Merits of curiosity: a simulation study

Lucas Gruaz[1,2,*,†], Alireza Modirshanechi[1,2,3,4,†], Johanni Brea[1,2]

[1] Brain-Mind Institute, School of Life Sciences, EPFL, Lausanne, Switzerland
[2] School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland
[3] Helmholtz Munich, Munich, Germany
[4] Max Planck Institute for Biological Cybernetics, Tübingen, Germany

[*] Corresponding author: lucas.gruaz@epfl.ch
[†] These authors contributed equally to this work

Abstract

'Why are we curious?' has been among the central puzzles of neuroscience and psychology in the past decades. Recent 'top-down' theories have hypothesized that curiosity, as a desire for some *intrinsically generated rewards* (e.g., novelty), is the *optimal* solution for survival in *complex environments* where we have evolved. To formalize and test this hypothesis, however, it is necessary to understand the relationship between (i) intrinsic rewards (as drives of curiosity), (ii) optimality conditions (as objectives of curiosity), and (iii) environment structures. Here, we demystify this relationship through a systematic simulation study. We first propose an algorithm for generating environments that capture key abstract features of different real-world situations. Then, within these environments, we simulate different artificial agents seeking six representative intrinsic rewards (novelty, surprise, information gain, empowerment, MOP and SPIE) and evaluate their performance regarding three potential objectives of curiosity (environment exploration, model accuracy and uniform state visitation). Our results show that the comparative performance of each intrinsic reward is highly dependent on the structural features of environments and the objective under consideration; this indicates that 'optimality' in the top-down theories of curiosity needs a precise formulation of the curiosity objective and the environment structure. Nevertheless, we found that agents seeking a combination of novelty and information gain always achieve a close-to-optimal performance; this proposes novelty and information gain as two principal axes of curiosity-driven behavior. These results, collectively, pave the way for the further development of computational models of curiosity and design of theory-informed experimental paradigms.

Introduction

Curiosity drives humans and animals to explore their environment and acquire knowledge about what appears to be new, puzzling, or strange (Berlyne, 1966; Gottlieb and Oudeyer, 2018; Kidd and Hayden, 2015; Modirshanechi et al., 2023b): Human babies prefer playing with toys that have surprising features (e.g., a car that passes through a solid wall) over normal toys (Stahl and Feigenson, 2015), monkeys look at novel visual stimuli longer than those they have seen before (Ghazizadeh et al., 2016; Ogasawara et al., 2022), rats prefer to explore mazes with complex structures than those with simple layouts (Montgomery, 1954), and mice have a higher breathing frequency when sniffing a new odor than a familiar one (Morrens et al., 2020). Mysteriously, the drive of curiosity can even occasionally overwrite primary needs such as for safety or food (FitzGibbon et al., 2020), e.g., human adults take the risk of receiving an electric shock only to know the secret of a magic trick (Lau et al., 2020), and monkeys give up juice rewards in return for the *information* of *future* reward (Bromberg-Martin et al., 2024). These observations have been among the central puzzles of neuroscience and psychology in the past decades[1], yet curiosity and its neuronal underpinning have remained mysterious and debated (see Forss et al. (2024); Modirshanechi et al. (2023b); Monosov (2024); Poli et al. (2024) for recent reviews).

From a theoretical perspective, there are two principal questions regarding curiosity: 'Why are humans and animals curious?' and 'What are they exactly curious about?' (Modirshanechi et al., 2023b). Modern theoretical attempts to address these questions use intrinsically motivated Reinforcement Learning (RL) framework (Baldassarre and Mirolli, 2013; Barto, 2013) and describe curiously-driven actions as those directed towards seeking an *intrinsically* generated 'reward' signal (Modirshanechi et al., 2023b; Murayama, 2022; Murayama et al., 2019; Oudeyer, 2018; Poli et al., 2024). In this framework, the answer to the 'What' question is given by the *intrinsic* reward (e.g., novelty or surprise of observations) that best describes the exploratory actions of a curious agent, as opposed to the *extrinsic* reward (e.g., the monetary or nutritional value of observations) that describes the exploitative actions (Aubret et al., 2019; Ladosz et al., 2022; Oudeyer and Kaplan, 2009). Given an intrinsic reward signal, the answer to the 'Why' question is often given by quantifying the benefits of the intrinsically motivated actions in terms of the agent's ability in, e.g., finding valuable sources of extrinsic reward (Gershman and Niv, 2015; Pathak et al., 2017; Singh et al., 2010a), gaining knowledge about the environment structure (Dubey and Griffiths, 2019), or unsupervised learning of complex skills (Mendonca et al., 2021; Oudeyer and Kaplan, 2009; Sekar et al., 2020).

In several experimental paradigms, intrinsically motivated RL algorithms have been successful in addressing the 'What' question and describing curiosity-driven and exploratory actions of human participants by considering novelty (Modirshanechi et al., 2023d; Xu et al., 2021), surprise (Kobayashi et al., 2019), information gain (Horvath et al., 2021; Nelson, 2005), progress rate (Poli et al., 2022; Ten et al., 2021a), or empowerment (Brändle et al., 2023; Klyubin et al., 2005) as the intrinsic reward signal. However, these studies do not address the paradoxical observation that the choice of intrinsic reward differs between different experimental paradigms (Modirshanechi et al., 2023b). A potential solution has been proposed by the 'top-down' models of curiosity (Modirshanechi et al., 2023b) that consider curiosity as the optimal mechanism for reaching a particular objective (the 'Why' of curiosity), e.g., finding the most valuable sources of extrinsic rewards in a class of environments (Alet et al., 2020; Dubey and Griffiths, 2019; Singh et al., 2010a; Zheng

[1]The seminal 1966 paper of Daniel Berlyne on curiosity (Berlyne, 1966) starts with the sentence 'Animals spend much of their time seeking stimuli whose significance raises problems for psychology.'

et al., 2020b). Instead of directly answering the 'What' question, these models characterize (i) the objective of curiosity and (ii) the class of environments where the curious agent lives. The 'What' of curiosity is determined by the reward signal reaching this objective in the specified class of environments. Hence, the observation that the 'What' of curiosity is experiment-dependent can be because of differences in the optimal strategies for reaching the curiosity objective in different experiments (Dubey and Griffiths, 2019, 2020). To advance our theoretical understanding of curiosity, it is hence necessary to understand the relationship between different (i) intrinsic rewards, (ii) objectives of curiosity, and, importantly, (iii) environment classes.

In this study, we aim to demystify this relationship. Specifically, we first design an algorithm for generating various environments with principally different characteristics, e.g., number of states, stochasticity of transitions, distribution of between-state connections, etc. We then formally define three performance measures as potential objectives of curiosity: (i) how fast a curious agent discovers all states of its environment, (ii) how accurately it learns the structure of the environment, and (iii) how uniformly it explores all the states. We then simulate different curious agents and quantify the merits of six representative intrinsic rewards (novelty, surprise, information gain, empowerment, maximum occupancy principle, and successor-predecessor intrinsic exploration) for maximizing these performance measures in different environments.

We show that, almost always, seeking information gain is the best strategy for the first two performance measures, whereas seeking novelty is the best strategy for the third. Building upon this observation, we show that an agent that seeks a combination of information gain and novelty can reach a close to the best performance for all three performance measures and in all classes of environments. This finding proposes information gain and novelty as two principal axes of curiosity-driven behavior (consistent with recent experimental findings, e.g., Dubey and Griffiths (2019); Monosov (2024); Poli et al. (2022)). Importantly, however, our results show that the relative performance of different intrinsic rewards is highly dependent on the structure of the environment. Finally, we show that our environment-generating algorithm proposes a novel approach to designing experimental paradigms where seeking different intrinsic rewards results in maximally different exploration strategies. These paradigms can be used in future experimental studies of curiosity in humans and animals (e.g., as in Modirshanechi et al. (2023d)).

Results

General framework

To study the behavior of curious agents, we use the intrinsically motivated RL framework. In this framework, each curious agent learns to navigate an environment represented by discrete states and transitions, where states represent specific locations within the environment, and transitions describe the agent's movement from one state to another as a result of its actions. Each transition is associated with a reward signal that guides the agent's action selection. Traditional RL relies on fixed, external rewards to shape the agent's behavior (Sutton and Barto, 2018). In contrast, intrinsically motivated RL uses internal reward signals that are non-stationary and evolve based on the agent's experience (Barto, 2013; Singh et al., 2010b). These intrinsic rewards encourage the agent to explore and learn from the environment without relying on external rewards.

We assume that the agent starts with no prior knowledge of the structure of the environment and builds a model of the environment by interacting with it. Specifically, we assume that the agent uses Bayesian inference (similar to Liakoni et al. (2022); Meyniel et al. (2016); Xu et al. (2021)) to estimate each transition probability $P(s'|s,a)$ (i.e., the probability of reaching state s' from state s by taking action a) for every state s, action a, and the next state s'. As a result, the agent counts transitions and constructs its environment model as

$$\hat{P}^{(t)}(s'|s,a) = \frac{C^{(t)}_{s,a\to s'} + \epsilon}{C^{(t)}_{s,a} + |S| \cdot \epsilon}\,, \tag{1}$$

where S denotes the set of all states, $|S|$ denotes the number of states, t is the current time step, $C^{(t)}_{s,a\to s'}$ is the count of the transition $s, a \to s'$ up to time t, and $C^{(t)}_{s,a}$ is the number of times action a has been taken from state s up to time t. The parameter $\epsilon > 0$ acts as a prior, preventing unseen transitions from being assigned a probability of zero (see Hyper-parameters selection for details). Then, using its model of the environment, the agent computes Q-value $Q(s,a)$ as an estimate of the expected future intrinsic rewards that the agent can collect, by taking action a at state s. The Q-values consider both immediate rewards and discounted future rewards and can be computed by solving the Bellman optimality equations (Sutton and Barto (2018))

$$Q^{(t)}(s,a) = \sum_{s'\in S} \hat{P}^{(t)}(s'|s,a)\Big(R^{(t)}(s,a,s') + \lambda \max_{a'\in A} Q^{(t)}(s',a')\Big)\,, \tag{2}$$

where $R^{(t)}(s,a,s')$ is the intrinsic reward for transitioning from s to s' via action a, determined by the agent's intrinsic motivation (detailed in Intrinsic motivations detailed), and $\lambda \in [0,1)$ represents the discount factor for the Q-values. The discount factor λ determines how much the agent values the future reward compared to the immediate rewards. These Q-values are updated using prioritized sweeping (Moore and Atkeson, 1993) with 100 iterations after each observed transition to iteratively converge to a solution of the Bellman equation.

At each time t, the agent's behavior in state s is described by the action policy $\pi^{(t)}_s$ which assigns probability $\pi^{(t)}_s(a)$ to selecting action a. We assume that the agent uses the Softmax of the Q-values as its action policy:

$$\pi^{(t)}_s(a) = \frac{e^{\beta Q^{(t)}(s,a)}}{\sum_{a'} e^{\beta Q^{(t)}(s,a')}} \in [0,1]\,, \tag{3}$$

where β is the Softmax inverse temperature (Sutton and Barto, 2018). This implies that the agent will strongly favor one action if it is clearly better than the others (i.e., if it has a much higher Q-value than the other actions), but the agent will choose all actions with almost equal probability if they all seem equally rewarding (i.e., if they have a similar Q-value).

Intrinsic motivations

We consider six types of intrinsic motivation, each defined by a reward function $R^{(t)}(s,a,s')$ that determines the Q-values (Eq. 2) and, accordingly, specifies the agent's action-policy (Eq. 3). Our first four choices of intrinsic rewards are well-established in the psychological literature (i.e., novelty (Modirshanechi et al., 2023d; Xu et al., 2021), surprise (Kobayashi et al., 2019), information gain (Horvath et al., 2021; Nelson, 2005) and empowerment (Brändle et al., 2023; Klyubin et al., 2005)), whereas the other two has been proposed only recently (Maximum Occupancy Principle (MOP) (Ramírez-Ruiz et al., 2024) and Successor-Predecessor Intrinsic Exploration (SPIE) (Yu et al., 2024)). In this section, we provide a brief and conceptual overview of each intrinsic motivation;

see Intrinsic motivations detailed for more detailed formulation and further theoretical analyses.

(i) Novelty rewards the agent for exploring rarely encountered states. Specifically, for a transition $s, a \to s'$, the agent receives a reward that is a decreasing function of the observation frequency of s', i.e., the less frequently the agent has visited s', the more rewarded it feels by visiting s'.

(ii) Surprise rewards the agent for experiencing unlikely transitions and encourages exploration of actions with uncertain or unexpected outcomes. Specifically, for a transition $s, a \to s'$, the agent receives a reward that is a decreasing function of $\hat{P}^{(t)}(s'|s,a)$ (Eq. 1), i.e., the less the agent expects to visit s' (conditioned on s and a), the more reward it feels by visiting s' (after taking a in s).

(iii) Information gain rewards the agent for reducing (the epistemic) uncertainty about the environment by acquiring new information. The reward for observing a transition $s, a \to s'$ is determined by the size of update of the agent's model of the environment, quantified using the KL divergence of the updated model from the previous model, i.e., the more the agent updates its estimated probabilities (Eq. 1) after transition $s, a \to s'$, the more rewarded it feels.

(iv) Empowerment rewards the agent for achieving states where its actions lead to a *diverse* set of *predictable* outcomes. The reward for observing a transition $s, a \to s'$ is the empowerment value of s', defined in Intrinsic motivations detailed, i.e., the more 'options' the agent has at state s', the more it feels rewarded by visiting s'.

(v) MOP can be seen as a regularized surprise that rewards the agent for experiencing unlikely transitions but also for maintaining a high-entropy policy. As a result, it motivates the agent to explore a wide range of states and actions and have diverse trajectories. The reward for observing a transition $s, a \to s'$ is a decreasing function of both $\hat{P}^{(t)}(s'|s,a)$ and $\pi_s^{(t)}(a)$. Details on how the policy is computed and integrated into the reward definition can be found in Intrinsic motivations detailed.

(vi) SPIE rewards the agent for visiting rare states as well as those that are critical for reaching isolated regions. Specifically, the reward for observing a transition $s, a \to s'$ is determined by the difficulty for the agent to reach s' from all other states except s. This encourages visiting s' if it is easy to reach from s but difficult from the other states; this is the case, e.g., if s' is in an isolated region or if s is a bottleneck state. Here, a state s' is considered difficult to reach from a state s if the agent rarely visits s' shortly starting from s.

Performance measures

While intrinsic motivations guide the agent's immediate and local behavior, they do not necessarily specify the long-term goal of curiosity. On the other hand, the curiosity outcome can be evaluated only after a series of actions and across the whole environment, hence it remains unclear what are the benefits of seeking different intrinsic rewards for a curious agent. To answer this question and quantify the merits of seeking different intrinsic rewards (the 'What' of curiosity), we define three performance measures that capture the potential ideal outcomes for a curious agent (the 'Why' of curiosity). Our definitions are inspired by previous literature and common intuition on the purpose of curiosity:

Measure 1: Environment exploration. Curiosity is closely linked to exploration (Kashdan et al., 2009; Modirshanechi et al., 2023b; Voss and Keller, 2013). Hence, one key goal of a curious agent can be to reach and visit all states in an environment. We measure the success of an agent, concerning this goal, by the fraction of unvisited states after a certain number of steps. A successful agent minimizes this fraction.

Measure 2: Model accuracy. Curiosity is often associated with gaining knowledge (Schmitt and Lahroodi, 2008; Szumowska and Kruglanski, 2020) and refining internal models (Pisula, 2009; Poli et al., 2024; Schmidhuber, 2010). Hence, another main goal of a curious agent can be to build the most accurate model of its environment. In our case, the internal model refers to the agent's estimation of the transition probabilities, which should closely approximate the true transition probabilities. We measure the success of an agent, concerning this goal, as the difference between the estimated transition probabilities $\hat{P}(s'|s,a)$ and the ground truth after a certain number of steps, using Root Mean Squared Error (RMSE). A successful agent minimizes this difference.

Measure 3: Uniform state visitation. It has been hypothesized that one main goal of curiosity is to find valuable sources of 'extrinsic' rewards (Bellemare et al., 2016; Modirshanechi et al., 2023b; Pathak et al., 2017). However, since the world is inherently changing (Liakoni et al., 2021; Nassar et al., 2010), the successful discovery of sources of rewards requires balanced and frequent visitation of all states. Hence, another main goal of a curious agent can be to achieve an even distribution of visits across the individual states, in order to avoid a disproportionate concentration in certain regions (similarly to Nedergaard and Cook (2023); Tolguenec et al. (2024)). This is also in line with observations that repetitive experiences induce boredom in humans (Geiwitz, 1966) and motivate them to seek new stimuli (Bench and Lench, 2013, 2019). A curious agent should similarly avoid staying in the same region for too long. We measure the success of an agent, concerning this goal, as the difference between the agent's state visitation frequency and the uniform distribution (using RMSE) after a certain number of steps. A successful agent minimizes this difference.

Environment generation

To systematically study the link between intrinsic rewards and curiosity objectives, we need a procedure for generating diverse environments with realistic features. In curiosity research, experimental paradigms are typically unique and hand-crafted, lacking standardized multi-step environments. Our goal in this section is to propose an environment generation algorithm that replicates the main relevant features of real-world environments as well as the environments commonly used in the experimental studies of curiosity (Fig. 1). Common environment structures in experimental studies of curiosity are mazes (Behrens et al., 2018; Kosoy et al., 2020; Tolman, 1948) and grid worlds (Botvinick et al., 2009; Dayan, 1993; de Tinguy et al., 2024; Piray and Daw, 2021; Singh et al., 2010a; Yu et al., 2024; Zheng et al., 2020b). These serve as the foundation for our generation algorithm. Additionally, some studies have highlighted the relevance of long-range connections (Viswanathan et al., 2016), sinks states (Modirshanechi et al., 2023d; Xu et al., 2021) and stochasticity (Mehlhorn et al., 2015; Modirshanechi et al., 2023d). Moreover, the number of available options has been shown to have an impact on human behavior (Fasolo et al., 2009; Mehlhorn et al., 2015; Scheibehenne et al., 2010). Taking these observations into account, our algorithm generates environments in three main steps (see Supplementary Section Environment generation for details): (i) It creates a maze with a branching structure, (ii) it integrates grid-like rooms within the maze, and finally, assigns each room to one (and exactly one) of the following properties:

- **Sink**: If a room is assigned to be a sink, then the algorithm introduces additional one-way connections from other parts of the environment *to this room*. A sink room is easy to reach from the rest of the environment. As a result, naive exploration strategies may struggle to navigate the entire environment without repeatedly falling into the sink. In video games, the starting point often acts as a sink state, as dying resets the player to the start. In real life, laying on a couch, watching TV, or scrolling on social media can be seen as sink states, as they are easy to engage in and may prevent agents from exploring other possibilities.
- **Source**: If a room is assigned to be a source, then the algorithm introduces additional one-way connections *from this room* to other parts of the environment. From a source room, it is easy to quickly reach any region of the environment. States in a source room have in general more available options than the rest of the environment. Real-life examples of source states are situations with a wide range of choices, which include being at an airport, choosing a dish at a restaurant, buying a house, planning a vacation, or moving to a new city.
- **Stochastic**: If a room is assigned to be stochastic, then transitions within the room are partly random. Specifically, when an agent selects an action a from a state s within a stochastic room, there is a fixed probability that the action will result in the agent moving to a random neighbor of s in the room instead of the intended destination of a. Unpredictability is common in everyday life, such as when watching TV, interacting with others, or engaging in activities where outcomes are not always certain (e.g., gambling or investing in the stock market).
- **Neutral**: If a room is assigned to be neutral, then none of the aforementioned modifications are applied to the room.

The algorithm receives, as input, a few parameters that specify the properties of the generated environments, such as the number of states, the number of intersections and rooms, the room sizes, the distribution of room types, and the intensity of the room properties. All parameters are described in the Supplementary Material (Table 1).

Environment types

Using our environment generation algorithm, we can create various types of environments. We focus on five types for most of our results, namely Neutral, Sink, Source, Stochastic and Mixed environments. Since the process is non-deterministic, many distinct environments can be produced within each type, but they are expected to exhibit similar properties. The environment types considered are detailed in Supplementary Table 1. In summary, each environment contains 100 states, including 4 rooms of 16 states each. Neutral environments contain 4 neutral rooms. Sink environments feature one sink room with 50 additional incoming connections. Source environments contain one source room with 50 additional outgoing connections. Stochastic environments include one stochastic room where actions lead to a random neighbor within the room. Finally, Mixed environments consist of one neutral room, one sink room (with 50 incoming connections), one source room (with 50 outgoing connections), and one stochastic room.

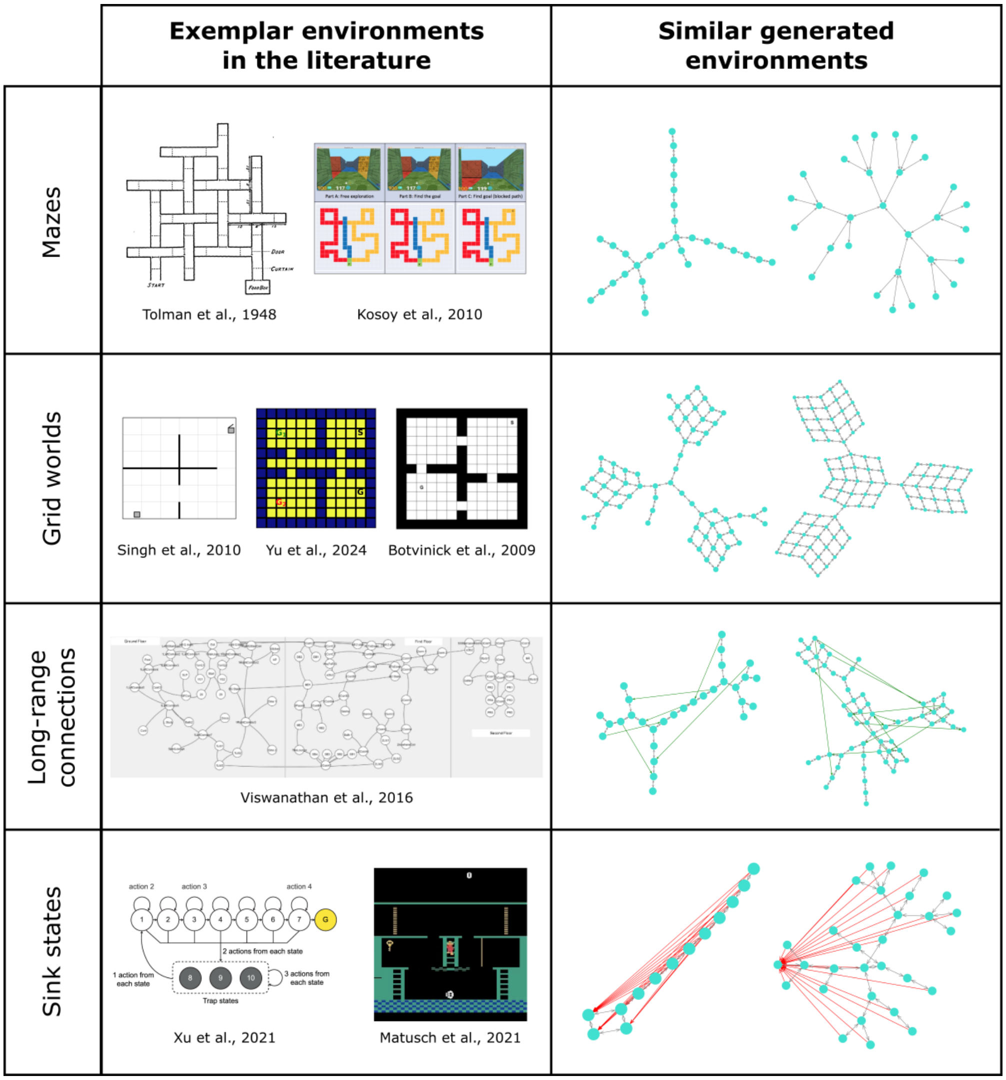


Figure 1: Comparison of environments from the exploration and curiosity literature with similar environments generated by our algorithm. The generated environments shown in the figure are exemplar realization that exhibit similar properties to the literature examples. However, due to the stochastic nature of the generation process, different instances with the same properties could also be produced. Blue nodes represent states, and edges indicate possible actions to transition between states. Gray edges are bidirectional. Green edges (originating from a source room) and red edges (leading to a sink room, see Environment generation) are unidirectional. Mazes are common in multi-step navigation tasks (Kosoy et al., 2020; Tolman, 1948) and are represented by complex, branching structures. Grid worlds, another common task type (Botvinick et al., 2009; Singh et al., 2010a; Yu et al., 2024), feature regular, grid-like structures. Long-range connections, highlighted as interesting in the literature (Viswanathan et al., 2016), are environments with states that have distant connections. Sink states are those that are easy to reach but hard to escape (Xu et al., 2021), similar to challenging game environments like Montezuma's Revenge (Matusch et al., 2020) where the starting state acts as a sink state since dying resets the player to the start.

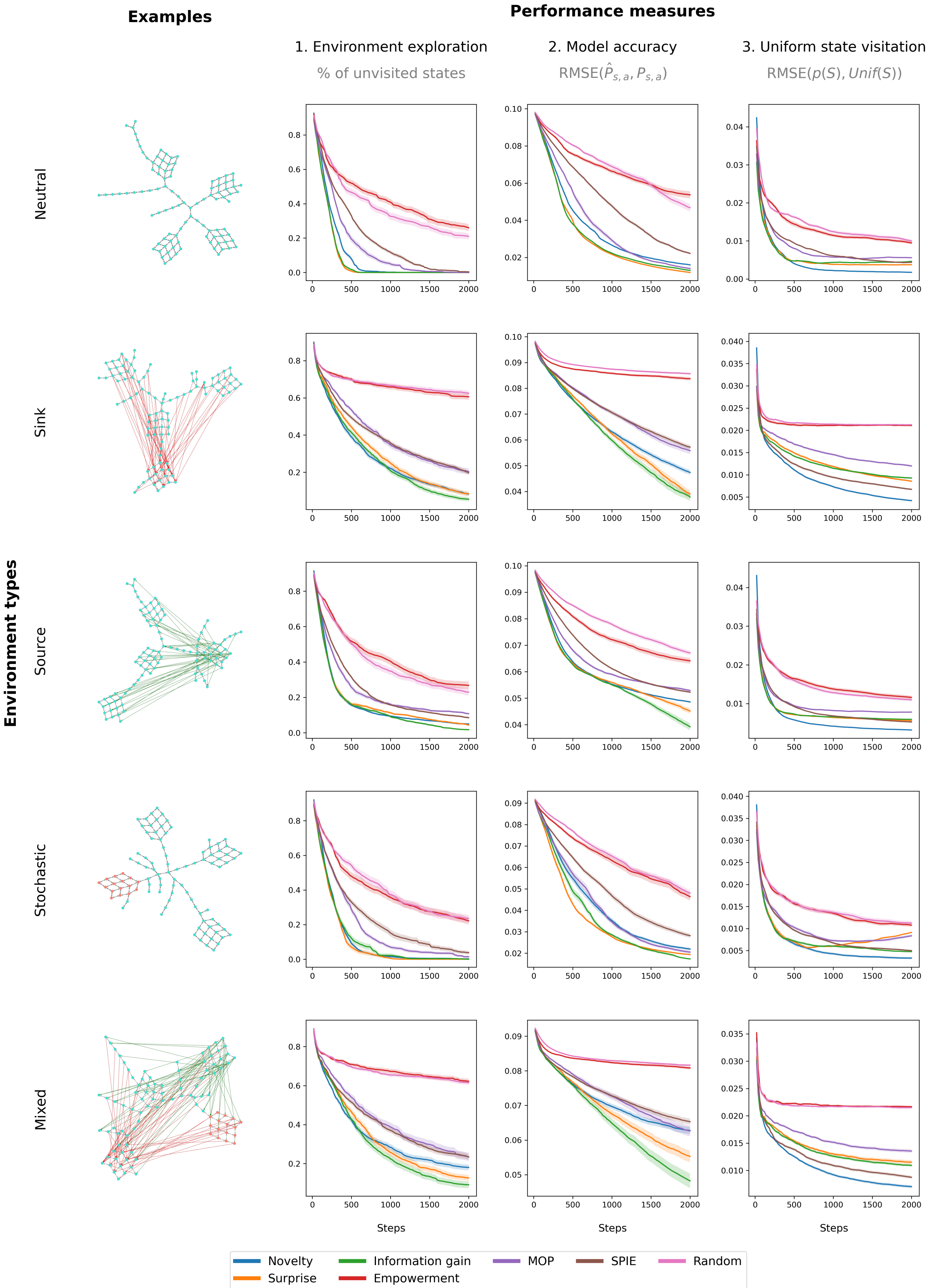


Figure 2: (Caption on the following page.)

Figure 2: Results for six intrinsic motivations (+ random), five environment types and three performance measures. Each subplot corresponds to the combination of one environment type and one performance measure. An exemplar environment is shown for each type. Blue nodes represent deterministic states, while red nodes correspond to stochastic states. The performance of each intrinsic motivation was evaluated over 50 different instances of each environment type, with the average displayed and the shaded areas representing the standard error of the mean. The first 2000 steps of simulation are shown. For performance measure 1, the y-axis represents the percentage of unvisited states. For measure 2, it displays the RMSE between the estimated transition probabilities and the ground truth. For measure 3, it shows the RMSE between the state visitation frequencies and the uniform distribution. In each case, a desirable performance is represented by a lower curve. In each case, the hyperparameter β was optimized for the first 500 steps only. **Environment types:** Neutral environments contain 4 neutral rooms. Sink environments contain one sink room with 50 additional connections leading to it. Source environments contains one source room with 50 additional connections originating from it. Stochastic environments include one stochastic room. Mixed environments consist of one neutral room, one sink room, one source room, and one stochastic room.

Performance analysis across different environments and measures

To quantify the merits of seeking different intrinsic rewards in different environments, we simulated model-based reinforcement learning agents and measured their performance (defined in Performance measures) in our five environment types (specified in Environment types).

Overall, we observe that the novelty-seeking agents (blue in Fig. 2) consistently have the best performance according to Measure 3 (uniform state visitation; Fig. 2, right) and are competitive on Measure 1 (environment exploration; Fig. 2, left), except in the Mixed environments. On the other hand, agents seeking Surprise (orange in Fig. 2) or Information Gain (green in Fig. 2) excel on Measures 1 and 2 (Fig. 2, left and middle) but perform consistently worse than novelty-seeking agents for Measure 3 (Fig. 2, right). Interstingly, agents seeking Empowerment (red in Fig. 2) perform poorly across all scenarios; this is essentially because they avoid unknown regions, which are perceived as non-empowering due to uncertainty. As a result, they avoid further exploration of the environment and remain in where they initially explored. Agents seeking either of the two recently proposed intrinsic rewards, MOP and SPIE (purple and brown in Fig. 2, respectively), perform worse than agents seeking surprise, information-gain, or even novelty on Measures 1 and 2. However, SPIE sometimes outperforms surprise and information-gain on Measure 3, while MOP is only better than random agents (pink in Fig. 2) and those seeking Empowerment on Measure 3.

While the performance of agents seeking each intrinsic reward is fairly consistent across multiple environments of the same kind (Supplementary Fig. 7), it varies strongly between environments of different types (different rows of Fig. 2). Different environment types affect performance in distinct ways: Neutral environments offer a good reference point. As sink rooms are challenging to escape, it is also more challenging to explore Sink environments than Neutral environments. As a result, Sink environments can more vividly show the differences in the performance of different agents (particularly for Measure 3; Fig. 2, row 2, column 3). On the other hand, in Source environments, building an accurate model of the environment (Measure 2) requires agents to repeatedly visit the source room to test all actions. This benefits Surprise and Information Gain agents, which are attracted to unknown actions, but is specifically detrimental for Novelty as it discourages state revisitation (Fig. 2, row 3, column 2). Interstingly, in Stochastic environments, Surprise and MOP tend to stay in the stochastic room after learning sufficiently about the environment, resulting in poor performance on Measure 3 (Fig. 2, row 4, column 3, see Intrinsic motivations detailed for a formal explanation of this asymptotic behavior), whereas the other algorithms do not show such an excessive attraction to stochasticity. Mixed environments combine features of previous types but display different behaviors. Notably, Novelty performs worse in these environments on measures 1 and 2 compared to others.

To go beyond the comparison across environment types, we next evaluated the impact of specific environment parameters on agent performance. Specifically, we manipulated the branching rate and the number of sink connections (Fig. 3) in an environment inspired by Xu et al. (2021). Specifically, we considered a class of environments with 100 states, where 4 states built a single *sink* room, i.e., 96 states were neutral and outside of the room. In this setting, the branching rate influences how these 96 states are arranged. At a branching rate of 0, the states are arranged in a straight line, whereas at a branching rate of 1, the states are arranged in a tree-like structure (see examples in Fig. 3a). Importantly, the performance of different algorithms drastically changes as the branching rate increases from 0 to 1 (Fig. 3a). Novelty and SPIE, initially top performers at a branching rate of 0, become among the worst as the branching rate increases to 1 in the first two measures. This could be explained by the tendency of novelty-seeking agents to choose actions that are known to lead to a relatively novel state, s, rather than taking an unknown action in some situations (where the expected novelty of the unknown action might be less than that of s). As a result, novelty-seeking agents may not explore all possible actions and could miss large parts of the environment, especially when the branching rate is 1. Similarly, increasing the number of sink connections generally benefits Novelty and SPIE comparatively to other motivations (Fig. 3b). This shows that the structure of the environment has a great influence on the comparative performance of intrinsic motivations, indicating that results from experiments in one specific environment may not generalize well to others. For example, in an environment very similar to the case with a branching rate of 0, Xu et al. (2021) found that Novelty to be dominant drive of human exploration. Whether this result is environment-independent can be, for example, tested by repeating the same experimental task in an environment with a branching rate of 1 (see Modirshanechi et al. (2023d) for an alternative replication of the results of Xu et al. (2021)).

Novelty and information gain as two main axes of curiosity

In the previous section, we saw that agents seeking different intrinsic rewards exhibit a diverse range of performance in different environment types. However, we also observed that the best performing intrinsic reward, for every environment type or performance measure, is either Novelty or Information Gain (Fig. 2 and Fig. 3). Specifically, by integrating over time (Fig. 4), we observe that Information Gain outperforms all other motivations in environment exploration (Measure 1) and model accuracy (Measure 2), whereas Novelty is the best reward signal in achieving uniform state visitation (Measure 3).

These results propose that Novelty and Information Gain are two key drives of exploration. To further this proposition, we simulated model-based RL agents that use a linear combination of Novelty and Information Gain as the reward signal (Fig. 5). Interestingly, we observe that, by even having a fixed and equal weight for Novelty and Information Gain ($\alpha = 0.5$ in (Fig. 5)), these ‘hybrid’ RL agents reached close-to-optimal performance in all environment types and for all performance measures (Fig. 5). This implies that an agent that can adaptively and on demand fine-tune its reward function will always reach the best performance (see Modirshanechi et al.

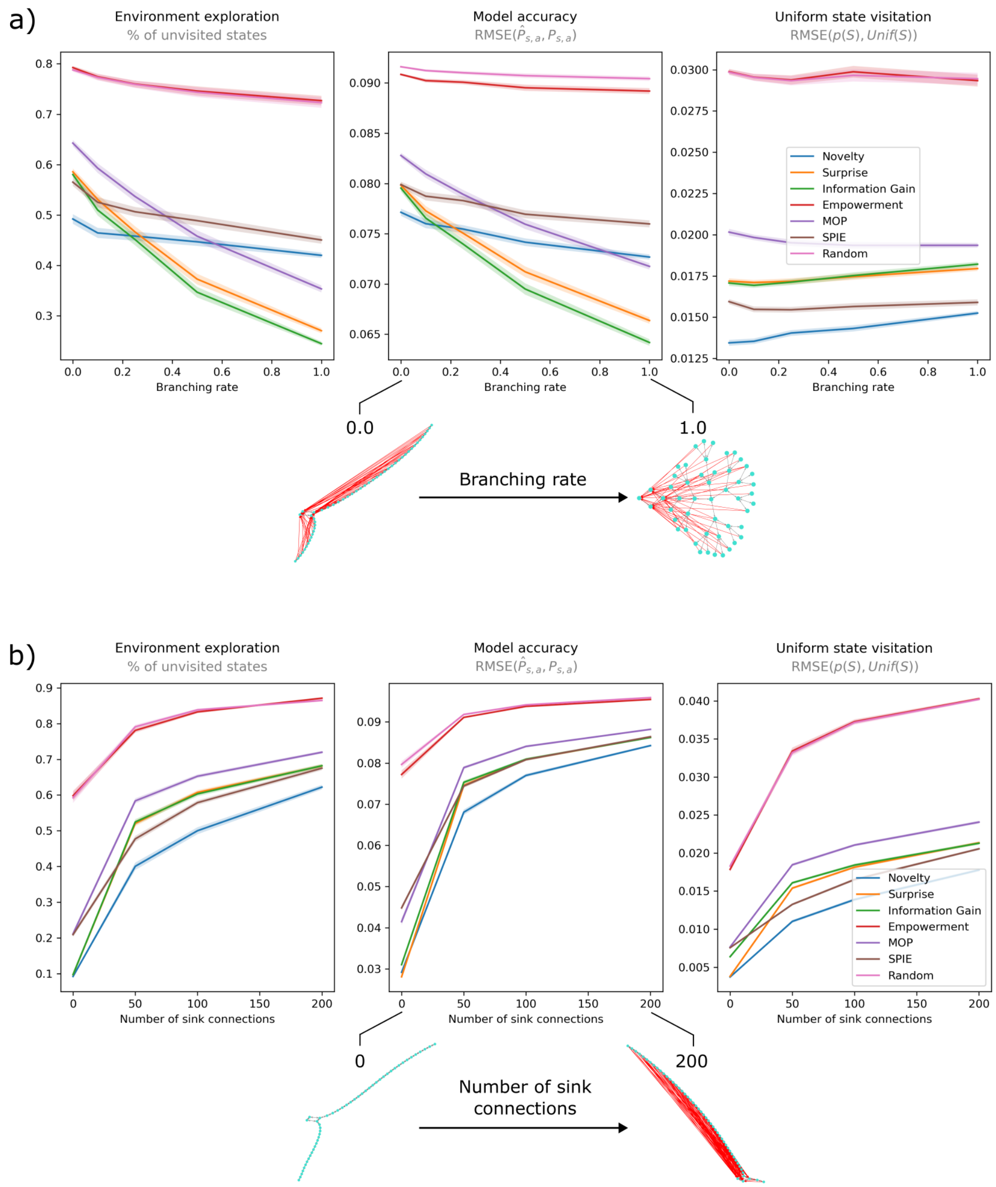


Figure 3: Performance variation of each intrinsic motivation as a single environment parameter is changed. The environment, inspired by Xu et al. (2021), contains one sink (trap) room with 4 states and 96 other states. The parameters used to generate the environments can be found in Table 1. The exemplar environments shown are smaller versions (50 states), for illustration purposes. To compute the score for a given environment, we run the agent as in Fig. 2 and calculate the area under the curve (AUC) of each measure over 2000 steps of simulations. The score for each environment type is obtained by averaging this value over 50 environment instances. (a) The parameter changed is the branching rate: at a branching rate of 0, the states are arranged in a straight line, while at a branching rate of 1, each state has multiple actions leading to distinct parts of the environment. In each case, 100 additional connections lead to the sink. (b) The parameter changed is the number of sink connections, while the branching rate is fixed to 0.

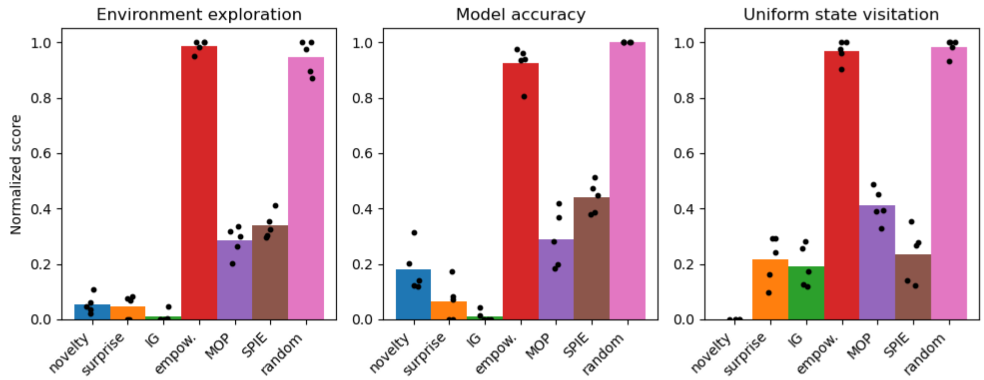


Figure 4: Average normalized score across environments for each intrinsic motivation, calculated as follows: for each setup (environment and measure), the score of each intrinsic motivation is computed as the area under the curve of Fig. 2. These scores are normalized, setting the best-performing intrinsic motivation to 0 and the worst to 1. Each dot represents the score on one environment type, and the average score over all environments is displayed. The same experiment was conducted using the KL divergence instead or RMSE for measure 2 and 3. The results are very similar and can be found in Supplementary Fig. 8.

(2023c) for a discussion). Importantly, this observation supports the hypothesis that Novelty and Information Gain are fundamental axes of curiosity, with each providing distinct benefits (in line with recent experimental studies on humans (Dubey and Griffiths, 2019; Monosov, 2024; Poli et al., 2022)).

Dissociating intrinsic motivations

To gain further insights into how different intrinsic motivations influence exploratory behavior, we analyzed exploration patterns of agents seeking different intrinsic rewards within the Mixed environment type (environments with one sink, one source, one stochastic, and one neutral room; see Environment types).

Specifically, we quantified the proportion of time that agents spend in different rooms of the environments (Fig. 6). Agents with a random policy predominantly remain in the sink room due to the difficulty of escaping it through random actions. Novelty-driven agents, on the other hand, quickly achieve a near-uniform state visitation frequency. Agents seeking SPIE follow the same trend as Novelty-seeking agents, but they learn more slowly than Novelty. After sufficient learning, Surprise-driven agents mostly spend time in the stochastic room, which has the highest transition uncertainty. Agents seeking MOP behave similarly to Surprise-driven agents, but they lean closer to random agents – as MOP also rewards policy entropy. As observed before (Fig. 2), agents driven by Information Gain learn effectively, but they eventually trend towards the random policy (as Information Gain converges to zero; see Intrinsic motivations detailed). Different from all other agents, Empowerment-driven agents do not explore the environment sufficiently to even discover all four rooms; they mainly stay within known regions (unknown regions are expected to be non-empowering due to uncertainty) which is most of the time the sink room as it acts like an attractor. However, once agents driven by Empowerment know the transition probabilities of the entire environment (i.e., are aware of the properties of all four rooms), they spend most of their time in the source room, which offers the highest empowerment.

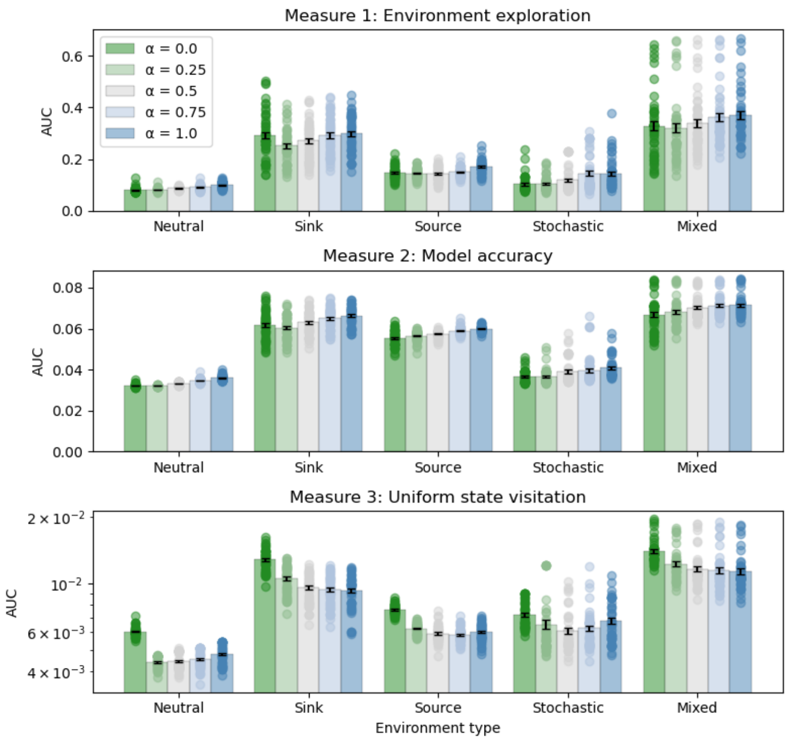


Figure 5: Combination of Information Gain and Novelty. At each step, the agent receives a weighted combination of information gain and novelty rewards, as $\alpha \cdot \text{Nov} + (1 - \alpha) \cdot \text{IG}$. In green, the agent is fully motivated by information gain; in blue, it only receives novelty rewards. For each value of α, the parameter β was optimized separately as in Hyper-parameters selection. The agents were run similarly as for Fig. 2, and the Area Under the Curve (AUC) after 2000 steps is reported. The results are averaged over 50 different instances of each environment type. The error bars represent the standard error of the mean. For each measure, a desirable performance is represented by a lower bar.

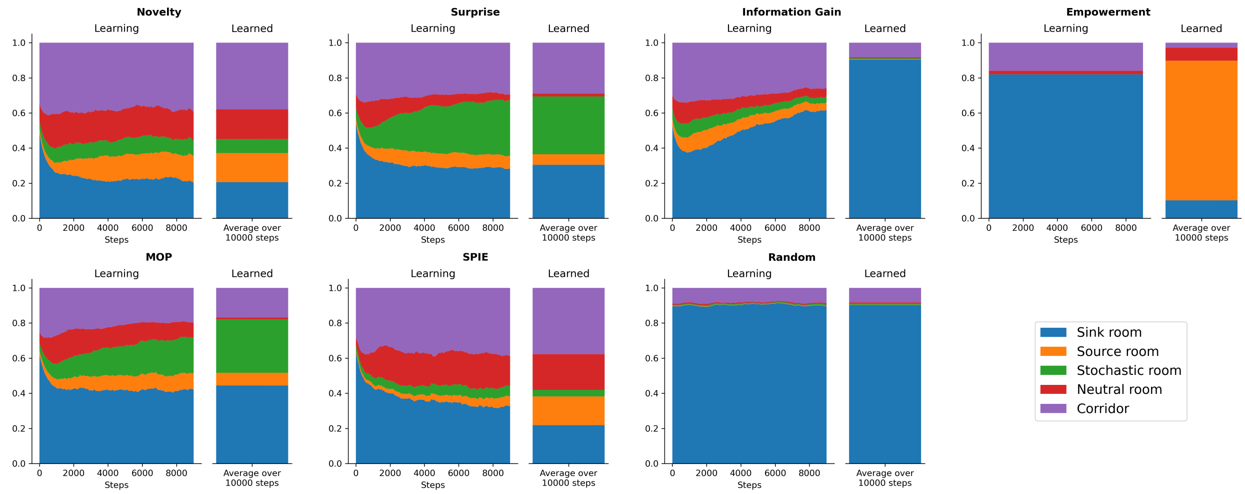


Figure 6: Proportion of time spent in each region of the environment. Agents were run in the Mixed environment (see Environment types) for 10'000 steps. Each room contains 16 states, and the corridor contain 36 states. In the "Learning" phase, agents start without knowledge of the environment and build a model of it, as in previous experiments. The evolution of the proportion of time spent in each region during learning is shown, with a window average of 1000 steps. In the "Learned" phase, the experiment is repeated but the agent's model of the environment is fixed to the ground truth to assess for asymptotic behavior. The proportion of time spent is averaged over the 10'000 steps. Both phases are repeated 50 times and averaged. To allow for a fair comparison, the hyperparameter β is computed as $\frac{1}{std(r)}$ where $std(r)$ is the standard deviation of the intrinsic reward r computed over 10'000 steps under a random policy. The behavior of each intrinsic motivation in the "Learned" case corresponds to the expected asymptotic behavior, derived from the reward formulation in Intrinsic motivations detailed.

Overall, these results show that our environment generation algorithm can help to differentiate and highlight essential features of various intrinsic motivations. The different exploration patterns confirm that different intrinsic motivations lead to unique behaviors, even within the same environment. This suggests that our algorithm for environment generation can be used to design experiments where behavioral differences between agents seeking different intrinsic rewards are most easily detectable. These experiment designs can be used for identifying exploration strategies in both humans and animals.

Discussion

In this study, we aimed to answer key questions about curiosity-driven behavior in humans and animals using simulated agents. Using a new environment generation algorithm, we assessed how different intrinsic motivations affect exploration in various environments. Our results show three main points: First, the performance of curiosity-driven agents depends highly on the structure of their environment. Second, information gain and novelty are the two most effective drivers of curiosity; information gain helps with exploring and understanding environments better, while novelty encourages a more even exploration of the environment. Third, different intrinsic motivations produce different exploratory patterns. Our environment generator creates settings where these differences are clear, making it easy to dissociate between different intrinsic motivations.

Our contributions can be summarized in two main points, which are developed in the following paragraphs: (i) we demonstrate the significant impact of environment structure on the performance of curious agents, and (ii) we introduce an environment generator to facilitate experimental design across multiple domains.

Our first main contribution is the evidence that environment structure significantly affects the performance of curious agents. Most recent studies on human curiosity use one or a few environments (Brändle et al., 2023; Horvath et al., 2021; Kobayashi et al., 2019; Poli et al., 2022; Ten et al., 2021b) to test hypotheses and draw conclusions. Interestingly, the conclusions often vary between experiments, suggesting that humans do not seek the same curiosity signals in all scenarios. We address this inconsistency by showing how the environment's structure influences the expected results. An optimal curious agent should not display the same behavior across different experiments. This may suggest that the simple strategies exhibited by humans in experiments are part of a more complex strategy with different assumptions about the task. While the importance of environment structure in exploration behavior has been acknowledged (Mehlhorn et al., 2015), it has to our knowledge not been highlighted with such precision and significance.

Our second major contribution is our proposed environment generation algorithm. This algorithm offers several advantages: (i) It simplifies the design of environments to test specific hypotheses. For instance, if we want to determine whether an agent (e.g., a human participant) behaves more similarly to novelty-seeking or surprise-seeking agents, the environment generator provides a rigorous framework for creating an environment that clearly distinguishes between the two. (ii) The algorithm allows for the creation of diverse environments to test agents in various scenarios while keeping a common ground for comparison. It helps isolate key parameters that significantly impact behavior. In many fields, it is common to use multiple environments to test a method, but these environments are often either very similar to one another (Kosoy et al., 2020; Yu et al., 2023; Zheng et al., 2020a), lacking generalization, or very different (Matusch et al., 2021; Piray and Daw, 2021; Singh et al., 2010b), making comparisons and interpretation difficult due to a lack of common ground. A parameterized environment representation helps generate various environments while maintaining a common basis for comparison. Additionally, the stochastic nature of the algorithm smooths out minor environmental details, ensuring that only relevant features significantly impact the results. For instance, in a fixed environment, we cannot be certain that observed results are due to the main feature of interest rather than an unrelated detail. With a stochastic environment generator, such details can be averaged out, ensuring that only relevant features significantly impact the results after multiple runs. (iii) The environment generator can serve as a valuable tool in other domains. For instance, it can be used for benchmarking in different areas, such as comparing model-based versus model-free approaches, or for developing and testing meta-learning algorithms. This flexibility enhances its utility across various research contexts, making it a powerful tool for experimental design and evaluation.

We used model-based RL to assess curiosity-driven behavior. It remains to be explored whether our findings hold true in other setups, such as model-free RL. Additionally, we did not consider scenarios where external rewards are present alongside intrinsic rewards. While it is expected that combining these two reward types would produce intuitive results, our simulations focused exclusively on intrinsic rewards, leaving this aspect unexplored. Another limitation is that all environments in our study were static, with no modifications occurring during the agent's navigation. Certain scenarios or hypotheses may require dynamic environments to better reflect real-world complexities. Furthermore, while our environment generation algorithm is expressive, it may not capture all real-life scenarios. It serves as an initial step that can be supplemented with additional factors that researchers find relevant. Future research should address these limitations.

Our findings are aligned with the intuition that humans adapt their exploration strategies to the task. Future studies could investigate the conditions under which this adaptation occurs. Such research could help clarify how people balance their curiosity-driven exploration with the specific goals of a task.

In conclusion, our study increases the understanding of curiosity by clarifying the roles of different intrinsic motivations and how they affect exploration behavior in different kinds of environments. Our environment generator is a tool for future research, specifically: experiment design, algorithm testing, and meta-learning.

Methods

Intrinsic motivations detailed

We consider six intrinsic motivations: novelty, surprise, information gain, empowerment, Maximum Occupancy Principle, and Successor-Predecessor. Each is described below.

Novelty

Novelty, as intrinsic motivation, rewards the agent for exploring unusual states—those encountered infrequently (Aubret et al., 2019; Bellemare et al., 2016; Ostrovski et al., 2017). We use the same mathematical formulation as Xu et al. (2021). We define the observation frequency of a state s as

$$p_N^{(t)}(s) = \frac{C_s^{(t)} + 1}{\sum_{s'} C_{s'}^{(t)} + |S|}$$

where $C_s^{(t)}$ represents the number of times state s has been encountered up to time t. The novelty of a state s is then expressed as a decreasing function of the observation frequency :

$$R_{Novelty}^{(t)}(s) = -\log p_N^{(t)}(s)$$

Asymptotic behavior: Let $P_\pi(s)$ be the long-term observation frequency achieved by a fixed policy π. The expected *average* novelty reward at each step for an agent following π is asymptotically equal to

$$\begin{aligned}
\mathbb{E}_{s\in S}[R_{Novelty}] &= \sum_s P_\pi(s) \cdot R_{Novelty}(s) && (4)\\
&= -\sum_s P_\pi(s) \cdot \log(P_\pi(s)) && (5)\\
&= \mathcal{H}(P_\pi), && (6)
\end{aligned}$$

where $\mathcal{H}(P_\pi)$ is the entropy of the state observation frequency. As discount factor λ gets close to 1, the policy π that maximizes Q-values in Eq. 2 becomes the same as the policy π that maximizes $\mathbb{E}_{s\in S}[R_{Novelty}]$ (Puterman, 1994). Hence, an agent focused on maximizing this reward will, intuitively and for large discount factors, adopt a policy π that increases the entropy of the state observation frequency. This should result in a close to uniform state visitation (Measure 3 in Performance measures).

Surprise

Surprise, as intrinsic motivation, rewards the agent when observing transitions that were anticipated to be unlikely. We follow (Achiam and Sastry, 2017; Barto et al., 2013) and define the surprise of a transition as its Shannon surprise or surprisal (mod, 2022; Modirshanechi et al., 2023a):

$$R_{Surprise}^{(t)}(s, a, s') = -\log \hat{P}^{(t)}(s'|s, a)$$

Here, $\hat{P}^{(t)}(s'|s, a)$ represents the estimated probability of the transition. Higher intrinsic rewards are granted for transitions the agent considers improbable.

Asymptotic behavior: Over time, the estimated transition probabilities $\hat{P}(s'|s,a)$ should converge to the true probabilities $P(s'|s,a)$. The expected surprise reward obtained for taking an action a in state s is

$$\begin{aligned}
\mathbb{E}_{s' \in S}[R_{Surprise}(s,a,\cdot)] &= \sum_{s'} \hat{P}(s'|s,a) \cdot R_{Surprise}(s,a,s') && (7) \\
&= -\sum_{s'} \hat{P}(s'|s,a) \cdot \log(\hat{P}(s'|s,a)) && (8) \\
&\underset{t \to \infty}{\approx} -\sum_{s'} P(s'|s,a) \cdot \log(P(s'|s,a)) && (9) \\
&= \mathcal{H}(S'|s,a), && (10)
\end{aligned}$$

where $\mathcal{H}(S'|s,a)$ is the entropy of the next state distribution given action a in state s. This implies that, in the long run, the agent will prefer actions that lead to stochastic (uncertain) outcomes, as deterministic actions will eventually yield no reward. Therefore, after learning sufficiently about the environment, the surprise-seeking agent will focus on stochastic areas of the environment.

Information gain

Information gain, as intrinsic motivation, rewards the agent based on the amount of information it acquires, equivalent to the decrease of uncertainty in the knowledge that the agent has of the environment (Itti and Baldi, 2009; Oudeyer and Kaplan, 2009; Storck et al., 1995). We use the formulation also referred to as Postdictive surprise (mod, 2022; Kolossa et al., 2015; Modirshanechi et al., 2023a). Following a transition, the agent updates its environment model, and the intrinsic reward is determined by the difference between the updated and previous models. In mathematical terms:

$$R_{IG}^{(t)}(s,a,s') = KL\left(\hat{P}^{(t)}(\,\cdot\,|s,a) \;||\; \hat{P}^{(t+1)}(\,\cdot\,|s,a,s_{t+1}=s')\right)$$

Where KL is the Kullback-Liebler divergence (Kullback, 1997). Here, $\hat{P}^{(t)}(\,\cdot\,|s,a)$ and $\hat{P}^{(t+1)}(\,\cdot\,|s, a, s_{t+1}=s')$ are the estimated probability distributions over next states before and after observing the transition $s,a \to s'$, respectively.

Asymptotic behavior: Over time, the estimated transition probabilities $\hat{P}(s'|s,a)$ will converge to the true probabilities $P(s'|s,a)$. Therefore, the information gain reward $R_{IG}^{(t)}(s,a,s')$ for every transition will tend to 0 as $t \to \infty$. This implies that the agent will converge to the uniformly random policy.

Empowerment

Empowerment is a measure of the degree of control or influence an agent has over its environment from a particular state (Klyubin et al., 2005; Salge et al., 2013). It's a way to quantify how much an agent can affect or change its surroundings (i.e. the future observed state) based on its actions from that state. Formally, the empowerment of a state s is defined as the channel capacity of the actuation channel, i.e. the maximum potential information transmission between the agent's actions and the subsequent impact of these actions after a certain duration. Here we consider

1-step empowerment, which is defined as:

$$E^{(t)}(s) = \max_{p(a)} \ I(S'; A|s) \tag{11}$$

$$= \max_{p(a)} (\mathcal{H}(S') - \mathcal{H}(S'|A)) \tag{12}$$

$$= \max_{p(a)} (\mathcal{H}(A) - \mathcal{H}(A|S')) \tag{13}$$

where A and S' are random variable for the action and next state, respectively. There are multiple ways to intuitively understand this formula. Examining eq.12, we note that in order to maximize empowerment, we aim to maximize the entropy of the next state S', implying a diversity of potential next states. Simultaneously, we seek to minimize $\mathcal{H}(S'|A)$, to reduce stochasticity in the process. This conceptually aligns with the desire to have control over the destination when selecting an action. An alternative interpretation is found in eq.13. To maximize empowerment, we want to maximize $\mathcal{H}(A)$ to enable numerous possible actions, while minimizing $\mathcal{H}(A|S')$ to account for the fact that multiple actions may lead to the same state. Essentially, this seeks to maximize the count of *effective* actions—those leading to diverse outcomes. In each case, we consider the maximum over all possible action distributions $p(a)$. For an agent driven by empowerment as intrinsic motivation, we set $R^{(t)}_{Empowerment}(s, a, s') = E^{(t)}(s')$.

Asymptotic behavior: An agent driven by empowerment will seek out states with a large number of available options, as these states offer the most control. In the long run, the agent's estimation of the transition probabilities will converge to the true probabilities. Therefore, the agent will tend to stay in the most empowering regions of the environment (e.g. source states) and avoid reaching isolated areas with fewer options.

Maximum Occupancy Principle (MOP)

Introduced in Ramírez-Ruiz et al. (2024), MOP as intrinsic motivation considers that the goal of an agent's behavior is to maximize the occupancy of future action-state paths. The agent aims to maximize the return

$$R_{MOP}(s, a, s') = -\log\left(\pi^{\alpha_{MOP}}(a|s)\hat{P}^{\beta_{MOP}}(s'|s, a)\right) \tag{14}$$

Where the subscript (t) has been omitted for clarity. An agent motivated by MOP is expected to favor high entropy policies and highly stochastic regions of the environment. In our experiments, we set $\alpha_{MOP} = \beta_{MOP} = 1$ to give equal weights to these two aspects. Unlike for other intrinsic motivations, we do not compute the policy by applying softmax on Q-values. Instead, we use a modified version of value iteration as in Moreno-Bote and Ramirez-Ruiz (2023); Ramírez-Ruiz et al. (2024) to consider the optimal policy at every step.

Asymptotic behavior: As detailed in Ramírez-Ruiz et al. (2024), MOP aims to find a policy π that maximizes the value function $V_\pi(s)$ defined as

$$V_\pi(s) = \alpha\mathcal{H}(A|s) + \beta\sum_a \pi(a|s)\mathcal{H}(S'|s, a) + \gamma\sum_{a,s'} \pi(a|s)P(s'|s, a)V_\pi(s'), \tag{15}$$

where $\mathcal{H}(A|s)$ is the policy in state s, and $\mathcal{H}(S'|s, a)$ is the entropy of the next state distribution given action a in state s. The first term favors states with multiple available actions, the second term encourages to experience stochastic transition and the last term accounts for the value of the next state. The agent will aim to reach the states where $V_\pi(s)$ is highest. Therefore, after learning sufficiently about the environment, we expect the agent to spend most of its time in stochastic areas and regions with many actions.

Successor-Predecessor Intrinsic Exploration (SPIE)

SPIE was introduced in Yu et al. (2024). Instead of only rewarding the agent for discovering new states like Novelty, SPIE also rewards it for visiting states that lead to isolated regions. The key idea is to use both forward-looking (successor) and backward-looking (predecessor) information to identify and navigate critical or "bottleneck" states. The reward is defined based on the successor representation (SR), which measures how often one state is expected to be visited in the future with the current policy, given that the agent is currently in a specific state. The reward is defined as:

$$R^{(t)}_{SPIE}(s,a,s') = \hat{M}^{(t)}[s,s'] - \|\hat{M}^{(t)}[\cdot,s']\|_1 \tag{16}$$

where $\hat{M}^{(t)}[s,s']$ is the learned SR for the state s' given state s, and $\|\hat{M}^{(t)}[\cdot,s']\|_1$ is the sum of the SRs of s' from all states. Intuitively, the reward is high when state s' is difficult to reach from all states except s. Therefore, if s is a bottleneck state, the reward is high, encouraging the agent to visit such states. Unlike the original paper, we do not approximate the matrix $\hat{M}^{(t)}[s,s']$ using an online TD-learning rule. Instead, we compute it exactly after each observed transition using the agent's environment model.

Asymptotic behavior: Yu et al. (2024) argues that the behavior of SPIE is non-trivial, even when the matrix M is known or fixed. However, since the reward is higher for rarely encountered states, we expect the agent to reach a close to uniform state visitation.

Hyper-parameters selection

The framework described in General framework contains three hyper-parameters: ϵ, λ and β. The parameter ϵ is a small positive constant added to transition counts to prevent zero probabilities for unseen transitions, λ is the discount factor that determines the weight of future rewards compared to immediate rewards, and β is the Softmax inverse temperature parameter that influences the randomness of the action selection based on the Q-values.

In all experiments, we set $\epsilon = 1/n$ and $\lambda = \sqrt[n/2]{0.5}$ where n is the number of states in the environment, so that a future reward that is $n/2$ step away is discounted to half its value. On the other hand, β is optimized in a more complex manner. Each combination of intrinsic motivation, performance measure, and environment type is referred to as a setup. The inverse temperature β was optimized separately for each setup. For instance, in Fig. 2, with 6 intrinsic motivations, 3 performance measures, and 5 environment regimes, there are 90 setups, requiring 90 optimized values for β. The optimization process for each setup is as follows: First, we generate 50 environments based on the chosen type. Then, we find the value of β that gives the best score using grid search. To compute the score for a specific choice of β, we run an agent for 500 steps on each environment. We evaluate the performance measure every 100 steps and calculate the average, resulting in a score for each environment. The overall score is calculated as the average score across the 50 environments.

Acknowledgement

We thank Wulfram Gerstner for feedback and discussions throughout the project and on related topics, and Sophia Becker for helpful discussions, in particular on the asymptotic behavior of novelty, surprise and information gain. This research was supported by Swiss National Science Foundation No. 200020_207426.

Author Contributions

Competing Interests statement

The authors declare no competing interests.

Code and data availability

All code and data needed to reproduce the results reported in this manuscript will be made publicly available after publication acceptance.

References

A taxonomy of surprise definitions. *Journal of Mathematical Psychology*, 2022. doi: 10.1016/j.jmp.2022.102712.

J. Achiam and S. Sastry. Surprise-Based Intrinsic Motivation for Deep Reinforcement Learning, Mar. 2017. URL http://arxiv.org/abs/1703.01732. arXiv:1703.01732 [cs].

F. Alet, M. F. Schneider, T. Lozano-Perez, and L. P. Kaelbling. Meta-learning curiosity algorithms. In *International Conference on Learning Representations*, 2020.

A. Aubret, L. Matignon, and S. Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019.

G. Baldassarre and M. Mirolli. *Intrinsically Motivated Learning Systems: An Overview*, pages 1–14. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-32375-1. doi: 10.1007/978-3-642-32375-1_1.

A. Barto, M. Mirolli, and G. Baldassarre. Novelty or Surprise? *Frontiers in Psychology*, 4, 2013. ISSN 1664-1078. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00907.

A. G. Barto. Intrinsic Motivation and Reinforcement Learning. In G. Baldassarre and M. Mirolli, editors, *Intrinsically Motivated Learning in Natural and Artificial Systems*, pages 17–47. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-32375-1. doi: 10.1007/978-3-642-32375-1_2. URL https://doi.org/10.1007/978-3-642-32375-1_2.

T. E. J. Behrens, T. H. Muller, J. C. R. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-Nelson. What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, 100(2):490–509, Oct. 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.10.002. URL https://www.sciencedirect.com/science/article/pii/S0896627318308560.

M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying Count-Based Exploration and Intrinsic Motivation, Nov. 2016. URL http://arxiv.org/abs/1606.01868. arXiv:1606.01868 [cs, stat].

S. W. Bench and H. C. Lench. On the Function of Boredom. *Behavioral Sciences*, 3(3):459–472, Sept. 2013. ISSN 2076-328X. doi: 10.3390/bs3030459. URL https://www.mdpi.com/2076-328X/3/3/459. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

S. W. Bench and H. C. Lench. Boredom as a seeking state: Boredom prompts the pursuit of novel (even negative) experiences. *Emotion*, 19(2):242–254, 2019. ISSN 1931-1516. doi: 10.1037/emo0000433. Place: US Publisher: American Psychological Association.

D. E. Berlyne. Curiosity and exploration. *Science*, 153(3731):25–33, 1966. doi: 10.1126/science.153.3731.25.

M. M. Botvinick, Y. Niv, and A. G. Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280, Dec. 2009. ISSN 0010-0277. doi: 10.1016/j.cognition.2008.08.011. URL https://www.sciencedirect.com/science/article/pii/S0010027708002059.

F. Brändle, L. J. Stocks, J. B. Tenenbaum, S. J. Gershman, and E. Schulz. Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*, 2023. doi: 10.1038/s41562-023-01661-2.

E. S. Bromberg-Martin, Y.-Y. Feng, T. Ogasawara, J. K. White, K. Zhang, and I. E. Monosov. A neural mechanism for conserved value computations integrating information and rewards. *Nature Neuroscience*, 27:159–175, 2024. doi: 10.1038/s41593-023-01511-4.

F. Brändle, L. J. Stocks, J. B. Tenenbaum, S. J. Gershman, and E. Schulz. Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*, 7(9):1481–1489, Sept. 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01661-2. URL https://www.nature.com/articles/s41562-023-01661-2. Number: 9 Publisher: Nature Publishing Group.

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

D. de Tinguy, T. Van de Maele, T. Verbelen, and B. Dhoedt. Spatial and Temporal Hierarchy for Autonomous Navigation Using Active Inference in Minigrid Environment. *Entropy*, 26(1):83, Jan. 2024. ISSN 1099-4300. doi: 10.3390/e26010083. URL https://www.mdpi.com/1099-4300/26/1/83. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

R. Dubey and T. L. Griffiths. Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, 127(3):455–476, 2019. doi: 10.1037/rev0000175.

R. Dubey and T. L. Griffiths. Understanding exploration in humans and machines by formalizing the function of curiosity. *Current Opinion in Behavioral Sciences*, 35:118–124, 2020. doi: 10.1016/j.cobeha.2020.07.008.

B. Fasolo, R. Hertwig, M. Huber, and M. Ludwig. Size, entropy, and density: What is the difference that makes the difference between small and large real-world assortments? *Psychology & Marketing*, 26(3):254–279, 2009. ISSN 1520-6793. doi: 10.1002/mar.20272. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mar.20272. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.20272.

L. FitzGibbon, J. K. L. Lau, and K. Murayama. The seductive lure of curiosity: information as a motivationally salient reward. *Current Opinion in Behavioral Sciences*, 35:21–27, 2020. ISSN 2352-1546. doi: 10.1016/j.cobeha.2020.05.014.

S. Forss, A. Ciria, F. Clark, C.-l. Galusca, D. Harrison, and S. Lee. A transdisciplinary view on curiosity beyond linguistic humans: animals, infants, and artificial intelligence. *Biological Reviews*, 99(3):979–998, 2024. doi: 10.1111/brv.13054.

P. J. Geiwitz. Structure of boredom. *Journal of Personality and Social Psychology*, 3(5):592–600, 1966. ISSN 1939-1315. doi: 10.1037/h0023202. Place: US Publisher: American Psychological Association.

S. J. Gershman and Y. Niv. Novelty and inductive generalization in human reinforcement learning. *Topics in cognitive science*, 7(3):391–415, 2015. doi: 10.1111/tops.12138.

A. Ghazizadeh, W. Griggs, and O. Hikosaka. Ecological origins of object salience: Reward, uncertainty, aversiveness, and novelty. *Frontiers in Neuroscience*, 10:378, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00378.

J. Gottlieb and P.-Y. Oudeyer. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19:758–770, 2018. doi: 10.1038/s41583-018-0078-0.

L. Horvath, S. Colcombe, M. Milham, S. Ray, P. Schwartenbeck, and D. Ostwald. Human belief state-based exploration and exploitation in an information-selective symmetric reversal bandit task. *Computational Brain & Behavior*, 2021. doi: 10.1007/s42113-021-00112-3.

L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, June 2009. ISSN 0042-6989. doi: 10.1016/j.visres.2008.09.007. URL https://www.sciencedirect.com/science/article/pii/S0042698908004380.

T. B. Kashdan, M. W. Gallagher, P. J. Silvia, B. P. Winterstein, W. E. Breen, D. Terhar, and M. F. Steger. The curiosity and exploration inventory-II: Development, factor structure, and psychometrics. *Journal of Research in Personality*, 43(6):987–998, Dec. 2009. ISSN 0092-6566. doi: 10.1016/j.jrp.2009.04.011. URL https://www.sciencedirect.com/science/article/pii/S0092656609001275.

C. Kidd and B. Y. Hayden. The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460, 2015. doi: 10.1016/j.neuron.2015.09.010.

A. Klyubin, D. Polani, and C. Nehaniv. Empowerment: a universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135 Vol.1, 2005. doi: 10.1109/CEC.2005.1554676.

K. Kobayashi, S. Ravaioli, A. Baranès, M. Woodford, and J. Gottlieb. Diverse motives for human curiosity. *Nature Human Behaviour*, 3:587–595, 2019. doi: 10.1038/s41562-019-0589-3.

A. Kolossa, B. Kopp, and T. Fingscheidt. A computational analysis of the neural bases of Bayesian inference. *NeuroImage*, 106:222–237, 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2014.11.007.

E. Kosoy, J. Collins, D. M. Chan, S. Huang, D. Pathak, P. Agrawal, J. Canny, A. Gopnik, and J. B. Hamrick. Exploring Exploration: Comparing Children with RL Agents in Unified Environments, July 2020. URL http://arxiv.org/abs/2005.02880. arXiv:2005.02880 [cs].

S. Kullback. *Information Theory and Statistics*. Courier Corporation, July 1997. ISBN 978-0-486-69684-3. Google-Books-ID: luHcCgAAQBAJ.

P. Ladosz, L. Weng, M. Kim, and H. Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022. ISSN 1566-2535. doi: 10.1016/j.inffus.2022.03.003.

J. K. L. Lau, H. Ozono, K. Kuratomi, A. Komiya, and K. Murayama. Shared striatal activity in decisions to satisfy curiosity and hunger at the risk of electric shocks. *Nature Human Behaviour*, 4(5):531–543, 2020. doi: 10.1038/s41562-020-0848-3.

V. Liakoni, A. Modirshanechi, W. Gerstner, and J. Brea. Learning in volatile environments with the Bayes factor surprise. *Neural Computation*, 33(2):1–72, 2021. doi: 10.1162/neco_a_01352.

V. Liakoni, M. P. Lehmann, A. Modirshanechi, J. Brea, A. Lutti, W. Gerstner, and K. Preuschoff. Brain signals of a surprise-actor-critic model: Evidence for multiple learning modules in human decision making. *NeuroImage*, 246:118780, 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2021.118780.

B. Matusch, J. Ba, and D. Hafner. Evaluating agents without rewards. *arXiv preprint arXiv:2012.11538*, 2020.

B. Matusch, J. Ba, and D. Hafner. Evaluating Agents without Rewards, Feb. 2021. URL http://arxiv.org/abs/2012.11538. arXiv:2012.11538 [cs].

K. Mehlhorn, B. R. Newell, P. M. Todd, M. D. Lee, K. Morgan, V. A. Braithwaite, D. Hausmann, K. Fiedler, and C. Gonzalez. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2(3):191–215, 2015. ISSN 2325-9973. doi: 10.1037/dec0000033. Place: US Publisher: Educational Publishing Foundation.

R. Mendonca, O. Rybkin, K. Daniilidis, D. Hafner, and D. Pathak. Discovering and achieving goals via world models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24379–24391. Curran Associates, Inc., 2021.

F. Meyniel, M. Maheu, and S. Dehaene. Human inferences about sequences: A minimal transition probability model. *PLOS Computational Biology*, 12:1–26, 2016. doi: 10.1371/journal.pcbi.1005260.

A. Modirshanechi, S. Becker, J. Brea, and W. Gerstner. Surprise and novelty in the brain. *Current Opinion in Neurobiology*, 82:102758, 2023a. ISSN 0959-4388. doi: 10.1016/j.conb.2023.102758.

A. Modirshanechi, K. Kondrakiewicz, W. Gerstner, and S. Haesler. Curiosity-driven exploration: foundations in neuroscience and computational modeling. *Trends in Neurosciences*, 46(12):1054–1066, 2023b. ISSN 0166-2236. doi: 10.1016/j.tins.2023.10.002.

A. Modirshanechi, K. Kondrakiewicz, W. Gerstner, and S. Haesler. Curiosity-driven exploration: foundations in neuroscience and computational modeling. *Trends in Neurosciences*, 46(12):1054–1066, Dec. 2023c. ISSN 0166-2236, 1878-108X. doi: 10.1016/j.tins.2023.10.002. URL https://www.cell.com/trends/neurosciences/abstract/S0166-2236(23)00240-0. Publisher: Elsevier.

A. Modirshanechi, W.-H. Lin, H. A. Xu, M. H. Herzog, and W. Gerstner. The curse of optimism: a persistent distraction by novelty. *bioRxiv*, 2023d. doi: 10.1101/2022.07.05.498835.

I. E. Monosov. Curiosity: primate neural circuits for novelty and information seeking. *Nature Reviews Neuroscience*, (25):195–208, 2024. doi: 10.1038/s41583-023-00784-9.

K. C. Montgomery. The role of the exploratory drive in learning. *Journal of Comparative and Physiological Psychology*, 47(1):60–64, 1954. doi: 10.1037/h0054833.

A. W. Moore and C. G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13(1):103–130, Oct. 1993. ISSN 1573-0565. doi: 10.1007/BF00993104. URL https://doi.org/10.1007/BF00993104.

R. Moreno-Bote and J. Ramirez-Ruiz. Empowerment, Free Energy Principle and Maximum Occupancy Principle Compared. Nov. 2023. URL https://openreview.net/forum?id=0cHrsQox0Z.

J. Morrens, Çağatay Aydin, A. Janse van Rensburg, J. Esquivelzeta Rabell, and S. Haesler. Cue-evoked dopamine promotes conditioned responding during learning. *Neuron*, 106(1):142–153.e7, 2020. ISSN 0896-6273. doi: 10.1016/j.neuron.2020.01.012.

K. Murayama. A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic–extrinsic rewards. *Psychological Review*, 129(1):175–198, 2022. doi: 10.1037/rev0000349.

K. Murayama, L. FitzGibbon, and M. Sakaki. Process account of curiosity and interest: A reward-learning perspective. *Educational Psychology Review*, pages 1–21, 2019.

M. R. Nassar, R. C. Wilson, B. Heasly, and J. I. Gold. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, 30(37): 12366–12378, 2010. doi: 10.1523/JNEUROSCI.0822-10.2010.

A. Nedergaard and M. Cook. k-Means Maximum Entropy Exploration, Nov. 2023. URL http://arxiv.org/abs/2205.15623. arXiv:2205.15623 [cs].

J. D. Nelson. Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 112(4):979–999, 2005. doi: 10.1037/0033-295X.112.4.979.

T. Ogasawara, F. Sogukpinar, K. Zhang, Y.-Y. Feng, J. Pai, A. Jezzini, and I. E. Monosov. A primate temporal cortex–zona incerta pathway for novelty seeking. *Nature Neuroscience*, 25, 2022. doi: 10.1038/s41593-021-00950-1.

G. Ostrovski, M. G. Bellemare, A. v. d. Oord, and R. Munos. Count-Based Exploration with Neural Density Models, June 2017. URL http://arxiv.org/abs/1703.01310. arXiv:1703.01310 [cs].

P.-Y. Oudeyer. Computational theories of curiosity-driven learning. *arXiv preprint arXiv:1802.10546*, 2018.

P.-Y. Oudeyer and F. Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1:6, 2009. doi: 10.3389/neuro.12.006.2007.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2778–2787. JMLR.org, 2017.

P. Piray and N. D. Daw. Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature Communications*, 12(1):4942, Aug. 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25123-3. URL https://www.nature.com/articles/s41467-021-25123-3. Publisher: Nature Publishing Group.

W. Pisula. *Curiosity and Information Seeking in Animal and Human Behavior*. Jan. 2009.

F. Poli, M. Meyer, R. B. Mars, and S. Hunnius. Contributions of expected learning progress and perceptual novelty to curiosity-driven exploration. *Cognition*, 225:105119, 2022. ISSN 0010-0277. doi: 10.1016/j.cognition.2022.105119.

F. Poli, J. X. O'Reilly, R. B. Mars, and S. Hunnius. Curiosity and the dynamics of optimal exploration. *Trends in Cognitive Sciences*, 28(5):441–453, 2024. doi: 10.1016/j.tics.2024.02.001.

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

J. Ramírez-Ruiz, D. Grytskyy, C. Mastrogiuseppe, Y. Habib, and R. Moreno-Bote. Complex behavior from intrinsic motivation to occupy action-state path space, Feb. 2024. URL http://arxiv.org/abs/2205.10316. arXiv:2205.10316 [cs, q-bio].

C. Salge, C. Glackin, and D. Polani. Empowerment – an Introduction, Oct. 2013. URL http://arxiv.org/abs/1310.1863. arXiv:1310.1863 [nlin].

B. Scheibehenne, R. Greifeneder, and P. M. Todd. Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload. *Journal of Consumer Research*, 37(3):409–425, Oct. 2010. ISSN 0093-5301. doi: 10.1086/651235. URL https://doi.org/10.1086/651235.

J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. doi: 10.1109/TAMD.2010.2056368.

F. F. Schmitt and R. Lahroodi. The Epistemic Value of Curiosity. 58(2):125–148, 2008. ISSN 00132004. URL https://www.proquest.com/docview/214139535/abstract/10AAF3F6C1DE407DPQ/1. Num Pages: 24 Place: Urbana, United Kingdom Publisher: Blackwell Publishing Ltd.

R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. Planning to explore via self-supervised world models. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8583–8592. PMLR, 2020.

S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82, 2010a. doi: 10.1109/TAMD.2010.2051031.

S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82, June 2010b. ISSN 1943-0612. doi: 10.1109/TAMD.2010.2051031. URL https://ieeexplore.ieee.org/document/5471106. Conference Name: IEEE Transactions on Autonomous Mental Development.

A. E. Stahl and L. Feigenson. Observing the unexpected enhances infants' learning and exploration. *Science*, 348(6230):91–94, 2015. doi: 10.1126/science.aaa3799.

J. Storck, S. Hochreiter, and J. Schmidhuber. Reinforcement Driven Information Acquisition In Non-Deterministic Environments. *ICANN'95*, 2, Jan. 1995.

R. S. Sutton and A. G. Barto. *Reinforcement Learning, second edition: An Introduction*. MIT Press, Nov. 2018. ISBN 978-0-262-35270-3. Google-Books-ID: uWV0DwAAQBAJ.

E. Szumowska and A. W. Kruglanski. Curiosity as end and means. *Current Opinion in Behavioral Sciences*, 35:35–39, Oct. 2020. ISSN 2352-1546. doi: 10.1016/j.cobeha.2020.06.008. URL https://www.sciencedirect.com/science/article/pii/S2352154620300966.

A. Ten, P. Kaushik, P.-Y. Oudeyer, and J. Gottlieb. Humans monitor learning progress in curiosity-driven exploration. *Nature Communications*, 12:5972, 2021a. doi: 10.1038/s41467-021-26196-w.

A. Ten, P. Kaushik, P.-Y. Oudeyer, and J. Gottlieb. Humans monitor learning progress in curiosity-driven exploration. *Nature Communications*, 12(1):5972, Oct. 2021b. ISSN 2041-1723. doi: 10.1038/s41467-021-26196-w. URL https://www.nature.com/articles/s41467-021-26196-w. Number: 1 Publisher: Nature Publishing Group.

P.-A. L. Tolguenec, Y. Besse, F. Teichteil-Konigsbuch, D. G. Wilson, and E. Rachelson. Exploration by Learning Diverse Skills through Successor State Measures, June 2024. URL http://arxiv.org/abs/2406.10127. arXiv:2406.10127 [cs].

E. C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. ISSN 1939-1471. doi: 10.1037/h0061626. Place: US Publisher: American Psychological Association.

V. Viswanathan, M. Lees, and P. M. A. Sloot. The influence of memory on indoor environment exploration: A numerical study. *Behavior Research Methods*, 48(2):621–639, June 2016. ISSN 1554-3528. doi: 10.3758/s13428-015-0604-1. URL https://doi.org/10.3758/s13428-015-0604-1.

H.-G. Voss and H. Keller. *Curiosity and Exploration: Theories and Results*. Elsevier, Oct. 2013. ISBN 978-1-4832-6307-6. Google-Books-ID: hXiLBQAAQBAJ.

H. A. Xu, A. Modirshanechi, M. P. Lehmann, W. Gerstner, and M. H. Herzog. Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLOS Computational Biology*, 17(6):e1009070, June 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009070. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009070. Publisher: Public Library of Science.

C. Yu, N. Burgess, M. Sahani, and S. Gershman. Successor-Predecessor Intrinsic Exploration, Sept. 2023. URL http://arxiv.org/abs/2305.15277. arXiv:2305.15277 [cs].

C. Yu, N. Burgess, M. Sahani, and S. J. Gershman. Successor-Predecessor Intrinsic Exploration, Jan. 2024. URL http://arxiv.org/abs/2305.15277. arXiv:2305.15277 [cs].

Z. Zheng, J. Oh, M. Hessel, Z. Xu, M. Kroiss, H. V. Hasselt, D. Silver, and S. Singh. What Can Learned Intrinsic Rewards Capture? In *Proceedings of the 37th International Conference on Machine Learning*, pages 11436–11446. PMLR, Nov. 2020a. URL https://proceedings.mlr.press/v119/zheng20b.html. ISSN: 2640-3498.

Z. Zheng, J. Oh, M. Hessel, Z. Xu, M. Kroiss, H. Van Hasselt, D. Silver, and S. Singh. What can learned intrinsic rewards capture? In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11436–11446. PMLR, 2020b.

Supplementary Material

Environment generation

All the parameters used for generating environments are described in Table 1. The environments are generated in three steps:

1. Maze generation: a maze is generated with a given number of states and branching rate. The branching rate determines the number of intersections in the environment. The algorithm for generating the maze is defined in Algorithm 1.

2. Room integration: some states in the maze are transformed into rooms. A room is a square grid, with each state having four actions to navigate up, down, left or right whenever these actions are available (when the state is not on a border). Neighbors of a transformed state are connected to the middle of the room borders (maximum 4 neighbors, one for each side of the square room). Parameters determine the fraction of states that are transformed into rooms and the size of the rooms.

3. Room properties: Each room is assigned one of sink, source, stochastic or neutral. For each sink room, we iteratively sample a state u in the room and a state v outside the room uniformly at random, and connect v to u. We repeat until the desired number of edges has been added. For each source room, we do the same process but inverse the direction of connections. The transition dynamics inside stochastic rooms are altered as follows: when an agent selects an action a from a state s within a the room, there is a fixed probability that the action will result in the agent moving to a random neighbor of s in the room instead of the intended destination of a. Finally, neutral rooms do not receive any modification.

Algorithm 1 Algorithm to generate the initial maze

```
Require: n > 0, branch_rate ∈ [0, 1]
  Q ← empty queue
  ENQUEUE(Q, 1)
  next_state ← 2
  while next_state ≤ n do
      cur_state ← DEQUEUE(Q)
      CONNECT(cur_state, next_state)
      CONNECT(next_state, cur_state)
      rand ∈ [0, 1] uniformly at random
      if rand < branch_rate and n_neighbors(cur_state) < 4 then
          ENQUEUE(Q, cur_state)    ▷ The current state is put back in the Queue if it does not
                                     already have 4 neighbors
      end if
      ENQUEUE(Q, next_state)
      next_state += 1
  end while
```

Parameter	Range	Short description	Environment types					
			Neutral	Sink	Source	Stochastic	Mixed	Trap (Fig. 3)
n_s	$[1, \infty]$	Number of states in the initial maze.	40	40	40	40	40	97
branch rate	[0,1]	Probability of creating a new intersection when adding a state.	0.2	0.2	0.2	0.2	0.2	$0 \to 1$
n_{room}	$[0, n_s]$	Number of rooms.	4	4	4	4	4	1
room size	$[1, \infty]$	Size of the side of rooms.	4	4	4	4	4	2
p_{sink}	[0,1]	Fraction of sink rooms.	0	0.25	0	0	0.25	1
p_{source}	$[0, 1 - p_{\text{sink}}]$	Fraction of source rooms.	0	0	0.25	0	0.25	0
$p_{\text{stochastic}}$	$[0, 1 - p_{\text{sink}} - p_{\text{source}}]$	Fraction of stochastic rooms.	0	0	0	0.25	0.25	0
$n_{\text{edges per sink}}$	$[0, \infty]$	Number of additional connection leading to each sink room.	0	50	0	0	50	$0 \to 200$
$n_{\text{edges per source}}$	$[0, \infty]$	Number of additional connection originating from each source room.	0	0	50	0	50	0
uncontrollability	[0,1]	Probability for an action taken in a stochastic room to lead to a random neighbor instead of the expected destination.	0	0	0	1	1	0

Table 1: Summary of all environment parameters used in the generation process. The right side shows the environment types considered with the corresponding parameter values.

Robustness of results

Robustness to change of metrics

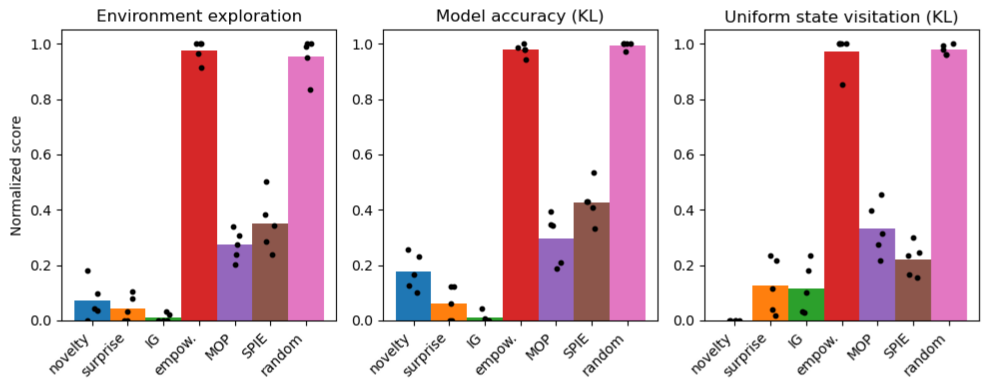
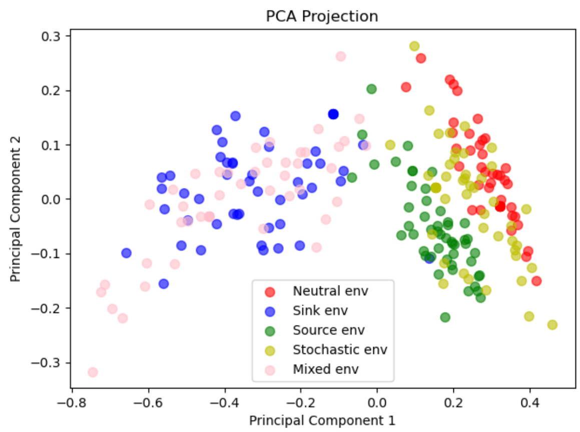
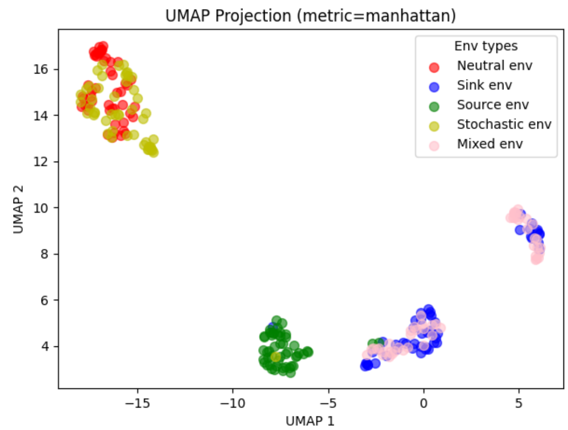


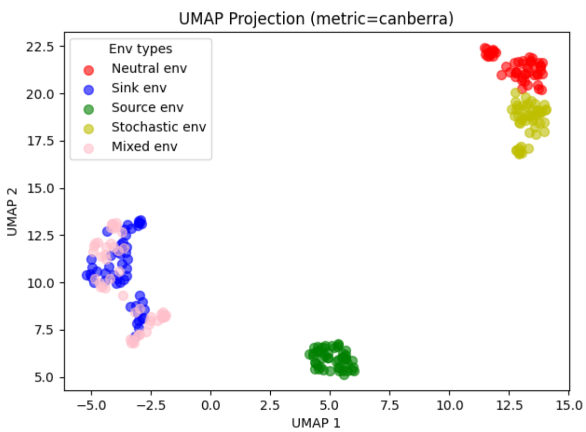
Figure 8: Average normalized score across environments for each intrinsic motivation, computed as in Fig. 4, but using the KL divergence instead of RMSE for measure 2 and 3. The results are very similar and the same conclusions can be drawn.



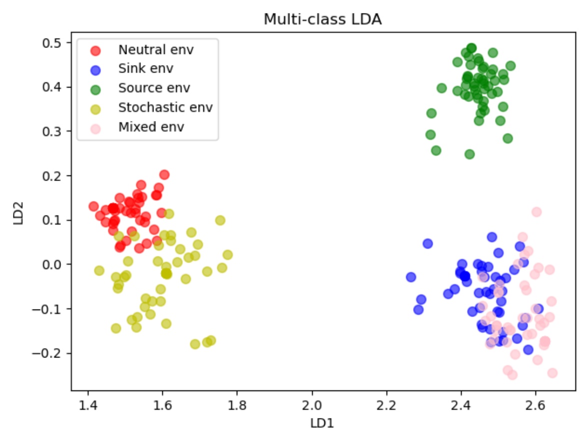
(a) PCA



(b) UMAP with Manhattan distance



(c) UMAP with Canberra distance



(d) Multi-class Linear Discriminant Analysis (LDA)

Figure 7: Consistency of performance within each environment type. The environment types are described in Environment types. Various projections of performance vectors for each environment are shown. Each dot corresponds to one environment sampled from one of the given types. For each such sample, a vector of performance is created as follows: we run each intrinsic motivation for 2000 steps and calculate the Area Under the Curve for each performance measure (same curve as in Fig. 2). For each environment, we obtain a performance vector of size $(n_{\mathrm{IM}} \cdot n_{\mathrm{measures}}) = (6 \cdot 3)$ where n_{IM} is the number of intrinsic motivations and n_{measures} is the number of measures. (a) We apply PCA and display the top two principal components. (b)-(c) We use UMAP with Manhattan and Canberra distances. (d) We apply multi-class LDA. Clusters are observed in each method. Sink and Mixed environments consistently overlap, probably due to the presence of sink rooms in both cases. Neutral and Stochastic environments sre also close, but remain distinguishable in (c) and (d). This similarity is probably due to the fact that a stochastic room doesn't change the environment dynamics as much as sink and source rooms.