits history up to each snapshot. Figure 12, depicting the histogram of correlation coefficients between the two measures, demonstrating strong positive correlations between empirical visit-counters and generalized counters represented by E-values. These results indicate that E-values are an effective way for counting effective visits in continuous MDPs. Note that the number of model parameters used to estimate $E(s, a)$ in this case is much smaller than the size of the table we would have to use in order to track state-action counters in such binning resolution.

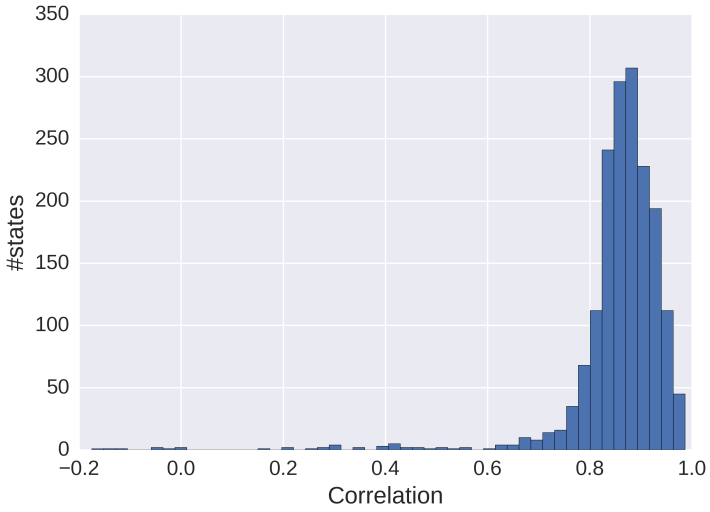


Figure 12: Histogram of correlation coefficients between empirical visit counters and C_E throughout training, per state ($\gamma_E = 0$).

D RESULTS ON CONTINUOUS MDPS – MOUNTAINCAR

To test the performance of E-values based agents, simulations were performed using the MountainCar environment. The version of the problem considered here is with sparse and delayed reward, meaning that there is a constant reward of 0 unless reaching a goal state which provides a reward of magnitude 1. Episode length was limited to 1000 steps. We used linear approximation with tile-coding features (Sutton & Barto, 1998), learning the weights vectors for Q and E in parallel. To guarantee that E-values are uniformly initialized and are kept between 0 and 1 throughout learning, we initialized the weights vector for E-values to 0 and added a logistic non-linearity to the results of the standard linear approximation. In contrast, the Q-values weights vector was initialized at random, and there was no non-linearity. We compared the performance of several agents. The first two used only Q-values, with a softmax or an ϵ-greedy action-selection rules. The other two agents are the DORA variants using both Q and E values, following the LLL determinization for softmax either with $\gamma_E = 0$ or with $\gamma_E = 0.99$. Parameters for each agent (temperature and ϵ) were fitted separately to maximize performance. The results depicted in Figure 13 demonstrate that using E-values with $\gamma_E > 0$ lead to better performance in the MountainCar problem

In addition we tested our approach using (relatively simple) neural networks. We trained two neural networks in parallel (unlike the two-streams single network used for Atari simulations), for predicting Q and E values. In this architecture, the same technique of 0 initializing and a logistic non-linearity was applied to the last linear of the E-network. Similarly to the linear approximation approach, E-values based agents outperform their ϵ-greedy and softmax counterparts (not shown).

Figure 13: Probability of reaching goal on MountainCar (computed by averaging over 50 simulations of each agent), as a function of training episodes. While Softmax exploration fails to solve the problem within 1000 episodes, LLL E-values agents with generalized counters ($\gamma_E > 0$) quickly reach high success rates.

Chapter 3

Learning Optimal Exploration: a Maximum Entropy Approach

כי כארבע רוחות השמים פרשתי אתכם

זכריה ב

Status: Unpublished

Learning Optimal Exploration: a Maximum Entropy Approach

Lior Fox[*]
lior.fox@mail.huji.ac.il

Yonatan Loewenstein[†]
yonatan@huji.ac.il

Abstract

Efficient exploration is crucial for the performance of agents in complex environments. Exploration has been traditionally studied in the framework of Reinforcement Learning, in which the objective is maximizing a reward function. Here, we consider the question of optimal exploration independently of a reward-maximizing goal. We define the exploratory fitness of a policy as the entropy of its induced discounted visitation distribution. We start with the *planning* scenario, showing how an optimal exploration policy can be found efficiently when the transition model is known, and study its properties. Then, we discuss *learning*, showing how the principles of this optimal exploration can be applied when the transition model is unknown, for both model-based and model-free learning.

1 Introduction

Exploration is typically considered in the framework of Reinforcement Learning (RL) as a way of finding the policy that maximize the expected long-term (possibly discounted) reward. In this framework, all policies are equally good in the absence of rewards. However intuitively, even in the absence of any rewards some policies are more "effective" – from an exploratory point of view – than others. The goal of this work is to present a framework that formalizes this intuition by defining an optimality criterion for exploration, and to study the properties of the resulting optimal exploration policies. We posit that an optimal exploratory policy is one which efficiently covers as much of the environment as possible. In what follows, we mathematically define both "efficiently" and "covers".

Most previous works attempting at similar goals suggested some form of "intrinsic motivation" (Storck et al., 1995; Chentanez et al., 2005; Oudeyer and Kaplan, 2009) which could serve as an internal reward signal for the agent. Such internal rewards could be based on visit counters (Bellemare et al., 2016; Tang et al., 2017; Ostrovski et al., 2017; Fox et al., 2018), prediction errors (Pathak et al., 2017; Burda et al., 2019), information gain (Little and Sommer, 2014; Still and Precup, 2012; Houthooft et al., 2016), and other related ideas to measure exploratory usefulness of states (Thrun, 1992). We could characterize most of these as using some measure of the agent's own learning (e.g. visits of novel states, or changes in the estimated parameters of the environment's model) as a reinforcement signal (Schmidhuber, 1991). By contrast, the quality of exploration in our framework is an intrinsic property of the policy, which is independent of the "internal state" (or the particular history) of the agent.

In §2 we define the settings and the objective function for optimal exploration, and show how the optimization problem of finding such optimal exploration policy can be solved efficiently when the model of the environment is known. In §3 we demonstrate this approach using several challenging environments and study the resulting policies. In §4 we discuss the case when the model is unknown, and the exploration policy has to be learned from observations – in both model-based and model-free frameworks. We conclude with a more detailed review of related work, and a further discussion.

[*] The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem

[†] The Edmond and Lily Safra Center for Brain Sciences, Department of Cognitive Sciences, The Alexander Silberman Institute of Life Sciences, and The Federmann Center for the Study of Rationality The Hebrew University, Jerusalem

2 Optimality criterion for exploration: Maximum-Entropy approach

2.1 Settings and the objective function

We consider the settings of a Markov-Decision-Process (MDP) $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, P, r, \rho_S^0\right)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $P\left(s'|s,a\right)$ is the transition model, $r\left(s,a\right)$ is the reward function (which will not play any role in this work), and $\rho_S^0\left(s\right)$ is a distribution over initial states. For clarity, we assume that the number of actions available at each state is equal, but our approach and results can be easily generalized when this is not the case.

Any stationary policy π – a (stochastic) mapping from states to actions – induces a distribution over trajectories $\tau = \left(s_0, a_0, s_1, a_1, \ldots\right)$, which factorizes according to the Markov property:

$$\mathbb{P}_\pi\left[\tau\right] = \rho_S^0\left(s_0\right)\pi\left(a_0|s_0\right)\prod_{t>0} P\left(s_t|s_{t-1}a_{t-1}\right)\pi\left(a_t|s_t\right) \tag{1}$$

Central to our work is the notion of the *discounted visitation distribution* induced by a policy π. This distribution, over state-action pairs, measures the occupancies of each such pair when trajectories are generated by the policy π (i.e, sampled according to Equation (1)), with future visits contributing less due to the exponential *temporal discounting* with factor $0 < \gamma < 1$. We denote this distribution, for a policy π and a discount factor γ, as $\boldsymbol{\rho}_\gamma^\pi$:

$$\rho_\gamma^\pi\left(s,a\right) = \left(1-\gamma\right)\sum_{t=0}^{\infty}\gamma^t \sum_{\tau: s_t=s, a_t=a} \mathbb{P}_\pi\left[\tau\right]$$

where the $\left(1-\gamma\right)$ pre-factor is required for proper normalization, such that $\sum_{s,a}\rho_\gamma^\pi\left(s,a\right) = 1$. Explicitly marginalizing time, this could be written as:

$$\rho_\gamma^\pi\left(s,a\right) = \left(1-\gamma\right)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}_\pi^{(t)}\left[s,a\right] \tag{2}$$

where $\mathbb{P}_\pi^{(t)}\left[s,a\right]$ is the probability of visiting $\left(s,a\right)$ at time t.

We postulate that an optimal-exploration policy is one that maximizes the entropy of the induced discounted visitation distribution over the state-action pairs. In other words, a policy that results in visits of all state-action pairs as uniformly as possible, and as early as possible. This poses the following optimization problem:

$$\pi^* \equiv \arg\max_\pi H\left[\boldsymbol{\rho}_\gamma^\pi\right] \tag{3}$$

where $H\left[\cdot\right]$ is the standard information-theoretic entropy of a distribution.

2.2 Basic properties of $\boldsymbol{\rho}_\gamma^\pi$

For a stationary policy π, it is clear from equation (1) that the sequence of visited state-action pairs forms a Markov Chain, with the transition matrix $\mathbf{W}^\pi$: $W^\pi_{sa,s'a'} = P\left(s|s',a'\right)\pi\left(a|s\right)$ (where we index entries by state-action pairs). Furthermore, the initial distribution over state-action pairs, which we define as $\boldsymbol{\rho}^{\mathbf{0}}$, is: $\rho^0\left(s,a\right) = \rho_S^0\left(s\right)\pi\left(a|s\right)$. Note that $\rho^0\left(s,a\right) = \mathbb{P}_\pi^{(0)}\left[s,a\right]$.

Lemma 1. *$\boldsymbol{\rho}_\gamma^\pi$ is the unique solution to equation $\boldsymbol{\rho} = \left(1-\gamma\right)\boldsymbol{\rho}^{\mathbf{0}} + \gamma\mathbf{W}^\pi\boldsymbol{\rho}$. That is:*

$$\boldsymbol{\rho}_\gamma^\pi = \left(1-\gamma\right)\left(\mathbf{I}-\gamma\mathbf{W}^\pi\right)^{-1}\boldsymbol{\rho}^{\mathbf{0}} \tag{4}$$

Lemma 2. *Let $\boldsymbol{\rho}_\gamma^\pi$ be the discounted visitation distribution induced by the policy π. Then for any state s with non-zero (marginal) probability under $\boldsymbol{\rho}_\gamma^\pi$: $\pi\left(a|s\right) = \frac{\rho_\gamma^\pi(s,a)}{\sum_{a'}\rho_\gamma^\pi(s,a')}$*

The proofs are standard and can be found in the literature (e.g (Wang et al., 2007; Puterman, 1990)). For completeness, we include proofs using our notation in the supplementary material.

2.3 Finding an optimal exploration policy

Solving the optimization problem of Equation (3) directly is challenging, because the induced discounted visitation distribution $\boldsymbol{\rho}_\gamma^\pi$ is a complicated, non-linear function of the policy π (Equation (4)). We therefore present an alternative approach, inspired by dual methods in classical RL (Wang et al., 2007).

The key observation is that instead of maximizing over policies, it is possible instead to optimize directly over state-action probability distributions, under appropriate constraints (which we define below). Again, this is analogous to solving standard RL problem by optimizing such probabilities rather than the value function or the policy. Clearly, an arbitrary probability distribution ρ over state-action pairs is not necessarily realizable as the discounted visitation distribution of any policy π. If, however, ρ satisfies the "self-consistency" condition(s) in the form appearing in Lemma 1, then it is guaranteed that a corresponding π exist such that $\rho = \rho_\gamma^\pi$. Moreover such π is easily constructed from ρ itself (in the same way as in Lemma 2).

Note that π explicitly appears in Lemma 1. Therefore, to complete the argument, the constraints on ρ must be re-written so as to omit the explicit dependencies on π, which are due to $\mathbf{W}^\pi$ and $\boldsymbol{\rho}^{\mathbf{0}}$ being functions of π. The following theorem establishes this, and form the basis for our main algorithm (Algorithm 1). This reduces the optimization problem in 3 to a standard, convex optimization

Algorithm 1 Maximum-Entropy optimal exploration

Input MDP $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, P, r, \rho_S^0\right)$, discount factor γ
Solve the following optimization problem to get ρ^*:

$$\begin{aligned}
\max_{\rho} \quad & -\sum_{s,a} \rho\left(s,a\right) \log \rho\left(s,a\right) \\
\text{s.t} \quad & \rho\left(s,a\right) \geq 0, \qquad \sum_{s,a} \rho\left(s,a\right) = 1, \\
& \sum_{a'} \rho\left(s,a'\right) = \left(1-\gamma\right) \rho_S^0\left(s\right) \\
& \qquad + \gamma \sum_{s',a'} \rho\left(s',a'\right) P\left[s|s',a'\right]
\end{aligned}$$

Output exploration policy $\pi^*\left(a|s\right) = \frac{\rho^*(s,a)}{\sum_{a'} \rho^*(s,a')}$

problem – finding a maximum-entropy distribution under linear constraints.

Theorem 3. *Assume $\rho\left(s,a\right)$ is a distribution over state-action pairs, such that for every state s, the following condition holds:*

$$\sum_{a'} \rho\left(s,a'\right) = \left(1-\gamma\right) \rho_S^0\left(s\right) + \gamma \sum_{s',a'} \rho\left(s',a'\right) P\left(s|s',a'\right) \tag{5}$$

Let $\pi\left(a|s\right) = \frac{\rho(s,a)}{\sum_{a'} \rho(s,a')}$. Then, $\boldsymbol{\rho} = \boldsymbol{\rho}_\gamma^\pi$.

The proof is provided in the supplementary material.

3 Properties of optimal exploration

To study the properties of the proposed optimal exploration, we apply it in several environments. We show that the resultant policies are able to implicitly overcome key challenges for effective exploration, such as balancing short-term and long-term outcomes, and achieving "determined" exploration which is temporally extended.

3.1 Temporally extended exploration

We consider a variant of the well-known N-chain MDP (Strens, 2000), depicted in Figure 1(a). The environment consists of N states $s_1, \ldots, s_N$. At each state, the agent can either "Step", advancing to the next state, or "Exit", returning to the initial state s_1. In s_N, both actions lead back to s_1.

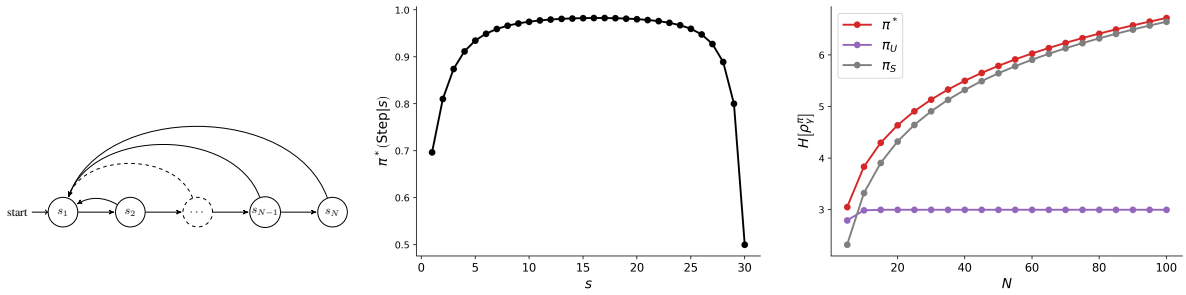
In this environment, a uniform policy explores very poorly, because its probability of reaching a state decreases exponentially fast with the number of "Steps" required to reach it. Indeed, as shown in Figure 1(b), the optimal exploration policy (Algorithm 1) is biased towards 'Step', allowing the agent to explore more uniformly the different state-actions. The tendency to take the "Step" action depends on the specific location in the chain, k in an inverted U-shape manner. For small k states, the probability to "Step" increases with k because the larger k, the more sparsely visited are the subsequent states. Near the end of the chain, the probability of "Step" decreases because there are not so many states left to explore ahead.

To quantify the effectiveness of the optimal exploration policy, Figure 1(c) depicts the entropy of the discounted visitation distributions of several policies, as a function of the environment size N. The entropy of $\boldsymbol{\rho}_\gamma^\pi$ is almost independent of the size of the chain for uniform exploration (purple) indicating that this policy fails to explore in but the smallest environments. By contrast, for optimal exploration (red), the entropy of $\boldsymbol{\rho}_\gamma^\pi$ increases with the size of the chain, indicating that optimal exploration can well-cover much larger environments. For comparison, the simple always "step" heuristic (gray) does substantially better than a uniform policy, but is still suboptimal. This is because we require maximal entropy of a distribution over state-actions, and not only states.

3.2 Diverse exploration

In general, the optimal exploration policy is stochastic, yielding a distribution of trajectories. To see that, consider the gridworld MDP depicted in Figure 2. The environment is a large gridworld, consisting of four rooms separated by walls, with only 1-tile sized "doorways" connecting different rooms. Reaching the lower-left corner from the top-left corner requires the agent to pass through 3 specific state-action pairs (namely, go through the aforementioned "doorways"), which are unlikely to be reached in random exploration. Indeed, the discounted visitation distribution of a uniform exploration policy (Figure 2(a)) decays exponentially fast with the geodesic distance (i.e. the distance of the shortest trajectory) from the initial state, and trajectories rarely get out of the first room, exhibiting dithering behavior around the initial state (Figure 2(b)). This exemplifies, again, the inefficiency of random exploration in MDPs in which a large number of specific actions is needed in order to reach some of the states. By contrast, our maximum-entropy approach finds a policy that achieves temporally extended exploration (Figures 2(c) and 2(d)). This is despite the fact that the policy is purely reactive and memoryless, and does not rely explicitly on any form of temporal-abstractions such as options (Sutton et al., 1999).

Comparing Figures 2(a) and 2(c), it is worthwhile noting that while in 2(a) the distribution decreases mono-



(a) (b) (c)

Figure 1: Results on N-chain MDP. (a) Illustration of the environment. The agent start at s_1, and at each state can choose between "step" and "exit". (b) The optimal exploration policies (probability of choosing "step" at each state) for different values of γ ($N = 30$, $\gamma = 0.999$). (c) The entropy of ρ^π_γ ($\gamma = 0.999$) for a uniform policy (π_U), a policy that always chooses "step" (π_S), and the optimal exploration policy (π^*) in different environment sizes.

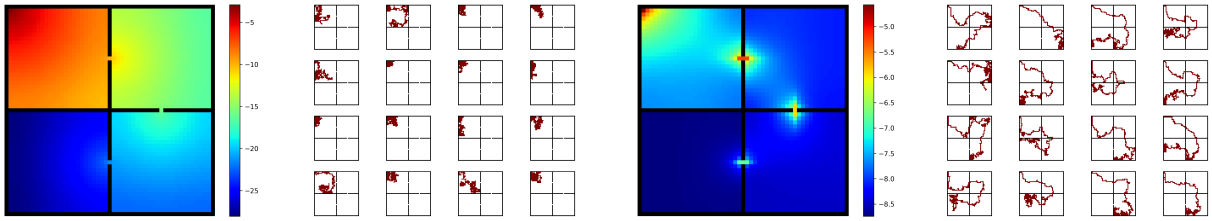
tonically with the (geodesic) distance from the initial state, the discounted visitation distribution of the optimal exploration policy (Figure 2(c)) has significant modes around the doorways. These modes result in a visitation distribution that locally, is less uniform than that of a random policy but allows for a more uniform global visitation by identifying the bottlenecks in the environment.

Importantly, while being sampled from the same stationary policy, different optimal policy trajectories exhibit diverse paths, typically visiting all four rooms (Figure 2(d)). This allows the policy to explore – on average – nearby locations (e.g, the first room) while also quickly reach remote states (e.g, the fourth room).

3.3 Discounted exploration

The parameter γ controls the effective length of the exploratory trajectories. In the limit of $\gamma = 0$, the discounted visitation distribution is reduced to the initial distribution (over state-actions), that is $\rho^\pi_\gamma = \rho^0$. In this case, an optimal solution is achieved by setting the policy to be uniform in each state which has non-zero initial probability (under ρ^0_S). This exploratory policy does not take into consideration the long-term consequences of exploration beyond the immediate action. In the limit of $\gamma = 1$ ρ^π_γ reduces to the stationary distribution of the Markov-chain induced by π, when this distribution exists. In particular, the stationary distribution is independent of the initial distribution ρ^0. Since there is no discounting, later visits of state-action pairs contribute just as much as early visits. When $0 < \gamma < 1$, the discounting sets an effective length for trajectories, beyond which contribution to the discounted visitation distribution is negligible. In this case, an optimal exploration policy has to "well cover" the state-action space – on average – with finite-length trajectories.

In some reinforcement-learning settings, the agent chooses between small rewards available immediately and larger rewards in the long-term. In such a case the optimal policy depends on the discount factor, which determines the trade-off between short- and long-term goals. Similarly, the optimal exploration policy will depend on the discount factor if there is a conflict between short-term and long-term exploratory goals. The results depicted in Figure 3 demonstrates that even in a relatively-simple MDP, the dependence of the optimal exploration policy on the discount factor is non-trivial. The agent in this environment has 4 possible actions. Specifically, from the initial state (red square) it can go left to explore the smaller but closer room, or go right to explore the bigger, but farther away room. In this case, the optimal exploration policy at the initial state (i.e, the preference for going right or left) depends in a non-monotonic way on γ, as depicted in Figure 3 (right). When $\gamma = 0$, all four actions have the same exploratory value because only the immediate exploration (initial visit) is considered. As γ increases, going left becomes favorable since it leads to many potential states (the small room to the left). As γ further increases, the contribution from reaching the bigger room (farther to the right) becomes significant, and the preference is switched towards going right. In the limit when γ reaches 1, all four actions are again equally preferable. This is because in such a domain, a random-walk policy will result in a uniform stationary distribution, due to the underlying diffusion-like dynamics.



(a) (b) (c) (d)

Figure 2: Gridworld MDP consists of four large rooms (25×25 tiles each), separated by walls and connected only through 1-tile sized "doorways". The initial state is the upper-left corner. We use $\gamma = 0.99$. All sampled trajectories are of length $T = 500$. **Left**: log-probability map of the discounted visitation distribution (actions marginalized out) (a) and randomly chosen examples of sampled trajectories (b) for a uniform policy, choosing actions with equal probability. **Right:** log-probability map of the discounted visitation distribution (actions marginalized out) (c) and randomly chosen examples of sampled trajectories (d) for the optimal exploration policy. Note the difference in color scales between (a) and (c).

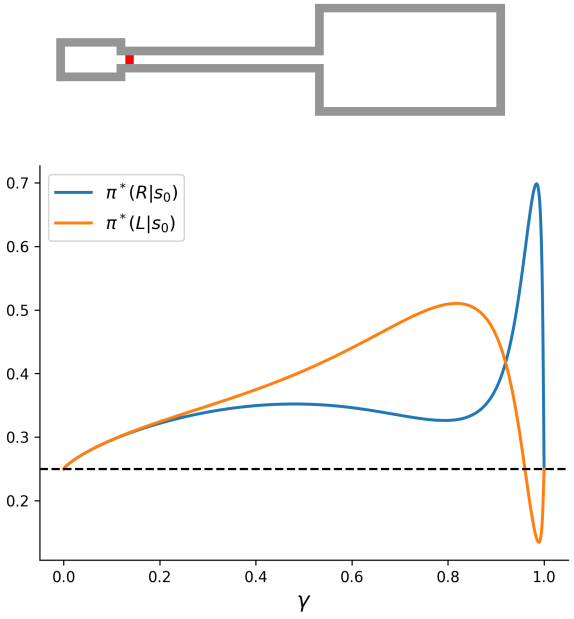


Figure 3: **Up**: Gridworld MDP. Red tile denotes the initial state s_0, and gray tiles are walls. The agent could move left to explore a relatively small room (6×3 tiles) that is nearby (one tile away), or move right to explore a bigger room (20×11 tiles) which is further away (22 tiles). **Down**: The optimal exploration policy for actions Right (blue), Left (orange) at s_0, as a function of γ. The probabilities for actions Up, Down (not shown) are equal to each other throughout. Dashed horizontal line denotes probability 0.25, where all actions are chosen uniformly.

4 Learning with optimal exploration

Optimal exploration when the parameters MDP are fully known is an interesting theoretical question. It is comparable to the planning problem of finding optimal policy (in the sense of maximizing reward) when the MDP is fully known. However typically, the goal of exploration is to learn the MDP or a good policy. Therefore, we now turn to discuss how optimal exploration policy can be learned when the parameters of the MDP are not known. We consider two approaches to this problem: one is model-based, in which the parameters of the MDP are learned concurrently with the learning of the optimal exploration policy. In the other, the exploration policy is learned directly from experience.

4.1 Model-Based learning of exploration

As shown in §2, the optimal exploration policy can be efficiently computed if the parameters of the MDP are known. Complementary to that, a good exploratory policy facilitates the learning of these parameters. We propose to learn the MDP parameters and the optimal exploration policy iteratively. At each "epoch", the agent computes the optimal exploration policy (using Algorithm 1) according to its current estimated model of the environment (which can also be the prior in the beginning of learning). The resultant policy is then executed in the environment to collect more samples, and the agent improve its model estimation.

The effectiveness of this procedure is exemplified in Figure 4(a) for the chain MDP (presented in Figure 1(a)). Initially, the policy is uniform. Within 30 episodes (80 steps per episode) the policy converges to the optimal exploration policy. We quantified the model learning us-

ing Missing-Information – the sum of KL-divergences $\sum_{s,a} D_{KL}\left[P\left(\cdot|s,a\right) \parallel \hat{P}\left[\cdot|s,a\right]\right]$ where $\hat{P}$ is the estimated model (Little and Sommer, 2014). Trivially, our procedure (blue) converges much faster than a random explorer (orange) which fails for this environment. Remarkably, learning was also faster than that of a similar iterative approach, in which the exploration policy was defined in terms of maximizing the (long-term) predictive information gain (VI-PIG, green; (Little and Sommer, 2014)).

4.2 Optimal exploration by Policy Gradient

Next, we consider a model-free scenario in which an agent learns a policy directly, without explicitly estimating the transition model. We rely on policy gradient methods to learn an approximated optimal exploration policy, by optimizing an objective function that we define below.

Let $\tau = (s_0, a_0, \ldots, s_T, a_T)$ be a trajectory. We define the discounted visit counters as $\tilde{C}\left(s,a\right) = \sum_{t=0}^{T} \gamma^t \mathbb{1}_{[s_t=s, a_t=a]}$. Note that for $T \to \infty$, we have $\left(1-\gamma\right)\mathbb{E}\left[\tilde{C}\left(s,a\right)\right] = \rho_\gamma^\pi\left(s,a\right)$. Based on these, we define the empirical visitation distribution, and use its entropy to define the return of a trajectory. Formally: $R_{\text{ent}}\left(\tau\right) = -\sum_{s,a} \tilde{p}_{sa} \log \tilde{p}_{sa}$, where $\tilde{p}_{sa} = \frac{\tilde{C}(s,a)}{\sum_{s'a'} \tilde{C}(s',a')}$ (See Algorithm 2).

Note that the defined objective ("reward") is a function of an entire trajectory, and cannot be decomposed to sum of Markovian rewards. Nevertheless, Policy Gradient methods can still be used to optimize such reward functions (see also (Shalev-Shwartz et al., 2016)). Learning is used to optimize $\mathbb{E}\left[R_{\text{ent}}\left(\tau\right)\right]$. Due to Jensen's inequality, $\mathbb{E}\left[R_{\text{ent}}\left(\tau\right)\right] \leq H\left[\boldsymbol{\rho}_\gamma^\pi\right]$ (Paninski, 2003), so by optimizing the expected return $\mathbb{E}\left[R_{\text{ent}}\left(\tau\right)\right]$ we optimize a lower-bound on the original objective – the entropy of the discounted visitation distribution. Note that this bound can be tightened if we calculate $\tilde{p}_{sa}$ based on averaging several trajectories.

We demonstrate this approach in the chain MDP, tracking R_{ent} and $H\left[\boldsymbol{\rho}_\gamma^\pi\right]$ throughout the learning process (of π). Note that neither the true distribution $\boldsymbol{\rho}_\gamma^\pi$ nor its entropy are known to the agent. Nevertheless, as depicted in Figure 4(b), learning indeed optimizes R_{ent} and R_{ent} is a lower bound of $H\left[\boldsymbol{\rho}_\gamma^\pi\right]$. Therefore, learning improves the desired objective of $H\left[\boldsymbol{\rho}_\gamma^\pi\right]$, in this example leading to a policy whose entropy is indistinguishable from that of the optimal exploration policy.

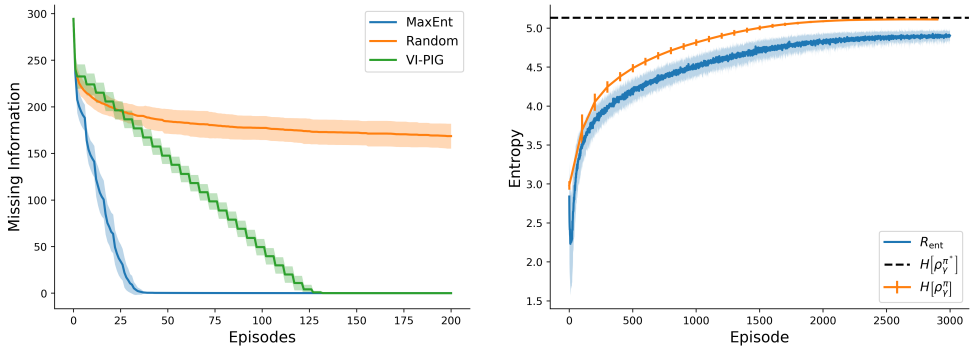
Algorithm 2 Policy Gradient for learning exploration

init policy parameters θ, Learning rate η, discount factor γ
repeat
Rollout π in the environment to sample trajectory τ
Define $\tilde{C}\left(s,a\right) = \sum_{t=0}^{T} \gamma^t \mathbb{1}_{[s_t=s, a_t=a]}$, $\tilde{p}_{sa} = \frac{\tilde{C}(s,a)}{\sum_{s'a'} \tilde{C}(s',a')}$, $R_{\text{ent}}\left(\tau\right) = -\sum_{s,a} \tilde{p}_{sa} \log \tilde{p}_{sa}$.
Update parameters: $\theta \leftarrow \theta + \eta\left(R_{\text{ent}}\left(\tau\right) - b\right)\sum_t \nabla \log \pi\left(a_t|s_t\right)$ $\{b$ is an optional baseline$\}$
until converged

4.3 Function approximation

For MDPs with continuous (or very large) state-space, one may try to apply Algorithm 2 by using density models (e.g (Bellemare et al., 2016; Fox et al., 2018; Tang et al., 2017; Ostrovski et al., 2017)) to approximate the (empirical) visitation distribution. However, there are several challenges. First, this approach relies on the existence of such counters or models. Second, it can be challenging to estimate $\tilde{p}_{sa}$ because the denominator ($\sum_{s'a'} \tilde{C}\left(s',a'\right)$) requires the integration over the entire state-action space, which can be intractable in many problems. Therefore, we considered an alternative simpler approach. The main idea is to apply algorithm 2 *in feature space*, rather than in the actual state-space. These features do not have to form a counting or a density model, and could be (but do not have to be) the set of features used by the policy.

We demonstrate this approach using the MountainCar MDP (Moore, 1990). The state-space (position and velocity) is featurized using random Fourier features (Rahimi and Recht, 2008; 2009). The learned policy is log-linear in these predefined features (see Supplementary Material). Our objective is to find the policy that maximizes the entropy of the discounted "visitation" distribution in feature space. We approximated this entropy by binarizing the features and computing the empirical discounted distribution (probability of being "on") of every feature. The return (R_{ent}) was then defined as the mean of binary entropies across all features. This is a rather crude approximation, ignoring both the analog nature of the features and the correlations between them. Nevertheless, as depicted in Figure 5, we find that in practice, this approach is very effective, yielding useful exploratory behavior in this problem. To test the generality of this approach, we also applied it to the Acrobot MDP (Sutton, 1996), using the same features and entropy approximation. As shown in the Figure 6, our approach yields effective and useful exploratory behavior in this



(a) (b)

Figure 4: Model-based and model-free learning in the chain-MDP introduced in Figure 1(a) ($N = 30$, $\gamma = 0.999$, Episode length limited to 80 steps). **(a)** Model-based learning. All agents stated from a uniform prior. For the non-random agents, the exploration policy was updated each 5 episode. **(b)** Model-free learning with policy-gradient. The agent starts with a random policy and is trained to optimize R_{ent} (see main text and Algorithm 2). This optimizes a lower-bound on the entropy of ρ_γ^π (which is unavailable to the agent). The policy-gradient algorithm used is vanilla REINFORCE (Williams, 1992) with baseline subtraction (average of R_{ent} over last 20 episodes). Average results over 60 simulations, shaded region and error bars denote standard deviations (**(a)** and **(b)**).

problem as well. In fact we find that the learned policies readily reach the goal state using pure exploration, despite the fact that this was not an explicit goal in the learning process, and the agents never observe the "external" reward of the environment (Figure 7). Together, these results demonstrate the potential of this approach for exploration in environments with sparse or delayed rewards, in which complex behavior(s) has to be learned before any external reward is observed.

5 Discussion and related work

We presented a novel approach for defining optimal exploration in the absence of reward signals. In our framework, the goal is to maximize the entropy of the discounted visitation distribution induced by the policy. We showed how the resultant optimal policies overcome key challenges of exploration in MDPs, namely achieving temporally extended exploration and balancing long-term and short-term exploratory outcomes. We presented an efficient algorithm for finding optimal exploration policy when the transition model of the MDP is known, as well as an approximating model-free algorithm that maximizes a lower-bound on the aforementioned entropy.

The discounted visitation distribution rises naturally in standard RL settings, where it can be used to compute the value-function of a given policy (Puterman, 1990). The reason for its centrality is the fact that this distribution encodes the statistics of all future (including temporal discounting) states visited by starting at a given initial condition. This fact underlies the motivation of using these *multiple* distributions – one for each "initial" state-action pair – as a useful *representation* of the state-action itself, a concept known as the Successor Representation (SR) (Dayan, 1993), making the value-function linear in the representation. However, one important limitation of this approach is that the representation itself is *policy-dependent*. Moreover, there is no natural method for simple, online, "improvement" procedure analogue to the policy-improvement step in value-based methods. By contrast, in our work, the goal is to find a specific optimal policy for which the specific induced visitation distribution has a desired property (of maximum entropy). Moreover, we take a particular initial-state distribution to be part of the MDP definition, and this distribution is not necessarily concentrated on a single state-action pair. As we have shown, in order to learn (or compute) that optimal policy, it is possible to use various forms of state representations, either tabular, or approximated.

There is a vast literature on exploration in RL. Particularly relevant for this work are approaches which disassociate exploration from external reward and rely on some sort of "intrinsic motivation", as discussed in §1 (and references therein).

Another aspect of exploration that is relevant to this work is the challenge of achieving temporally-extended (sometimes referred to as "deep") exploration (Osband et al., 2016a;b). Independently of reward, this can be addressed in the framework of model-based learning (Little

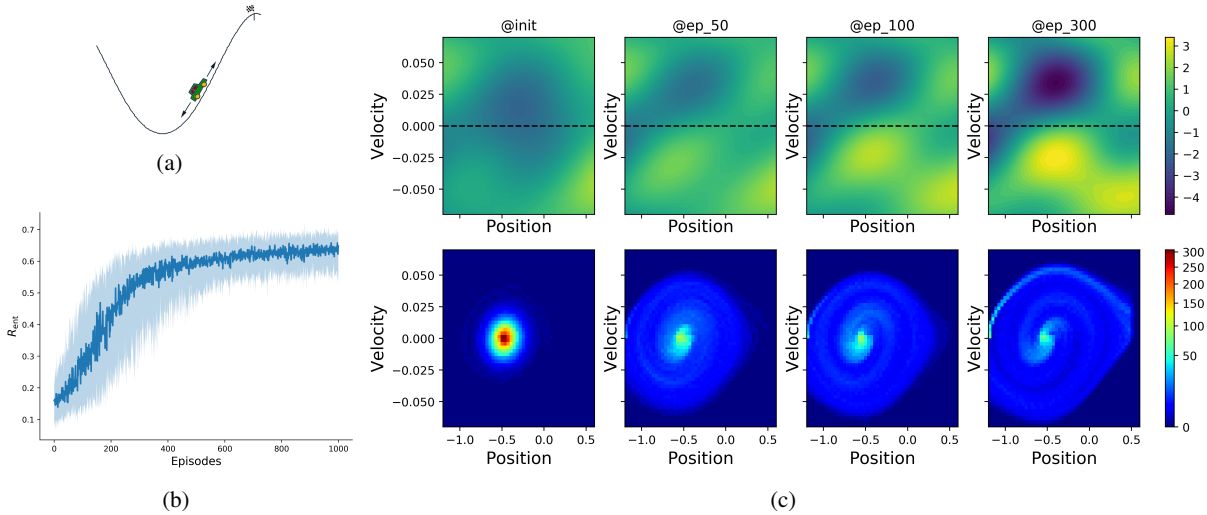


Figure 5: **(a)** The MountainCar MDP. **(b)** Median "reward" R_{ent} (solid line; 60 simulations) increases with learning (see text). Shaded region denotes $25-75$ percentile. **(c)** Illustration of the learning process, showing a policy and its induced visitation distribution at learning episodes 0 (initilaization, random policy), 50, 100, and 300. Top row depicts $\log \frac{\pi(\text{Left}|s)}{\pi(\text{Right}|s)}$. Horizontal dashed line denotes states with velocity $= 0$. Bottom row depicts histograms (normalized; undiscounted) of visitations in 500 sampled trajectories from the policy. With the learned policy, the agent is able to cover the (reachable) state-space much more uniformly, comparing to a random policy. Notably, the learned policy generate diverse trajectories rather than choosing a particular trajectory.

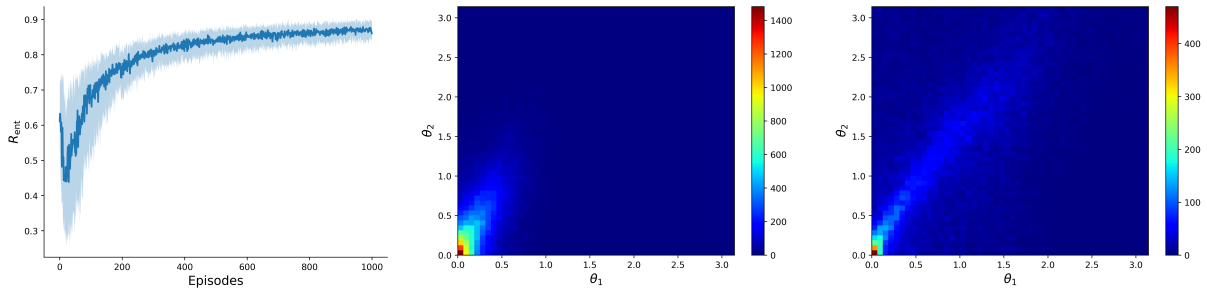
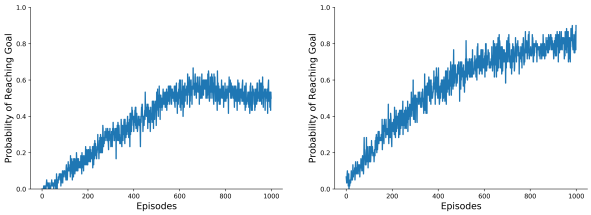


Figure 6: Results on the Acrobot MDP (Sutton, 1996). **(a)** Median "reward" R_{ent} (solid line; 60 simulations) increases with learning. Shaded region denotes $25-75$ percentile. **(b)**, **(c)** Histograms of (undiscounted) state visits in 100 sampled trajectories from a random policy and from a trained policy, respectively. Visits are shown in the $\theta_1\theta_2$ plane for visualization only – the original state-space include angular velocities $\dot{\theta}_1, \dot{\theta}_2$ as well.

and Sommer, 2014). Recently, generalized counters have been proposed (Fox et al., 2018) as a method accounting for long-term consequences of exploration. These generalized counters have been effectively implemented in model-free RL (Fox et al., 2018; Oh and Iyengar, 2018). In our work, temporally-extended exploration is achieved by defining a global objective which depends on the interaction of the policy with the MDP dynamics. This yields policies that take into account the future consequences of actions, where the relevant future is controlled by γ.

Entropy-based objectives for exploration has been previously proposed. However, most of those studies considered the entropy of *the policy* itself (Mnih et al., 2016; Schulman et al., 2017; Haarnoja et al., 2018). This approach, however, does not take into account the exploratory long-term consequences of actions. As a result, in the absence of (external) reward, these methods imply that an optimal exploration policy is to choose ac-



(a) (b)

Figure 7: Proportion of agents (out of 60 simulated agents) to reach the goal state per episode for the MountainCar MDP **(a)** and the Acrobot MDP **(b)**. Agents learn to reach goal by pure exploration, without ever observing any external rewards.

tions with equal probability. Independently of our work, a recent study has proposed the same objective of maximizing the entropy of the discounted visitation distribution for exploration (Hazan et al., 2018). There are several notable differences between that study and our work. First, when the MDP is known, we show how a single stationary policy can be efficiently computed from the MDP (§2.3), rather than finding a mixture policy, as proposed there. This approach also enabled us to study the properties of the optimal exploration policy (§3). Second, we propose a different solution in the case of unknown MDPs. One advantage of our algorithm is that it can be readily combined with standard RL methods for maximizing average reward. For example, one can apply Policy Gradient learning on a linear combination of the external reward and the exploration objective function R_{ent} (Algorithm 2). Finally, we propose a method for maximizing exploration in the feature space. This method is applicable for continuous-state MDPs without relying on density models for the state-space.

Learning complex behavior in the absence of reward have also been studied from the perspective of options, and in particular options discovery (Sutton et al., 1999; Machado and Bowling, 2016; Machado et al., 2017). Our approach yields policies which diversely cover the state-space while also identifying important bottleneck states (e.g. Figure 2), however it does not explicitly rely on temporal-abstractions or non-stationary policies.

In the context of standard RL problems, effective exploration should also be sensitive to the reward signal, because it is typically more useful to explore around the state-actions that are estimated to be more valuable. In such tasks our pure exploration approach can be combined with reward-based learning. Optimal exploration is particularly relevant to tasks in which the reward function is not stationary. Such tasks are common in neuroscience, where animals are often trained to repeatedly search for food pallets placed in random locations in a known environments (Knierim et al., 1995; Spiers et al., 2013). Other relevant settings are task-agnostic RL scenarios, in which the goal is to learn an environment in the absence of rewards, in order to later quickly solve (possibly multiple) reward-related tasks.

References

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1471–1479. Curran Associates, Inc., 2016.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Nuttapong Chentanez, Andrew G Barto, and Satinder P Singh. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 1281–1288, 2005.

Peter Dayan. Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation*, 5(4):613–624, 07 1993. ISSN 0899-7667. doi: 10.1162/neco.1993.5.4.613. URL https://doi.org/10.1162/neco.1993.5.4.613.

Lior Fox, Leshem Choshen, and Yonatan Loewenstein. DORA the explorer: Directed outreaching reinforcement action-selection. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ry1arUgCW.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/haarnoja18b.html.

Elad Hazan, Sham M Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. *arXiv preprint arXiv:1812.02690*, 2018.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in*

Neural Information Processing Systems, pages 1109–1117, 2016.

James J Knierim, Hemant S Kudrimoti, and Bruce L McNaughton. Place cells, head direction cells, and the learning of landmark stability. *Journal of Neuroscience*, 15(3):1648–1659, 1995.

Daniel Y Little and Friedrich T Sommer. Learning and exploration in action-perception loops. *Closing the Loop Around Neural Systems*, page 295, 2014.

Marlos C Machado and Michael Bowling. Learning purposeful behaviour in the absence of rewards. *arXiv preprint arXiv:1605.07700*, 2016.

Marlos C Machado, Marc G Bellemare, and Michael Bowling. A laplacian framework for option discovery in reinforcement learning. In *International Conference on Machine Learning*, pages 2295–2304, 2017.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

Andrew William Moore. Efficient memory-based learning for robot control. 1990.

Min-hwan Oh and Garud Iyengar. Directed exploration in pac model-free reinforcement learning. *arXiv preprint arXiv:1808.10552*, 2018.

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

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2377–2386, New York, New York, USA, 20–22 Jun 2016b. PMLR.

Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Rémi Munos. Count-based exploration with neural density models. *arXiv preprint arXiv:1703.01310*, 2017.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. doi: 10.1162/089976603321780272. URL https://doi.org/10.1162/089976603321780272.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008. URL http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.pdf.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1313–1320. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3495-weighted-sums-of-random-kitchen-sinks-replacing-minimization-with-randomization-in-learning.pdf.

Jürgen Schmidhuber. Curious model-building control systems. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, pages 1458–1463. IEEE, 1991.

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Hugo J Spiers, Robin MA Hayman, Aleksandar Jovalekic, Elizabeth Marozzi, and Kathryn J Jeffery. Place field repetition and purely local remapping in a multicompartment environment. *Cerebral Cortex*, 25(1): 10–25, 2013.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Jan Storck, Sepp Hochreiter, and Jürgen Schmidhuber. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the international conference on artificial neural networks, Paris*, volume 2, pages 159–164. Citeseer, 1995.

Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, pages 943–950, 2000.

Richard S Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding.

In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 1038–1044. MIT Press, 1996. URL http://papers.nips.cc/paper/1109-generalization-in-reinforcement-learning-successful-examples-using-sparse-coarse-coding.pdf.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2753–2762, 2017.

Sebastian B. Thrun. Efficient exploration in reinforcement learning, 1992.

Tao Wang, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *Approximate Dynamic Programming and Reinforcement Learning, 2007. ADPRL 2007. IEEE International Symposium on*, pages 44–51. IEEE, 2007.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

A Proofs of lemmas and theorem

Lemma 1. *$\boldsymbol{\rho}_\gamma^\pi$ is the unique solution to equation $\boldsymbol{\rho} = (1-\gamma)\,\boldsymbol{\rho^0} + \gamma \mathbf{W}^\pi \boldsymbol{\rho}$. That is:*

$$\boldsymbol{\rho}_\gamma^\pi = (1-\gamma)\left(\mathbf{I} - \gamma \mathbf{W}^\pi\right)^{-1} \boldsymbol{\rho^0} \tag{1}$$

Proof. By definition,

$$\rho_\gamma^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi^{(t)}[s,a] \tag{2}$$

Separating the sum in equation (2) to $t=0$ and $t>0$ yields:

$$\rho_\gamma^\pi(s,a) = (1-\gamma)\,\mathbb{P}_\pi^{(0)}[s,a] + \gamma(1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi^{(t+1)}[s,a]$$

The first summand is by definition $(1-\gamma)\,\rho^0(s,a)$. For the second summand, because the visited state-actions form a Markov chain, then:

$$\mathbb{P}_\pi^{(t+1)}[s,a] = \sum_{s',a'} W_{sa,s'a'}^\pi \mathbb{P}_\pi^{(t)}[s',a']$$

and therefore

$$\begin{aligned}(1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi^{(t+1)}[s,a] &= \sum_{s'a'} W_{sa,s'a'}^\pi (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi^{(t)}[s',a'] \\ &= \sum_{s'a'} W_{sa,s'a'}^\pi \rho_\gamma^\pi(s'a')\end{aligned}$$

Combining both terms yields the following recursive equation (in vector form):

$$\boldsymbol{\rho}_\gamma^\pi = (1-\gamma)\,\boldsymbol{\rho^0} + \gamma \mathbf{W}^\pi \boldsymbol{\rho}_\gamma^\pi \tag{3}$$

Rearranging terms completes the proof. Note that $\mathbf{I} - \gamma\mathbf{W}^\pi$ is non-singular because $\mathbf{W}^\pi$ is a stochastic matrix and therefore its largest eigenvalue (in absolute value) is equal to 1. Because $0 < \gamma < 1$, ρ_γ^π is the unique solution to equation (3). □

Lemma 2. *Let $\boldsymbol{\rho}_\gamma^\pi$ be the discounted visitation distribution induced by the policy π. Then for any state s with non-zero (marginal) probability under $\boldsymbol{\rho}_\gamma^\pi$: $\pi(a|s) = \frac{\rho_\gamma^\pi(s,a)}{\sum_{a'} \rho_\gamma^\pi(s,a')}$*

Proof. Since the policy is stationary, for every t we have $\mathbb{P}_\pi^{(t)}[s,a] = \mathbb{P}_\pi^{(t)}[s]\,\pi(a|s)$. Using the definition of $\rho_\gamma^\pi(s,a)$ (Equation (2)) and substituting the former identity yields the result. □

Theorem 3. *Assume $\rho(s,a)$ is a distribution over state-action pairs, such that for every state s, the following condition holds:*

$$\sum_{a'} \rho(s,a') = (1-\gamma)\,\rho_S^0(s) + \gamma \sum_{s',a'} \rho(s',a')\,P(s|s',a') \tag{4}$$

Let $\pi(a|s) = \frac{\rho(s,a)}{\sum_a' \rho(s,a')}$. Then, $\boldsymbol{\rho} = \boldsymbol{\rho}_\gamma^\pi$.

Proof. We assume $\sum_{a'} \rho(s,a') > 0$ for all states. If this is not the case for a particular state s, then we can arbitrarily define $\pi(a|s) = \frac{1}{|\mathcal{A}|}$.

Multiplying Equation (4) by $\pi(a|s) = \frac{\rho(s,a)}{\sum_a' \rho(s,a')}$ yields

$$\rho(s,a) = (1-\gamma)\,\rho_S^0(s)\,\pi(a|s) + \gamma \sum_{s',a'} P(s|s',a')\,\pi(a|s)\,\rho(s',a')$$

where the first summand is $(1-\gamma)\,\rho^0(s,a)$ and the second summand is $\gamma \sum_{s',a'} W_{sa,s'a'}^\pi \rho(s',a')$. Therefore, we get the following vector equation:

$$\boldsymbol{\rho} = (1-\gamma)\,\boldsymbol{\rho^0} + \gamma \mathbf{W}^\pi \boldsymbol{\rho}$$

Using Lemma 1 completes the proof, since $\boldsymbol{\rho}_\gamma^\pi$ is the unique solution to the last equation. □

B Maximum-Entropy exploration in feature-space: linear policies with random features

Here we provides the technical details regrading the implementation of our approach to the MountainCar and Acrobot problems (§4.3). We standardized the observations (so that entries of the state vectors $\mathbf{s}$ has 0 mean and unit variance), and constructed random features of the state-space using a variant of Random Kitchen Sinks with Fourier features[1] [2, 3]:

$$\phi_i(\mathbf{s}) = \frac{\sqrt{2}}{\sqrt{d}} \cos\left(\mathbf{a}_i^\top \mathbf{s} + \mathbf{b}_i\right)$$

where $\mathbf{a}_i, \mathbf{b}_i$ are random vectors with $\mathbf{a} \sim \mathcal{N}(0, 2\alpha I)$ and $\mathbf{b} \sim \mathcal{U}[0, 2\pi]$. α and d are parameters (where d is the number of features). All reported results are using $d = 20$. The feature space was constructed by concatenating two such feature vectors with $\alpha = 1$ and $\alpha = 0.5$. We denote the feature-vector for a given state as $\boldsymbol{\phi}(\mathbf{s})$.

Figure 1 illustrate a particular set of such random features in the MountainCar MDP (in which the original state space is two dimensional, position and velocity):

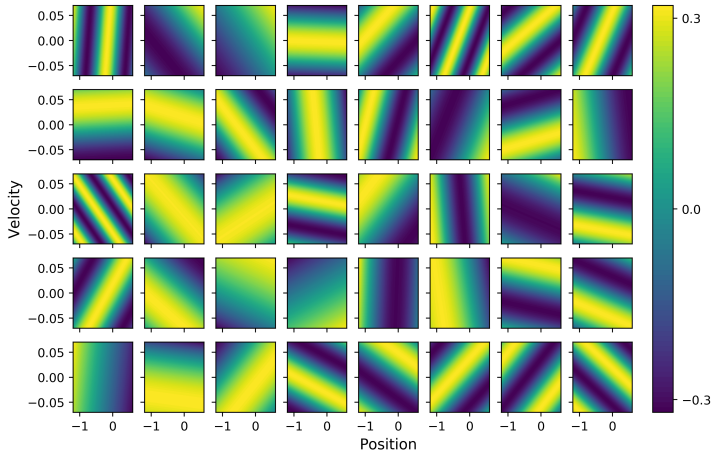


Figure 1: Example of 40 random features for the MountainCar MDP.

The policy is parameterized by a matrix $\mathbf{W}$ of dimensions $|\mathcal{A}| \times 2d$, such that $\pi(\cdot|s) \propto \exp\{\mathbf{W}\boldsymbol{\phi}(\mathbf{s})\}$. The score function for this model is:

$$\frac{\partial}{\partial W_{ij}} \log \pi(a|\mathbf{s}) = \phi_j(\mathbf{s}) \left(\delta_{ia} - \pi(a_i|\mathbf{s})\right) \tag{5}$$

Given a sampled trajectory (episode) $\tau = (s_1, a_1, \ldots, s_T, a_T)$, we calculate, for each feature, the (discounted) probability of being positive:

$$q_i = \frac{\sum_{t=0}^{T} \gamma^t \mathbb{1}_{[\phi_i(s_t)>0]}}{\sum_{t=0}^{T} \gamma^t}$$

The return of the trajectory is then defined to be the average (over features) of the binary entropies of these probabilities:

$$R_{\text{ent}}(\tau) = \frac{1}{2d} \sum_{i=1}^{2d} H_2[q_i] \tag{6}$$

where $H_2[q] = -q \log_2 q - (1-q) \log_2 (1-q)$. Note that $0 \leq R_{\text{ent}} \leq 1$.

Taken together, Equations (5) and (6) specify the required quantities for the Policy Gradient learning rule discussed in the main text (§4.2 and §4.3).

[1] Implemented by the RBFSAMPLER function in Scikit [1]

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[2] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008. URL `http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.pdf`.

[3] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1313–1320. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3495-weighted-sums-of-random-kitchen-sinks-replacing-minimization-with-randomization-in-learning.pdf`.

Chapter 4

On the Computational Principles Underlying Human Exploration

פעם אחת הייתי מהלך בדרך וראיתי תינוק יושב על פרשת דרכים ואמרתי לו: "באיזה דרך נלך לעיר?" אמר לי "זו קצרה וארוכה וזו ארוכה וקצרה" והלכתי בקצרה וארוכה, כיון שהגעתי לעיר מצאתי שמקיפין אותה גנות ופרדיסין. חזרתי לאחורי אמרתי לו "בני, הלא אמרת לי קצרה" אמר לי: "ולא אמרתי לך ארוכה?"

עירובין נג,ב

Status: Submitted, under review

Citation (preprint): Fox, L., Dan, O., and Loewenstein, Y. (2023). On the computational principles underlying human exploration. *PsyArxiv preprint*

https://doi.org/10.31234/osf.io/s96b2

On the computational principles underlying human exploration

Lior Fox* lior.fox@mail.huji.ac.il
Ohad Dan† ohad.dan@gmail.com
Yonatan Loewenstein*‡ yonatan@huji.ac.il

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.

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 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

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

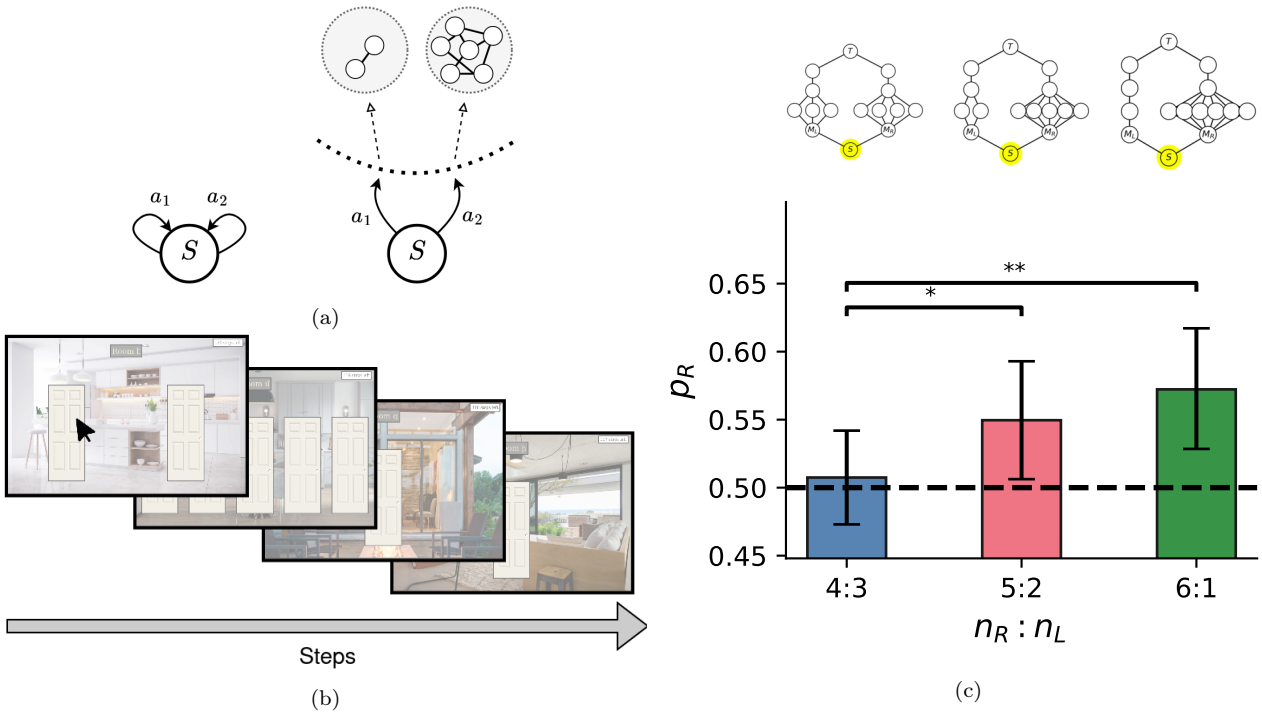


Figure 1: **Directed exploration in complex environments**. (**a**) In a bandit problem (left), actions have no long-term consequences. In complex environments (right), actions have long-term consequences as particular actions might lead, in the future, to different parts of the state-space. In this 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: $^{*}: p < 0.05$, $^{**}: p < 0.01$; $^{***}: p < 0.001$).

visited S (of the one particular environment they were assigned to) 20 times.

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 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.

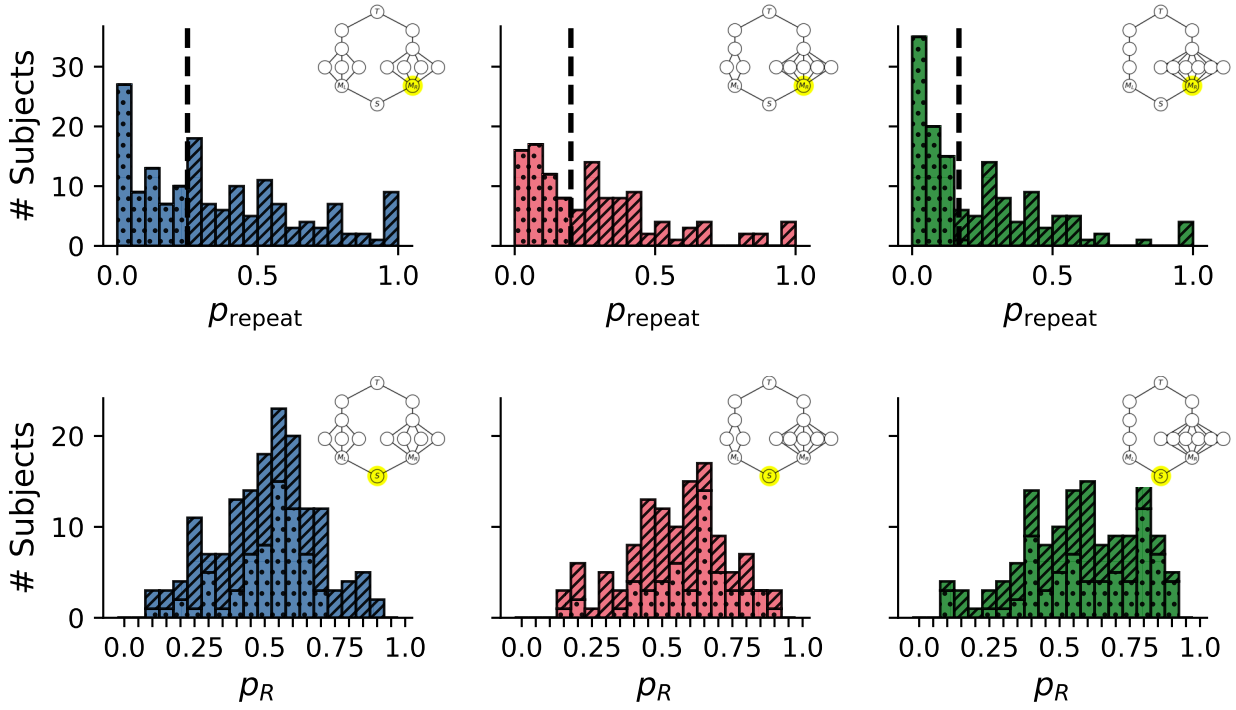
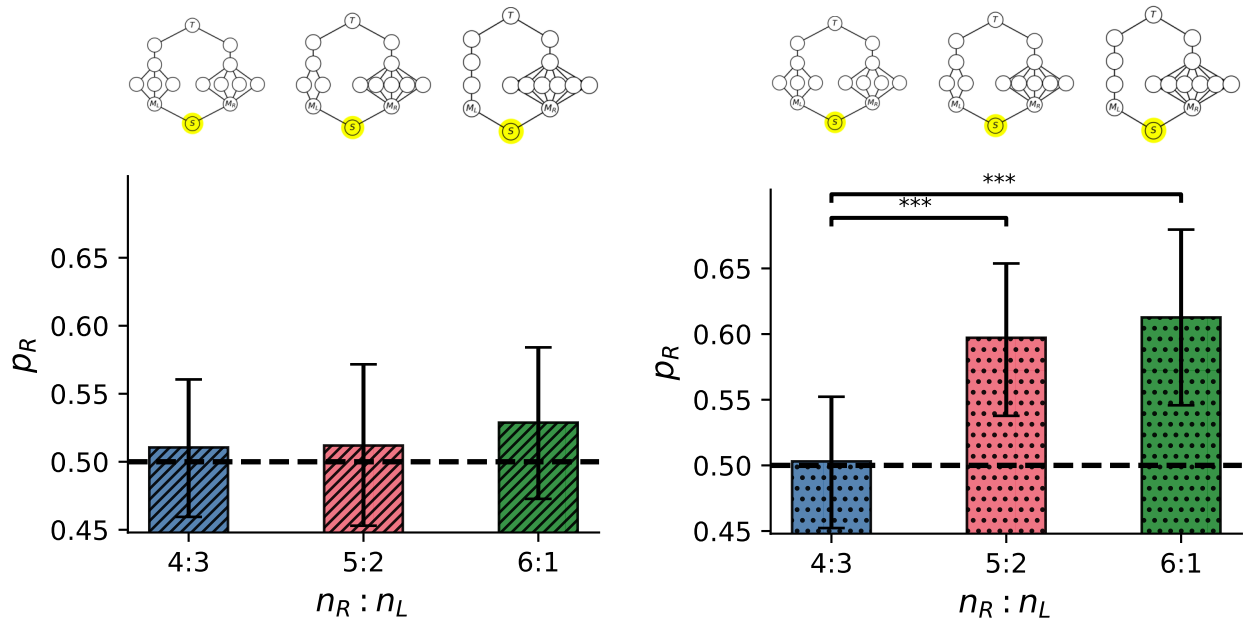


Figure 2: **Heterogeneity in exploration strategies**. **Top**: Histograms of p_{repeat} at state M_R (highlighted in yellow) for participants in the three conditions of Experiment 1 (left to right: $n_R = 4, 5, 6$). Dashed vertical line represents the value expected by chance, $1/n_R$. Based on their 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.

Temporal discounting

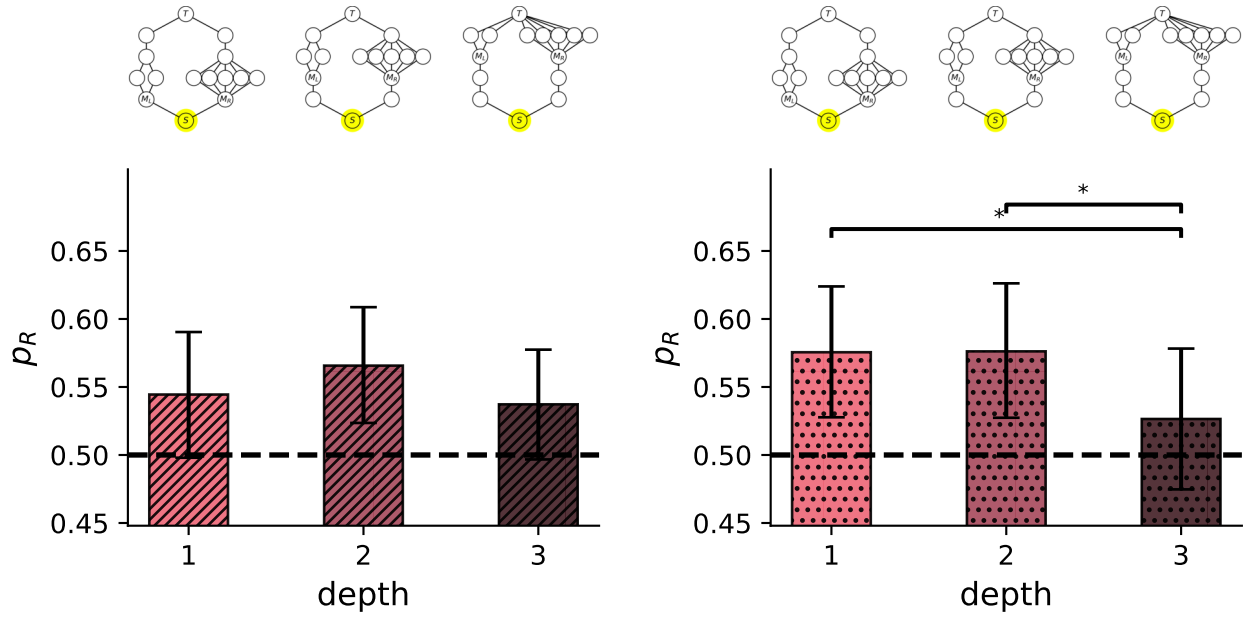
In the previous section we showed that if the future exploratory consequences of the actions are one trial ahead, humans are sensitive to these consequences. It is well known that in humans and animals, the value of a reward is discounted with its delay (Vanderveldt et al., 2016). We hypothesized that 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.



(a) (b)

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").



(a) (b)

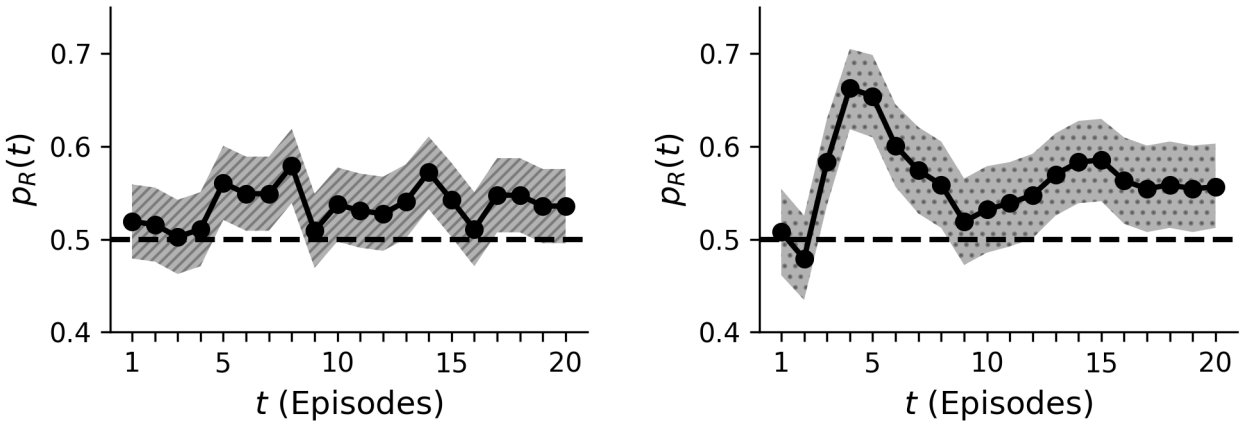
Figure 4: **Temporal discounting of exploratory consequences**. The three mazes in Experiment 2 (Top) had the same imbalance ($n_R = 5$, $n_L = 2$), however we varied the depth of M_R (and M_L) relative to the root state S (left to right: depth $= 1, 2, 3$). "Poor" and "good" directed explorers (striped 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").

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 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 ($p_R^* = 0.5$). 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 3$^{\text{rd}}$ 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).



(a) (b)

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 $p_R(t)$.

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:

$$E(s,a) \leftarrow E(s,a) + \eta\left(-E(s,a) + \gamma E(s',a')\right) \tag{1}$$

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, $\gamma E\left(s',a'\right)$ represents future uncertainty propagating back to (s,a). This second part is weighted by a discount-factor parameter, $0 \leq \gamma \leq 1$. 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.

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:

$$\pi\left(a|s\right) = \frac{e^{\beta E(s,a)}}{\sum_{a'} e^{\beta E(s,a')}} \tag{2}$$

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:

$$E\left(s,a\right) \leftarrow \left(1-\eta\right) E\left(s,a\right) \tag{3}$$

and hence, after n visits of (s,a), the associated E-value is $E\left(s,a\right) = \left(1-\eta\right)^n$, 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 $E\left(s,a\right)$ following a visit of (s,a) is more complex. In addition to the decay term, a term that is proportional to $E\left(s',a'\right)$ is added to $E\left(s,a\right)$. Notably, $E\left(s',a'\right)$ 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

[1] because $\eta < 1$, we have that $\log\left(1-\eta\right) < 0$.

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 $E\left(s = S, a = \text{right}\right)$ per a single visit of $(s = S, a = \text{right})$. 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 $E\left(s, a\right)$ 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 $E\left(s = S, a = \text{right}\right) = E\left(s = S, a = \text{left}\right)$, 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 $p_R\left(t\right)$, 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

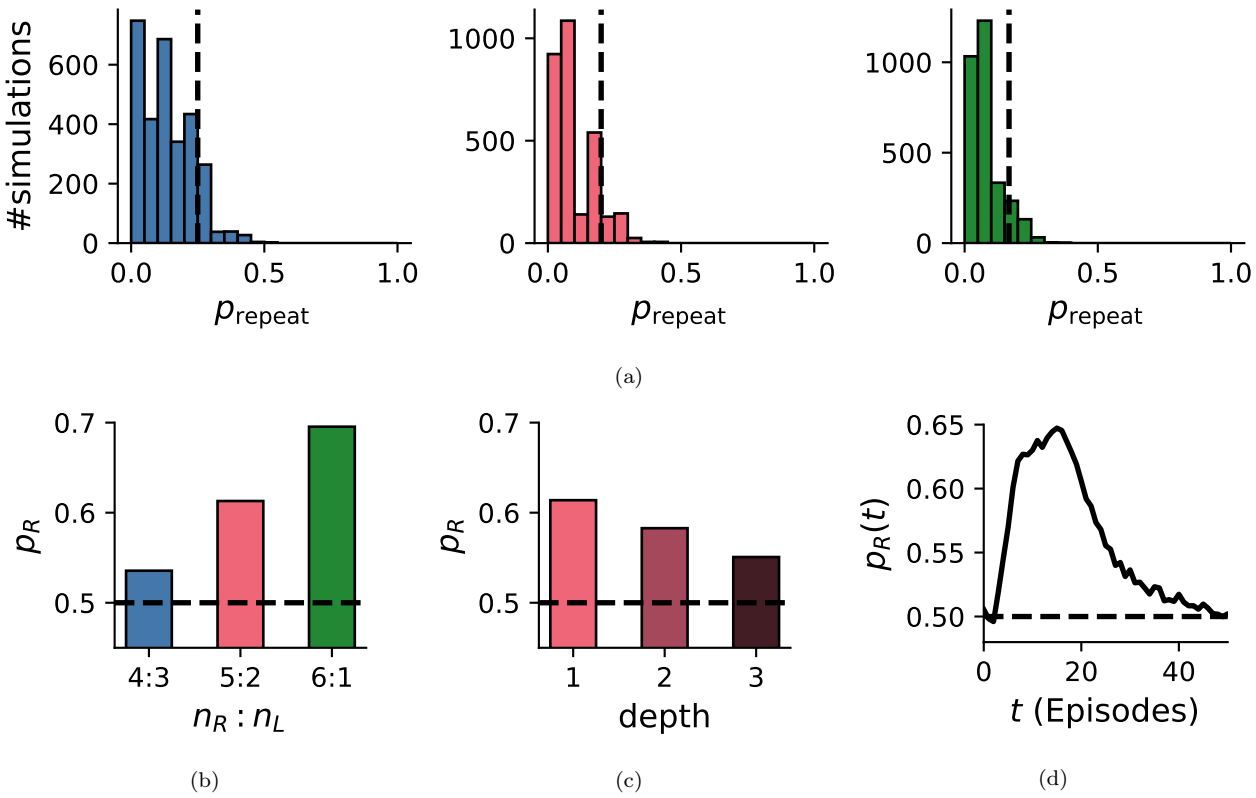


Figure 6: **Simulations results**. Simulating behavior of the E-values model (Equation 1–2) reproduces the main findings of directed exploration in the maze task. **(a)** In M_R, the model exhibits directed exploration which manifests in low values of p_{repeat} (shown for the 3 conditions of Experiment 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: $\eta = 0.9, \beta = 5, \gamma = 0.6$.

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.

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.

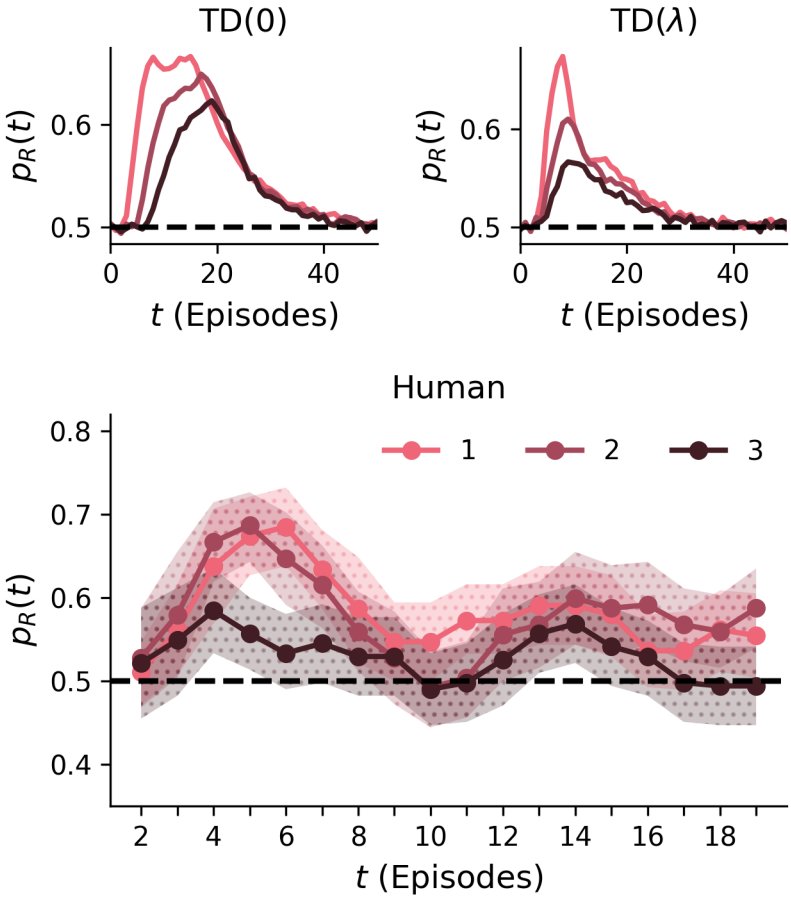


Figure 7: **1-step backups and trajectory-based updates**. Learning dynamics simulated by the E-values model using the 1-step backup learning rule of TD (0) (Equation 1–2; **top left**) and the trajectory-based learning rule TD (λ) (Methods, Algorithm 1; **top right**) in the 3 environments of Experiment 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).

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.

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 long-term 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.

Finally, non-stationary components can be incorporated within the context of uncertainty-driven exploration and the E-values model to account for steady-state exploratory behavior. For example, a different model might posit that rather than only decaying with visits, E-values of specific state-actions are also increased with each step in which these state-actions were *not* visited. This will lead to a "recency" drive that might be important in non-stationary environments, and will promote a non-trivial steady-state exploratory behavior. A more delicate option has to do with the functional form of the action-selection function. In particular, changing from softmax to "hard" max (i.e., greedy E-value) will lead to a convergence to a non-trivial steady-state policy. The intuition is that while all E-values decay to 0, the rate of decay is different for different state-actions, and a "greedy E" policy converges to the one that attempts at equalizing these rates for different actions (within each state separately). Practically, this can also be achieved by gradually increasing the gain parameter β as a function of time.

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, but the *number* of available actions was not included as an explicit feature in the state representation available for the agent. 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.

The role of computational models in studying human exploration The E-values model(s) are inconsistent with some important features of human behavior, most notably (as mentioned before) learning speed. We therefore believe that an attempt to fit the model parameters using the collected data is misleading. This, however, does not imply that the model is useless. We view the role of our theoretical model as a mathematical, algorithmic (and hence, concrete) manifestation of the psychological principles (which are inevitably more vague and open to interpretation) that we hypothesized to be underlying human exploration. As such, the model is viewed as a theory that can qualitatively *reproduces* key aspects of human behavior in this task, from a set of minimal, well-understood, and theoretically justified, computational principles. The failure modes of the model are just as informative, as they provide insight into aspects of human behavior that cannot be explained by these principles alone, at least in a generic manifestation.

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.

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

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 $p_R(\text{A}) - p_R(\text{B})$ 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

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**
        $E(s_t,a_t) \leftarrow E(s_t,a_t) + \eta\varepsilon(s_t,a_t)\left(\gamma E(s',a') - E(s,a)\right)$ ▷ 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

for updating the 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.

Acknowledgments

We thank Lea Kaplan for technical support, and Gal Yarden for discussions. This work was supported by the Israel Science Foundation (Grant 757/16), the Deutsche Forschungsgemein-

[2]Implicitly assuming an episodic setting, in which every episode terminates after a finite number of steps.

schaft Grant (CRC 1080), the David and Inez Myers Chair in Neural Computation, and the Gatsby Charitable Foundation.

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

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

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior*, 22(1):231–242.

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

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

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. (2019). Exploration by random network distillation. In *International Conference on Learning Representations*.

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

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):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., and Efros, A. (2018). Investigating human priors for playing video games. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1349–1357. PMLR.

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

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., et al. (2019). Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons. *Nature*, 570(7762):509–513.

Fox, L., Choshen, L., and Loewenstein, Y. (2018). DORA the explorer: Directed outreaching reinforcement action-selection. In *International Conference on Learning Representations*.

Fox, L., Dan, O., Elber-Dorozko, L., and Loewenstein, Y. (2020). Exploration: from machines to humans. *Current Opinion in Behavioral Sciences*, 35:104–111. Curiosity (Explore vs Exploit).

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

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, 6(3):277.

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. and Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12):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., and Knight, Z. A. (2022). Dopamine subsystems that track internal states. *Nature*, 608(7922):374–380.

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

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019). Provably efficient maximum entropy exploration. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2681–2691, Long Beach, California, USA. PMLR.

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

Kolling, N. and Akam, T. (2017). (reinforcement?) learning to forage optimally. *Current Opinion in Neurobiology*, 46:162–169. Computational Neuroscience.

Laquitaine, S., Piron, C., Abellanas, D., Loewenstein, Y., and Boraud, T. (2013). Complex population response of dorsal putamen neurons predicts the ability to learn. *PLOS ONE*, 8(11):null.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lebovich, L., Darshan, R., Lavi, Y., Hansel, D., and Loewenstein, Y. (2019). Idiosyncratic choice bias naturally emerges from intrinsic stochasticity in neuronal dynamics. *Nature Human Behaviour*, 3(11):1190–1202.

Little, D. Y. and Sommer, F. T. (2014). Learning and exploration in action-perception loops. *Closing the Loop Around Neural Systems*, page 295.

Machado, M. C. and Bowling, M. (2016). Learning purposeful behaviour in the absence of rewards. *arXiv preprint arXiv:1605.07700*.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., and Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2(3):191.

Meuleau, N. and Bourgine, P. (1999). Exploration of multi-state environments: Local measures and back-propagation of uncertainty. *Machine Learning*, 35(2):117–154.

Mobbs, D., Trimmer, P. C., Blumstein, D. T., and Dayan, P. (2018). Foraging for foundations in decision neuroscience: insights from ethology. *Nature Reviews Neuroscience*, 19(7):419–427.

Mongillo, G., Shteingart, H., and Loewenstein, Y. (2014). The misbehavior of reinforcement learning. *Proceedings of the IEEE*, 102(4):528–541.

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

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

Osband, I., Roy, B. V., and Wen, Z. (2016b). Generalization and exploration via randomized value functions. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2377–2386, New York, New York, USA. PMLR.

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

Oudeyer, P.-Y. and Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6.

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

Rummery, G. A. and Niranjan, M. (1994). *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering.

Schmidhuber, J. (1991). Curious model-building control systems. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, pages 1458–1463. IEEE.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Schulz, E., Franklin, N. T., and Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive Psychology*, 119:101261.

Shteingart, H., Neiman, T., and Loewenstein, Y. (2013). The role of first impression in operant learning. *Journal of Experimental Psychology: General*, 142(2):476.

Smith, K. S., Berridge, K. C., and Aldridge, J. W. (2011). Disentangling pleasure from incentive salience and learning signals in brain reward circuitry. *Proceedings of the National Academy of Sciences*, 108(27):E255–E264.

Still, S. and Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148.

Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the international conference on artificial neural networks, Paris*, volume 2, pages 159–164. Citeseer.

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

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

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

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

Urai, A. E., de Gee, J. W., Tsetsos, K., and Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *eLife*, 8:e46331.

Vanderveldt, A., Oliveira, L., and Green, L. (2016). Delay discounting: pigeon, rat, human—does it matter? *Journal of Experimental Psychology: Animal learning and cognition*, 42(2):141.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

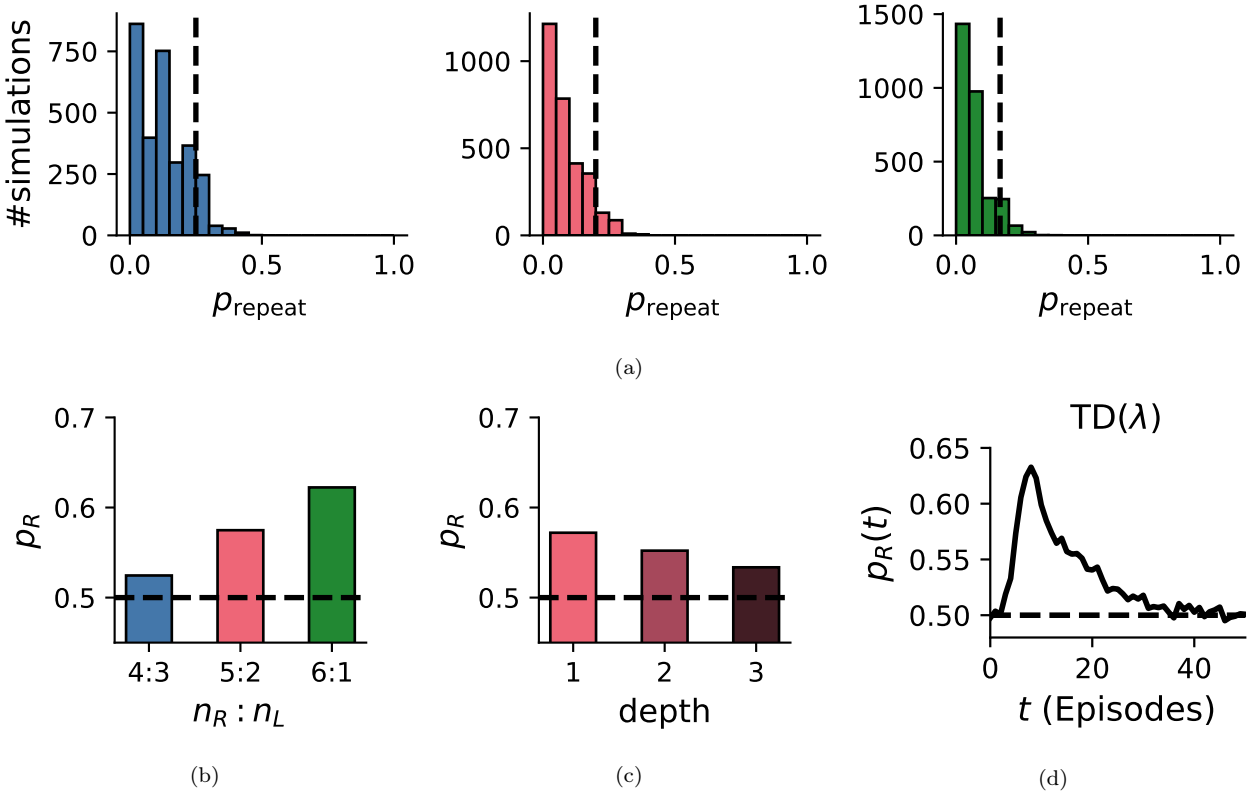
Wilson, R., Wang, S., Sadeghiyeh, H., and Cohen, J. D. (2020). Deep exploration as a unifying account of explore-exploit behavior. *PsyArXiv preprint*.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., and Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074.

Zahavy, T., O' Donoghue, B., Desjardins, G., and Singh, S. (2021). Reward is enough for convex mdps. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25746–25759. Curran Associates, Inc.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc.

Supplementary material



(a) (b) (c) (d)

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: $\eta = 0.9, \beta = 5, \gamma = 0.6, \lambda = 0.6$.

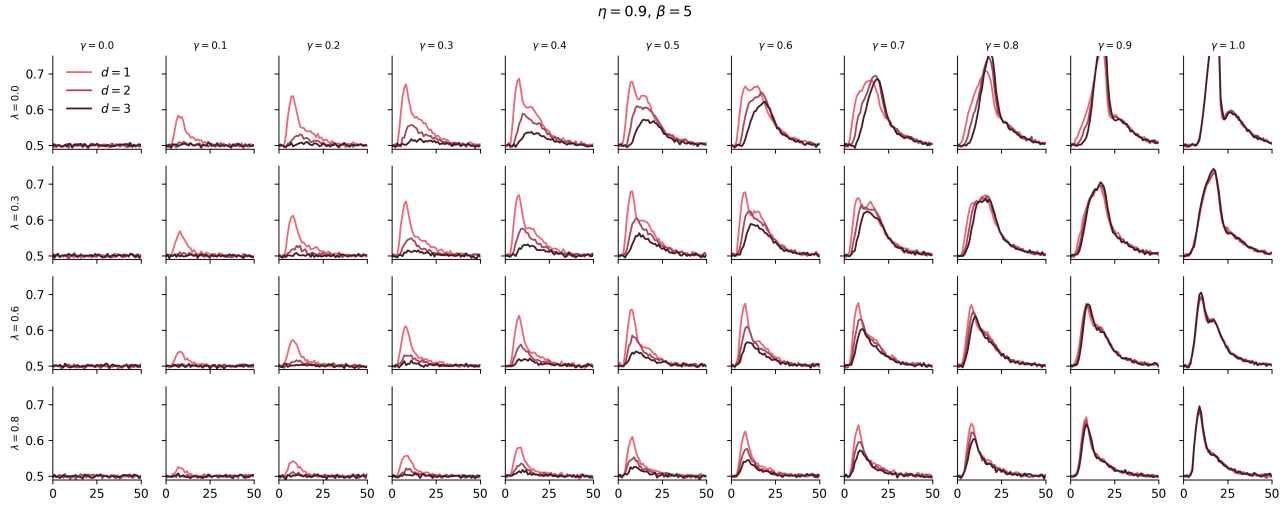


Figure S2: **Model parameters**. Learning curves of the TD (λ) model in the 3 environments of Experiment 2 for different values of γ,λ (with fixed $\eta = 0.9$, $\beta = 5$). With infinite discounting ($\gamma = 0$), future consequences are neglected, resulting in a uniform (counter-based like) policy 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.

Chapter 5

Discussion

רחוק מה שהיה ועמק
עמק מי ימצאנו

קהלת ז

This thesis has focused on the unique challenges of exploration in complex, structured, environments. A particular emphasis was given to the fact that in such environments, the long-term exploratory consequences of actions matter, and should be taken into account – from an algorithmic-theoretical perspective, and from a behavioral perspective alike. To do that, there is a need for both appropriate models and experimental paradigms, as well as the conceptual way to connect between them.

The theoretical framework that we have used for approaching these questions here is Reinforcement Learning (RL), with a particular (though not exclusive) focus on *model-free* RL. Model-free methods are conceptually simple, can be relatively easily scaled to large domains, and have long been used as a minimal model of operant learning in the behavioral sciences. Here, we have shown that model-free methods can also be pushed quite far in the context of exploratory tasks. Specifically, they can be used to form useful measures of uncertainty for directed exploration in complex environments (Chapter 2), as well as learn and generate non-trivial exploratory behaviors in the complete absence of external rewards (Chapter 3). Even more so, they provide successful

predictions for human behavior in structured exploration tasks (Chapter 4).

Yet, model-free methods are by no means the end of the story. From an algorithmic perspective, model-free methods can be slow, and suffer from large sample complexity. One potential reason for that is that a model-free algorithm only uses reward observations for learning, which are typically sparser, and of a much lower dimension, compared to state observations which are used for model-based learning (Recht, 2019). Moreover, they are inherently limited in their generalization ability, for example when the reward structure changes but the environmental dynamics remain unchanged (Dayan, 1993). Building artificial agents that can learn abstract or "latent" models directly from sensory observations, and use them for planning and reasoning, remains a frontier topic in RL and Artificial Intelligence in general (Ha and Schmidhuber, 2018; Hafner et al., 2019; Moerland et al., 2020; LeCun, 2022).

From a cognitive perspective, simple model-free methods are also an incomplete description of learning and behavior. Indeed, the fact that animals can generalize and plan based on their prior experience has historically been an important argument for the existence of structured mental representations such as cognitive maps (Tolman, 1948), and today it is well established that RL in humans involves model-based components (Daw et al., 2011). Recent studies further demonstrated the importance of model-based techniques in accounting for some aspects of human exploration in complex environments, beyond the bandit task (Xu et al., 2021; Brändle et al., 2022). It is worth mentioning that the general question of model-based versus model-free learning is tightly related to the question of behavior in complex environments. This because in a bandit task, the two alternatives cannot – in principle – be differentiated, as there is no "model" of the environment to be learned, other than the rewards model. In the context of this work, while model-free methods have proved a useful description of human learning and exploration in complex environments, we interpret the shortcoming of the model in explaining the rapid learning observed in humans as a strong indication for mechanisms beyond pure model-free learning (see more in the discussion section of Chapter 4). Overall, characterizing the role that planning, and model-based techniques in general, play in human exploration remains an important open topic for future research.

While we have focused on studying exploration at the behavioral level, the theoretical

and experimental frameworks developed in this thesis could further inform the study of the neural basis of exploratory behavior. This, in fact, might be another advantage of the model-free methods: their conceptual simplicity makes is easier to identify potential neural signatures of their implementation. The ever-increasing development of experimental technology in neuroscience, which enables the recording of neural activity in behaving animals and in complex environments (simulated or physical), opens up new possibilities and opportunities for novel RL theory to inform studies of richer behaviors than reward maximization alone.

The analogy between reward and "information" (in the broader sense of the term) has been a recurring theme in this work. It has informed us on multiple levels: from the conceptual overall problems (of "learning to explore", and even "planning to explore"), through the intermediate building blocks (such as specific objective functions, temporal discounting), to the tools and techniques that can be used (various model-free RL algorithms, including both value-based and policy-gradient approaches). Recent studies have further established this analogy more formally, for other behavioral objectives beyond exploration (Zhang et al., 2020; Zahavy et al., 2021). But despite the usefulness of this analogy, there are important differences between the problem of learning to maximize reward and learning to explore. Perhaps most notably are the two issues of stationarity and ergodicity.

Exploration is inherently non-stationary in the sense that it is often interpreted as part of a learning process. This poses a technical limitation of applying standard RL algorithms – which often assume a stationary, Markovian environments and rewards – to the "learning to explore" problem. We have approached this issue from two complementary angles. First, the E-values method (Chapter 2) relies on tracking the Temporal Differences learning dynamics itself, without positing an actual "exploration reward". This results, technically, in a stationary target for the exploration learning, but comes at the cost that the steady-state values are trivial (in the sense that they are not sensitive to the environmental structure). The maximum-entropy approach (Chapter 3), on the other hand, is driven by a stationary objective which will generate a non-trivial steady-state behavior, but this comes at the expense that learning has to track a non-stationary (and in some sense non-Markov) "reward" targets.

This non-stationarity is the main reason for our choice of Policy Gradient methods

to learn the optimal exploration policy, as these methods are somewhat more tolerant to violations of the standard MDP assumptions compared to Value-based methods (Shalev-Shwartz et al., 2016). Indeed, related approaches that build on Policy Gradient have more recently been proposed by several authors (Zhang et al., 2020; Mutti et al., 2021; Zhang et al., 2021). Continually optimizing the policy (against a changing "reward" function) can be seen as an alternative to related methods in which the complete solving of a progression of different MDPs, each with a different (stationary) reward function, is required (Hazan et al., 2019; Lee et al., 2019; Zahavy et al., 2021). Overall, the usefulness of standard RL techniques in solving problems beyond the classical expected reward maximization raises the question of whether value-based techniques can also be applied for that purpose. As of writing these lines, this question remains largely open (though see Geist et al., 2022).

Another, somewhat related, difference between the "standard" reward maximization and non-standard objectives (such as the maximum-entropy exploration) is in the very nature of optimality they imply. In virtually all cases – for both reward-maximization and "non-standard" objectives – the optimization objective is defined in terms of the visitation distribution (over state-actions) induced by the policy. As such, the objective is some *measure F of the expected behavior*, where the "expected behavior" is over the ensemble average of many trajectories. In the standard RL case, the measure F is a linear functional, and therefore (due to the linearity of expectation) the problem is mathematically equivalent to the *expectation of the measure F of the behavior*. This, however, no longer holds in the "non-standard" case, for example where F is the entropy functional. In that case, the single-trajectory performance can be, even in expectation, far from optimal. For example, single trajectories generated by the optimal exploration policy (in the sense of Chapter 3) do not well-cover the entire state space of the MDP – it is only the statistical ensemble of trajectories that does.

This formalism is mathematically convenient, as it makes the optimization problem tractable, particularly in guaranteeing that an optimal solution can be realized by a fully reactive, Markovian policy (i.e., a memoryless policy that only depends on the current state). By contrast, requiring optimality on the single-trajectory level is much more demanding, and reactive Markovian policies are strictly suboptimal in that case compared to non-Markovian policies (Mutti et al., 2022). From the perspective of hu-

man (and animal) behavior, this gap is a potential concern, as it is unclear that such "optimality by ensemble" (as opposed to "optimality on a single expected episode") is behaviorally relevant. Indeed, similar issues have been recognized in the broader context of economical decision making (Peters, 2019). This gap suggests that notions of optimality that do not rely as heavily on Markovian assumptions, in contrast to those commonly used in RL, might be an important extension for the study of exploration. For example, theories of optimal foraging which incorporate the idea of diminishing returns and rely on the Marginal Value Theorem (Charnov, 1976) might better account for some aspects of natural "exploration" compared to standard RL models (Kolling and Akam, 2017).

Bibliography (for Introduction and Discussion)

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

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

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

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

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. (2019). Exploration by random network distillation. In *International Conference on Learning Representations*.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, 9(2):129–136.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.

Dayan, P. (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation*, 5(4):613–624.

Dayan, P. and Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185–196. Cognitive neuroscience.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., et al. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419.

Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R., and Pietquin, O. (2022). Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 489–497, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

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.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3389–3396.

Ha, D. and Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. (2019). Learning latent dynamics for planning from pixels. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565. PMLR.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019). Provably efficient maximum entropy exploration. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of*

the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 2681–2691, Long Beach, California, USA. PMLR.

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

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Kolling, N. and Akam, T. (2017). (reinforcement?) learning to forage optimally. *Current Opinion in Neurobiology*, 46:162–169. Computational Neuroscience.

Kolter, J. Z. and Ng, A. Y. (2009). Near-bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 513–520. ACM.

LeCun, Y. (2022). A path towards autonomous machine intelligence version. *Open Review*, 62.

Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S., and Salakhutdinov, R. (2019). Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*.

Little, D. Y. and Sommer, F. T. (2014). Learning and exploration in action-perception loops. *Closing the Loop Around Neural Systems*, page 295.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Moerland, T. M., Broekens, J., Plaat, A., and Jonker, C. M. (2020). Model-based reinforcement learning: A survey.

Mongillo, G., Shteingart, H., and Loewenstein, Y. (2014). The misbehavior of reinforcement learning. *Proceedings of the IEEE*, 102(4):528–541.

Mutti, M., De Santi, R., and Restelli, M. (2022). The importance of non-markovianity in maximum state entropy exploration. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16223–16239. PMLR.

Mutti, M., Pratissoli, L., and Restelli, M. (2021). Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9028–9036.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154. Special Issue: Dynamic Decision Making.

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

Oudeyer, P.-Y. and Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6.

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

Peters, O. (2019). The ergodicity problem in economics. *Nature Physics*, 15(12):1216–1221.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition.

Recht, B. (2019). A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):253–279.

Schmidhuber, J. (1991). Curious model-building control systems. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, pages 1458–1463. IEEE.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

Still, S. and Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148.

Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the international conference on artificial neural networks, Paris*, volume 2, pages 159–164. Citeseer.

Sutton, R. S. (1992). Introduction: The challenge of reinforcement learning. *Machine Learning*, 8(3):225–227.

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

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

Tesauro, G. (1992). Practical issues in temporal difference learning. *Machine Learning*, 8(3):257–277.

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

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

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, 55(4):189.

Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., and Herzog, M. H. (2021). Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLOS Computational Biology*, 17(6):1–32.

Zahavy, T., O' Donoghue, B., Desjardins, G., and Singh, S. (2021). Reward is enough for convex mdps. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25746–25759. Curran Associates, Inc.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc.

Zhang, J., Ni, C., Yu, z., Szepesvari, C., and Wang, M. (2021). On the convergence and sample efficiency of variance-reduced policy gradient method. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2228–2240. Curran Associates, Inc.

Appendix A

Exploration: From Machines to Humans

עצביהם כסף וזהב מעשה ידי אדם

תהלים קטו

Status: Published

Citation: Fox, L.*, Dan, O.*, Elber-Dorozko, L.*, and Loewenstein, Y. (2020). Exploration: from machines to humans. *Current Opinion in Behavioral Sciences, 35*, 104-111.

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

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Exploration: from machines to humans

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

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

Addresses
[1] The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Israel

[2] Dept. of Cognitive Sciences, The Hebrew University, Jerusalem, Israel
[3] The Federmann Center for the Study of Rationality, The Hebrew University, Jerusalem, Israel
[4] The Alexander Silberman Institute of Life Sciences, The Hebrew University, Jerusalem, Israel

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

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

This review comes from a themed issue on **Curiosity (Exolore versus Exploit)**

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

For a complete overview see the Issue and the Editorial

Available online 10th October 2020

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

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

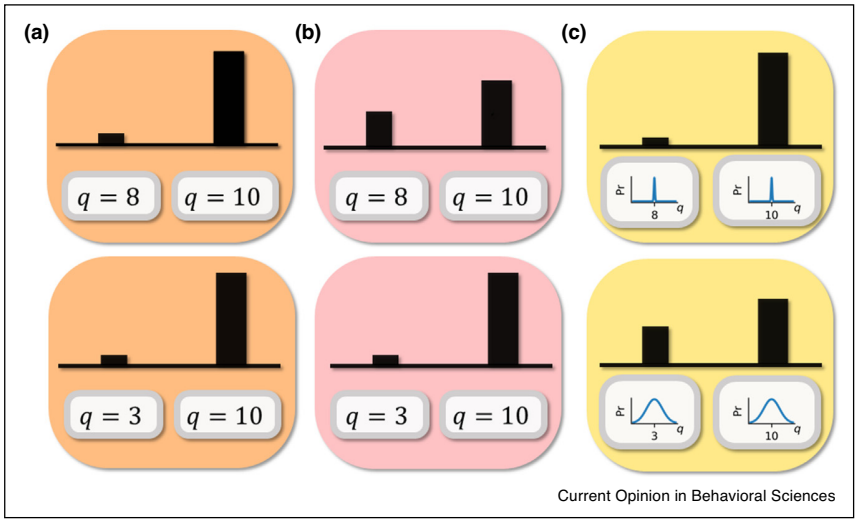
Introduction

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

Random exploration

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

Figure 1



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

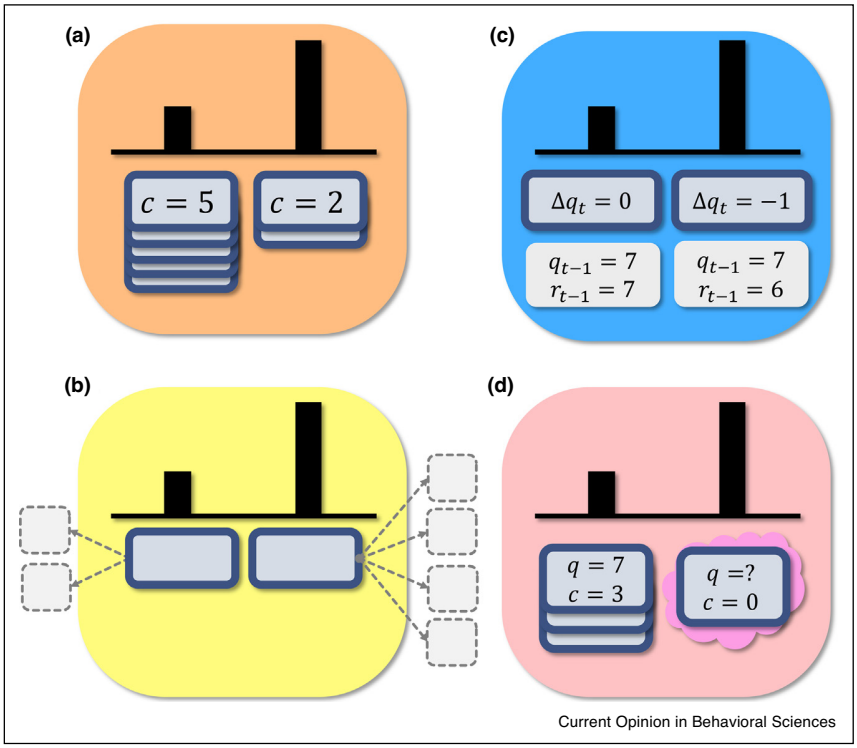
with respect to rewards (greedy). With a low probability (ε), the agent explores by randomly selecting another action. Exploration this way, however, does not distinguish between the non-greedy alternative actions. Therefore, a more graded approach, in which alternative actions that are deemed better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

Directed exploration

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

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

Figure 2



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

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

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

sufficiently many times. This approach underlies several algorithms that are theoretically guaranteed to efficiently explore [25,26].

Studying exploration in humans using the bandit task

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

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

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

Challenges in identifying human exploration in the bandit task

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

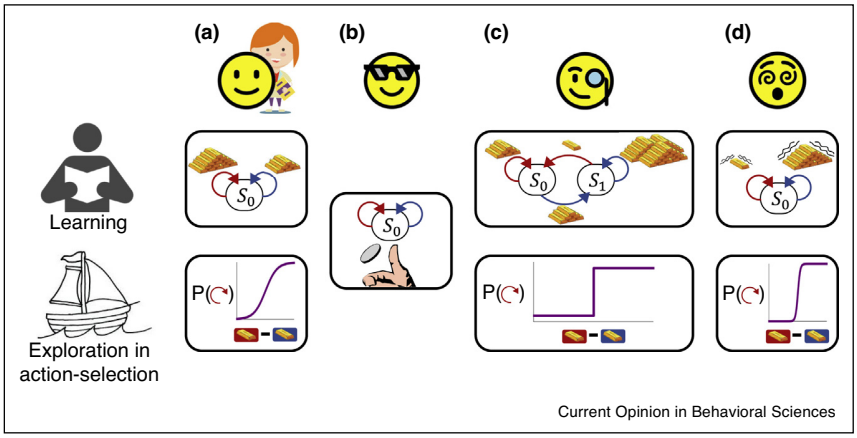
Moreover, in the framework of action-value estimation in the bandit task, it is typically assumed that the participants estimate action-values as if they are in a one-state MDP. However, it is well known that humans 'detect' temporal structures even in random sequences [50,51]. This result suggests that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

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

Ecological exploration

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

Figure 3



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

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

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

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

in a similar way that 'optimism in the face of uncertainty' contributes to exploration in ML.

Exploration and curiosity

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

Concluding remarks

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

Conflict of interest statement

Nothing declared.

Funding

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

Acknowledgement

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

References and recommended reading

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

- • of special interest

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

27. • Gershman SJ: **Deconstructing the human algorithms for exploration**. *Cognition* 2018, **173**:34-42.
Do participants utilize directed exploration in two-armed bandit tasks? To address this question, the effects of uncertainties in the estimated action-values on participants' choice behavior were studied. The paper reports that uncertainty in an action-value affects both the slope of the action-selection function– an indication for sampling based random exploration, as well as the bias of the action-selection function – an indication for directed exploration. The conclusion is that participants' utilize both directed-exploration and random-exploration in their learning behavior.

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

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

30. • Shteingart H, Neiman T, Loewenstein Y: **The role of first impression in operant learning**. *J Exp Psychol Gen* 2013, **142**:476-488.
In this paper, the behavior of human participants in a two-armed bandit task is analyzed. It characterizes non-parametrically the action-selection function in humans, underlying random exploration. Specifically, while humans are sensitive to the difference in action-values (as in softmax), they exhibit substantial exploration even when the difference between these values is large.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

52. • Morse EB, Runquist WN: **Probability-matching with an unscheduled random sequence**. *Am J Psychol* 1960, **73**:603-607.
This paper describes participants' repeated choice behavior in two, very similar, two-alternative choice tasks. In the first, participants were instructed to predict whether a rod that is dropped would cross lines drawn on the floor. In the second, they had to predict which of two bulbs would turn on in the trial. Despite the fact that both groups of participants observed the exact same sequence of binary events, their behaviors differed. They tended to maximize in the first task and to probability-

match in the second. These results highlight the importance of a world model in learning.

53. • Findling C, Skvortsova V, Dromnelle R, Palminteri S, Wyart V: **Computational noise in reward-guided learning drives behavioral variability in volatile environments**. *Nat Neurosci* 2019, **22**:2066-2077.
Non-greedy choices in a two-armed bandit task experiment are typically interpreted as reflecting exploration. By comparing variability in a partial-feedback task to that in a full-feedback task (in which no exploration is expected), it is shown that the majority of non-greedy decisions stem from learning noise, rather than reflecting exploration.

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

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

56. Dubey R, Agrawal P, Pathak D, Griffiths TL, Efros AA: **Investigating human priors for playing video games**. *Proceedings of the 35th International Conference on Machine Learning* 2018:1349-1357. PMLR 80.
The authors systematically modified video-games' environment in order to mask visual information that could be used as priors. It turns out the human participants' learning in the game heavily relies on such priors. Specifically, they exhibit different patterns of learning and exploration in the 'masked' conditions. By contrast, artificial agents are largely unaffected by the masking of almost all visual priors.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

תקציר

אקספלורציה היא מרכיב מרכזי בלמידה בעזרת ניסוי וטעיה. בסביבות מורכבות, אקספלורציה עשויה להיות מאתגרת באופן מיוחד, כיוון שלהחלטות המתקבלות על-ידי הסוכן יש, במקרים רבים, השלכות ארוכות טווח אשר צריכות להילקח בחשבון על מנת להבטיח אקספלורציה יעילה. כך, על מנת להצליח לחקור את הסביבה באופן יעיל, נדרשת למידה של הסביבה, בעוד שכדי ללמוד את הסביבה יש צורך להכיר אותה מלכתחילה.

החלק הראשון של עבודת דוקטורט זו מציג שתי גישות אלגוריתמיות להתמודדות עם אתגר האקספלורציה. הראשונה היא אקספלורציה המונעת מ "אי-וודאות", בהתבסס על האינטואיציה לפיה האתגר ב"למידה עבור אקפסלורציה" הוא אנלוגי לאתגר שבלמידת פונקציית-ערך בבעיות של למידת חיזוק. הגישה השניה היא גישה נורמטיבית, ובה אקספלורציה אופטימלית מוגדרת ככזו שממקסמת פונקציית מטרה מסויימת: האנטרופיה של התפלגות הביקורים במרחב המצבים-פעולות של הסביבה, כפי שמושרית על-ידי ההתנהגות של הסוכן. החלק השני של עבודה זו הוא מחקר התנהגותי, ובו נחקרה אקספלורציה אנושית לאור העקרונות החישוביים אשר זוהו בעזרת המודלים. במחקר זה מודגם כי בני-אדם רגישים לתוצאות ארוכות הטווח של בחירותיהם במטלות אקספלורציה. תוצאות אלו מעידות על אסטרטגיות אקספלורציה אשר "מפעפעות" אי-וודאות לאורך מסלולים במרחב המצבים-פעולות, בניגוד לשימוש באומדים מקומיים בלבד של אי-וודאות.

עבודה זו נעשתה בהדרכתו של

פרופסור יונתן לוינשטיין

אקספלורציה בסביבות מורכבות:

מידול חישובי והתנהגות אנושית

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

ליאור פוקס

הוגש לסנאט האוניברסיטה העברית בירושלים

אוגוסט 2023