OPEN MIND

Discoveries in Cognitive Science

an open access journal

Citation: Gruaz, L., Modirshanechi, A., Becker, S., & Brea, J. (2025). Merits of Curiosity: A Simulation Study. *Open Mind: Discoveries in Cognitive Science*, *9*, 1037–1065. https://doi.org/10.1162/opmi.a.9

DOI:
https://doi.org/10.1162/opmi.a.9

Supplemental Materials:
https://doi.org/10.1162/opmi.a.9

Received: 15 November 2024
Accepted: 27 May 2025

Competing Interests: The authors declare no conflict of interests.

Corresponding Authors:
Lucas Gruaz
lucas.gruaz@epfl.ch
Alireza Modirshanechi
alireza.modirshanechi@helmholtz-munich.de

The MIT Press

REPORT

Merits of Curiosity: A Simulation Study

Lucas Gruaz[1,2*], Alireza Modirshanechi[1,2,3,4*], Sophia Becker[1,2], and 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
*These authors contributed equally to this work.

Keywords: curiosity, RL, reinforcement learning, algorithm, computational, empowerment, environment, exploration, generation, information gain, MOP, neuroscience, novelty, SPIE, structure, surprise

ABSTRACT

'Why are we curious?' has been among the central puzzles of neuroscience and psychology in the past decades. A popular hypothesis is that curiosity is driven by *intrinsically generated reward signals*, which have evolved to support survival in *complex environments*. To formalize and test this hypothesis, we need to understand the enigmatic 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. First, we propose an algorithm to generate environments that capture key abstract features of different real-world situations. Then, we simulate artificial agents that explore these environments by seeking one of six representative intrinsic rewards: novelty, surprise, information gain, empowerment, maximum occupancy principle, and successor-predecessor intrinsic exploration. We evaluate the exploration performance of these simulated agents regarding three potential objectives of curiosity: state discovery, model accuracy, and uniform state visitation. Our results show that the comparative performance of each intrinsic reward is highly dependent on the environmental features and the curiosity objective; this indicates that 'optimality' in top-down theories of curiosity needs a precise formulation of assumptions. Nevertheless, we found that agents seeking a combination of novelty and information gain always achieve a close-to-optimal performance on objectives of curiosity as well as in collecting extrinsic rewards. This suggests that novelty and information gain are two principal axes of curiosity-driven behavior. These results pave the way for the further development of computational models of curiosity and the 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 & Oudeyer, 2018; Gruber & Ranganath, 2019; Kidd & Hayden, 2015; Modirshanechi, Kondrakiewicz, et al., 2023): Human babies prefer playing with toys that have surprising features (e.g., a car that seemingly passes through a solid wall) over normal toys (Stahl & 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). Strikingly, the drive of curiosity can occasionally even override 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 *information* about *future* rewards (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 debated (see Jirout et al. (2024); Modirshanechi, Kondrakiewicz, et al. (2023); 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 exactly are they curious about?' (Modirshanechi, Kondrakiewicz, et al., 2023). Modern theoretical attempts to address these questions use the intrinsically motivated Reinforcement Learning (RL) framework (Baldassarre & Mirolli, 2013; Barto, 2013) and describe curiosity-driven actions as those directed towards seeking an *intrinsically* generated 'reward' signal (Modirshanechi, Kondrakiewicz, et al., 2023; Murayama, 2022; Murayama et al., 2019; Oudeyer, 2018; Poli et al., 2024). In this framework, the answer to the 'What' question of curiosity 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., 2023; Ladosz et al., 2022; Oudeyer & Kaplan, 2007). Given an intrinsic reward signal, the 'Why' question of curiosity is often answered by quantifying how much the intrinsically motivated actions improve the agent's ability to, e.g., find valuable sources of extrinsic reward (Gershman & Niv, 2015; Pathak et al., 2017; Singh, Lewis, Barto, & Sorg, 2010), gain knowledge about the environment structure (Dubey & Griffiths, 2020a), or learn complex skills without supervision (Mendonca et al., 2021; Oudeyer & Kaplan, 2007; Sekar et al., 2020).

In several experiments, intrinsically motivated RL algorithms have successfully described curiosity-driven behavior of human participants by considering novelty (Modirshanechi et al., 2025; Xu et al., 2021), surprise (Kobayashi et al., 2019), information gain (Ghilardi et al., 2024; Horvath et al., 2021; Nelson, 2005), progress rate (Poli et al., 2022, 2025; Ten et al., 2021), or empowerment (Brändle et al., 2023) as the intrinsic reward signal. It may seem paradoxical that different experimental studies provide different answers to the 'What' question of curiosity. However, there may be a simple explanation: 'top-down' curiosity models identify the intrinsic motivational signal that best satisfies the objective of curiosity (the 'Why' of curiosity) in the class of environments where the curious agent lives (Alet et al., 2020; Dubey & Griffiths, 2020a; Modirshanechi, Kondrakiewicz, et al., 2023; Singh, Lewis, Barto, & Sorg, 2010; Zheng et al., 2020); hence, the 'What' of curiosity can be experiment-dependent, because the optimal strategies to reach the curiosity objective can differ across experiments (Dubey & Griffiths, 2020a, 2020b). To advance our theoretical understanding of curiosity, it is therefore 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 (Figure 1A). To this end, we first design an algorithm for generating various environments with different key characteristics, e.g., number of states, stochasticity of transitions, distribution of between-state connections, etc. We

[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.'

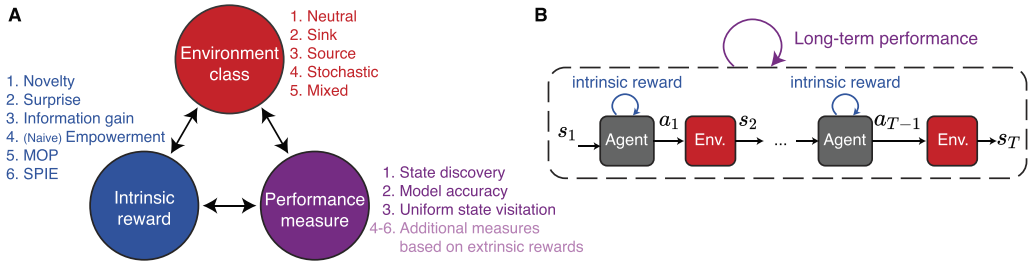

Figure 1. Environment-Reward-Performance triangle. A. Top-down theories hypothesize that curiosity is the optimal solution for survival in complex environments with sparse extrinsic rewards. To precisely formalize and test these theories, we need to understand the relationship between *intrinsic rewards* (as drives of curiosity; see Intrinsic rewards), (ii) potential measures of *performance* (to quantify optimality; see Performance measures and Performance on curiosity measures correlates with collection of extrinsic rewards), and (iii) *environment* structures (see Environment generation). **B.** Intrinsic rewards guide the agent's immediate and local behavior throughout its life, interacting with its environment. The objective of curiosity is then evaluated as a long-term performance measure based on the agent's sequence of states and actions. Our goal is to study how the long-term performance depends on the locally computed intrinsic rewards, given different environment structures.

then formally define three performance measures as potential objectives of curiosity: (i) how fast a curious agent discovers all states of its environment ('state discovery'), (ii) how accurately it learns the structure of the environment ('model accuracy'), and (iii) how uniformly it explores all the states ('uniform state visitation'). We then simulate different curious agents and quantify the merits of six representative intrinsic rewards: novelty (Modirshanechi et al., 2025; Xu et al., 2021), surprise (Kobayashi et al., 2019), information gain (Horvath et al., 2021; Nelson, 2005), empowerment (Brändle et al., 2023; Klyubin et al., 2005), maximum occupancy principle (MOP) (Ramírez-Ruiz et al., 2024), and successor-predecessor intrinsic exploration (SPIE) (Yu et al., 2024).

We show that, with few exceptions, seeking information gain is the best strategy for state discovery and model accuracy, whereas seeking novelty is the best strategy for uniform state visitation. Building upon this observation, we show that an agent seeking a combination of information gain and novelty can reach a close-to-optimal performance across all three performance measures and in all classes of environments. Additionally, we show that the same combination of information gain and novelty enables agents to achieve excellent performance in collecting *extrinsic* reward under varying conditions. Importantly, however, our results also reveal that the relative performance of different intrinsic rewards is highly dependent on the structure of the environment. We show that these environment-dependent differences can be leveraged by our environment-generating algorithm to design experiments in which seeking different intrinsic rewards leads to maximally different exploration strategies. These experiments can be used in future studies of curiosity with humans and animals.

RESULTS

To study the behavior of curious agents, we use the intrinsically motivated RL framework (Baldassarre & Mirolli, 2013; Barto, 2013). In this framework, each curious agent learns to move through an environment with 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 the agent's 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, i.e., the agents learn to take actions that maximize their future external rewards (Sutton & Barto, 2018). In contrast, intrinsically motivated RL uses internal reward signals that evolve based on the agent's experience (Barto, 2013; Singh, Lewis,

Barto, & Sorg, 2010). These intrinsic rewards encourage the agent to explore and learn from the environment without relying on external rewards or supervision.

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 consider tabular RL (Sutton & Barto, 2018) and assume that the agent uses Bayesian inference (similar to Liakoni et al. (2022); Meyniel et al. (2016); Xu et al. (2021)) to estimate transition probabilities $P(s' \mid s,\ a)$ of reaching state s' from state s by taking action a. The estimates are given by

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

where N_s denotes the number of states, $C^{(t)}_{s,a}$ is the number of times action a has been taken in state s up to time t, and $C^{(t)}_{s,a,s'}$ is the count of transitions where the agent moved to state s' by taking action a in state s. The factor $\frac{1}{N_s}$ acts as an uninformative prior that (i) assigns uniform transition probabilities to action-state pairs that have never been experienced and (ii) assumes that this prior is worth one observation (Efron & Hastie, 2016). Using its model of the environment (i.e., the estimated transition probabilities in Equation 1), the agent computes Q-values $\hat{Q}(s,a)$ as an estimate of the expected future intrinsic rewards that it can collect after taking action a in 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); see Methods)

$$\hat{Q}^{(t)}(s,a) = \sum_{s'} \hat{P}^{(t)}(s'|s,a)\left(R^{(t)}(s,a,s') + \gamma \max_{a'} \hat{Q}^{(t)}(s',a')\right)\ , \tag{2}$$

where $R^{(t)}(s,\ a,\ s')$ is the intrinsic reward for transitioning from s to s' via action a (see Intrinsic rewards), and $\gamma \in [0,1)$ represents the discount factor for the Q-values. The discount factor γ determines how much the agent values future rewards compared to the immediate rewards.

At each time t, the agent's behavior in state s is described by an action policy that assigns the 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:

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

where β is the softmax inverse temperature and controls the stochasticity in decision-making (Sutton & Barto, 2018). This implies that the agent will strongly favor one action if the action 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). The parameter β determines how strongly the agent's action probabilities are influenced by differences in Q-values: a higher β makes the agent more likely to select the action with the highest Q-value, while a lower β results in more evenly distributed probabilities across actions.

Our goal is to identify the environment-dependent benefits of seeking different types of intrinsic reward $R^{(t)}(s,\ a,\ s')$ for achieving different objectives of curiosity, where curiosity objectives are defined as the long-term performance of curious agents (Figure 1).

Intrinsic Rewards

We consider six types of intrinsic reward (Figure 1A), each defined by a reward function $R^{(t)}(s, a, s')$ that determines the Q-values (Equation 2) and, accordingly, specifies the agent's action-policy (Equation 3): Novelty (Modirshanechi et al., 2025; Xu et al., 2021), Surprise (Kobayashi et al., 2019), Information gain (Horvath et al., 2021; Nelson, 2005), Empowerment (Brändle et al., 2023; Klyubin et al., 2005), 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 reward; see Intrinsic rewards detailed for more detailed formulation and further theoretical analyses.

(i) **Novelty** rewards the agent for exploring rarely encountered states. Specifically, for a transition to state 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 higher the reward of 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 to state s' after taking action a in state s, the agent receives a reward that is a decreasing function of $\hat{P}^{(t)}(s'|s,a)$ (Equation 1), i.e., the less the agent expects to visit s' (conditioned on s and a), the higher the reward of visiting s' (after taking a in s).

(iii) **Information gain** rewards the agent for reducing its (epistemic) uncertainty about the environment's transition probabilities by acquiring new information. For a transition to state s' after taking action a in state s, the reward is determined by the size of the update of the agent's model of the environment, i.e., the more the agent updates its estimated probabilities (Equation 1) after the transition, the higher the reward. The amount of update is quantified using the KL divergence between the updated model and the previous model.

(iv) **(Naive) Empowerment** rewards the agent for visiting states where available actions have a *diverse* set of *predictable* outcomes. For a transition to state s' after taking action a in state s, the reward is equal to the 'naive empowerment' value of s' (defined in Intrinsic rewards detailed), i.e., the more 'options' the agent has in state s', the higher the reward of visiting s'. We use the term 'naive empowerment' instead of 'empowerment' to emphasize the difference between our definition, which relies on the *estimated* transition dynamics $\hat{P}^{(t)}(s'|s,a)$, and prior works on empowerment, which assumed that the true transition probabilities $P(s'|s, a)$ are known to the agent (Klyubin et al., 2005; Salge et al., 2014; Volpi & Polani, 2020); see Discussion for further links to related works.

(v) **MOP** can be seen as a regularized surprise that rewards the agent for experiencing unlikely transitions and, simultaneously, 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. For a transition to state s' after taking action a in state s, the reward 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 rewards detailed.

(vi) **SPIE** rewards the agent for visiting rare states as well as those that are critical for reaching isolated regions. Specifically, for a transition to state s' after taking action a in state s, the reward 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, for example, if s' is isolated or if s is a bottleneck state.

Performance Measures

While intrinsic rewards guide the agent's immediate and local behavior, the objective of curiosity in the top-down theories can be evaluated only after a series of actions and across the whole environment (Figure 1B; see also Dubey et al. (2022); Hadfield-Menell et al. (2020); Singh, Lewis, and Barto (2010); Sorg et al. (2010) for similar discussions). How seeking different intrinsic rewards (the 'What' of curiosity) helps agents achieve the curiosity objectives (the 'Why' of curiosity) has remained unclear. To quantitatively answer this question, we define three performance measures that capture the potential long-term goals of a curious agent. Inspired by the existing curiosity literature, our definitions formalize common intuitions about the purpose of curiosity:

Measure 1: State discovery. Curiosity is closely linked to exploration and discovering otherwise unknown states (Kashdan et al., 2009; Modirshanechi, Kondrakiewicz, et al., 2023; Voss & Keller, 2013). Hence, one key goal of a curious agent can be to reach and visit all states in an environment. We measure the performance of an agent, with respect to this goal, by the fraction of unvisited states after a certain number of steps (T in Figure 1B). The lower this fraction, the more successful the agent.

Measure 2: Model accuracy. Curiosity is often associated with gaining knowledge (Schmitt & Lahroodi, 2008; Szumowska & 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 setup, the internal model refers to the agent's estimate $\hat{P}$ of the transition probabilities (Equation 1), which should closely approximate the true transition probabilities P. We measure the performance of an agent, with respect to this goal, as the difference between the estimated transition probabilities $\hat{P}$ and the ground truth P after a certain number of steps (T in Figure 1B), using the Root Mean Squared Error (RMSE). The lower this difference, the more successful the agent.

Measure 3: Uniform state visitation. It has been hypothesized that curiosity facilitates finding valuable sources of 'extrinsic' rewards (Bellemare et al., 2016; Modirshanechi, Kondrakiewicz, et al., 2023; Pathak et al., 2017). However, since the world is inherently changing (Liakoni et al., 2021; Nassar et al., 2010; Piray & Daw, 2021a), the successful discovery of sources of rewards requires persistent exploration that enables a balanced and frequent visitation of all states. Our first two measures do not account for such persistence as they focus on early-stage exploratory behavior. For instance, if an agent initially explores the entire environment but then remains confined to a small region, it performs perfectly according to state discovery and might also have a good model accuracy; however, the agent will inevitably miss out on any valuable information that may appear later in other parts of the environment. To avoid such a disproportionate visit of certain regions, another main goal of a curious agent can be to achieve an even distribution of visits across the individual states (similarly to Nedergaard & Cook (2023); Tolguenec et al. (2024)). This is mathematically equivalent to minimizing the expected state re-visitation time so that the agent never leaves a state unattended for too long (see Equivalence of uniform state visitation and minimal state re-visitation time). We measure the performance of an agent, with respect to this goal, as the difference between the agent's state visitation frequency and the uniform distribution (using RMSE) after a certain number of steps. The lower this difference, the more successful the agent.

There are theoretical connections in asymptotic regimes between certain intrinsic rewards and performance measures, notably between novelty and uniform state visitation and between

information gain and model accuracy (see Intrinsic rewards detailed for an analysis of the asymptotic behavior of each intrinsic reward). However, under realistic constraints, including incomplete knowledge and temporal discounting, the relationship between intrinsic rewards and performance measures is less clear. In our experiments, we examine how each intrinsic reward performs with such constraints across all metrics and environmental conditions. This will provide us with insights beyond what theoretical asymptotic relationships suggest.

Environment Generation

To systematically study the link between intrinsic rewards and curiosity objectives, we need a procedure to generate diverse environments with realistic features (Figure 1). In curiosity research, experimental paradigms are typically task-specific and hand-crafted. Importantly, most experimental paradigms lack standardized multistep environments that resemble real-world situations (Modirshanechi, Kondrakiewicz, et al., 2023). 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 experimental studies of curiosity (Figure 2).

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 & Daw, 2021b; Singh, Lewis, Barto, & Sorg, 2010; Yu et al., 2024; Zheng et al., 2020). These serve as the foundation for our environment-generating algorithm. This foundation will be subsequently used for the integration of other features that have been previously reported relevant, such as long-range connections (Viswanathan et al., 2016), sink states (Modirshanechi et al., 2025; Xu et al., 2021), stochasticity (Mehlhorn et al., 2015; Modirshanechi et al., 2025), and variability in the number of available actions (Fasolo et al., 2009; Mehlhorn et al., 2015; Scheibehenne et al., 2010).

Specifically, our algorithm generates environments in three steps (Figure 2A; see Supplementary Material: Environment generation for details and pseudocode): **First**, it creates a maze with a branching structure connecting individual states (Maze generation in Figure 2A). **Second**, it randomly samples a certain number of states (red nodes in Figure 2A, left) and transforms them into grid-like rooms (Room integration in Figure 2A). By these two steps alone, we can produce a diverse series of environments consisting of grid-like rooms connected via randomly generated corridors (Figure 2B1–B2). Designing our environments in terms of 'rooms' and 'corridors' imposes a structure where moving between groups of densely connected states (i.e., rooms) is possible only by going through some specific groups of sequentially connected states (i.e., corridors). While 'rooms' and 'corridors' are associated with spatial navigation, we also use these terms to refer to abstract spaces; specifically, states in our environments do not necessarily correspond to physical locations, but they can also refer to abstract states such as stages of one's career in real life or the set of collected goods in a computer game. Finally, in the **third** and last step, the algorithm integrates long-range connections, stochasticity, and variable action availability into the environment by assigning exactly one of the following properties to each of the grid-like rooms within the maze (Room-property assignment in Figure 2A):

- **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 states in *this room* (Figure 2A and B4). 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,

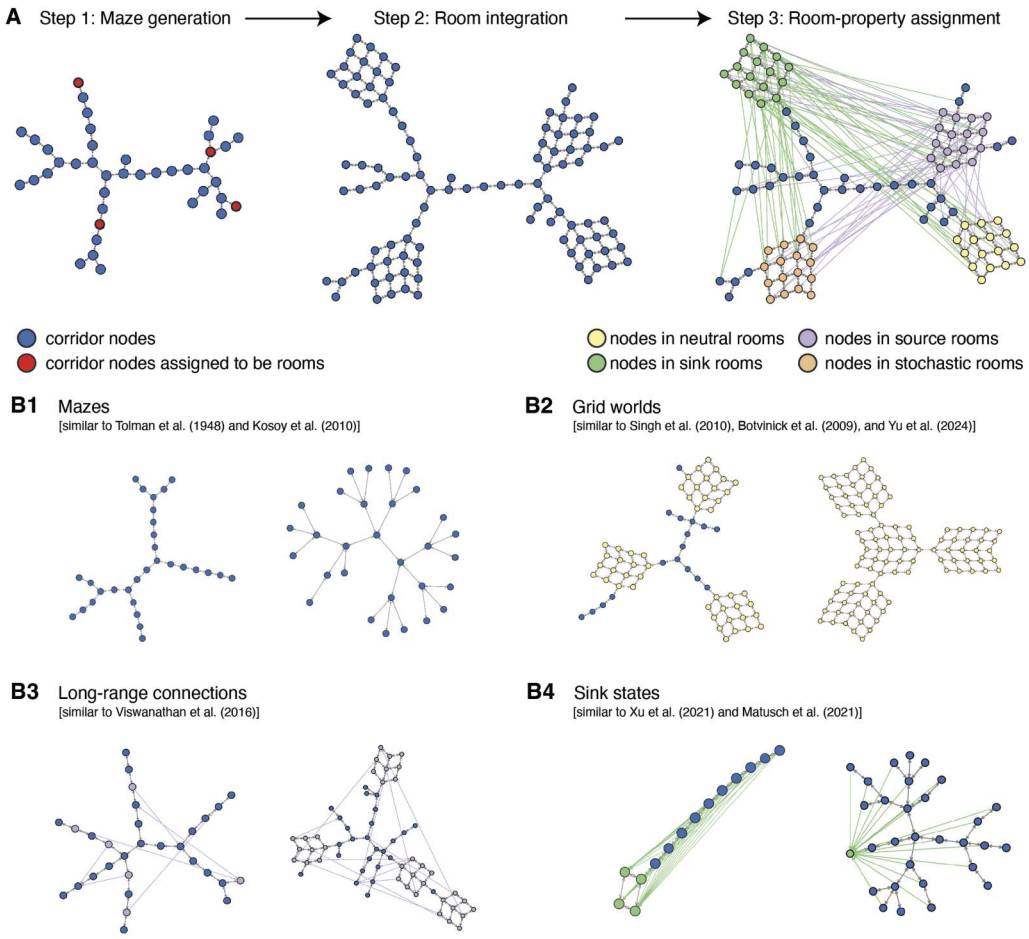

Figure 2. Environment generation. A. Schematic of environment generation process in three steps (see Environment generation for details). Nodes represent states, and edges the possible actions to transition between states. Gray edges are bidirectional. Purple edges (originating from a source room) and green edges (leading to a sink room) are unidirectional. **B**. Examples of generated environments by our algorithm corresponding to representative environments from the exploration and curiosity literature. **B1**. Multistep spatial mazes (Kosoy et al., 2020; Tolman, 1948) are represented by complex branching structures. **B2**. Grid worlds (Botvinick et al., 2009; Singh, Lewis, Barto, & Sorg, 2010; Yu et al., 2024) feature regular, grid-like structures. **B3**. Long-range connections (Viswanathan et al., 2016) are present in environments where states have distant connections. **B4**. 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., 2021) where the starting state acts as a sink state since dying resets the player to the start.

as dying resets the player to the start. In real life, social media addiction can result in 'scrolling on the phone' becoming a sink 'room,' as it is easy to engage with 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 the states in *this room* to other parts of the environment, creating long-range connections. From a source room, it is easy to quickly reach any region of the environment. States in a source room generally have 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, or using online portals to buy a house or plan a vacation.

- **Stochastic**: If a room is assigned to be stochastic, then transitions in 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* instead of the intended destination of *a*. Such unpredictability is common in everyday life, for example, when making financial investments, relying on the weather forecast to plan a weekend, or buying a lottery ticket.
- **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 set of parameters that specifies the properties of the generated environments, such as the number of states, the number of branching points, the number of rooms in the maze structure, the room sizes, the distribution of room types, and the intensity of the room properties (see Supplementary Material: Table 1 for details). By varying these parameters, we can create various classes of environments. For most of our results, we focus on five classes of environments with 100 states, including 4 rooms of 16 states each: *Neutral, Sink, Source, Stochastic*, and *Mixed* environments (see Figure 3 left column for examples and Supplementary Material: Table 1 for details). *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. *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. Since the environment-generating process is non-deterministic, many distinct environments can be produced within each class, but different environments of the same class are expected to exhibit similar properties. See Supplementary Material: Figure S3 for an analysis of environments with different numbers of states.

Ranking of Intrinsic Rewards Sensitively Depends on the Environment Class and Curiosity Objective

To quantify the merits of seeking different intrinsic rewards in different environments, we simulated model-based reinforcement learning agents (including a random agent as baseline) and measured their performance (Performance measures) in our five environment classes (specified in Environment generation).

Overall, novelty-seeking agents (blue in Figure 3) consistently have the best performance according to the uniform state visitation objective (Figure 3, right). This is because they are drawn to rarely visited states, which helps them to balance state counts (uniform state visitation), while also being competitive on state discovery (Figure 3, left). On the other hand, agents seeking surprise (orange in Figure 3) or information gain (green in Figure 3) excel on state discovery and model accuracy (Figure 3, left and middle) due to their incentive to select unknown actions, which is key for discovering all regions of the environment (state discovery) and building accurate estimates of transition probabilities (model accuracy). However, they perform consistently worse than novelty-seeking agents for uniform state visitation (Figure 3, right). Interestingly, agents seeking naive empowerment (red in Figure 3) perform poorly across all scenarios, because they avoid unknown regions, which they perceive as non-empowering due to epistemic uncertainty (i.e. the uninformative prior $\frac{1}{N_s}$ of $\hat{P}^{(t)}(s'|s,a)$ in Equation 1). As a result, empowerment-seeking agents prefer to remain where they have already explored, which harms exploration (see Discussion). Agents seeking MOP and SPIE (purple and brown in Figure 3, respectively) perform worse than agents seeking surprise, information gain, or novelty on state discovery and model accuracy. MOP's performance, as expected, lies between

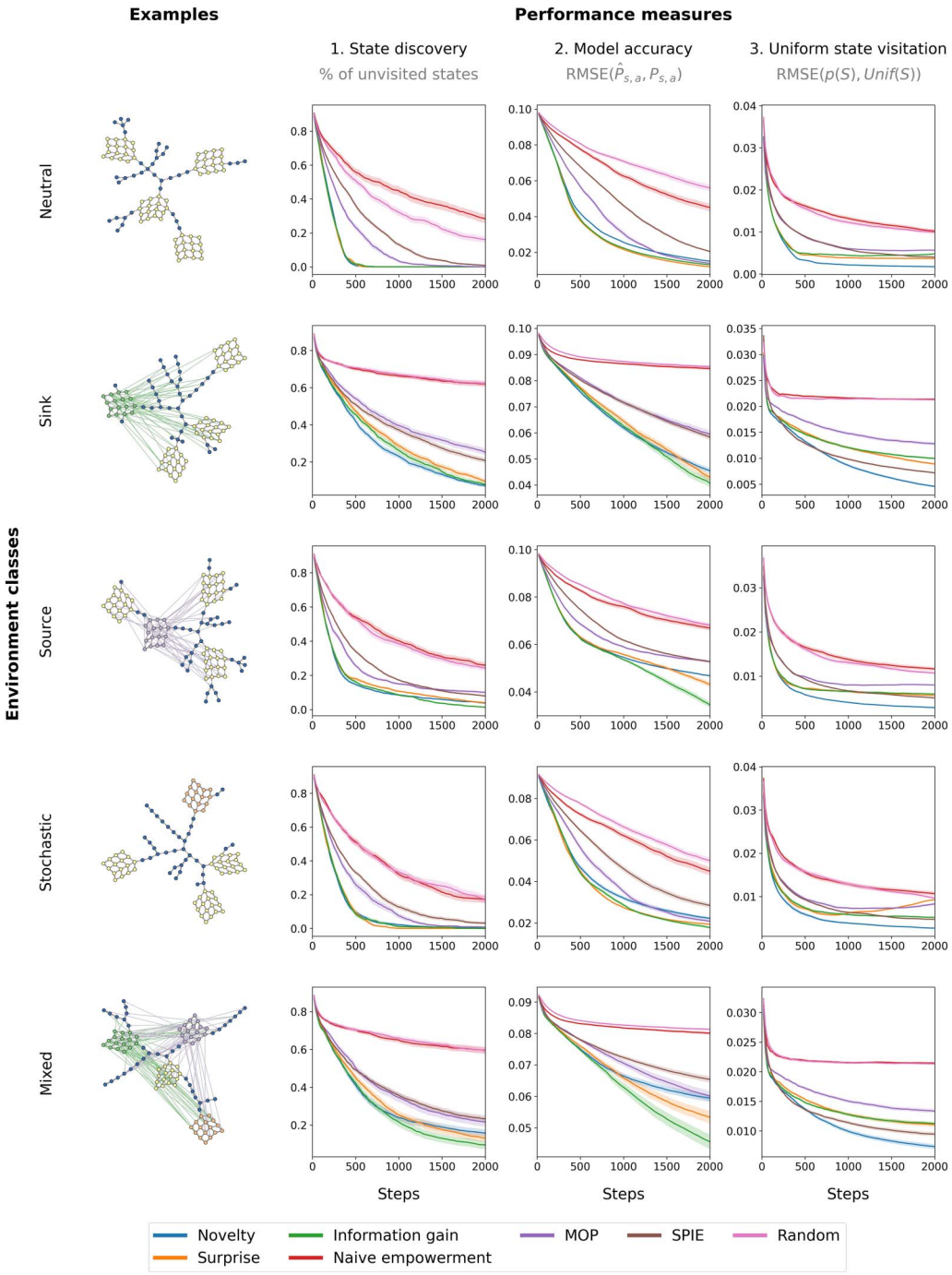

Figure 3. Comparative performance of seeking different intrinsic rewards in different classes of environments. Each subplot corresponds to the combination of one environment class (rows) and one performance measure (columns). An exemplar environment is shown for each class, where each node represents a state (same color code as in Figure 2). The performance of seeking each intrinsic reward was evaluated over 50 different instances of each environment class and for 2000 steps: the solid curves show the average, and the shaded areas show the standard error of the mean. For all measures, lower values correspond to better performance. The softmax inverse temperature β was optimized for the first 500 steps in each case (see Methods for details). The same experiment was conducted with the agents running for 4000 steps instead of 2000 (Supplementary Material: Figure S4), and using the KL divergence instead of RMSE for model accuracy and uniform state visitation (Supplementary Material: Figure S5), both yielding similar results. *Environment classes:* Neutral environments contain 4 neutral rooms. Sink environments contain one sink room with 50 additional connections leading to it. Source environments contain 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.

surprise and random agents (pink in Figure 3), as it combines surprise rewards with a high-entropy policy. On the other hand, SPIE-seeking agents are attracted to rarely visited states (similarly to novelty-seeking agents), sometimes outperforming surprise and information gain on uniform state visitation (Figure 3, right).

While the performance of agents seeking each intrinsic reward is fairly consistent across multiple environments of the same class (Supplementary Material: Figure S2), it varies strongly between environments of different classes (rows of Figure 3). Different environment classes affect performance in distinct ways: Neutral environments offer a good reference point where, for example, agents seeking novelty, surprise, or information gain perform very similarly on state discovery (Figure 3, row 1). However, in Sink environments, which are harder to explore due to the challenge of escaping sink rooms, performance differences become more pronounced, particularly on uniform state visitation (Figure 3, row 2, column 3). In these environments, novelty-seeking agents perform best on state discovery (Figure 3, row 2, column 1) because they prioritize exploring rarely visited states, whereas agents seeking surprise or information gain focus on testing all available actions, and thus frequently encounter the sink states (see also Figure 7), slowing their state discovery and leading to an imbalanced state visitation. On the other hand, in Source environments, building an accurate model of the environment (model accuracy) requires agents to repeatedly visit the source room. This benefits agents seeking surprise or information gain, which are attracted to unknown actions, but is detrimental for novelty-seeking agents which avoid revisiting the same states several times (Figure 3, row 3, column 2). Stochastic environments pose a particular challenge for agents seeking surprise or MOP as they tend to stay in the stochastic room after learning sufficiently about the environment. This leads to poor performance of these agents on uniform state visitation (Figure 3, row 4, column 3, see Intrinsic rewards detailed for a formal explanation of this asymptotic behavior), compared to other agents that are less sensitive to action stochasticity. Mixed environments combine features of all available classes, which results in a large total number of actions in the environments. This makes model building generally difficult and benefits agents seeking surprise or information gain.

To better understand the impact of selected environment features, we systematically varied their intensity and evaluated their impact on the performance of different agents. Specifically, we manipulated the branching rate (Figure 4A) as well as the number of sink connections (Figure 4B) in an environment inspired by Xu et al. (2021). We considered a class of environments with 100 states, where 4 states are in a single *sink* room, and the other 96 states are neutral and outside 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 Figure 4A). Importantly, the performance of different algorithms changes drastically as the branching rate increases from 0 to 1 (Figure 4A). Agents seeking novelty or SPIE, top performers at a branching rate of 0, become among the worst in state discovery and model accuracy as the branching rate increases to 1. Since novelty-seeking agents sometimes prefer actions leading to novel states over unknown actions, they may not explore all possible actions in a given state and, thus, miss parallel branches in environments with a high branching rate. On the other hand, increasing the number of sink connections generally benefits novelty- and SPIE-seeking agents more than other agents (Figure 4B). This can be explained by the fact that agents seeking surprise, information gain, or MOP are incentivized to try all available actions, often leading them to the sink and preventing efficient exploration of the whole environment. This shows that the environment's structure significantly influences the comparative performance of intrinsic rewards, indicating that results from experiments in one specific environment may not

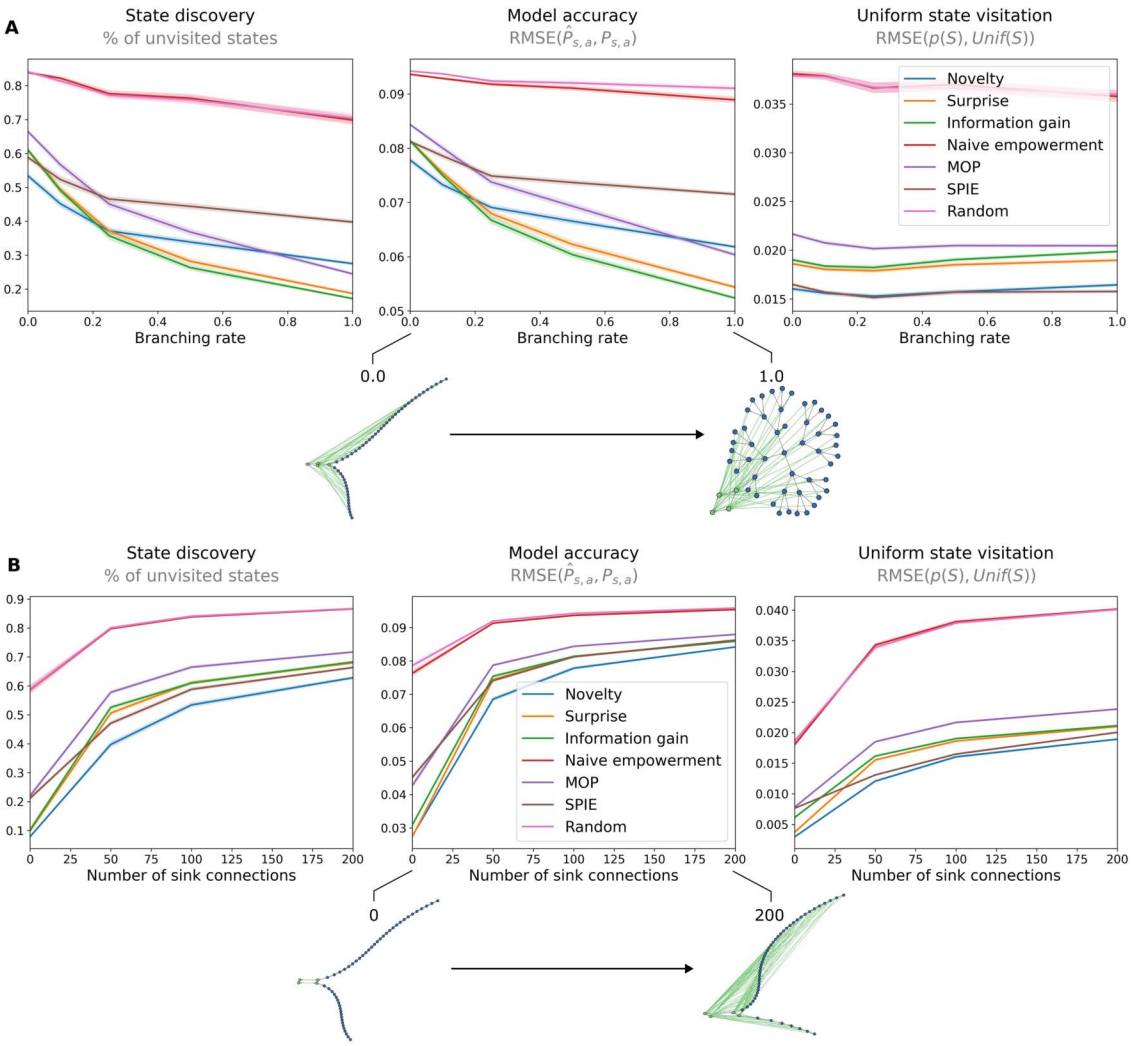

Figure 4. Comparative performance of seeking different intrinsic rewards as a function of environment parameters. The environment, inspired by Xu et al. (2021), contains one sink room with 4 states and 96 other states (see Supplementary Materials: Table 1). We simulated agents seeking different intrinsic rewards (as in Figure 3) and defined the performance score as the area under the curve of each measure over 2000 steps of simulations. The performance score for each environment class (i.e., a specific set of environment parameters) is obtained by averaging this value over 50 environment instances. **A**. Comparative performance as a function of 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. The exemplar environments shown are smaller versions (50 states) for illustration (same color code as in Figure 2). In each case, 100 additional connections lead to the sink. **B**. Comparative performance as a function of the number of sink connections with a fixed branching rate of 0.

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 novelty to be the dominant drive of human exploration. To test whether these results generalize across environments, our findings suggest that the experiment of Xu et al. (2021) should be repeated in an environment with a branching rate of 1 (also see Modirshanechi et al. (2025) for an alternative way to test this).

Novelty and Information Gain Are Two Main Axes of Curiosity

In the previous section, we observed that the ranking of different intrinsic rewards is highly dependent on the environment class and the curiosity objective. However, we also observed

that the best performing intrinsic reward was, with a few exceptions, either novelty or information gain (Figure 3 and Figure 4). We further confirmed this observation by evaluating the performance of different agents in 135 environment classes with different features (Figure 5A; see Variants of environment classes for details): Agents seeking information gain outperform all other agents in state discovery and model accuracy, whereas agents seeking novelty are the best-performing agents in achieving uniform state visitation.

Therefore, we hypothesized that novelty and information gain are two key drivers of exploration, and that exploration based on a combination of these rewards may achieve close-to-optimal performance across all objectives and environmental classes. To test this hypothesis, we simulated model-based RL agents that use a linear combination of novelty and information gain as the reward signal (Figure 5B). Interestingly, we observe that even with a fixed and equal weight for novelty and information gain ($\alpha = 0.5$ in Figure 5B), these

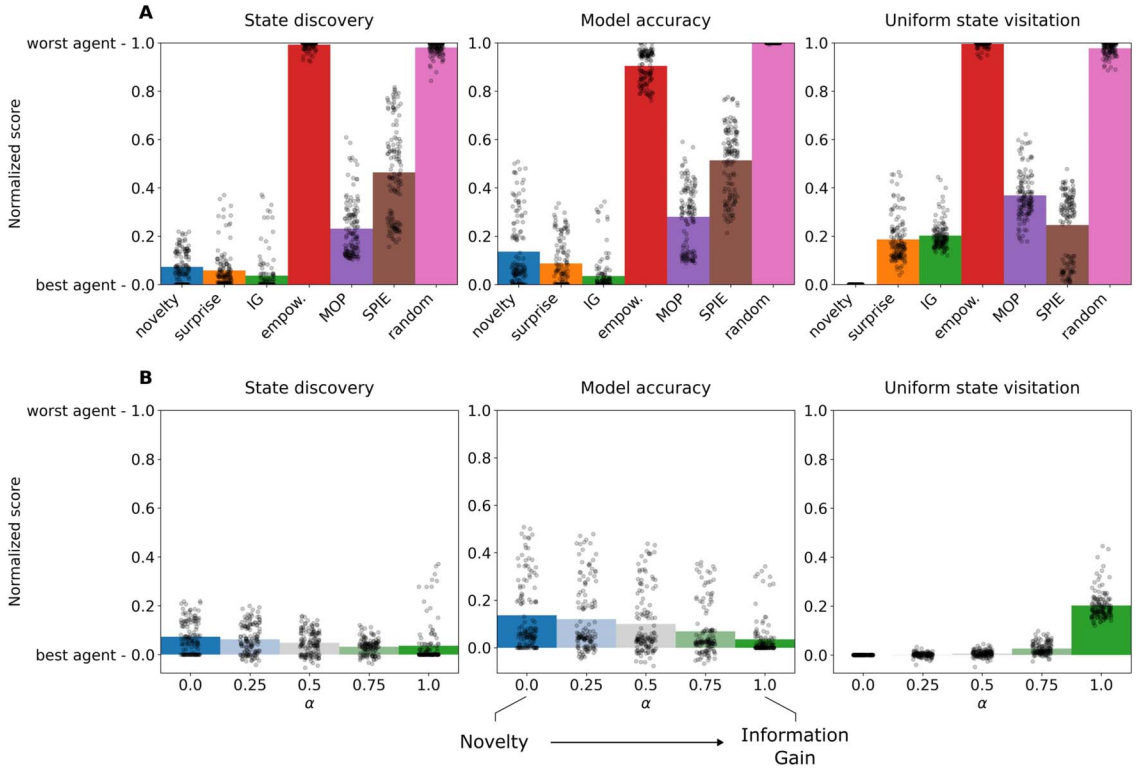

Figure 5. Average normalized performance score across environments for each intrinsic reward. A. To have a broad range of environments, we included variants of each of our different environment classes by systematically varying the branching rate, number of rooms, and room sizes (resulting in a total of 135 new environment classes; see Variants of environment classes for details). For each variant class and intrinsic reward, the raw performance score was computed as the average area under the curve (similarly to Figure 4) over 100 environment instances. Raw scores were then normalized separately for each variant class by setting the best-performing intrinsic reward to 0 and the worst to 1. Each dot represents the normalized score for one variant class, and the bars show the average scores across all environment classes. **B**. Combination of novelty and information gain. We simulated a series of new agents (in the same environments as in A) by using reward functions that linearly combined novelty and information gain: $R_\alpha^{(t)}(s,a,s') = \alpha \cdot \mathrm{IG}^{(t)}(s,a,s') + (1-\alpha) \cdot \mathrm{Nov}^{(t)}(s')$ with $\alpha \in \{0.25, 0.5, 0.75\}$. We also report the results from A for pure novelty-seeking and pure information-gain-seeking, corresponding to $\alpha = 0$ and $\alpha = 1$, respectively. The agents' performance was evaluated using the same normalization as in A. Agents seeking the combination of novelty and information gain with a mixture ratio of $\alpha = 0.5$ or 0.75 are almost always as good as the best-performing agents. Some dots fall below zero, because hybrid agents are occasionally better than the best agents in A.

'hybrid' RL agents reached close-to-optimal performance in all environment classes and for all performance measures. Interestingly, for state discovery, adding a small amount of novelty reward ($\alpha = 0.75$ in Figure 5B, left) improves the performance of agents purely seeking information gain in some environment classes (high cluster of dots for information gain in Figure 5B, left).

This improvement is much more pronounced for uniform state visitation (Figure 5B, right), where we observe that adding any amount of novelty reward brings the agents very close to the optimal score. Collectively, these results support the hypothesis that novelty and information gain are two fundamental axes of curiosity, with each providing distinct benefits, in line with recent experimental studies on humans (Dubey & Griffiths, 2020a; Monosov, 2024; Poli et al., 2022).

Performance On Curiosity Measures Correlates With Collection of Extrinsic Rewards

Collecting '*extrinsic*' rewards is often the ultimate goal of exploration in many real-world settings (Aubret et al., 2023; Cohen et al., 2007; Ladosz et al., 2022; Schulz & Gershman, 2019; Brant & Kavanau, 1964) and is also the natural long-term objective of curiosity in most top-down theories (Alet et al., 2020; Modirshanechi, Kondrakiewicz, et al., 2023; Singh, Lewis, Barto, & Sorg, 2010; Zheng et al., 2020). Our performance measures (Performance measures) were intuitively motivated by how they can benefit a curious agent in collecting extrinsic rewards, but the relationship was only *implicit*. Hence, an obvious question would be whether we would find the same comparative results as in the previous section if we consider performance measures that are *explicitly* designed for collecting extrinsic rewards.

To answer this question, we simulated agents in different environments and measured their efficiency in locating and reaching these extrinsic rewards. Specifically, we considered three scenarios for collecting extrinsic rewards, where each scenario intuitively corresponds to one of our initial performance measures (Figure 6):

1. **Persistent extrinsic reward:** The extrinsic reward is available at a fixed but unknown state from the beginning of the episode. The agent's goal is to discover the reward location. Intuitively, an agent that excels in **state discovery** should find the reward quickly.
2. **Announced late extrinsic reward:** The extrinsic reward is introduced after 2000 steps, and the agent is explicitly notified of its location. At this point, the agent switches to an exploitation mode, where it is motivated solely by extrinsic rewards, selects actions greedily, and has to use its learned model of the environment to locate the reward efficiently. Intuitively, an agent with a highly **accurate model** of the environment should leverage this information to navigate efficiently.
3. **Unannounced late extrinsic reward:** The extrinsic reward is introduced after 2000 steps, but the agent is not informed of its location. The goal of the agent is to discover the reward location. Intuitively, an agent that **explores uniformly** should be more likely to encounter the reward promptly.

Our results confirm these intuitions (compare Figure 5 with Figure 6): state discovery aligns with the ability to find a reward that is present from the start; model accuracy is crucial for efficiently locating an extrinsic reward when its position is announced; and uniform state visitation facilitates finding an extrinsic reward when no location information is provided. Indeed, the comparative performance patterns closely mirror our earlier findings: agents driven by information gain excel in the first two scenarios, while those driven by novelty perform best

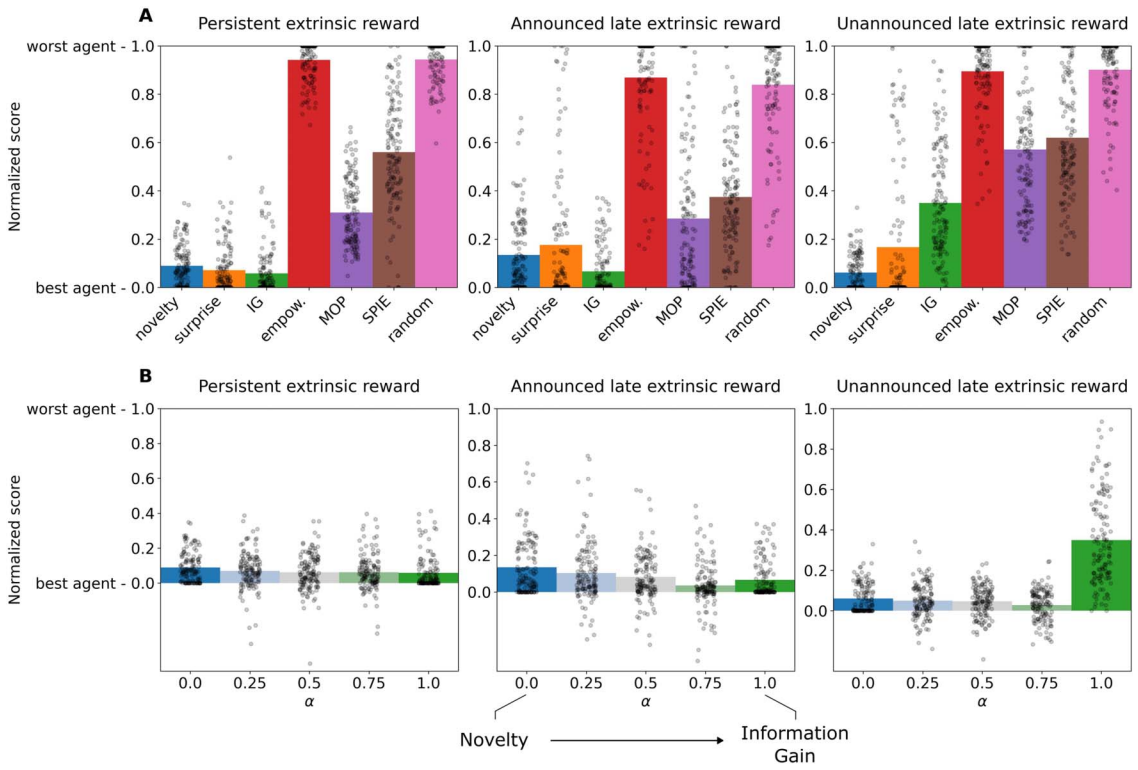

Figure 6. Average performance in collecting extrinsic rewards. *Persistent extrinsic reward* corresponds to an extrinsic reward present from the beginning, *announced late extrinsic reward* corresponds to an extrinsic reward introduced after 2000 steps when the agent is also notified of its location, and *unannounced late extrinsic reward* corresponds to an extrinsic reward introduced after 2000 steps without the agent knowing about it. As in Figure 5, we included variants of each of our different environment classes by varying the branching rate, number of rooms, and room sizes. **A.** For each variant class and intrinsic reward, the raw performance score was computed as the average time to reach the extrinsic rewards after it was introduced, cropped to 1000 steps, over 100 environment instances. Raw scores were then normalized separately for each variant class by setting the best-performing intrinsic reward to 0 and the worst to 1. Each dot represents the normalized score for one variant class, and the bars show the average scores across all environment classes. **B**. Combination of novelty and information gain (same design as in Figure 5). The agents' performance was evaluated in the same environments as in A and using the same normalization. Some dots fall below zero because hybrid agents are occasionally better than the best agents in A.

in the third (Figure 6A). Moreover, combining novelty and information gain again leads to near-optimal performance across all three conditions (Figure 6B, $\alpha = 0.75$).

Mixed Environments Allow Behavioral Dissociation of Intrinsic Rewards

Thus far, we compared different intrinsic rewards based on their performance in achieving the potential objectives of curiosity, i.e., their long-term behavior. What we have not delved into yet is (i) how exactly agents seeking different intrinsic rewards explore, and (ii) how these exploration patterns evolve. In this section, we study this link between structural features and different exploration strategies by analyzing the exploration patterns of agents seeking different intrinsic rewards within the Mixed environment class (environments with one sink, one source, one stochastic, and one neutral room; see Environment generation).

Specifically, we quantified the proportion of time that agents spend in different rooms of the environments (Figure 7). As a baseline, agents with a random policy predominantly remain in the sink room due to the difficulty of escaping it through random actions. On the other hand,

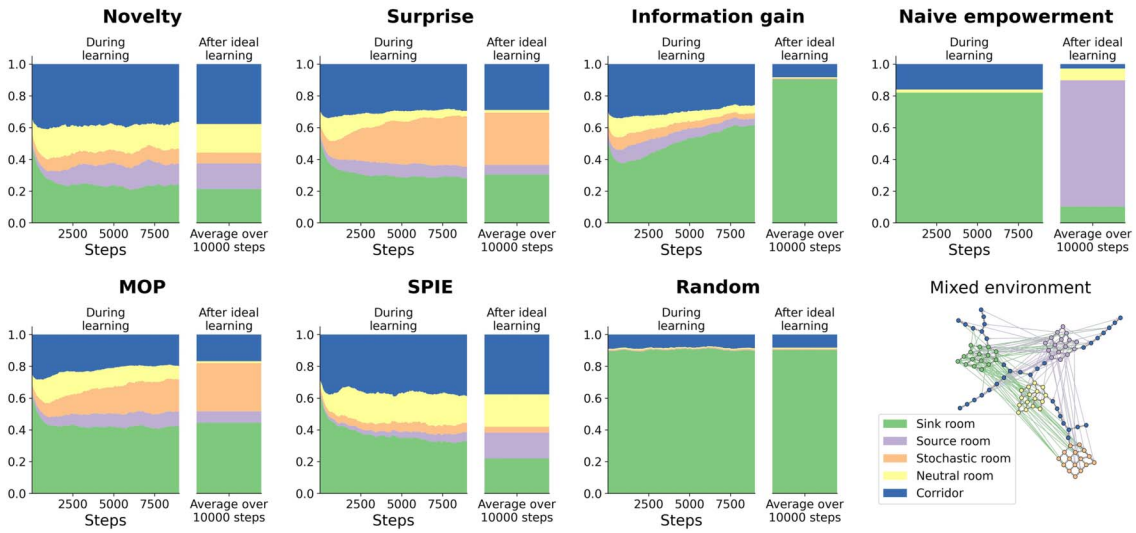

Figure 7. Agents seeking different intrinsic rewards are attracted to different rooms in mixed environments. Agents were simulated for 10'000 steps in 50 instances of Mixed environments with four rooms (see Environment generation), where 16 states were assigned to each room and 36 states to the corridor. We studied the proportion of time agents spend in different regions of the environment during and after learning. In the 'During learning' phase, agents started without knowledge of the environment and built a model of it, as in previous experiments. Each panel shows the evolution of the proportion of time spent in each region during learning, with an averaging window of 1000 steps (same color code as in Figure 2). In the 'After ideal learning' phase, the experiment was repeated, but the agent's model of the environment was fixed to the ground truth to assess asymptotic behavior. The proportion of time spent is averaged over 10'000 steps. Both phases were repeated 50 times and averaged. To have the same level of randomness in decision-making, the softmax inverse temperature β was fixed at $\frac{1}{\text{std}(r)}$ 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 reward in the 'After ideal learning' case corresponds to the expected asymptotic behavior derived from the reward formulation in Intrinsic rewards detailed.

novelty-seeking agents quickly achieve a near-uniform state visitation frequency. Agents seeking SPIE follow the same trend as novelty-seeking agents, but they have a slower convergence rate to their asymptotic behavior than novelty-seeking agents. After sufficient learning, surprise-seeking agents spend a large fraction of their time in the stochastic room, which has the highest transition uncertainty. Agents seeking MOP behave similarly to surprise-seeking agents, with a slight inclination towards the behavior of random agents, as MOP also rewards policy entropy. As observed before (Figure 3), agents seeking information gain explore different parts of the environment effectively, but they eventually tend towards the random policy (as information gain converges to zero; see Intrinsic rewards detailed). Different from all other agents, naive empowerment-driven agents do not explore the environment sufficiently. They stay mainly within known regions because unknown regions are expected to be non-empowering due to uncertainty (see Discussion); this is usually the sink room, as it acts as an attractor. However, once agents driven by naive empowerment know the transition probabilities of the entire environment (i.e., are aware of the properties of all four rooms, in the 'After ideal learning' phase in Figure 7), they spend most of their time in the source room, which offers the highest empowerment. This shows that, in mixed environments, the specific pattern of exploration gives strong indications about which intrinsic reward drives curiosity.

The distinct exploration patterns confirm that different intrinsic rewards 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 experimental designs can identify exploration strategies of curious agents, such as humans or animals, based on their action choices.

Correlation Between Intrinsic Rewards in the Mixed Environments

To further analyze intrinsic rewards beyond overall performance and behavioral decisions, we conducted a more detailed comparison at the level of reward time-series. We simulated a random agent in instances of Mixed environments and evaluated all six intrinsic rewards at every time step (Figure 8; see Supplementary Material: Figure S1 for the same analysis under other policies). These intrinsic rewards were used solely for correlation analysis and did not influence the agent's actions.

During learning (Figure 8A), surprise, information gain, and MOP exhibit consistently high correlations (≥0.87). This is surprising given the substantial behavioral differences between agents seeking these intrinsic rewards (Figure 7). However, the reward time-series can still have high alignments because all three rewards take high values when agents take unknown or stochastic actions, and they all take low values when agents take well-known and deterministic actions. This shows the importance of focusing on specific transitions when dissociating time-series for surprise and information gain (see, e.g., Visalli et al. (2019, 2021)). Similarly, novelty and SPIE show a relatively strong correlation (~0.7), as both reward signals take high values when agents transition to rarely visited states. A moderate (<0.5) correlation exists between the surprise-like and novelty-like groups, reflecting the intuitive link between exploring uncertain actions and encountering rarely visited states. In contrast, naive empowerment displays a negative correlation with all other rewards. When the agent is aware of the true transition probabilities (i.e., after ideal learning; Figure 8B), the correlations are different. The correlation between surprise and MOP is weaker, likely because the surprise rewards for deterministic actions vanish in the absence of epistemic uncertainty, whereas the MOP rewards still take high values when the policy has a high entropy, which, under the random policy, is the case in states with many available actions. However, the correlation between novelty and SPIE

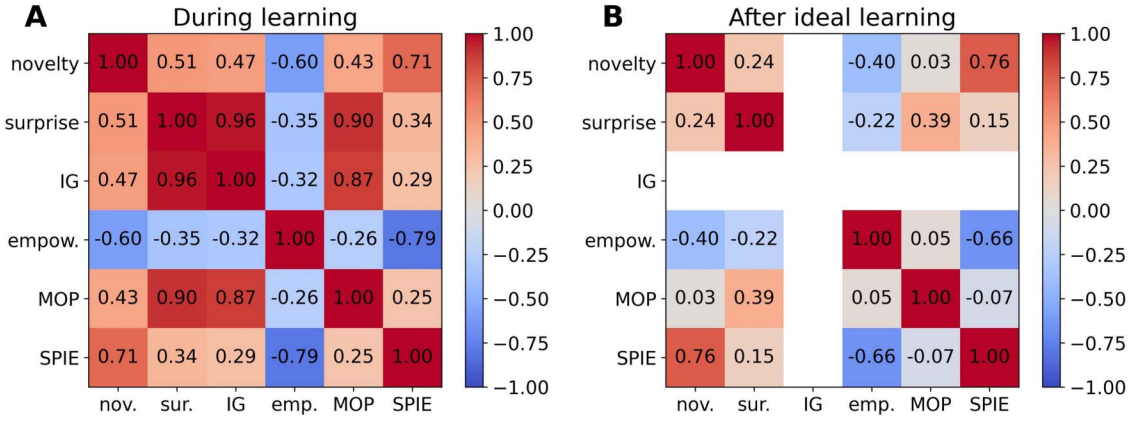

Figure 8. Correlations between intrinsic rewards under random policy. We simulated a random agent under the same conditions as in Figure 7, i.e., for 10,000 steps in 50 instances of Mixed environments with four rooms (see Environment generation). Each room contained 16 states, while the corridor had 36 states. All intrinsic rewards were computed at every step, but they were not used for action-selection (see Supplementary Material: Figure S1 for the same analysis under a different policy). **A.** The agent did not have prior knowledge of transition probabilities and had to learn them throughout exploration (same condition as 'During learning' in Figure 7). **B.** The agent had full knowledge of the environment (same condition as 'After ideal learning' in Figure 7).

remains high. In the absence of epistemic uncertainty, information gain is always zero, so its correlation with other reward signals is ill-defined.

DISCUSSION

We developed a systematic environment generation algorithm and assessed how environmental features such as stochasticity and transition structure affect the performance of exploration based on different intrinsic rewards. In summary, we made three main observations: 1. The best exploration strategy is highly dependent on the environment's structure and the assumed curiosity objective. 2. Despite this dependency, seeking a combination of information gain and novelty is almost always close to optimal. 3. Our environment-generating algorithm enables the design of complex environments that allow behavioral dissociation of different exploration strategies.

One main implication of our observations is that making justified assumptions about environment structure is crucial for defining any notion of optimality in top-down theories of curiosity (Modirshanechi, Kondrakiewicz, et al., 2023; Poli et al., 2024; Singh, Lewis, Barto, & Sorg, 2010). These theories characterize curiosity by defining a curiosity objective (the 'Why') and deriving the intrinsic reward signal that best fulfills this objective (the 'What'). Our results highlight that the objective-reward relationship is not straightforward and that the optimal reward signal depends critically on the environmental context. For example, our simulations show that in environments where exploring all possible actions is beneficial, such as a highly branched maze, intrinsic rewards that aim to reduce transition uncertainty are favored. Conversely, in environments where many actions lead to undesirable outcomes, such as sink states, intrinsic rewards that prioritize discovering rare states are more effective than those that encourage exhaustive exploration. Hence, we suggest that any optimality-based conclusion on the nature of curiosity should be made carefully. While the importance of environment structure on human exploratory behavior has been previously acknowledged (Dubey & Griffiths, 2020b; Meder & Nelson, 2012; Mehlhorn et al., 2015), a rigorous quantification of these effects on the exploration with different intrinsic rewards has so far been missing. Our study is a step towards systematically quantifying these effects. Importantly, our systematic analysis relies on our environment-generating algorithm, which enables us to generate environments based on concrete assumptions about the underlying environmental structure. Our environment-generating algorithm can be extended, for example, to incorporate non-stationary dynamics as in Behrens et al. (2007); Liakoni et al. (2021); Nassar et al. (2010), ecological priors as in Jagadish et al. (2024), or other realistic and complex features. The algorithm and its extensions can be further used for benchmarking of curiosity-driven exploration, comparing model-based versus model-free RL algorithms (Daw et al., 2011; Kool et al., 2016), or developing and testing meta-learning algorithms (Alet et al., 2020; Binz et al., 2023; Wang et al., 2016).

Our environment generation algorithm can also be used to design theory-guided experiments that enable identifying the dominant drive of exploration of a curious agent, such as a human participant. The algorithm's stochastic nature allows one to test a specific hypothesis in multiple environments that share key structural features, going beyond the typical studies of curiosity in single fixed environments (Brändle et al., 2023; Horvath et al., 2021; Kobayashi et al., 2019; Poli et al., 2022; Ten et al., 2021). While it is common in RL studies to test different algorithms across multiple environments, these environments are often either very similar to one another (Kosoy et al., 2020; Yu et al., 2024; Zheng et al., 2020), which limits their ability to assess generalizability, or very different (Jach et al., 2024; Matusch et al., 2021; Piray

& Daw, 2021b; Singh, Lewis, Barto, & Sorg, 2010), which complicates comparison and interpretation due to the lack of a consistent foundation. Using a parameterized environment representation, our environment-generating algorithm proposes a new perspective for generating varied environments while maintaining a common framework for comparison.

We observed a persistently poor performance of agents seeking naive empowerment. While this may appear surprising at first glance, we note that empowerment as an intrinsic reward was historically intended for guiding an agent toward environmental regions with a high level of control rather than motivating exploration and knowledge acquisition (Klyubin et al., 2005; Salge et al., 2014). Importantly, the majority of works on empowerment even explicitly assume that agents are a priori aware of the true transition probabilities or some accurate estimates of them (Brändle et al., 2023; Klyubin et al., 2005; Salge et al., 2014; Volpi & Polani, 2020). The poor performance of empowerment in our work is due to this difference and must be interpreted accordingly: when the transition probabilities are unknown and initialized with a uniform prior, seeking a naive estimation of empowerment discourages an agent from exploring the unseen parts of the environment. This interpretation is consistent with prior works on empowerment with unknown transition probabilities (Aubret et al., 2023; Becker-Ehmck et al., 2021; Bharadhwaj et al., 2022; Cao et al., 2025; Jung et al., 2011; Leibfried et al., 2019). These works have either reported inconclusive benefits of empowerment-seeking (Bharadhwaj et al., 2022; Leibfried et al., 2019) or used further manipulations that result in a reward signal that resembles a combination of information gain and empowerment (Becker-Ehmck et al., 2021; Cao et al., 2025; Jung et al., 2011). However, while there appears to be converging evidence that a naive version of empowerment is not useful for exploration (and hence curiosity), this evidence by no means implies that empowerment is not a useful intrinsic reward for purposes other than exploration, e.g., for seeking control (Brändle et al., 2023; Salge et al., 2014) or representation learning (Bharadhwaj et al., 2022).

Finally, we emphasize our use of efficient model-based RL with perfect memory to assess the performance of different exploration strategies. The choice of efficient model-based RL enabled us to pinpoint the cause of inefficiency and subpar performance to the specific type of intrinsic reward utilized for exploration. However, while humans can have an almost perfectly model-based behavior in simple environments (da Silva & Hare, 2020), they appear to behave mainly model-free in complex environments (Xu et al., 2021). In principle, putting constraints on an agent's resources, for example, by limiting it to use model-free policies or having limited memory, as in Bhui et al. (2021); Lieder and Griffiths (2019), can influence its conditionally optimal exploration strategies (Binz & Schulz, 2022). How our findings change depending on such constraints remains to be explored in future studies.

In conclusion, our results provide novel insights into curiosity-driven behavior by clarifying the link between intrinsic rewards, curiosity objectives, and environmental structure (Figure 1). Moreover, our environment generator provides a systematic methodology for future research, for designing experimental paradigms, benchmarking RL algorithms, and meta-learning novel RL algorithms.

METHODS

Intrinsic Rewards Detailed

We considered six intrinsic rewards: novelty, surprise, information gain, empowerment, Maximum Occupancy Principle, and Successor-Predecessor. We use $\mathcal{S}$ to denote the set of all states, and N_s for the number of states.

Novelty. Novelty-seeking agents are rewarded for exploring unusual states, i.e., those encountered infrequently (Aubret et al., 2023; Bellemare et al., 2016; Modirshanechi, Becker, et al., 2023; Ostrovski et al., 2017). Since we work with tabular RL, we use the same mathematical formulation as Xu et al. (2021) and define the observation frequency of a state s as

$$p_N^{(t)}(s) = \frac{C_s^{(t)} + \frac{1}{N_s}}{\sum_{s'} C_{s'}^{(t)} + 1}, \tag{4}$$

where $C_s^{(t)}$ represents the number of times state s has been encountered up to time t, and $\frac{1}{N_s}$ acts as an uninformative prior (c.f. Equation 1). The novelty of a state s is then expressed as a decreasing function of the observation frequency:

$$R_{Novelty}^{(t)}(s) = -\log p_N^{(t)}(s). \tag{5}$$

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

$$\mathbb{E}_{s\in\mathcal{S}}\left[R_{Novelty}\right] = \sum_s P_\pi(s) \cdot R_{Novelty}(s) \tag{6}$$

$$= -\sum_s P_\pi(s) \cdot \log P_\pi(s) = H_{P_\pi}(S), \tag{7}$$

where $H_{P_\pi}(S)$ is the entropy of the state observation frequency. As the discount factor γ gets close to 1, i.e., in the limit of not discounting future rewards, the policy π that maximizes Q-values in Equation 2 becomes the same as the policy π that maximizes $\mathbb{E}_{s\in\mathcal{S}}\left[R_{Novelty}\right]$ (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 will result in a close to uniform state visitation (Measure 3 in Performance measures).

Surprise. Surprise-seeking agents are rewarded when observing transitions that were anticipated to be unlikely. We follow Achiam and Sastry (2017); Burda et al. (2019); Pathak et al. (2017) and define the surprise of a transition as its Shannon surprise or surprisal (Modirshanechi et al., 2022; Modirshanechi, Becker, et al., 2023):

$$R_{Surprise}^{(t)}(s,a,s') = -\log \hat{P}^{(t)}(s'|s,a) \tag{8}$$

Here, $\hat{P}^{(t)}(s'|s,a)$ represents the estimated probability of the transition (see Equation 1). Higher intrinsic rewards are received for transitions the agent considers improbable.

Asymptotic behavior: Given sufficient exploration of all state-action pairs, the estimated transition probabilities $\hat{P}(s'|s,a)$ should converge over time to the true probabilities $P(s'|s,a)$. Hence, after convergence, the expected surprise reward obtained for taking an action a in state s is

$$\mathbb{E}_{s'\in\mathcal{S}}\left[R_{Surprise}(s,a,\cdot)\right] = -\sum_{s'} P(s'|s,a) \cdot \log P(s'|s,a) = H(S'|s,a), \tag{9}$$

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

Information Gain. Agents seeking information gain are rewarded based on the amount of information they acquire from a given transition, equivalent to the reduction in their uncertainty about the environment (Itti & Baldi, 2009; Nelson, 2005; Oudeyer & Kaplan, 2007; Storck et al., 1995). We use a formulation of information gain known as Postdictive surprise (Kolossa et al., 2015; Modirshanechi et al., 2022; Modirshanechi, Becker, et al., 2023). Following a transition, the agent updates its environment model, and the difference between the updated and previous model determines the intrinsic reward:

$$R_{IG}^{(t)}(s,a,s') = KL\Big(\hat{P}^{(t)}(\,\cdot\,|s,a)\;\|\;\hat{P}^{(t+1)}(\,\cdot\,|s,a,s_{t+1}=s')\Big) \tag{10}$$

Here, KL is the Kullback-Liebler divergence (Kullback, 1997), and $\hat{P}^{(t)}(\,\cdot\,|s,a)$ 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: Given sufficient exploration of all state-action pairs, the estimated transition probabilities $\hat{P}(s'|s,a)$ should converge over time 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 information-gain-seeking policy will converge to the uniformly random policy.

(Naive) Empowerment. Empowerment is a measure of the degree of control an agent has over its environment from a particular state (Klyubin et al., 2005; Salge et al., 2014) and quantifies how much an agent can influence future states. 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 number of steps. Here, we evaluate this channel capacity based on the agent's *estimate* of transition probabilities (Equation 1) rather than the true transition probabilities (see Equation 12 below). Hence, we refer to this quantity as 'naive empowerment' instead of 'empowerment' to emphasize the difference between our formulation and prior works that evaluated empowerment based on the true transition probabilities (Klyubin et al., 2005; Salge et al., 2014). Precisely, we consider 1-step naive empowerment (similarly to Leibfried et al. (2019)) defined as:

$$Emp^{(t)}(s) = \max_{P_A}\ I(S';A|s) \tag{11}$$

$$= \max_{P_A}\ \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}'} P_A(a|s)\hat{P}(s'|s,a)\log\left(\frac{P_A(a|s)\hat{P}(s'|s,a)}{P_A(a|s)\sum_{a\in\mathcal{A}}P_A(a|s)\hat{P}(s'|s,a)}\right) \tag{12}$$

$$= \max_{P_A}\left(H(S'|s) - H(S'|A,s)\right) \tag{13}$$

$$= \max_{P_A}\left(H(A,s) - H(A|S',s)\right) \tag{14}$$

where A and S' are random variables for the action and next state, respectively, $I(S';A|s)$ is the mutual information of S' and A from state s, $H(X)$ is the entropy of X, $\hat{P}(s'|s,a)$ is the agent's estimation of the transition probabilities defined in 1 and P_A is a dummy action policy for optimization. The optimization over P_A is done using the Blahut-Arimoto algorithm (Blahut, 1972). For an agent driven by naive empowerment as an intrinsic reward, we set $R_{Emp}^{(t)}(s,a,s') = Emp^{(t)}(s')$.

Naive empowerment $Emp^{(t)}(s)$ is zero in a given state s as long as the the estimated transition probabilities are given by the uninformative prior $\hat{P}(s'|s,a) = \frac{1}{N_s}$ for all actions a. Once the agent has a good estimate of the transition probabilities, then seeking naive empowerment can lead to interesting behavior. Examining Equation 13, naive empowerment quantifies the maximal difference between the entropy $H(S'|s)$ of the next state S' (i.e., diversity of potential next states) and the conditional entropy $H(S'|A, s)$ of the next state S' given action A (i.e., outcome uncertainty). This conceptually quantifies the amount of control over the destination when selecting an action in state s. Hence, agents seeking naive empowerment are drawn towards states with many actions and a high level of control. An alternative interpretation can be found by focusing on Equation 14: naive empowerment quantifies the maximal difference between the entropy $H(A|s)$ of current action A (i.e., a proxy of the number of actions) and the conditional entropy $H(A|S', s)$ of the current action A given the next state S' (i.e., a proxy of distinguishability of actions in light of their outcome). Hence, agents seeking naive empowerment are drawn towards a state with a high number of distinguishable actions.

Asymptotic behavior: An agent driven by naive empowerment will seek out states with many available actions, ideally with deterministic outgoing transition probabilities, as these states offer the most control. Given sufficient exploration of all state-action pairs, the estimated transition probabilities $\hat{P}(s'|s,a)$ should converge over time to the true probabilities $P(s'|s, a)$. In this case, the naive estimation of the empowerment of every state also converges to the true empowerment value of that state. 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. However, this may not be the case in early stages of exploration because seeking naive empowerment (based on a not-yet accurate estimate of transition probabilities) disfavors exploring unknown regions of the environment; this prevents having an accurate estimate of transition probabilities (as in Figure 7; see Discussion).

Maximum Occupancy Principle (MOP). Introduced in Ramírez-Ruiz et al. (2024), MOP as an intrinsic reward considers that an agent's behavior aims to maximize the occupancy of future action-state paths. Formally, the agent aims to maximize the return

$$R_{MOP}(s,a,s') = -\alpha_{MOP}\log\pi(a|s) - \beta_{MOP}\log\hat{P}(s'|s,a) \tag{15}$$

where α_{MOP} and β_{MOP} are weighting factors that determine the reward at each time step t (the subscript t has been omitted for clarity). An agent motivated by MOP is expected to favor high entropy policies because of the contribution of $\log\pi(a|s)$ as well as highly stochastic regions of the environment because of the contribution of $\log\hat{P}(s'|s,a)$ in 15. In our experiments, we set $\alpha_{MOP} = \beta_{MOP} = 1$ to give equal weights to these two aspects. Unlike other intrinsic rewards, we do not compute the policy by applying softmax on Q-values. Instead, we use the modified version of value iteration proposed by Moreno-Bote and Ramirez-Ruiz (2023); Ramírez-Ruiz et al. (2024) to consider the optimal policy at every step.

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

$$V_\pi(s) = \alpha_{MOP}H_\pi(A|s) + \beta_{MOP}\sum_a \pi(a|s)H(S'|s,a) + \gamma\sum_{a,s'}\pi(a|s)P(s'|s,a)V_\pi(s'), \tag{16}$$

where $H_\pi(A|s)$ is the policy in state s under policy π, and $H(S'|s, a)$ 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 experiencing stochastic transition, and the last term accounts for the value of the next state. The agent will aim to reach the states where 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 rewarding the agent only for discovering new states, like novelty, SPIE also rewards the agent for visiting states that lead to isolated regions of the environment. 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; Dayan (1993)), 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:

$$R^{(t)}_{SPIE}(s,a,s') = \hat{M}^{(t)}[s,s'] - \|\hat{M}^{(t)}[\,\cdot\,,s']\|_1, \tag{17}$$

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 using the agent's estimate of transition probabilities.

Asymptotic behavior: Yu et al. (2024) argue 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.

Equivalence of Uniform State Visitation and Minimal State Re-Visitation Time

This section gives a second perspective on why maintaining uniform state visitation (Measure 3 in Performance measures) is relevant for a curious agent. Specifically, we ask: After leaving a state s, how long does it take the agent to return to state s? If the agent never returns to state s, it will miss out on any potential reward that may appear there after its last visit. On the other hand, if it returns to state s too soon, then it will miss the opportunity to visit the other states in the environment. Hence, a main goal of a curious agent can be to minimize the expected re-visitation time for *all* states simultaneously. We show that this goal is equivalent to maintaining a uniform state visitation.

Formally, consider an agent that follows a stationary policy π to continuously explore its environment. We use $P_\pi(s)$ to denote the long-term state observation frequency achieved by this policy (the same as in Equation 6). For every given state $s \in \mathcal{S}$, the expected time $\mu_\pi(s)$ that it takes the agent to return to state s is equal to the inverse of its stationary observation frequency (Durrett, 2019)

$$\mu_\pi(s) \coloneqq \mathbb{E}_\pi[T_s | S_0 = s] = \frac{1}{P_\pi(s)}, \tag{18}$$

where T_s denotes the first time $t > 0$ when $S_t = s$. Accordingly, we can write the average expected re-visitation time as

$$\bar{\mu}_\pi \coloneqq \frac{1}{N_s}\sum_{s\in\mathcal{S}}\mu_\pi(s) = \frac{1}{N_s}\sum_{s\in\mathcal{S}}\frac{1}{P_\pi(s)} = N_s + N_s\underbrace{\sum_{s\in\mathcal{S}}\frac{\left(P_\pi(s) - \frac{1}{N_s}\right)^2}{P_\pi(s)}}_{=d_{\chi^2}(P_\pi, Unif(\mathcal{S}))}, \tag{19}$$

where the last equality is the result of a few lines of algebra, and $d_{\chi^2}(P_\pi, Unif(\mathcal{S}))$ is the χ^2 divergence of the state observation frequency P_π from the uniform state visitation. This implies

that finding a policy that minimizes the average expected re-visitation time $\bar{\mu}_\pi$ is equivalent to finding a policy that minimizes the divergence of the state visitation distribution from the uniform distribution.

Simulation Protocol

This section provides the details of our simulation protocol. We used prioritized sweeping (Moore & Atkeson, 1993) to solve the system of equations in 2 for finding Q-values after each observed transition; we used 100 iterations of prioritized sweeping to have as many iterations as the total number of states in each instance of the environment classes in Environment generation. In all experiments, unless stated otherwise, we generated 50 randomly sampled instances of each environment class. Agents were simulated for 2000 steps in each sampled environment, and the results were averaged across the 50 sampled environments to provide a single score for each environment class.

Hyper-Parameters Selection. The framework described in Equation 1–Equation 3 has two hyper-parameters: $\gamma \in [0,1)$ and $\beta \in \mathbb{R}^+$. In all experiments, we set $\gamma = 0.5^{\frac{2}{N_s}}$, where $N_s = N_s$ is the number of states in the environment; with this choice, a future reward that is $N_s/2$ steps away is discounted to half its value. The softmax inverse temperature β was separately optimized for each 'setup,' where a setup refers to a combination of intrinsic reward, performance measure, and environment class. For instance, in Figure 3, with 6 intrinsic rewards, 3 performance measures, and 5 environment regimes, there are 90 setups, requiring 90 optimized values for β. For each setup, we first generated 50 environments based on the chosen class. Then, we found the value of β that gives the best score using a grid search. To compute the score for a specific choice of β, we simulated an agent for 500 steps in each environment. We evaluated the performance measure every 100 steps and calculated the average, resulting in a score for each environment. The overall score was calculated as the average score across the 50 environments.

Variants of Environment Classes. To enable a more comprehensive evaluation of each intrinsic reward across diverse environments (Figure 5), we generated different variants of each of the environment classes (Neutral, Sink, Source, Stochastic, Mixed) in Environment generation. These variants were created by varying the branching rate, the number of rooms, and the number of states in each room (i.e., the room size). The parameter n_s is adjusted so that all resulting environments contain 100 states. The other environment parameters are fixed to the values in Table 1. Specifically, the branching rate takes one of the three values in {0,0.5,1}, the parameter n_{rooms} takes one of the three values in {1,2,4}, and the room size takes one of the three values in {2,3,4}. With this procedure, $3^3 = 27$ combinations are produced for each original environment class, resulting in a total of 135 variants overall. These variants of environment classes were used for the analysis shown in Figure 5 and Figure 6.

Experiments With Extrinsic Rewards. To evaluate whether our findings extend to the collection of extrinsic rewards, we conducted a series of experiments in which an external reward was introduced into the environment, either from the beginning of the episode or after 2000 steps (see Performance on curiosity measures correlates with collection of extrinsic rewards). We considered three conditions:

1. **Persistent extrinsic reward:** At the beginning of the episode, the agent's starting state and the reward location were selected at random with uniform distribution. The

agent was simulated for a maximum of 3000 steps, and its performance was measured by the number of steps taken to reach the goal state for the first time (or 3000 if the goal was never reached). For fair comparison, the parameter β for each environment and intrinsic reward was set to match the value used in Figure 5 under state discovery.

2. **Announced late extrinsic reward:** The agent explored the environment without any extrinsic reward for the first 2000 steps. Then, a goal state was randomly chosen with uniform distribution, and the agent was explicitly notified of the goal location. Accordingly, the agent's reward function was updated such that reaching the goal yielded a reward of 1, and all other states yielded 0. Then, the agent was simulated for an additional 1000 steps by using a greedy (instead of softmax) policy for maximizing this reward. The performance was measured by the number of steps taken to reach the goal after the goal state was introduced (or 1000 if the goal was never reached). For fair comparison, the parameter β for each environment and intrinsic reward was set to match the value used in Figure 5 under model accuracy.
3. **Unannounced late extrinsic reward:** As in the previous condition, the agent explored without extrinsic rewards for the first 2000 steps. Then, a goal state was randomly chosen with uniform distribution, but the agent received no information about it. The agent was simulated for 1000 additional steps after the appearance of the goal state, by seeking the same intrinsic reward as in the first 2000 steps and with no influence of extrinsic rewards. The performance was measured as the number of steps to reach the goal after it was introduced (or 1000 if not reached). For fair comparison, the parameter β for each environment and intrinsic reward was set to match the value used in Figure 5 under uniform state visitation.

Except for these modifications, all other aspects of the simulation remain consistent with the earlier experiments.

ACKNOWLEDGMENT

We thank Wulfram Gerstner for feedback and discussions throughout the project and on related topics. This research was supported by the Swiss National Science Foundation No. 200020_207426.

AUTHOR CONTRIBUTIONS

AM designed the study and developed the original theoretical results. All authors contributed to further conceptualization and development of the theoretical results. LG designed the environment-generating algorithm and performed the simulations. LG and AM made the visualization and wrote the first draft.

CODE AVAILABILITY

All code necessary to reproduce the results presented in this manuscript is available at https://github.com/gruaz-lucas/Merits_of_curiosity_julia, implemented in Julia. For broader accessibility, we have also developed a user-friendly Python implementation, available at https://github.com/gruaz-lucas/Merits_of_curiosity.

REFERENCES

Achiam, J., & Sastry, S. (2017). Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv*. https://doi.org/10.48550/arXiv.1703.01732

Alet, F., Schneider, M. F., Lozano-Perez, T., & Kaelbling, L. P. (2020). Meta-learning curiosity algorithms. In *International conference on learning representations*.

Aubret, A., Matignon, L., & Hassas, S. (2023). An information-theoretic perspective on intrinsic motivation in reinforcement learning: A survey. *Entropy*, *25*(2), 327. https://doi.org/10.3390/e25020327, PubMed: 36832693

Baldassarre, G., & Mirolli, M. (2013). Intrinsically motivated learning systems: An overview. In *Intrinsically motivated learning in natural and artificial systems* (pp. 1–14). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32375-1_1

Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In G. Baldassarre & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems* (pp. 17–47). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-32375-1_2

Becker-Ehmck, P., Karl, M., Peters, J., & van der Smagt, P. (2021). Exploration via empowerment gain: Combining novelty, surprise and learning progress. In *ICML 2021 workshop on unsupervised reinforcement learning*.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*, 1214–1221. https://doi.org/10.1038/nn1954, PubMed: 17676057

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002, PubMed: 30359611

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. *arXiv*. https://doi.org/10.48550/arXiv.1606.01868

Berlyne, D. E. (1966). Curiosity and exploration. *Science*, *153*(3731), 25–33. https://doi.org/10.1126/science.153.3731.25, PubMed: 5328120

Bharadhwaj, H., Babaeizadeh, M., Erhan, D., & Levine, S. (2022). Information prioritization through empowerment in visual model-based RL. In *International conference on learning representations*.

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21. https://doi.org/10.1016/j.cobeha.2021.02.015

Binz, M., Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. (2023). Meta-learned models of cognition. *Behavioral and Brain Sciences*, *47*, e147. https://doi.org/10.1017/S0140525X23003266, PubMed: 37994495

Binz, M., & Schulz, E. (2022). Modeling human exploration through resource-rational reinforcement learning. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems*.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, *18*(4), 460–473. https://doi.org/10.1109/TIT.1972.1054855

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280. https://doi.org/10.1016/j.cognition.2008.08.011, PubMed: 18926527

Brändle, F., Stocks, L. J., Tenenbaum, J. B., Gershman, S. J., & Schulz, E. (2023). Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*, *7*(9), 1481–1489. https://doi.org/10.1038/s41562-023-01661-2, PubMed: 37488401

Brant, D., & Kavanau, J. L. (1964). 'Unrewarded' exploration and learning of complex mazes by wild and domestic mice. *Nature*, *204*, 267–269. https://doi.org/10.1038/204267a0, PubMed: 14212425

Bromberg-Martin, E. S., Feng, Y.-Y., Ogasawara, T., White, J. K., Zhang, K., & Monosov, I. E. (2024). A neural mechanism for conserved value computations integrating information and rewards. *Nature Neuroscience*, *27*, 159–175. https://doi.org/10.1038/s41593-023-01511-4, PubMed: 38177339

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., & Efros, A. A. (2019). Large-scale study of curiosity-driven learning. In *International conference on learning representations*.

Cao, H., Feng, F., Fang, M., Dong, S., Yang, T., Huo, J., & Gao, Y. (2025). Towards empowerment gain through causal structure learning in model-based reinforcement learning. In *The thirteenth international conference on learning representations*.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). 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. https://doi.org/10.1098/rstb.2007.2098, PubMed: 17395573

da Silva, C. F., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, *4*, 1053–1066. https://doi.org/10.1038/s41562-020-0905-y, PubMed: 32632333

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/j.neuron.2011.02.027, PubMed: 21435563

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

de Tinguy, D., Van de Maele, T., Verbelen, T., & Dhoedt, B. (2024). Spatial and temporal hierarchy for autonomous navigation using active inference in minigrid environment. *Entropy*, *26*(1). https://doi.org/10.3390/e26010083, PubMed: 38248208

Dubey, R., & Griffiths, T. L. (2020a). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, *127*(3), 455–476. https://doi.org/10.1037/rev0000175, PubMed: 31868394

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

Dubey, R., Griffiths, T. L., & Dayan, P. (2022). The pursuit of happiness: A reinforcement learning perspective on habituation and comparisons. *PLOS Computational Biology*, *18*(8), e1010316. https://doi.org/10.1371/journal.pcbi.1010316, PubMed: 35925875

Durrett, R. (2019). *Probability: Theory and examples*. Cambridge University Press. https://doi.org/10.1017/9781108591034

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*. Cambridge University Press. https://doi.org/10.1017/CBO9781316576533

Fasolo, B., Hertwig, R., Huber, M., & Ludwig, M. (2009). 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. https://doi.org/10.1002/mar.20272

FitzGibbon, L., Lau, J. K. L., & Murayama, K. (2020). The seductive lure of curiosity: Information as a motivationally salient reward. *Current Opinion in Behavioral Sciences*, *35*, 21–27. https://doi.org/10.1016/j.cobeha.2020.05.014

Gershman, S. J., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science*, *7*(3), 391–415. https://doi.org/10.1111/tops.12138, PubMed: 25808176

Ghazizadeh, A., Griggs, W., & Hikosaka, O. (2016). Ecological origins of object salience: Reward, uncertainty, aversiveness, and novelty. *Frontiers in Neuroscience*, *10*, 378. https://doi.org/10.3389/fnins.2016.00378, PubMed: 27594825

Ghilardi, T., Poli, F., Meyer, M., Colizoli, O., & Hunnius, S. (2024). Early roots of information-seeking: Infants predict and generalize the value of information. *eLife*, 13. https://doi.org/10.7554/eLife.92388.1

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*, 758–770. https://doi.org/10.1038/s41583-018-0078-0, PubMed: 30397322

Gruber, M. J., & Ranganath, C. (2019). How curiosity enhances hippocampus-dependent memory: The prediction, appraisal, curiosity, and exploration (PACE) framework. *Trends in Cognitive Sciences*, *23*(12), 1014–1025. https://doi.org/10.1016/j.tics.2019.10.003, PubMed: 31706791

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S., & Dragan, A. (2020). Inverse reward design. *arXiv*. https://doi.org/10.48550/arXiv.1711.02827

Horvath, L., Colcombe, S., Milham, M., Ray, S., Schwartenbeck, P., & Ostwald, D. (2021). Human belief state-based exploration and exploitation in an information-selective symmetric reversal bandit task. *Computational Brain & Behavior*, *4*(4), 442–462. https://doi.org/10.1007/s42113-021-00112-3, PubMed: 34368622

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. https://doi.org/10.1016/j.visres.2008.09.007, PubMed: 18834898

Jach, H. K., Cools, R., Frisvold, A., Grubb, M. A., Hartley, C. A., Hartmann, J., Hunter, L., Jia, R., de Lange, F. P., Larisch, R., Lavelle-Hill, R., Levy, I., Li, Y., van Lieshout, L. L. F., Nussenbaum, K., Ravaioli, S., Wang, S., Wilson, R., Woodford, M., … Gottlieb, J. (2024). Individual differences in information demand have a low dimensional structure predicted by some curiosity traits. *Proceedings of the National Academy of Sciences*, *121*(45), e2415236121. https://doi.org/10.1073/pnas.2415236121, PubMed: 39467138

Jagadish, A. K., Coda-Forno, J., Thalmann, M., Schulz, E., & Binz, M. (2024). Human-like category learning by injecting ecological priors from large language models into neural networks. In *Forty-first international conference on machine learning*.

Jirout, J. J., Evans, N. S., & Son, L. K. (2024). Curiosity in children across ages and contexts. *Nature Reviews Psychology*, *3*(9), 622–635. https://doi.org/10.1038/s44159-024-00346-5

Jung, T., Polani, D., & Stone, P. (2011). Empowerment for continuous agent—Environment systems. *Adaptive Behavior*, *19*(1), 16–39. https://doi.org/10.1177/1059712310392389

Kashdan, T. B., Gallagher, M. W., Silvia, P. J., Winterstein, B. P., Breen, W. E., Terhar, D., & Steger, M. F. (2009). The curiosity and exploration inventory-II: Development, factor structure, and psychometrics. *Journal of Research in Personality*, *43*(6), 987–998. https://doi.org/10.1016/j.jrp.2009.04.011, PubMed: 20160913

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460. https://doi.org/10.1016/j.neuron.2015.09.010, PubMed: 26539887

Klyubin, A., Polani, D., & Nehaniv, C. (2005). Empowerment: A universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation* (Vol. 1, pp. 128–135). https://doi.org/10.1109/CEC.2005.1554676

Kobayashi, K., Ravaioli, S., Baranès, A., Woodford, M., & Gottlieb, J. (2019). Diverse motives for human curiosity. *Nature Human Behaviour*, *3*(6), 587–595. https://doi.org/10.1038/s41562-019-0589-3, PubMed: 30988479

Kolossa, A., Kopp, B., & Fingscheidt, T. (2015). A computational analysis of the neural bases of Bayesian inference. *NeuroImage*, *106*, 222–237. https://doi.org/10.1016/j.neuroimage.2014.11.007, PubMed: 25462794

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLOS Computational Biology*, *12*(8), e1005090. https://doi.org/10.1371/journal.pcbi.1005090, PubMed: 27564094

Kosoy, E., Collins, J., Chan, D. M., Huang, S., Pathak, D., Agrawal, P., Canny, J., Gopnik, A., & Hamrick, J. B. (2020). Exploring exploration: Comparing children with RL agents in unified environments. *arXiv*. https://doi.org/10.48550/arXiv.2005.02880

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Ladosz, P., Weng, L., Kim, M., & Oh, H. (2022). Exploration in deep reinforcement learning: A survey. *Information Fusion*, *85*, 1–22. https://doi.org/10.1016/j.inffus.2022.03.003

Lau, J. K. L., Ozono, H., Kuratomi, K., Komiya, A., & Murayama, K. (2020). Shared striatal activity in decisions to satisfy curiosity and hunger at the risk of electric shocks. *Nature Human Behaviour*, *4*(5), 531–543. https://doi.org/10.1038/s41562-020-0848-3, PubMed: 32231281

Leibfried, F., Pascual-Díaz, S., & Grau-Moya, J. (2019). A unified bellman optimality principle combining reward maximization and empowerment. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Liakoni, V., Lehmann, M. P., Modirshanechi, A., Brea, J., Lutti, A., Gerstner, W., & Preuschoff, K. (2022). Brain signals of a surprise-actor-critic model: Evidence for multiple learning modules in human decision making. *NeuroImage*, *246*, 118780. https://doi.org/10.1016/j.neuroimage.2021.118780, PubMed: 34875383

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

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. https://doi.org/10.1017/S0140525X1900061X, PubMed: 30714890

Matusch, B., Ba, J., & Hafner, D. (2021). Evaluating agents without rewards. *arXiv*. https://doi.org/10.48550/arXiv.2012.11538

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148. https://doi.org/10.1017/S1930297500002977

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191–215. https://doi.org/10.1037/dec0000033

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

Meyniel, F., Maheu, M., & Dehaene, S. (2016). Human inferences about sequences: A minimal transition probability model. *PLOS Computational Biology, 12*(12), e1005260. https://doi.org/10.1371/journal.pcbi.1005260, PubMed: 28030543

Modirshanechi, A., Becker, S., Brea, J., & Gerstner, W. (2023). Surprise and novelty in the brain. *Current Opinion in Neurobiology, 82*, 102758. https://doi.org/10.1016/j.conb.2023.102758, PubMed: 37619425

Modirshanechi, A., Brea, J., & Gerstner, W. (2022). A taxonomy of surprise definitions. *Journal of Mathematical Psychology, 110*, 102712. https://doi.org/10.1016/j.jmp.2022.102712

Modirshanechi, A., Kondrakiewicz, K., Gerstner, W., & Haesler, S. (2023). Curiosity-driven exploration: Foundations in neuroscience and computational modeling. *Trends in Neurosciences, 46*(12), 1054–1066. https://doi.org/10.1016/j.tins.2023.10.002, PubMed: 37925342

Modirshanechi, A., Lin, W.-H., Xu, H. A., Herzog, M. H., & Gerstner, W. (2025). Even if suboptimal, novelty drives human exploration. *bioRxiv*. https://doi.org/10.1101/2022.07.05.498835

Monosov, I. E. (2024). Curiosity: Primate neural circuits for novelty and information seeking. *Nature Reviews Neuroscience, 25*, 195–208. https://doi.org/10.1038/s41583-023-00784-9, PubMed: 38263217

Montgomery, K. C. (1954). The role of the exploratory drive in learning. *Journal of Comparative and Physiological Psychology, 47*(1), 60–64. https://doi.org/10.1037/h0054833, PubMed: 13130734

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

Moreno-Bote, R., & Ramirez-Ruiz, J. (2023). Empowerment, free energy principle and maximum occupancy principle compared. In *NeurIPS 2023 Workshop: Information-Theoretic Principles in Cognitive Systems*. https://openreview.net/forum?id=OcHrsQox0Z

Morrens, J., Aydin, Ç., van Rensburg, A. J., Rabell, J. E., & Haesler, S. (2020). Cue-evoked dopamine promotes conditioned responding during learning. *Neuron, 106*(1), 142–153. https://doi.org/10.1016/j.neuron.2020.01.012, PubMed: 32027824

Murayama, K. (2022). A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic-extrinsic rewards. *Psychological Review, 129*(1), 175–198. https://doi.org/10.1037/rev0000349, PubMed: 35099213

Murayama, K., FitzGibbon, L., & Sakaki, M. (2019). Process account of curiosity and interest: A reward learning perspective. *Educational Psychology Review*, 1–21. https://doi.org/10.1007/s10648-019-09499-9

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience, 30*(37), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010, PubMed: 20844132

Nedergaard, A., & Cook, M. (2023). k-means maximum entropy exploration. *arXiv*. https://doi.org/10.48550/arXiv.2205.15623

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112*(4), 979–999. https://doi.org/10.1037/0033-295X.112.4.979, PubMed: 16262476

Ogasawara, T., Sogukpinar, F., Zhang, K., Feng, Y.-Y., Pai, J., Jezzini, A., & Monosov, I. E. (2022). A primate temporal cortex-zona incerta pathway for novelty seeking. *Nature Neuroscience, 25*(1), 50–60. https://doi.org/10.1038/s41593-021-00950-1, PubMed: 34903880

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

Oudeyer, P.-Y. (2018). Computational theories of curiosity-driven learning. *arXiv*. https://doi.org/10.48550/arXiv.1802.10546

Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics, 1*, 6. https://doi.org/10.3389/neuro.12.006.2007, PubMed: 18958277

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

Piray, P., & Daw, N. D. (2021a). A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications, 12*(1), 6587. https://doi.org/10.1038/s41467-021-26731-9, PubMed: 34782597

Piray, P., & Daw, N. D. (2021b). Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature Communications, 12*(1), 4942. https://doi.org/10.1038/s41467-021-25123-3, PubMed: 34400622

Pisula, W. (2009). *Curiosity and information seeking in animal and human behavior*. Boca Raton, FL: Brown Walker Press.

Poli, F., Meyer, M., Mars, R. B., & Hunnius, S. (2022). Contributions of expected learning progress and perceptual novelty to curiosity-driven exploration. *Cognition, 225*, 105119. https://doi.org/10.1016/j.cognition.2022.105119, PubMed: 35421742

Poli, F., Meyer, M., Mars, R. B., & Hunnius, S. (2025). Exploration in 4-year-old children is guided by learning progress and novelty. *Child Development, 96*(1), 192–202. https://doi.org/10.1111/cdev.14158, PubMed: 39223863

Poli, F., O'Reilly, J. X., Mars, R. B., & Hunnius, S. (2024). Curiosity and the dynamics of optimal exploration. *Trends in Cognitive Sciences, 28*(5), 441–453. https://doi.org/10.1016/j.tics.2024.02.001, PubMed: 38413257

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons. https://doi.org/10.1002/9780470316887

Ramírez-Ruiz, J., Grytskyy, D., Mastrogiuseppe, C., Habib, Y., & Moreno-Bote, R. (2024). Complex behavior from intrinsic motivation to occupy action-state path space. *arXiv*. https://doi.org/10.48550/arXiv.2205.10316

Salge, C., Glackin, C., & Polani, D. (2014). Empowerment—An introduction. *Guided self-organization: Inception* (pp. 67–114), https://doi.org/10.1007/978-3-642-53734-9_4

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research, 37*(3), 409–425. https://doi.org/10.1086/651235

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

Schmitt, F. F., & Lahroodi, R. (2008). The epistemic value of curiosity. *Education Theory, 58*(2), 125–148. https://doi.org/10.1111/j.1741-5446.2008.00281.x

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology, 55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003, PubMed: 30529148

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., & Pathak, D. (2020). Planning to explore via self-supervised world models. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (pp. 8583–8592). PMLR.

Singh, S., Lewis, R. L., & Barto, A. G. (2010). Where do rewards come from? In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 2601–2606).

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

Sorg, J., Lewis, R. L., & Singh, S. (2010). Reward design via online gradient ascent. In *Advances in neural information processing systems* (Vol. 23). Curran Associates, Inc.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94. https://doi.org/10.1126/science.aaa3799, PubMed: 25838378

Storck, J., Hochreiter, S., & Schmidhuber, J. (1995). Reinforcement driven information acquisition in non-deterministic environments. In *ICANN'95* (pp. 159–164).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning, second edition: An introduction*. MIT Press.

Szumowska, E., & Kruglanski, A. W. (2020). Curiosity as end and means. *Current Opinion in Behavioral Sciences*, *35*, 35–39. https://doi.org/10.1016/j.cobeha.2020.06.008

Ten, A., Kaushik, P., Oudeyer, P.-Y., & Gottlieb, J. (2021). Humans monitor learning progress in curiosity-driven exploration. *Nature Communications*, *12*(1), 5972. https://doi.org/10.1038/s41467-021-26196-w, PubMed: 34645800

Tolguenec, P.-A. L., Besse, Y., Teichteil-Konigsbuch, F., Wilson, D. G., & Rachelson, E. (2024). Exploration by learning diverse skills through successor state measures. *arXiv*. https://doi.org/10.48550/arXiv.2406.10127

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626, PubMed: 18870876

Visalli, A., Capizzi, M., Ambrosini, E., Kopp, B., & Vallesi, A. (2021). Electroencephalographic correlates of temporal Bayesian belief updating and surprise. *NeuroImage*, *231*, 117867. https://doi.org/10.1016/j.neuroimage.2021.117867, PubMed: 33592246

Visalli, A., Capizzi, M., Ambrosini, E., Mazzonetto, I., & Vallesi, A. (2019). Bayesian modeling of temporal expectations in the human brain. *NeuroImage*, *202*, 116097. https://doi.org/10.1016/j.neuroimage.2019.116097, PubMed: 31415885

Viswanathan, V., Lees, M., & Sloot, P. M. A. (2016). The influence of memory on indoor environment exploration: A numerical study. *Behavior Research Methods*, *48*(2), 621–639. https://doi.org/10.3758/s13428-015-0604-1, PubMed: 26170049

Volpi, N. C., & Polani, D. (2020). Goal-directed empowerment: Combining intrinsic motivation and task oriented behavior. *IEEE Transactions on Cognitive and Developmental Systems*, *15*(2), 361–372. https://doi.org/10.1109/TCDS.2020.3042938

Voss, H.-G., & Keller, H. (2013). *Curiosity and exploration: Theories and results*. Elsevier.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv*. https://doi.org/10.48550/arXiv.1611.05763

Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., & Herzog, M. H. (2021). Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLOS Computational Biology*, *17*(6), e1009070. https://doi.org/10.1371/journal.pcbi.1009070, PubMed: 34081705

Yu, C., Burgess, N., Sahani, M., & Gershman, S. J. (2024). Successor-predecessor intrinsic exploration. *arXiv*. https://doi.org/10.48550/arXiv.2305.15277

Zheng, Z., Oh, J., Hessel, M., Xu, Z., Kroiss, M., Van Hasselt, H., Silver, D., & Singh, S. (2020). What can learned intrinsic rewards capture? In H. D. III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (pp. 11436–11446). PMLR.