Exploration Beyond Bandits

Brändle, Franziska*
franziska.braendle@tue.mpg.de
Max Planck Institute for Biological Cybernetics

Binz, Marcel*
marcel.binz@tue.mpg.de
Max Planck Institute for Biological Cybernetics

Schulz, Eric
eric.schulz@tue.mpg.de
Max Planck Institute for Biological Cybernetics

February 19, 2021

1 Introduction

Reinforcement learning is the study of how an agent – be it human, animal, or a machine – can learn to choose actions that maximize rewards [53]. To maximize long-term rewards, the agent must seek out information about the environment, even if it comes at the cost of temporarily missing out on more rewarding actions. How to strike the balance between maximizing immediate and long-term rewards is referred to as the *exploration-exploitation dilemma* [11]. On the one hand, the agent should focus on gaining as much rewards as currently possible. The maximization of rewards given the agent's current knowledge is called exploitation. On the other hand, the agent should search for further information to increase their knowledge, which could help to generate more rewards later on. The search for information in the context of reinforcement learning is called exploration.

Human exploration has been predominately studied in multi-armed bandit tasks [29]. The term multi-armed bandit stems from a colorful casino metaphor, in which the agent interacts with a row of slot machines, each associated with an unknown reward distribution. It is the agent's goal to maximize rewards by repeatedly sampling arms and collecting the resulting rewards. Ideal agents should explore by combining the immediate reward and the value of information for each action by thinking through future actions and calculating how

*Equal contribution.

much rewards would increase if more knowledge about the reward distributions is collected. However, such optimal exploration strategies are wildly intractable beyond a few special cases [56]. This is because the value of information depends on how information affects choices later on, which may also lead to new information, creating a scenario in which the complexity of planning increases as an exponential function of the agent's planning horizon.

Because optimal solutions to the exploration-exploitation dilemma are computationally intractable, humans, as well as other intelligent agents, must employ heuristics. Research on human exploration strategies has been centered around two such heuristics [43, 17]. These heuristics use the uncertainty about arms' rewards to guide exploratory choices. The first uncertainty-guided heuristic is to engage in *directed exploration*, seeking out options that are highly informative about the underlying reward distribution. Directed exploration can, for example, be implemented by adding an information bonus to the estimates of expected reward [1]. This bonus will then encourage the agent to explore arms with high uncertainty. The second uncertainty-guided heuristic is *random exploration*, i.e. to inject stochasticity into one's sampling behavior. One widely-adopted instantiation of random exploration applies a fixed source of stochasticity without caring about arms' uncertainty [13]. More sophisticated random exploration strategies, however, are uncertainty-guided and sample options relative to their probability of being optimal [55]. This approach may be viewed as a form of hypothesis testing, where the agent keeps track of multiple hypotheses and acts at each point in time as if a particular hypothesis was true.

We argue that the repertoire of human exploration strategies has itself not been well explored. We believe that there are two reasons for this, opening up two paths toward extending current theories. The first one is that studies on human exploration have almost exclusively focused on multi-armed bandit tasks. However, multi-armed bandits only constitute a small part of the problems that people typically encounter. For example, bandits do not include a mechanism to control the state of one's environment; yet this very mechanism is omnipresent in everyday exploration scenarios. Therefore, we believe that extending studies on human exploration to more complex paradigms can bring scientific experiments closer to the real world. To this end, we suggest that future work should move towards exploration problems that can be modeled as Markov Decision Processes (MDPs), in which an agent can control the state of the environment. The second reason is that past studies have focused almost exclusively on the two exploration strategies of random and directed exploration. However, people can engage in exploratory behaviors that cannot easily be captured by these two simple mechanisms; for example, when children are freely exploring how to build block towers, or when scientists explore theories to create better explanations of the observed data. Thus, we propose to study more sophisticated algorithms of exploration, such as empowerment and goal-conditioned exploration, in their ability to describe human behavior. Importantly, many of these strategies cannot be expressed within multi-armed bandit problems, but require the more expressive setting offered by MDPs.

This chapter is divided into three parts. In the first part, we review a

Paper	Bandit Type	Optimal	Uncertainty	Directed	Random
Steyvers et al. [46]	Simple	✗	?	?	?
Zhang and Yu [62]	Simple	✗	✓	✓	?
Gershman [17]	Simple	?	✓	✓	✓
Binz and Endres [6]	Simple	✗	✓	✓	✓
Wilson et al. [57]	Simple	?	?	✓	✓
Daw et al. [13]	Restless	?	✗	✗	✓
Speekenbrink et al. [44]	Restless	?	✓	✗	✓
Wimmer et al. [58]	Correlated	?	?	?	✓
Borji and Itti [7]	Correlated	?	✓	✓	?
Wu et al. [61, 60, 59]	Correlated	?	✓	✓	✓
Stojic et al. [47, 48]	Contextual	?	✓	✓	✓
Frank et al. [15]	Contextual	?	✓	✓	?

Table 1: Overview of past studies on human exploration the the multi-armed bandit setting. Red crosses (✗) mark the absence of empirical evidence for a particular exploration strategy. Green check marks (✓) indicate that evidence for a particular exploration strategy was obtained by a study. Gray question marks (?) indicate that a particular exploration strategy was not investigated.

subset of past studies on human exploration in multi-armed bandit tasks, with a particular focus on random and directed exploration. In the second part, we describe the shortcomings of multi-armed bandits and argue that we need to move toward more expressive tasks, i.e. MDPs, to understand the full breadth of human exploration. In this part, we outline the challenges that arise when extending random and directed exploration strategies to MDPs and discuss several new exploration strategies that can be studied in MDPs. In the final part, we speculate about novel paradigms to chart a path toward studying exploration beyond bandits.

2 Prior work on multi-armed bandits

Given its notorious difficulty, how do people actually cope with the exploration-exploitation dilemma? As previously mentioned, past studies on human exploration have mostly focused on the multi-armed bandit case, in which participants can sample between different options to maximize monetary rewards. We review a subset of these studies below.

In a simple variant of multi-armed bandit tasks, the reward distributions are stationary and independent of each other. In this setting, Steyvers et al. analyzed the data of 451 participants [46]. Their results showed that –rather than following an optimal exploration strategy– people largely applied simple heuristics. Zhang and Yu also used a stationary bandit task to compare human behavior with models of different degrees of sophistication, including the

optimal exploration strategy [62]. Their results showed that a non-optimal but "forgetful" Bayesian iterative learning model described human behavior best. Thus, people seem to follow heuristic strategies of exploration, even in simple bandit tasks.

As mentioned before, two of the most frequently-studied exploration strategies are random and directed exploration, which use the uncertainty of the arms' rewards to guide exploration. Whereas directed exploration applies an information bonus to seek out options with higher uncertainty, random exploration predicts that choice stochasticity increases with higher uncertainty across all arms. Gershman tested these predictions in a stationary two-armed bandit task [17]. In his task, rewards were generated from a Gaussian distribution with a fixed mean and standard deviation. This allowed for the manipulation of the total and relative uncertainty of the two arms by increasing the variance of either both or only one of the arms. The results of these experiments showed that participants applied a mix of both random and directed exploration. Binz and Endres demonstrated that participants exhibit individual differences in how they explore in the same two-armed bandit task, and that traces of both random and directed exploration can emerge from optimal reasoning under limited computational resources [6]. In the canonical "Horizon task" [57], Wilson et al. manipulated the number of samples participants could draw from a two-armed stationary bandit on each round. The results of these experiments showed that participants increased their exploration in the long-horizon condition and applied both directed as well as random exploration strategies. Together, these studies indicate that participants seem to rely on a mix of both random and directed exploration in simple, stationary multi-armed bandit tasks.

All bandit tasks described so far involved a stationary distribution of rewards. However, in plenty of real-life scenarios the reward distribution can change over time; for example, if your favorite restaurant is continuously decreasing in its quality. Researchers have therefore looked at human behavior in another class of paradigms called "restless bandits". In these paradigms, the mean of an arm's reward distribution changes during the experiment. Daw et al. investigated the underlying strategies and cortical substrates of exploration in a restless bandit task [13]. In their task, participants had to choose one of four arms whose expected values changed over time, following a decaying Gaussian random walk. They found no evidence for directed or uncertainty-guided, random exploration. In contrast to this finding, a study by Speekenbrink et al. found evidence for uncertainty-guided exploration in a restless bandit task [44]. In their experiment, participants also had to choose between four arms in a restless bandit task. Their results showed that subjects followed a random exploration strategy by choosing arms according to their probabilities of producing the maximum reward. Evidence for directed and random exploration in restless bandits can therefore be described as mixed.

The previously described paradigms assumed independent distributions of rewards between all available arms. However, naturally occurring scenarios often involve options whose rewards co-vary, for example when ordering food online from restaurants in the same district. A natural extension of past paradigms is

therefore to consider scenarios with correlated arms. Wimmer et al. let participants perform a four-armed bandit task with binary rewards [58]. Unknown to participants, the reward probabilities for pairs of arms were correlated across trials. Results showed that participants learned to take into account this correlational structure and built up an "acquired equivalence" between arms. Borji and Itti studied how people searched for the maximum of a one-dimensional function [7]. Functions provide an interesting set-up in which nearby options (inputs) co-vary naturally since they will produce similar outputs. Borji and Itti found that human behavior was in line with a Bayesian optimization algorithm that used a mechanism of generalization combined with an uncertainty-guided search strategy to find high functional outputs. Wu et al. extended this paradigm further [61, 60], studying one and two-dimensional functions in a spatially-correlated multi-armed bandit. In these tasks, nearby arms produced similar rewards, which provided traction for generalization to speed up participants' search for highly-rewarding options. They found that the same Bayesian model of generalization together with upper confidence bound sampling, i.e. a directed exploration strategy, predicted participants' search behavior best. In a follow-up study [59], Wu et al. used two correlated bandit paradigms to research commonalities and differences in spatial and conceptual information search. In the spatial task, participants had to samples arms on a grid in which rewards were correlated according to their position. In the conceptual task, participants were shown Gabor patches with different numbers of stripes and tilts, and patches with similar features produced similar rewards. As before, they found that exploration was guided by participants' ability to generalize over similar arms. Additionally, whereas participants employed directed exploration in the spatial task, they explored more randomly in the conceptual task. Taken these results together, there is substantial evidence that participants engage in both random and directed exploration in correlated bandit tasks. This is intuitive because the presence of correlational structure enhances the benefits of these exploration strategies [9].

The concept of contextual bandits extends these paradigms further. In contextual bandits, different arms can come with features that relate to an arm's expected rewards. This paradigm was used in several experiments and implemented in a diverse set of tasks. Stojic et al. created a task that displayed options as red boxes and used vertical and horizontal lines as the features of each arm: the shorter the lines, the higher the rewards [47]. They found that participants indeed took these contextual features into account to direct their exploration to more promising options. Moreover, they found evidence for directed exploration, since participants preferred options with the same expected average reward but higher relative uncertainty [48]. A different version of a contextual bandit task was put forward by Frank et al. [15]. In a so-called "clock task", participants could stop a clock running down to gain rewards. The rewards varied as a function of the position of the clock's arm and, depending on the condition, either increased, decreased, or stayed constant with time. This study produced strong evidence for directed exploration strategies. Another version of a contextual bandit was studied by Rich and Gureckis [36]. In their foraging

task, participants had to decide whether or not to sample different species of mushrooms. Each species had different probabilities of containing edible mushrooms, i.e. positive rewards, or poisonous mushrooms, i.e. negative rewards. Participants explored more given a longer horizon and took the frequency of encountered mushrooms into account. Taking the results of past studies using contextual bandits together, there is strong evidence that participants apply directed and random exploration strategies in such tasks. Moreover, they seem to combine these strategies with more elaborate mechanisms of learning and generalization.

Summarizing past work on human exploratory behavior in multi-armed bandits, we can see that the following two main results emerge:

1. Even in the simplest bandit problems –i.e. two-armed bandits with stationary reward distributions– people deviate from the optimal exploration strategy.

2. People often use uncertainty estimates to guide their exploratory behavior using a combination of directed and random strategies.

3 Extending multi-armed bandit tasks

Multi-armed bandits have served as a fertile ground for past studies on human exploration. Even though they can be extended to incorporate non-stationary rewards, correlated arms, and contextual features, the resulting paradigms might still fall short to describe the rich repertoire of human exploration strategies. We argue that one reason for the dearth of evidence for more sophisticated exploration strategies could be that multi-armed bandits do not contain a mechanism to control the state of one's environment. Many real-world problems, however, require a deliberate manipulation of the environment to achieve success. Imagine, for example, a child playing with differently sized and shaped building blocks. By constructing new objects, they are clearly able to influence their environment, changing not only the current state but perhaps even what options are available. It is not possible to capture this example in a bandit paradigm.

Markov Decision Processes (MDPs) offer a natural extension to multi-armed bandits that *can* capture problems which involve the manipulation of an environment [5]. Formally, an MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, T, p)$. $\mathcal{S}$ is a set of states, which – in our building block example – describe the current assembly of the blocks. $\mathcal{A}$ is a set of actions, which express how the child can act on the environment. T is the planning horizon of the agent and p defines a probability distribution $p(s_{t+1}, r_t | s_t, a_t)$ over the next state and an associated reward given that the agent has executed action a_t in state s_t. In the building blocks example, this probability distribution describes what happens when the child adds a new part to an existing assembly. From the joint distribution over transition

and reward probabilities, we can extract several other useful quantities:

$$p(s_{t+1}|s_t, a_t) = \int p(s_{t+1}, r_t|s_t, a_t)dr_t \tag{1}$$

$$r(s_t, a_t) = \int r_t \sum_{s_{t+1}} p(s_{t+1}, r_t|s_t, a_t)dr_t \tag{2}$$

The goal of an agent is then to find the policy $\pi^*(a_t|s_t)$ that maximizes the expected sum of rewards obtained over its planning horizon:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi,p}\left[\sum_{t=1}^{T} r(s_t, a_t)\right] \tag{3}$$

If p is known Equation 3 can be solved using dynamic programming [53]. However, it is typically assumed that the agent does not have access to the true distribution over transition and reward probabilities. It is this uncertainty in the agent's knowledge that causes the need for exploration.

MDPs can be viewed as a generalization of the multi-armed bandit paradigm, which means that each bandit problem may be formulated as an MDP. For example, a stationary bandit with independent reward functions can be expressed as an MDP with a single state, and therefore adheres to trivial transition probabilities (from the single state to itself with probability one). A contextual bandit can be expressed as an MDP in which the agent has no control over transition. In this case $p(s_{t+1}|s_t, a_t)$ simplifies to $p(s_{t+1}|s_t)$.

3.1 New challenges

Exploration algorithms for MDPs have been extensively studied in computer science. Below, we review several of these algorithms. First, we describe how algorithms of random and directed exploration can be extended to MDPs. Afterwards, we discuss how MDPs allow us to capture even richer forms of exploration. The discussed algorithms are illustrated in Figure 1.

Random exploration

Let us first look at random exploration. Osband et al. [32] discussed how Thompson sampling – i.e. exploration based on randomly-drawn beliefs – can be implemented in MDPs. To this end, they suggested an algorithm called *posterior sampling for reinforcement learning* (PSRL). PSRL keeps track of a posterior distribution over environment parameters θ, which is constantly updated via Bayes' rule as the agent interacts with the environment:

$$p(\theta|s_{1:t+1}, a_{1:t}, r_{1:t}) \propto p(s_{t+1}, r_t|s_t, a_t, \theta)p(\theta|s_{1:t}, a_{1:t-1}, r_{1:t-1}) \tag{4}$$

At the beginning of each episode, the agent draws a random sample from this posterior distribution and computes the optimal policy for the sampled

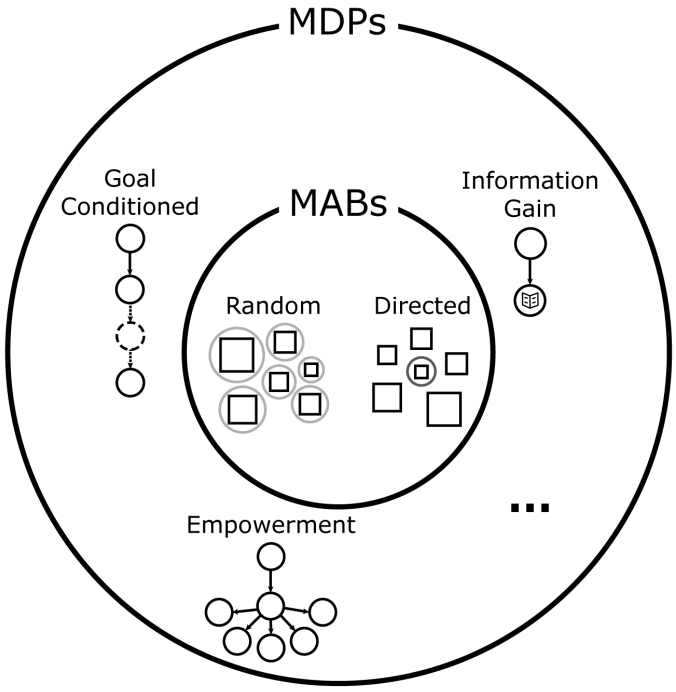

Figure 1: Overview of exploration strategies. While multi-armed bandits (MABs) can only capture directed and random exploration strategies, Markov Decision Processes (MDPs) are able to describe a wider range of strategies.

parameters by solving Equation 3. An agent applying PSRL assumes a randomly sampled hypothesis for an entire episode, and tests whether that particular hypothesis is true by using an exploration strategy that is consistent over a period of time [51]. Coming back to our building block example, if a child would apply PSRL as an exploration strategy, then they would maintain different hypotheses about what happens when you stack one block onto another. They might believe that one can stack only two blocks on top of each other, or that one can put pieces on top of each other indefinitely. The child would then sample one of these hypotheses and act as if the sampled hypothesis was true for an extended period. If, for example, the second hypothesis is sampled, the child might end up stacking blocks on top of each other until the resulting building crashes and a different hypothesis will be considered.

Re-sampling the parameters only at the beginning of an episode comes with advantages and disadvantages. It avoids erratic behavior that could arise if parameters were instead constantly resampled. However, it also implies that even if the agent obtains new information during an interaction with an environment, it has to wait until the end of the episode to actually use this information. If

people applied a PSRL-like strategy, they would presumably also need a good set of heuristics to decide when to sample a new hypothesis. This could happen when they have a moment of insight [26] or more gradually as parts of a hypothesis are proven incorrect [8].

Directed exploration

What about directed exploration? There exist several implementations of the principle of uncertainty-directed exploration in the context of MDPs [50, 2, 20, 14, 3, 49, 21]. Here, we focus on a particular example called *model based interval estimation with exploration bonus* (MBIE-EB) [50]. MBIE-EB keeps track of point estimates of environment parameters that are constantly updated. Based on these estimates it constructs, and solves, a new MDP with an optimistic reward function:

$$r_{\text{MBIE-EB}}(s_t, a_t) = r(s_t, a_t) + \frac{\beta}{\sqrt{N(s_t, a_t)}} \tag{5}$$

where $N(s_t, a_t)$ denotes the number of times the agent has taken action a_t in state s_t and β is a hyperparameter that controls the degree of exploration. An agent that applies MBIE-EB assigns higher rewards to rarely encountered state-action pairs, essentially directing it to explore situations in which uncertainty is high. When playing with building blocks, this would encourage the child to modify a building in a way they have never done before. If, for example, a child has already built many towers with four blocks, but has never considered putting a fifth block on top of it, they would be encouraged to do so under an MBIE-EB-based exploration strategy.

MBIE-EB comes with its own challenges when considering it as a model of human exploration. Most importantly, it assumes that the optimistic reward function $r_{\text{MBIE-EB}}$ changes after each interaction with the environment. In turn, this means that an agent would have to run an expensive reinforcement learning algorithm to solve Equation 3 on each time-step, which seems like an unrealistically strong requirement when considering that human processing power is limited. One way to potentially implement a cognitively more plausible version of MBIE-EB could be to assume that people approximate the reward function by a finite number of mental samples [38].

3.2 New opportunities

So far, we have described possible implementations of random and directed exploration in MDPs. PSRL keeps track of a posterior distribution over transition and reward probabilities and acts greedily with respect to sampled beliefs, thereby implementing a form of random exploration. MBIE-EB on the other hand keeps track of point estimates over transition and reward probabilities and uses these to construct and solve an optimistic MDP, thereby implementing a form of directed exploration. However, MDPs can also be approached using other exploration strategies, some of which we will discuss next.

Information gain

Imagine a child who just got their first set of building blocks as a birthday present. When they start playing with the blocks, they first need to figure out how they work: How do blocks stack on top of each other? What determines the stability of a tower? How much does stepping on a block hurt your parents? Children are able to learn about all of these questions by exploring a new toy. How do they accomplish this?

Intuitively, children are such great explorers, because they reward themselves for learning new things about the environment; frequently, learning itself is the reward for curious agents. Being rewarded for learning *per se* is also at the core of several theories of curiosity, including computational accounts of "learning progress" [23] and "learning as fun" [41]. The core idea behind the "Learning Progress Hypothesis" put forward by Oudeyer and colleagues is that agents should be most curious about, and therefore most likely to sample, options that lead to maximal learning progress. This is well-aligned with Schmidhuber's "Theory of Fun", which argues that options produce the most fun if they create maximal learning progress. Maximizing learning progress naturally leads to a preference for problems of medium complexity [33]: if a problem is too easy, then there is nothing to be learned from it; if a problem is too difficult, people cannot solve it and also will not learn from it. A preference for options of medium complexity is known to be present in children [24] and adults [16].

Multiple algorithmic approaches have been proposed to further formalize this idea in MDPs [19, 52, 40, 28, 34]. The main idea behind all of these approaches is to reward the agent for taking actions that maximize the reduction of uncertainty about the dynamics of the environment. To illustrate how an agent could find the reduction of uncertainty rewarding in an MDP set-up, we will focus on one such approach here, which has been put forward by Houthooft et al. [19]. Their algorithm assumes that the agent expresses beliefs about environment parameters through a probability distribution, which is constantly updated via Bayes' rule as the agent interacts with the environment (as described in Equation 4). How uncertain the agent is about the true environment parameters can be expressed in terms of the conditional entropy $H\left[\theta|s_{1:t}, a_{1:t-1}, r_{1:t-1}\right]$. This means that the expected reduction in uncertainty can be expressed as the difference in entropies across two successive time-steps:

$$H\left[\theta|s_{1:t}, a_{1:t-1}, r_{1:t-1}\right] - \mathbb{E}_{s_{t+1}, r_t \sim p(s_{t+1}, r_t|s_t, a_t)}\left[H\left[\theta|s_{1:t+1}, a_{1:t}, r_{1:t}\right]\right] \qquad (6)$$

Equation 6 can be interpreted as a measure of the agent's information gain about the environment's dynamics. It is common practice to use this term as an intrinsic bonus reward, which then encourages the agent to take actions that maximize its learning progress [19]. In the context of the example from before, such a mechanism could offer an explanation for how a child explores after getting their first set of building blocks as a birthday present. Naturally, the child wants to figure out how the new toy works and does so by seeking to construct things that reduce their uncertainty about the toy's mechanics. This process then enables the child to find out how blocks stack on top of each other

and what determines the stability of a tower after playing with the toy for a while.

Empowerment

After a child has learned the simple rules of how to combine building blocks, how could they continue learning about how to build more complex things? They could, for example, decide that knowing how to build robust walls will help them to construct many different buildings, such as homes, castles, or bridges. By figuring out how to construct elements that can be used in many different assemblies, children can empower themselves to further improve their abilities.

Inspired by examples from the animal kingdom, social sciences, and games, researchers have suggested the concept of *empowerment* to capture this kind of behavior [25]. Honey bees, for example, try to be mobile because it allows them to forage at multiple sides, people strive for money because it enables them to do a variety of activities, and board game players often play in a way that keeps their options open. In all of these cases, empowerment "[motivates] an agent to move to states of maximum influence" [30]. Empowerment is also a useful signal for exploration because it enables the agent to explore large parts of the state space in a short time. In simple multi-armed bandit problems, an agent cannot apply an empowerment-based strategy, because states do not exist or there is no control over them. Thus, studying empowerment necessitates the use of MDP scenarios.

Mathematically, one can construct an agent that implements empowerment-based exploration by encouraging the maximization of the information contained in actions about future states [37, 30]. Leibfried et al. [27] suggested to use one-step empowerment as an intrinsic bonus added to external rewards.[1] The one-step empowerment is defined as the mutual information between actions and future states conditioned on the current state; i.e., $I\left[a_t; s_{t+1}|s_t\right]$. To gain an intuition of why this leads to the desired behavior, it is useful to consider the decomposition of the mutual information in terms of the marginal and the conditional entropy:

$$I\left[a_t; s_{t+1}|s_t\right] = H\left[s_{t+1}|s_t\right] - \mathbb{E}_{a_t \sim \pi(a_t|s_t)}\left[H\left[s_{t+1}|a_t, s_t\right]\right] \tag{7}$$

The first term in Equation 7 encourages the agent to visit states that lead to a diverse set of future states. The second term encourages it to take actions for which it can predict the outcome. Therefore, incorporating Equation 7 as an intrinsic bonus reward causes the agent to visit states of maximum influence. However, it also leads to a challenging optimization problem:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi,p}\left[\sum_{t=1}^{T} r(s_t, a_t) + \beta I\left[a_t; s_{t+1}|s_t\right]\right] \tag{8}$$

[1] Empowerment is typically defined in terms of multi-step policies [37, 30]. Leibfried et al. [27] demonstrated that maximizing the cumulative one-step empowerment leads to similar behavior without the necessity to maintain multi-step policies.

Maximizing Equation 8 is difficult, because the augmented reward function depends on the optimal policy, and vice versa the optimal policy depends on the augmented reward function. How people would solve such a problem is an interesting avenue for future research.

Goal-conditioned exploration

Now that the child has acquired a basic set of construction skills, they might have a bigger goal – building the biggest castle the world has ever seen. While this goal is clearly not attainable, the child might still learn useful things along the way. They, for example, could discover how to built towers, rooftops, or draw-bridges.

Goal-conditioned reinforcement learning equips an agent with the ability to reach arbitrary goals [54, 39, 35, 12]. In this framework, the agent learns a policy $\pi(a|s, g)$ that is not only conditioned on the current state s but also on a goal g. The additional conditioning on a goal enables the agent to exhibit different behaviors, depending on what goal is currently attempted. Typically, goals are defined in terms of the state space. In the simplest case, goals are just elements of the state space itself, i.e., $g \in \mathcal{S}$. More sophisticated implementations learn a goal embedding as a function of the state space, i.e., $g \in \phi(\mathcal{S})$, and perform goal-directed reasoning in latent space defined by the embedding [64, 31]. In both cases, the agent is rewarded for reaching a given goal instead of following the original reward function. As this definition of goals relies heavily on the notion of states, such a strategy is not available in the multi-armed bandit paradigm.

While goal-conditioned reinforcement learning is not a method for exploration on its own, reasoning about goals can facilitate exploration. It has, for example, been demonstrated that the combination of goal-conditioned reinforcement learning with random exploration can speed up the time it takes to visit all states in the environment [22]. If, for example, the child's goal is to build a big castle but they do not know how to get there, it will be useful to explore how potential sub-components work. This type of exploration does not happen purely at random, because the child has a particular goal in mind. It is also not purely directed towards situations with high uncertainty, but instead attempts to explore things that are useful for the goal you are trying to accomplish. Future studies on human exploration could therefore assess if giving participants unobtainable but useful goals can improve their overall task performance later on.

4 Paradigms beyond bandits

We have argued that the standard multi-armed bandit setting is not rich enough to study the large repertoire of human exploration strategies. But how can we test whether people actually use the described types of exploration strategies? Advancing the study of human exploration will require the use of novel experimental paradigms. Here, we present some examples of such paradigms.

A straightforward extension to multi-armed bandits are grid-world problems. In a grid-world environment, an agent needs to navigate on a two-dimensional grid in the attempt to solve a specific task, for example to find a goal or to escape from an intricate maze. Grid-world environments have been frequently used as paradigms to compare artificial reinforcement learning agents [53, 10], and could therefore help to disentangle more complex human exploration strategies. For example, Zheng et al. showed that different exploration strategies lead to intrinsic reward functions with varying properties [63]. In particular, they found that directed exploration strategies can lead to over-exploration even after a goal has been found. This and other predictions could be easily tested in human participants.

Video games can provide another interesting direction for future studies on human exploration. Video games can easily incorporate different levels of complexity, ranging from simple Atari games, over modern physical game engines, all the way to realistic virtual reality environments. Frequently, data sets of people's behavior in video games can be accessed via the internet, and available data sets are much bigger than data sets collected in standard in-lab experiments [18, 45]. Furthermore, video-games are rich enough to capture all exploration strategies discussed in the last section. To illustrate this point, let us take a look at a classic role-playing game example: You are a hero, traveling through a fantasy world, completing missions by fighting against monsters. To choose which mission to complete next, you may apply different strategies. You could try to improve your sword fighting skills by combating a monster with a difficulty level that provides just the right challenge – not too easy and not too hard. This corresponds to an exploration strategy based on information gain. Alternatively, you could buy a horse to explore new areas faster. This corresponds to an empowerment-based exploration strategy. Lastly, you could decide to set yourself the goal of fighting against a dangerous vampire king. While you are not able to beat him at the moment, you could try to find out a lot about vampires and start by training against weaker ones to prepare yourself for the big fight. This is an example of a goal-conditioned exploration strategy. Of course, it might still be a while until psychologists could reliably study human exploration in such scenarios. Moreover, the sheer complexity of the available action spaces makes it hard to trace model player's behavior back to individual factors in such games leading to a loss of internal validity. However, it is possible to study human exploration in simpler games already. For example, Matusch et al. [4] used a pre-collected data-set of different Atari games and looked at how strongly intrinsic reward functions of different exploration algorithms correlated with human behavior. Their results showed that intrinsic objectives like information gain and empowerment correlated more strongly with human behavior than just the simple reward in each task, thereby providing initial evidence that more complex exploration strategies govern human game play.

In an ideal world, we would also like to directly study human exploration in realistic scenarios, including our running example of a child playing with building blocks. However, measuring human behavior in such settings constitutes a highly non-trivial challenge, especially since it is not always clear when a new

state or action has occurred. Nevertheless, it might still be possible to gain insights into how people explore in everyday situations by studying large-scale data sets. For example, Schulz et al. looked at 1.5 million orders from an online food delivery service and analyzed the customers' exploration behavior. They found that customers used uncertainty-guided exploration to decide where to order next [42]. One drawback with large online data sets is that they lack clear control over the factors that can influence people's behavior. This limits the conclusions that can be drawn from these settings. However, we believe that one way to partially address this concern is to study quasi-experiments in which changes to different users happened randomly, for example because new options were introduced to different users at different times.

While all these paradigms have their unique benefits and drawbacks, they can jointly allow us to look for more sophisticated strategies than just uncertainty-guided exploration. Eventually, we believe that these paradigms could be added into the experimentalist's toolkit and –together with more traditional paradigms– enrich our understanding of human information search in the context of reinforcement learning.

5 Conclusion

The attempt to find actions that maximize long-term rewards is a powerful tool to describe intelligent behavior. Yet any sufficiently complex reinforcement learning problem is also an information-seeking problem in disguise. This is because the drive to reap immediate rewards is always juxtaposed with the drive to seek out knowledge about one's environment that can lead to higher rewards later on. Finding the right balance between information-seeking and maximizing rewards according to one's current knowledge frames the exploration-exploitation dilemma, a canonical problem studied in humans and machines.

In this chapter, we have reviewed past studies on human exploration, which have primarily focused on multi-armed bandit tasks. In the multi-armed bandit paradigm, people seem to use a mix of two heuristic strategies. The first strategy is random exploration which induces some form of stochasticity in the decision-making process. The latter is directed exploration which optimistically seeks out options with higher relative uncertainty.

We then argued that using multi-armed bandits to study human exploration behavior can be unnecessarily restrictive. This could explain why past studies have only ever found evidence for random and directed exploration, and why –even for these two rather simple strategies– the evidence has occasionally been mixed. We have therefore proposed to extend current paradigms to study human exploration by including scenarios in which people can also affect the state of their environment. This leads to the set-up of MDPs, which have been widely-studied in the machine learning community.

The two classic exploration strategies can easily be extended to MDPs. Moreover, MDPs lend themselves well to study other, more sophisticated exploration strategies as well. These strategies include, but are not limited to,

strategies driven by information gain, algorithms that try to empower themselves to explore even more, and goal-conditioned exploration. We believe that all of these strategies could be considered in richer environments that more closely resemble the real world such as video games, real world behavior such as tasks of physical construction, as well as online consumer behavior.

We hope that this chapter can inspire future work focused on more advanced exploration strategies, and enables new insights on the human drive to seek out knowledge in reinforcement learning problems. In the end, further extending our descriptions of human exploration will also require us to extend our own exploration of experimental paradigms.

References

[1] Peter Auer. Using confidence bounds for exploitation-exploration tradeoffs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[2] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pages 89–96, 2009.

[3] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.

[4] Brendon Matusch Jimmy Ba and Danijar Hafner. Evaluating agents without rewards. *ArXiv*, abs/2012.11538, 2020.

[5] Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

[6] Marcel Binz and Dominik Endres. Where do heuristics come from? In *CogSci*, pages 1402–1408, 2019.

[7] Ali Borji and Laurent Itti. Bayesian optimization explains human active search. *Advances in neural information processing systems*, 26:55–63, 2013.

[8] Neil R Bramley, Peter Dayan, Thomas L Griffiths, and David A Lagnado. Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, 124(3):301, 2017.

[9] Franziska Brändle, Charley M Wu, and Eric Schulz. What are we curious about? *Trends in Cognitive Sciences*, 24(9):685–687, 2020.

[10] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

[11] Jonathan D Cohen, Samuel M McClure, and Angela J Yu. Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, 2007.

[12] Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Intrinsically motivated goal-conditioned reinforcement learning: a short survey. *arXiv preprint arXiv:2012.09830*, 2020.

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

[14] Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and kullback-leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122. IEEE, 2010.

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

[16] Andra Geana, Robert C. Wilson, Nathaniel Daw, and Jonathan D. Cohen. Boredom, information-seeking and exploration. *Cognitive Science*, 2016.

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

[18] Thomas Griffiths. Manifesto for a new (computational) cognitive revolution. *Cognition*, 135, 12 2014.

[19] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29:1109–1117, 2016.

[20] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[21] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[22] Yuu Jinnai, Jee Won Park, David Abel, and George Konidaris. Discovering options for exploration by minimizing cover time. *arXiv preprint arXiv:1903.00606*, 2019.

[23] Frédéric Kaplan and Pierre-Yves Oudeyer. Maximizing learning progress: an internal reward system for development. In *Embodied artificial intelligence*, pages 259–270. Springer, 2004.

[24] Celeste Kidd, Steven T Piantadosi, and Richard N Aslin. The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, 7(5):e36399, 2012.

[25] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. All else being equal be empowered. In *European Conference on Artificial Life*, pages 744–753. Springer, 2005.

[26] Wolfgang Köhler. *The Mentality of Apes*, volume 74. K. Paul, Trench, Trubner & Company, Limited, 1925.

[27] Felix Leibfried, Sergio Pascual-Díaz, and Jordi Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. *Advances in Neural Information Processing Systems*, 32:7869–7880, 2019.

[28] Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. *Advances in neural information processing systems*, 25:206–214, 2012.

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

[30] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 2125–2133, 2015.

[31] Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *arXiv preprint arXiv:1807.04742*, 2018.

[32] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[33] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.

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

[35] Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine. Temporal difference models: Model-free deep rl for model-based control. *arXiv preprint arXiv:1802.09081*, 2018.

[36] Alexander S Rich and Todd M Gureckis. Exploratory choice reflects the future value of information. *Decision*, 5(3):177, 2018.

[37] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer, 2014.

[38] Adam N Sanborn and Nick Chater. Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893, 2016.

[39] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pages 1312–1320. PMLR, 2015.

[40] Jürgen Schmidhuber. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pages 1458–1463, 1991.

[41] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[42] Eric Schulz, Rahul Bhui, Bradley C. Love, Bastien Brier, Michael T. Todd, and Samuel J. Gershman. Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, 116(28):13903–13908, 2019.

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

[44] Maarten Speekenbrink and Emmanouil Konstantinidis. Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, 7(2):351–367, 2015.

[45] Tom Stafford and Michael Dewar. Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science*, 25(2):511–518, 2014.

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

[47] Hrvoje Stojic, Pantelis P Analytis, and Maarten Speekenbrink. Human behavior in contextual multi-armed bandit problems. In *CogSci*. Citeseer, 2015.

[48] Hrvoje Stojić, Eric Schulz, Pantelis P Analytis, and Maarten Speekenbrink. It's new, but is it good? how generalization and uncertainty guide the exploration of novel options. *Journal of Experimental Psychology: General*, 149(10):1878, 2020.

[49] 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*, pages 881–888, 2006.

[50] 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.

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

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

[53] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[54] Richard S Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M Pilarski, Adam White, and Doina Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 761–768, 2011.

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

[56] Peter Whittle. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):143–149, 1980.

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

[58] G Elliott Wimmer, Nathaniel D Daw, and Daphna Shohamy. Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience*, 35(7):1092–1104, 2012.

[59] Charley M Wu, Eric Schulz, Mona M Garvert, Björn Meder, and Nicolas W Schuck. Similarities and differences in spatial and non-spatial cognitive maps. *PLoS computational biology*, 16(9):e1008149, 2020.

[60] Charley M Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D Nelson, and Björn Meder. Mapping the unknown: The spatially correlated multi-armed bandit. *bioRxiv*, page 106286, 2017.

[61] Charley M Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D Nelson, and Björn Meder. Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, 2(12):915–924, 2018.

[62] Shunan Zhang and J Yu Angela. Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *NIPS*, pages 2607–2615, 2013.

[63] Zeyu Zheng, Junhyuk Oh, Matteo Hessel, Zhongwen Xu, Manuel Kroiss, Hado Van Hasselt, David Silver, and Satinder Singh. What can learned intrinsic rewards capture? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11436–11446. PMLR, 13–18 Jul 2020.

[64] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017.