

# **Merits of curiosity — A simulation study**

**Lucas Gruaz, Alireza Modirshanechi, Johanni Brea**

**Presenter: Hanqi Zhou**

**2024.10.25**

# Main contribution

- Understand the relationship between
  - Intrinsic rewards (drives of curiosity)
  - Optimality conditions (objectives of curiosity)
  - Environment structures

# Main results

- Performance of each *intrinsic reward* is highly dependent on the *structural features of environments*
  - This indicates that ‘optimality’ in the top-down theories of curiosity needs a precise formulation of the curiosity objective and the environment structure
- Agents seeking a combination of *novelty* and *information gain* always achieve a close-to-optimal performance
  - This proposes novelty and information gain as two principal axes of curiosity-driven behavior

# Background

# Two principal questions

## Regarding curiosity

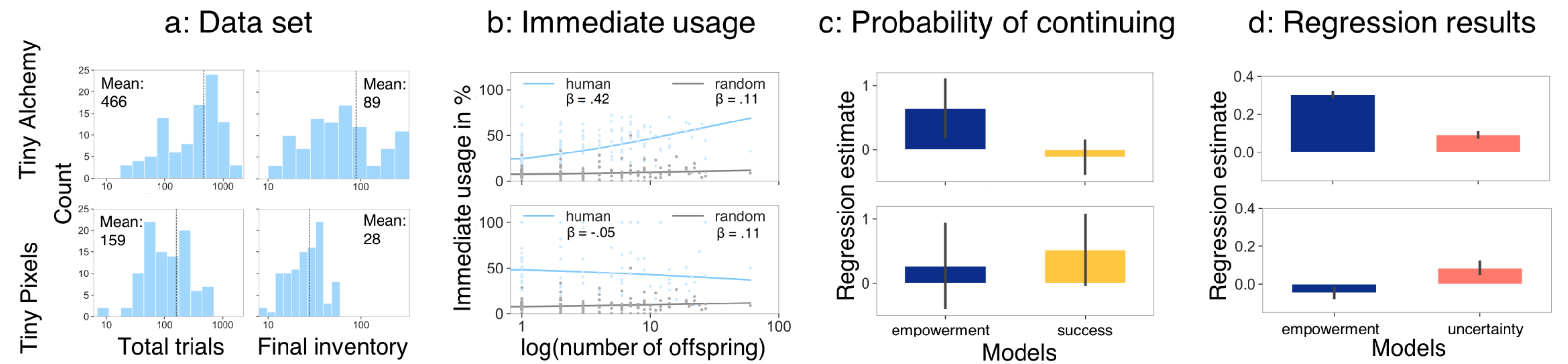
- **What** are humans exactly curious about?
- **Why** are humans and animals curious?

# Two principal questions

## Regarding curiosity

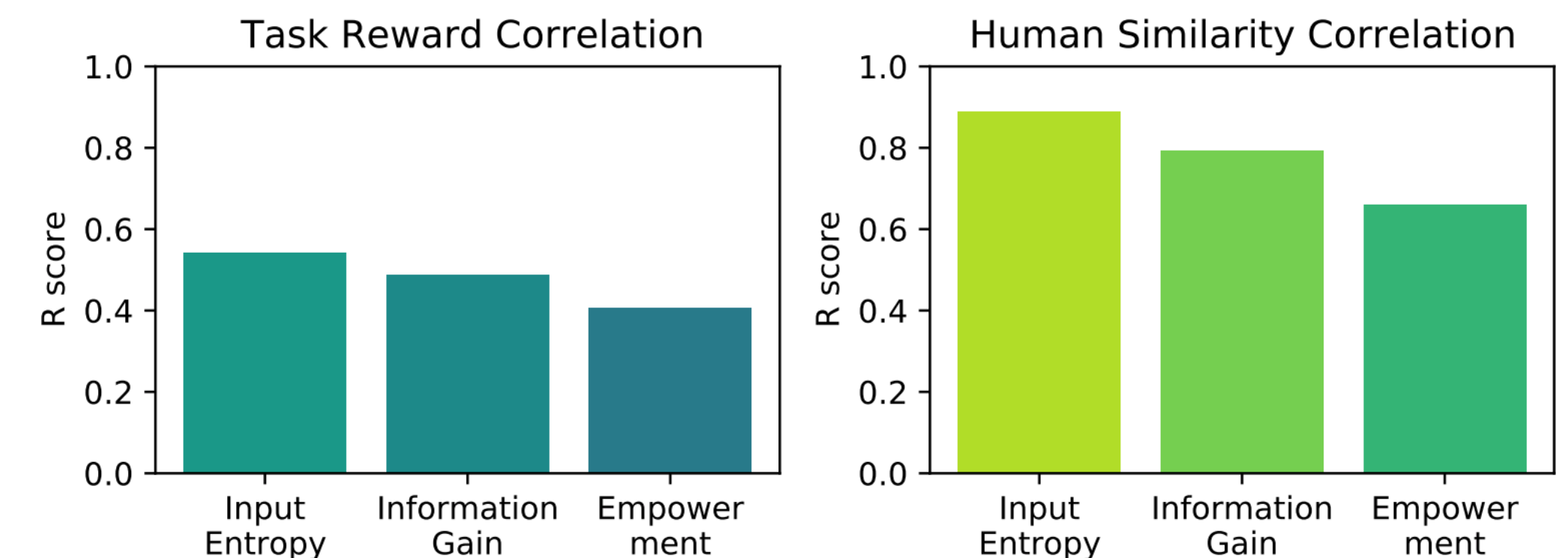
- **What** are they exactly curious about?
  - Intrinsic reward, e.g., novelty
  - Opposed to extrinsic reward, e.g., monetary value
- **Why** are humans and animals curious?
  - Quantifying the benefits of the intrinsically motivated actions in terms of the agent's ability, e.g., gaining knowledge about the environment

# Motivation



Brändle et al., 2023

- There are different definitions of intrinsic rewards in existing works (*what*)
  - But they are environment-dependent
- Solution: top-down models of curiosity, as the optimal mechanism for reaching a particular objective (*why*)
  - They are also environment-dependent



Matusch et al., 2021

# Method overview

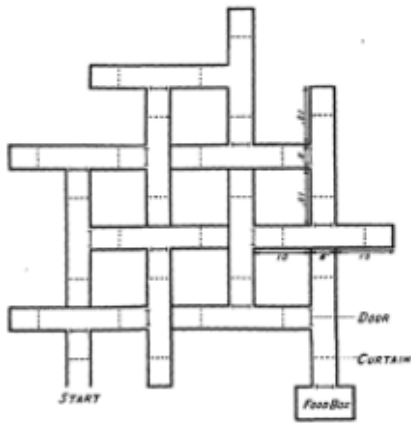
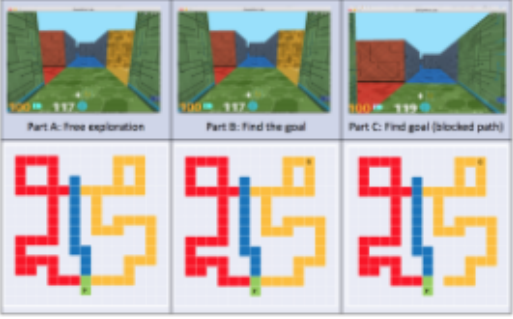
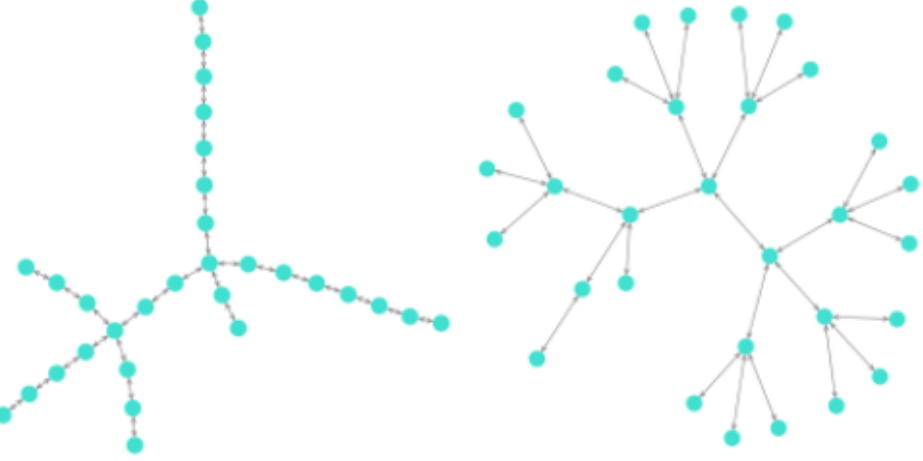


# Method overview

1. Generate environments systematically
2. Compare six intrinsic rewards
  - Novelty, surprise, information gain, empowerment, MOP and SPIE
3. Compare three potential objectives
  - Environment exploration, model accuracy and uniform state visitation

# 0. RL agent

- Environment
  - Discrete states and transitions
- Agent
  - Learns to navigate
  - Start with no prior knowledge of the structure

	Exemplar environments in the literature	Similar generated environments
Mazes	  <p>Tolman et al., 1948</p> <p>Kosoy et al., 2010</p>	

# 0. RL agent

Environment feature: transition function  $\hat{P}^{(t)}(s' | s, a) = \frac{C_{s,a \rightarrow s'}^{(t)} + \epsilon}{C_{s,a}^{(t)} + |S| \cdot \epsilon}$

State feature: Q-value learning

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

$$\text{Policy: } \pi_s^{(t)}(a) = \frac{e^{\beta Q^{(t)}(s, a)}}{\sum_{a'} e^{\beta Q^{(t)}(s, a')}} \in [0, 1]$$

# 1. Intrinsic motivation

- Novelty
- Surprise
- Information Gain
- Empowerment
- Maximum Occupancy Principle (MOP)
- Successor-Predecessor Intrinsic Exploration (SPIE)

# 1. Intrinsic motivation

## 1.1 Novelty

Reward infrequent states

The less frequently the agent has visited  $s'$ , the more rewarded it feels by visiting  $s'$

- Reward:  $R_{\text{Novelty}}^{(t)}(s) = -\log p_N^{(t)}(s)$
- The frequency is  $p_N^{(t)}(s) = \frac{C_s^{(t)} + 1}{\sum_{s'} C_{s'}^{(t)} + |S|}$

# 1. Intrinsic motivation

## 1.2 Surprise

Experience unlikely transitions

The less the agent expects to visit  $s'$  (conditioned on  $s$  and  $a$ ), the more reward it feels by visiting  $s'$  (after taking  $a$  in  $s$ )

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

# 1. Intrinsic motivation

## 1.3 Information gain

Reduce *epistemic uncertainty* about the environment

KL divergence of the updated model from the previous model, the more the agent updates its estimated probabilities after transition  $s, a \rightarrow s'$ , the more rewarded it feels.

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

# 1. Intrinsic motivation

## 1.4 Empowerment

Achieve states where its actions lead to a diverse set of predictable outcomes

the more ‘options’ the agent has at state  $s'$ , the more it feels rewarded by visiting  $s'$ .

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

- $$= \max_{p(a)} \left( \mathcal{H}(S') - \mathcal{H}(S' \mid A) \right)$$



# 1. Intrinsic motivation

## 1.5 MOP

A regularized surprise

Experiencing unlikely transitions but also for maintaining a high-entropy policy.

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

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

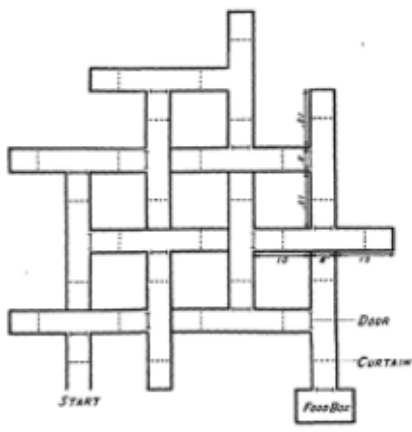
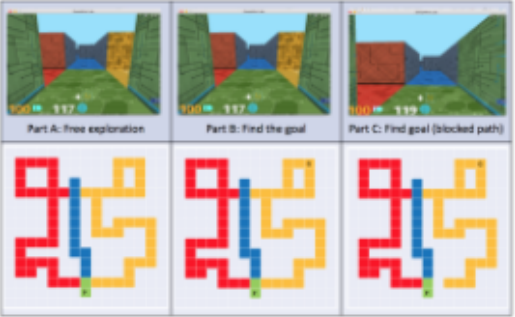
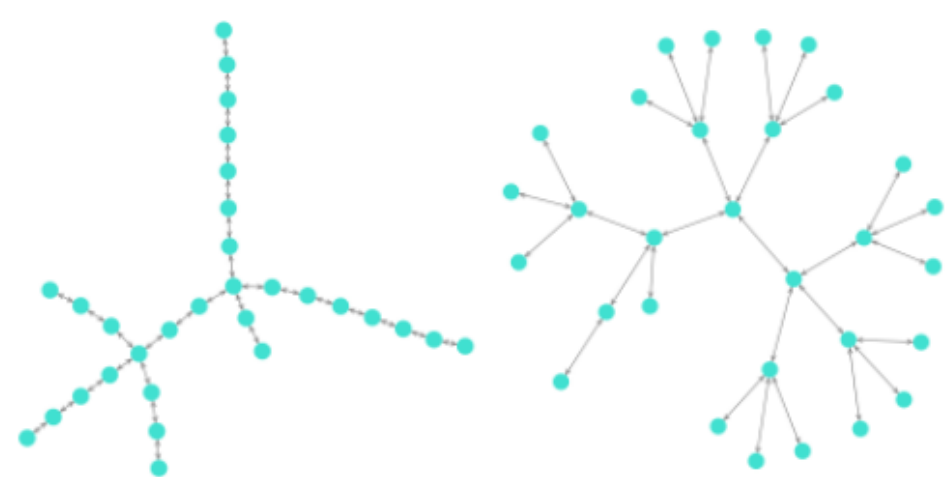

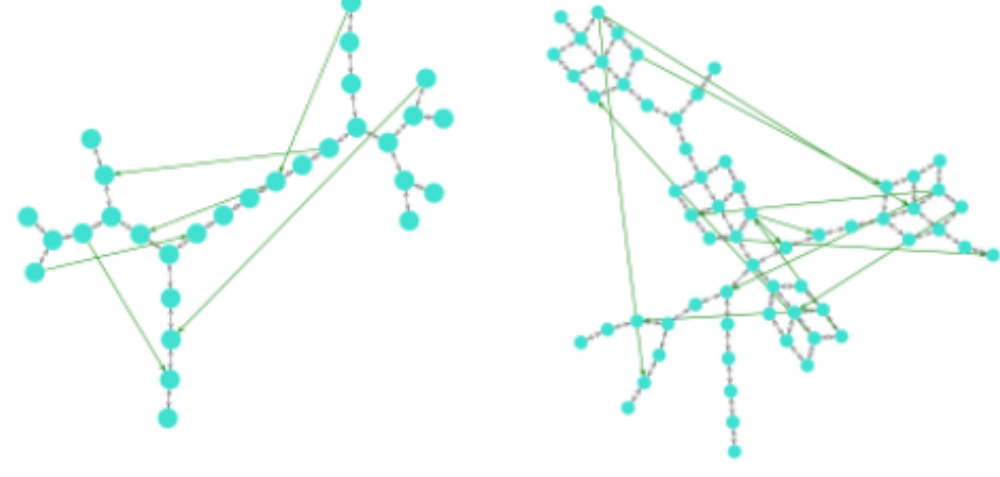
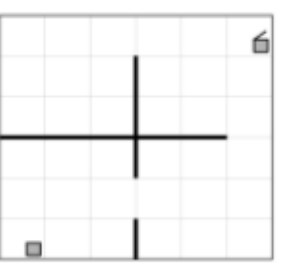
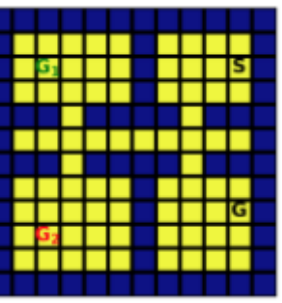
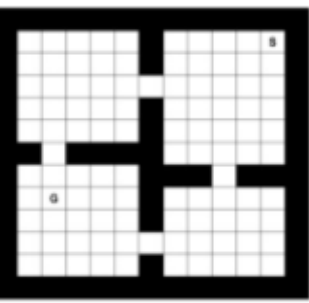
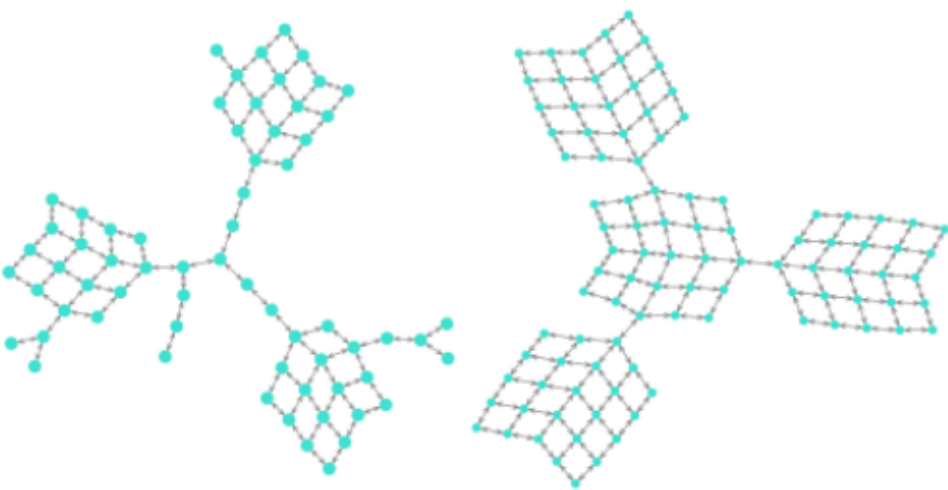
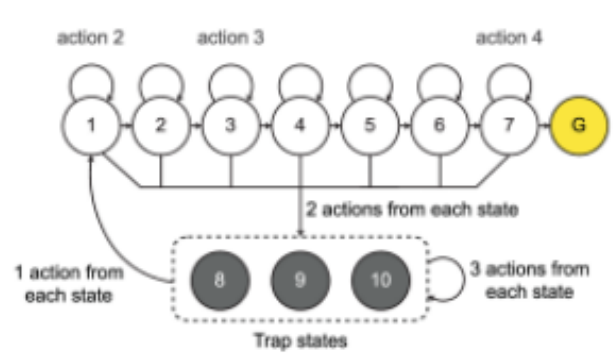
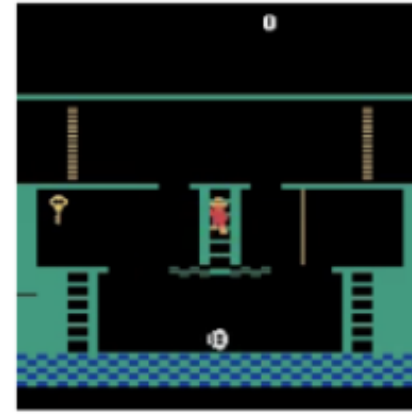
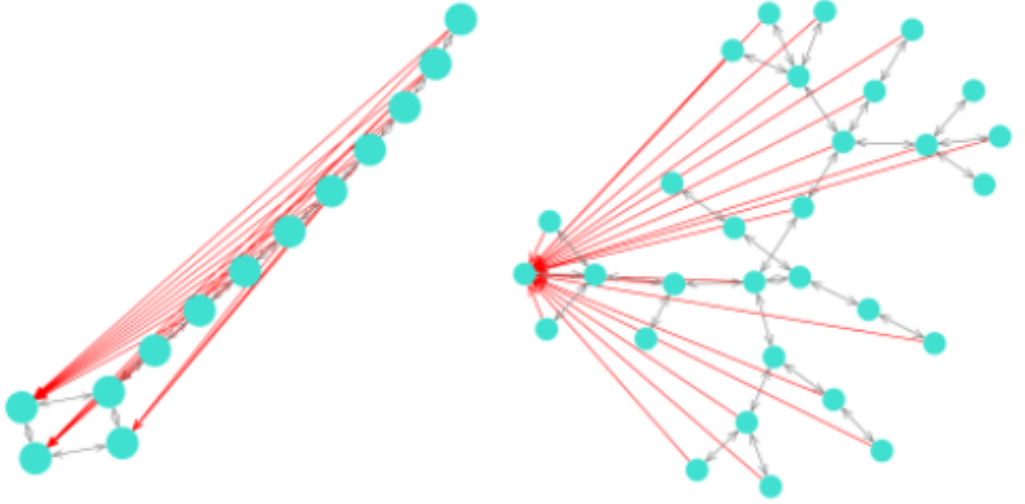
# 1. Intrinsic motivation

## 1.6 SPIE

Visiting rare states as well as those that are critical for reaching isolated regions

- 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
- $R_{SPIE}^{(t)}(s, a, s') = \hat{M}^{(t)}[s, s'] - \left\| \hat{M}^{(t)}[\cdot, s'] \right\|_1$

# 2. Environment generation

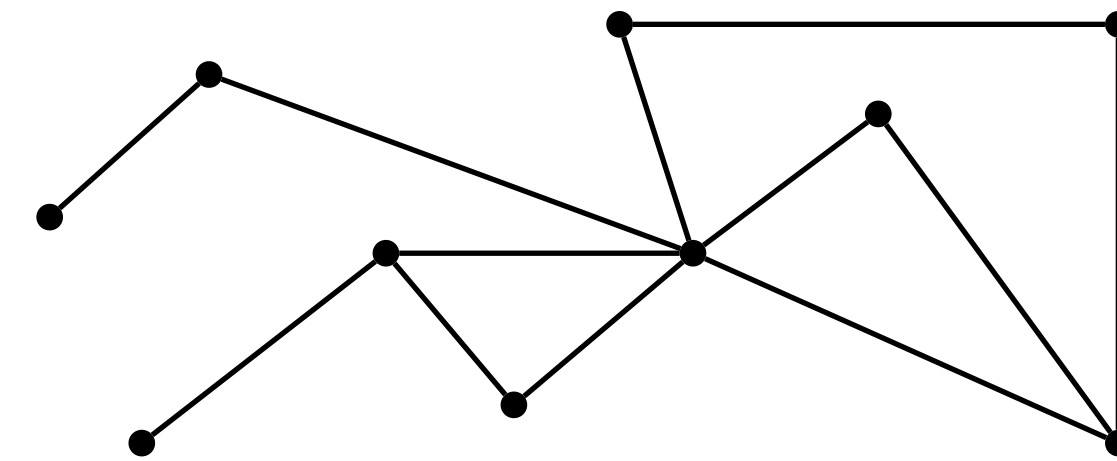
	Exemplar environments in the literature	Similar generated environments			
Mazes	<div><p>Tolman et al., 1948</p></div> <div><p>Kosoy et al., 2010</p></div>		Long-range connections	 <p>Viswanathan et al., 2016</p>	
Grid worlds	<div><p>Singh et al., 2010</p></div> <div><p>Yu et al., 2024</p></div> <div><p>Botvinick et al., 2009</p></div>		Sink states	<div><p>Xu et al., 2021</p></div> <div><p>Matusch et al., 2021</p></div>	

## 2. Environment generation

1. Create a maze with a branching structure
2. Integrate grid-like rooms within the maze
3. Assign each room to one (and exactly one) of the following properties:
  - Sink, source, stochastic, neural



Given *number of states* & *branching rate*  
E.g., 10 & 0.2

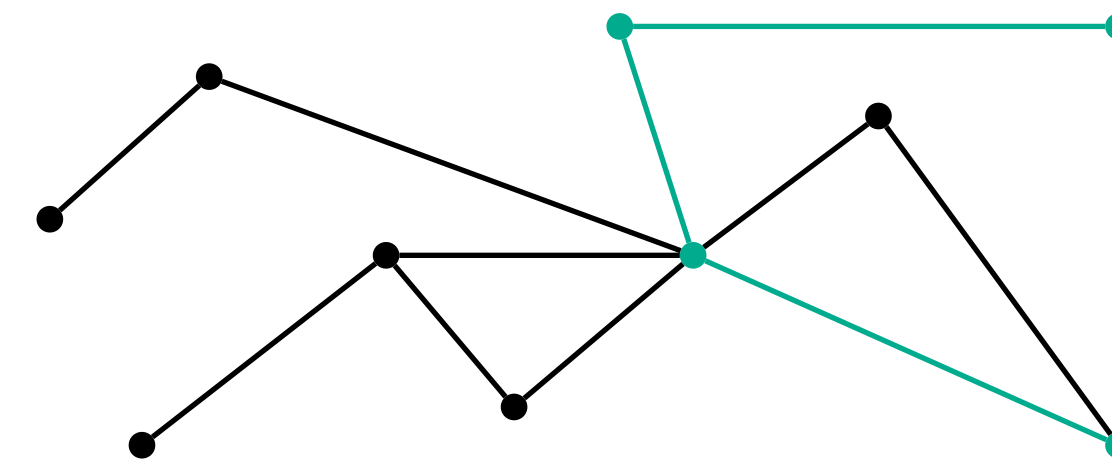


## 2. Environment generation

1. Create a maze with a branching structure
2. Integrate grid-like rooms within the maze
3. Assign each room to one (and exactly one) of the following properties:
  - Sink, source, stochastic, neural



Given *fraction of states to be chosen*  
& *size of the rooms*  
Transform them into a room

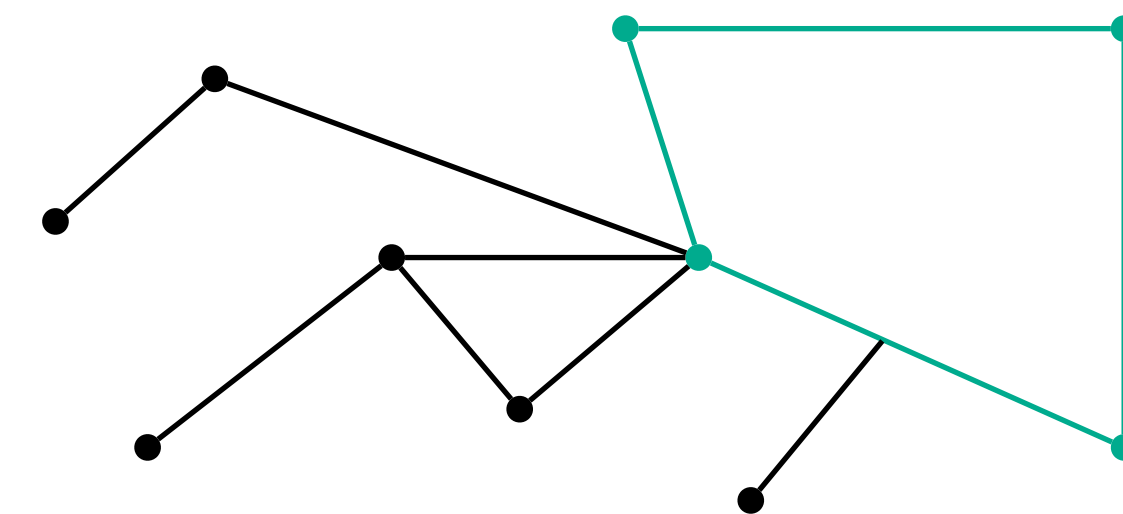
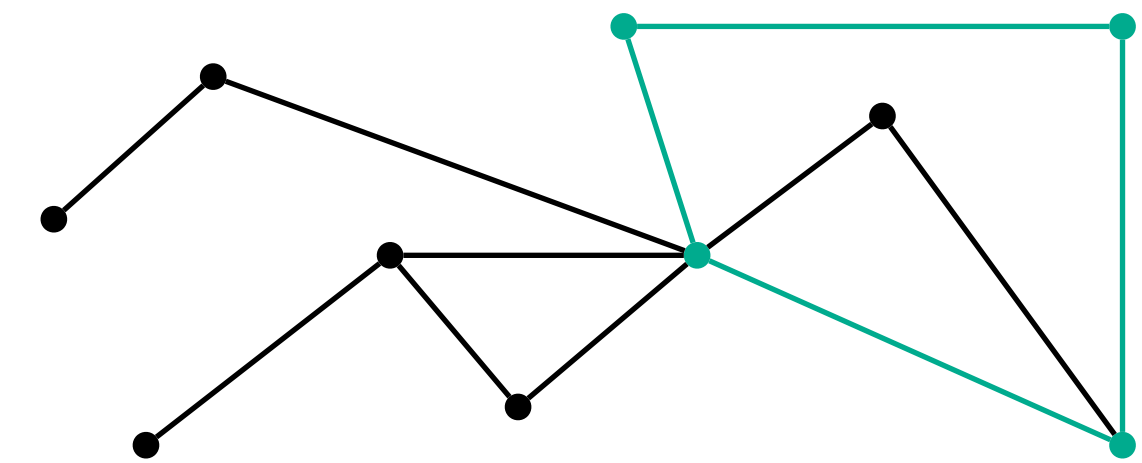


## 2. Environment generation

1. Create a maze with a branching structure
2. Integrate grid-like rooms within the maze
3. Assign each room to one (and exactly one) of the following properties:
  - Sink, source, stochastic, neural



Given *fraction of states to be chosen*  
& *size of the rooms*  
Transform them into a room



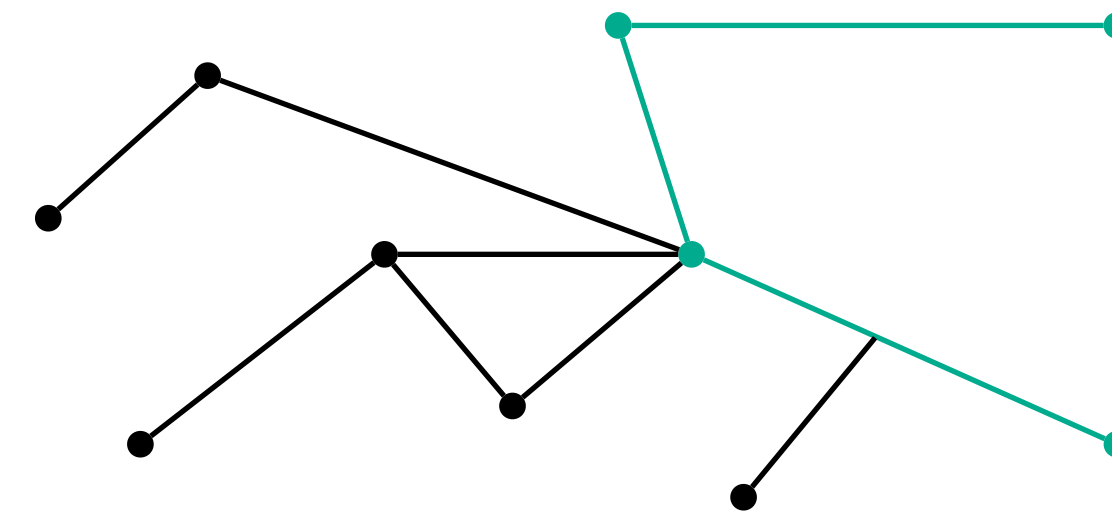
Neighbors of a transformed state are connected to the middle of the room borders (maximum 4 neighbors, one for each side of the square room).

## 2. Environment generation

1. Create a maze with a branching structure
2. Integrate grid-like rooms within the maze
3. Assign each room to one (and exactly one) of the following properties:
  - Sink, source, stochastic, **neural**



Given *room property*  
Transform them into a room

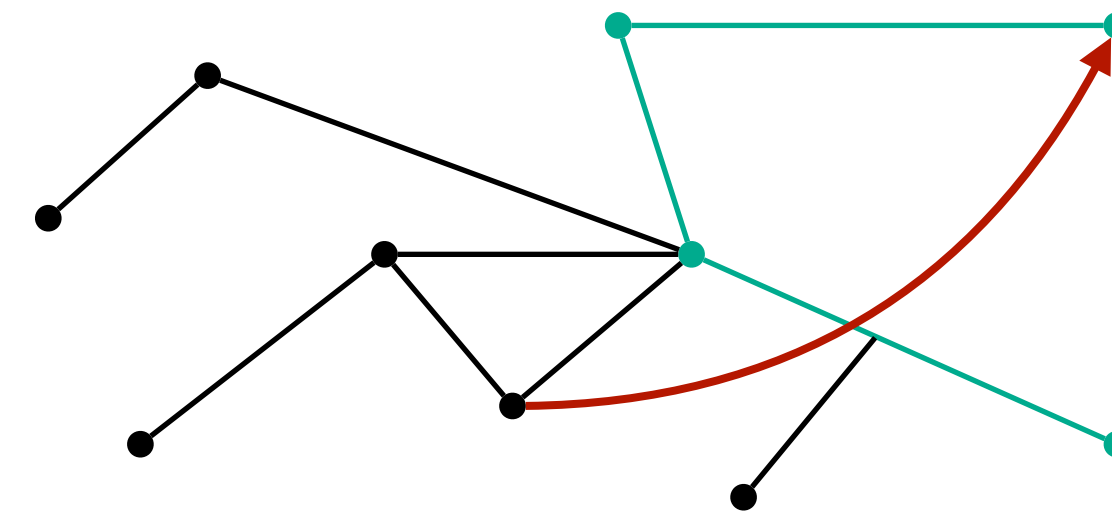


## 2. Environment generation

1. Create a maze with a branching structure
2. Integrate grid-like rooms within the maze
3. Assign each room to one (and exactly one) of the following properties:
  - **Sink**, source, stochastic, neural



Given *room property*  
Transform them into a room



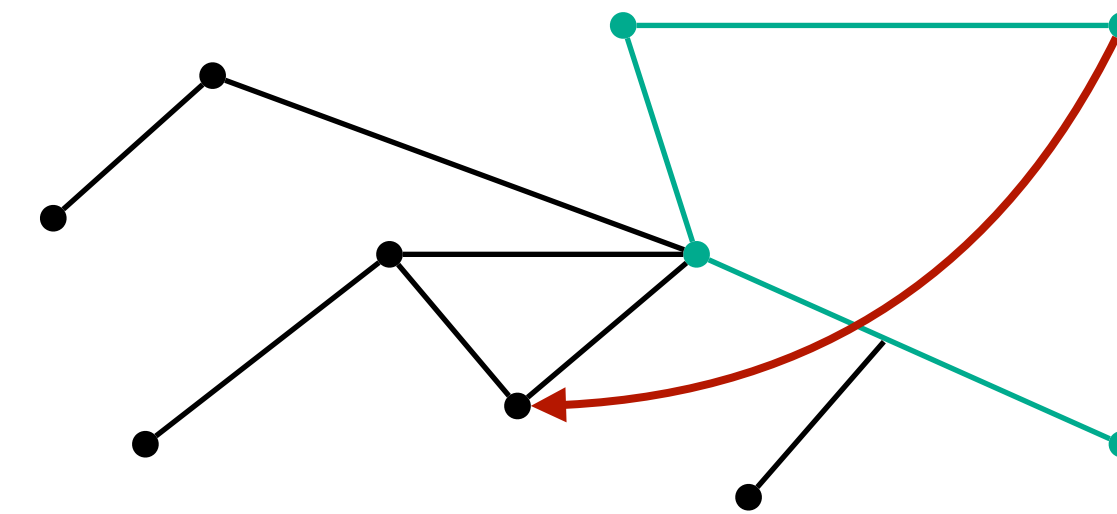


## 2. Environment generation

1. Create a maze with a branching structure
2. Integrate grid-like rooms within the maze
3. Assign each room to one (and exactly one) of the following properties:
  - Sink, **source**, stochastic, neural



Given *room property*  
Transform them into a room

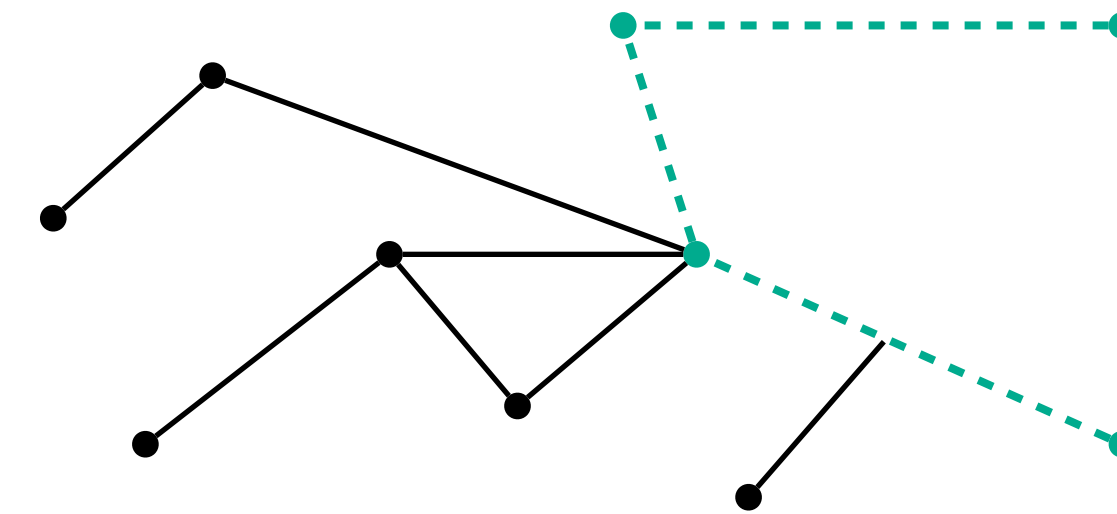


## 2. Environment generation

1. Create a maze with a branching structure
2. Integrate grid-like rooms within the maze
3. Assign each room to one (and exactly one) of the following properties:
  - Sink, source, **stochastic**, neural

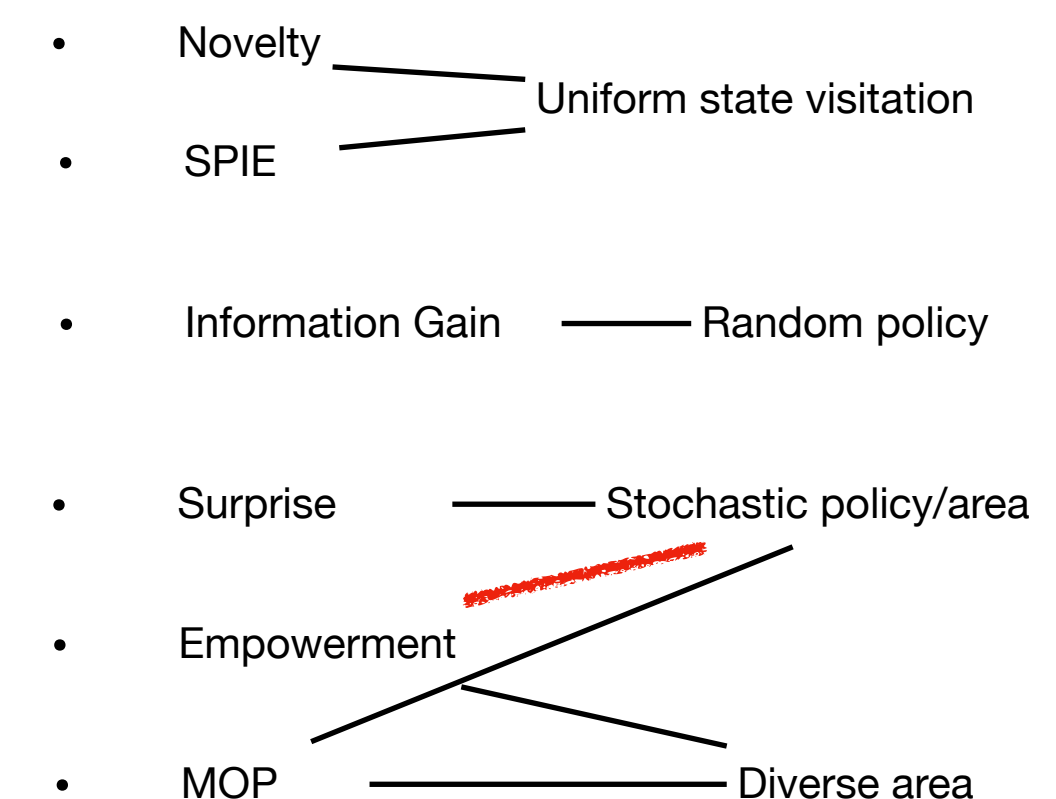


Given *room property*  
Transform them into a room



# 3. Objectives

1. **Environment exploration:** the fraction of unvisited states
2. **Model accuracy:** difference between the estimated transition probabilities and the ground truth
3. **Uniform state visitation:** difference between the agent's state visitation frequency and the uniform distribution



# Results

# Environment exploration

## % of unvisited exploration

Novelty

SPIE

Information  
Gain

Surprise

MOP

Empowerment

Uniform state  
visitation

Random policy

Stochastic  
policy/area

Diverse area

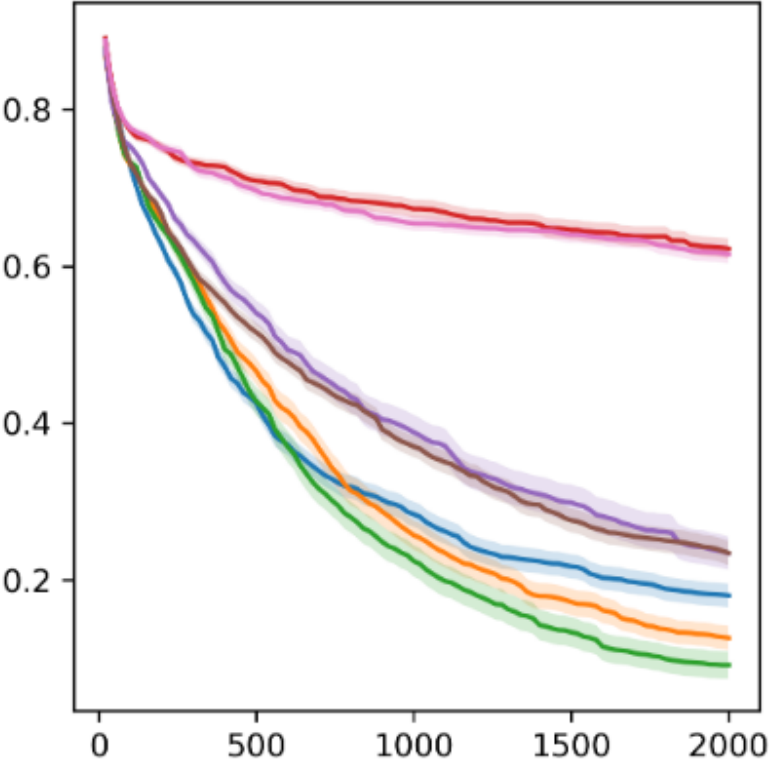
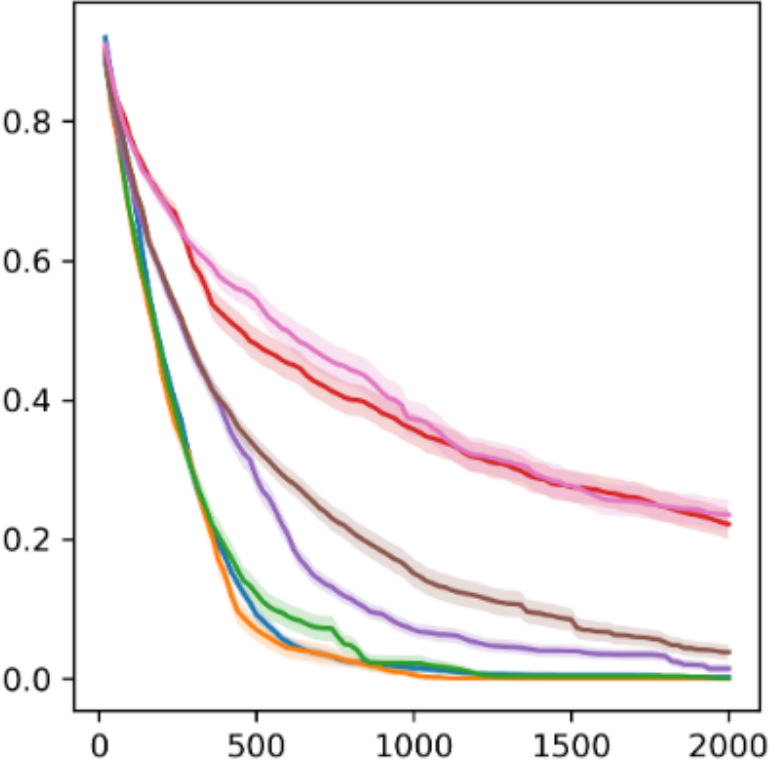
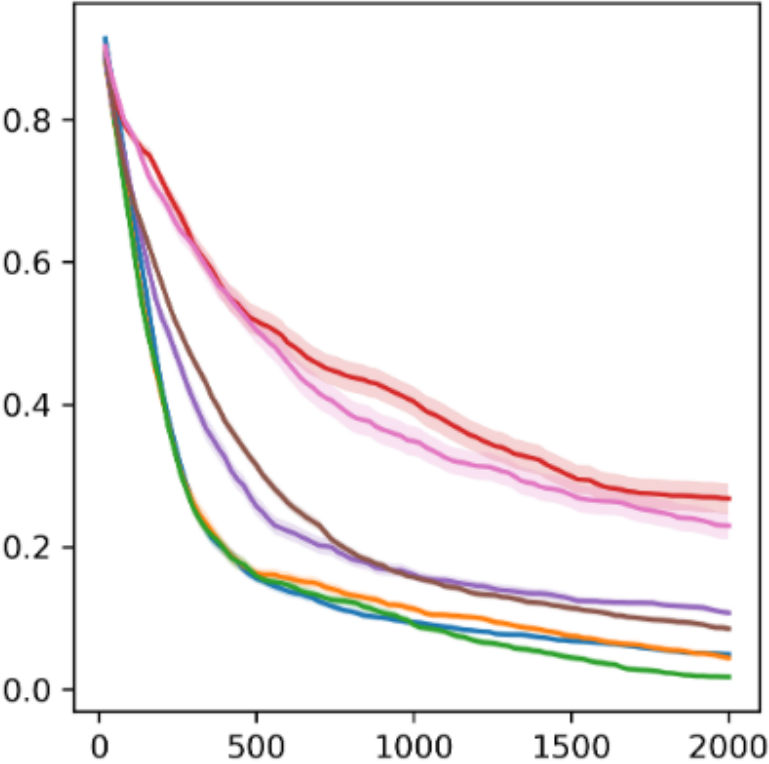
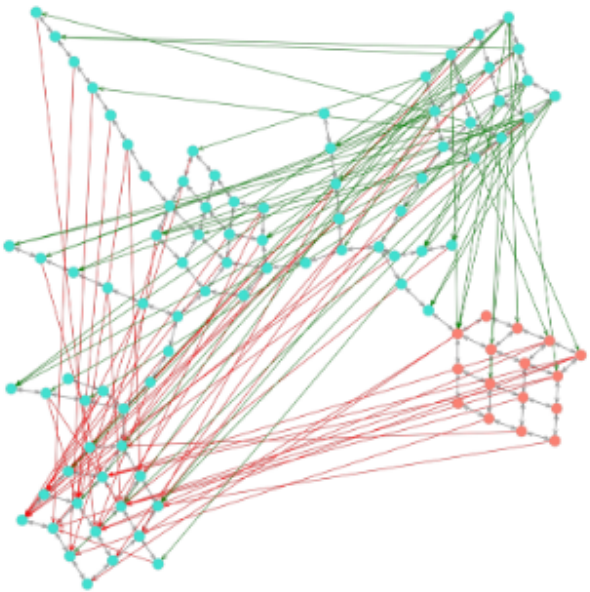
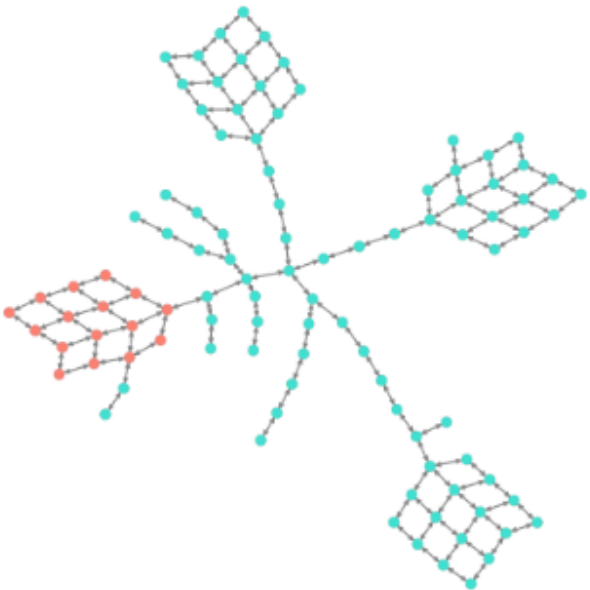
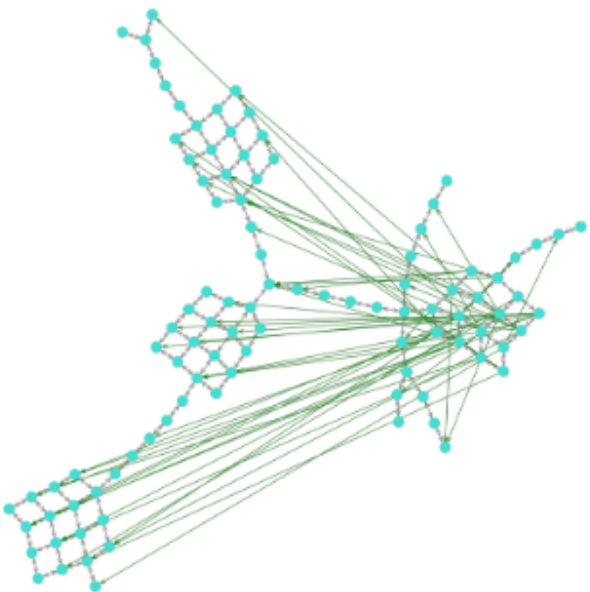
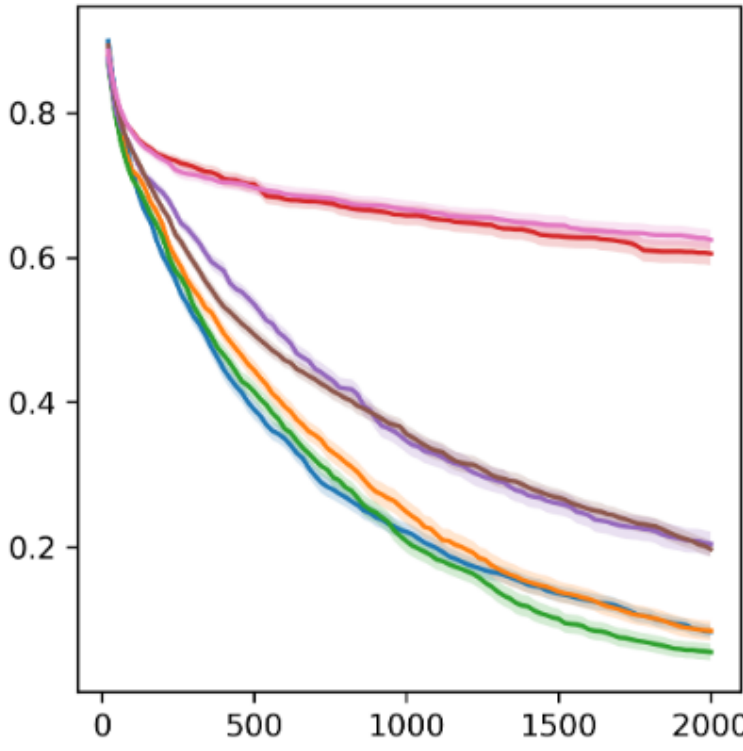
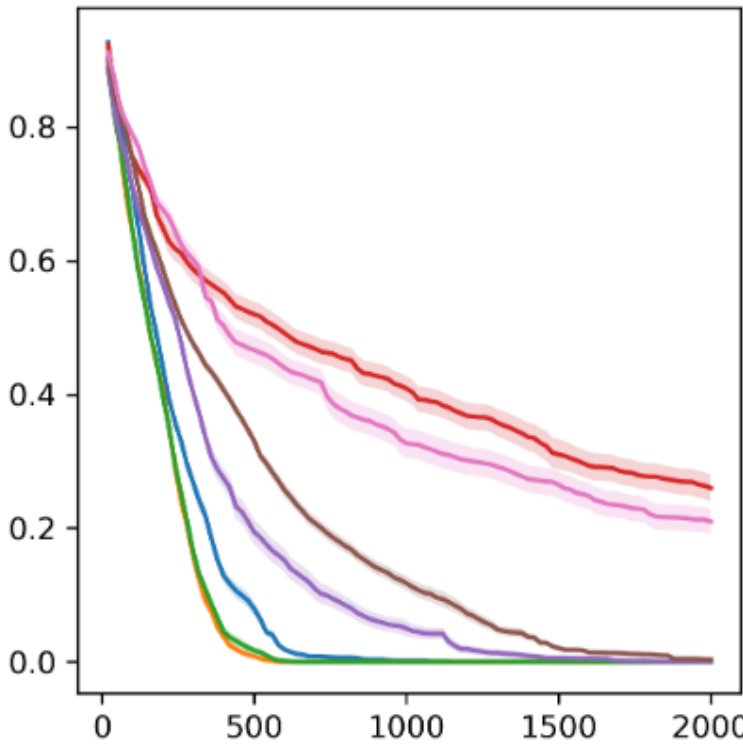
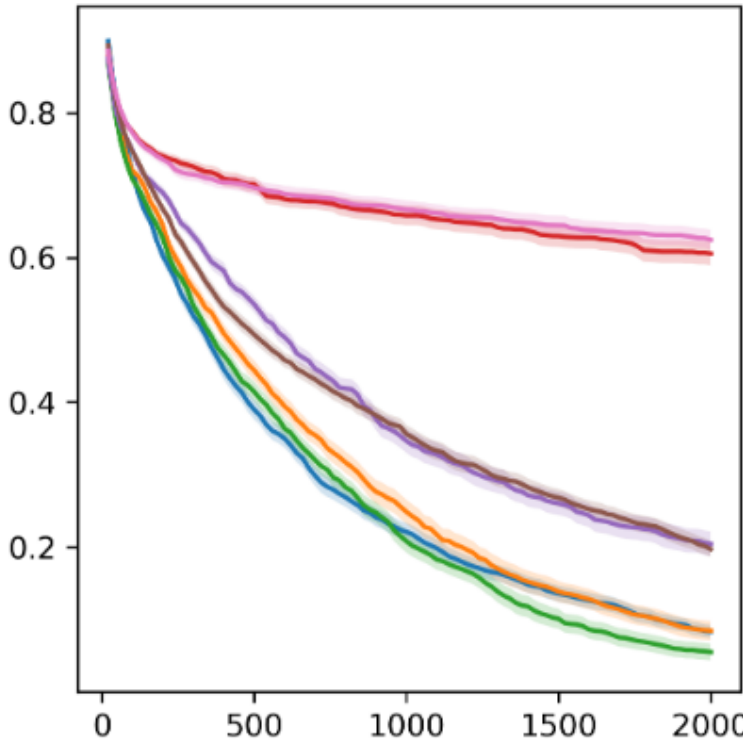
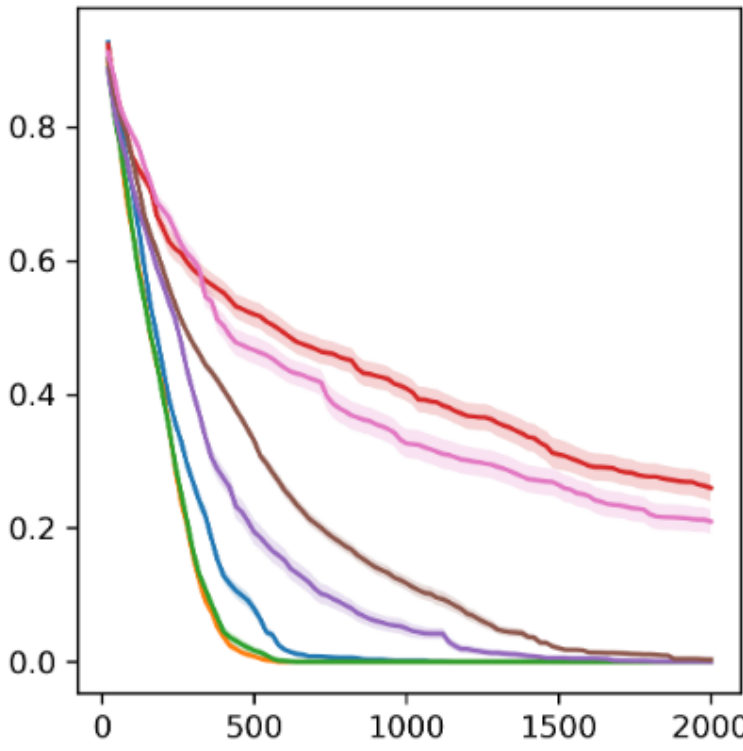
Neutral

Sink

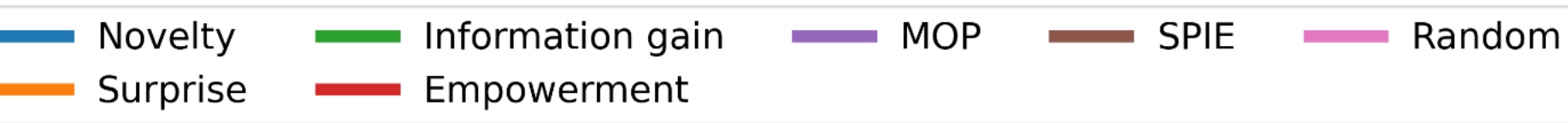
Source

Stochastic

Mixed



Lower is better





# Model accuracy

$$\text{RMSE} \left( \hat{P}_{s,a}, P_{s,a} \right)$$

Novelty

SPIE

Information  
Gain

Surprise

MOP

Empowerment

Uniform state

visitation

Random policy

Stochastic  
policy/area

Diverse area

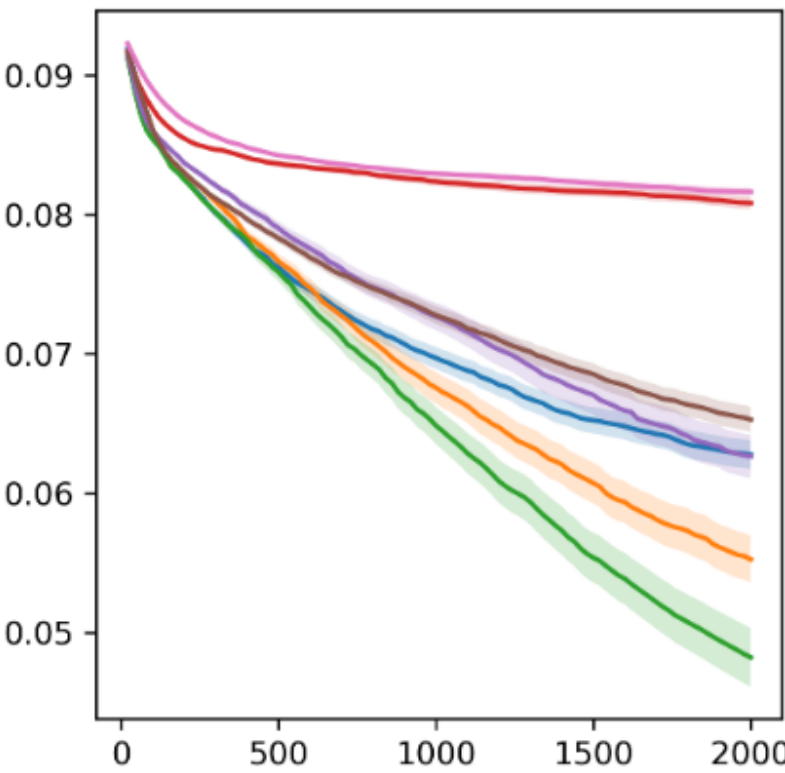
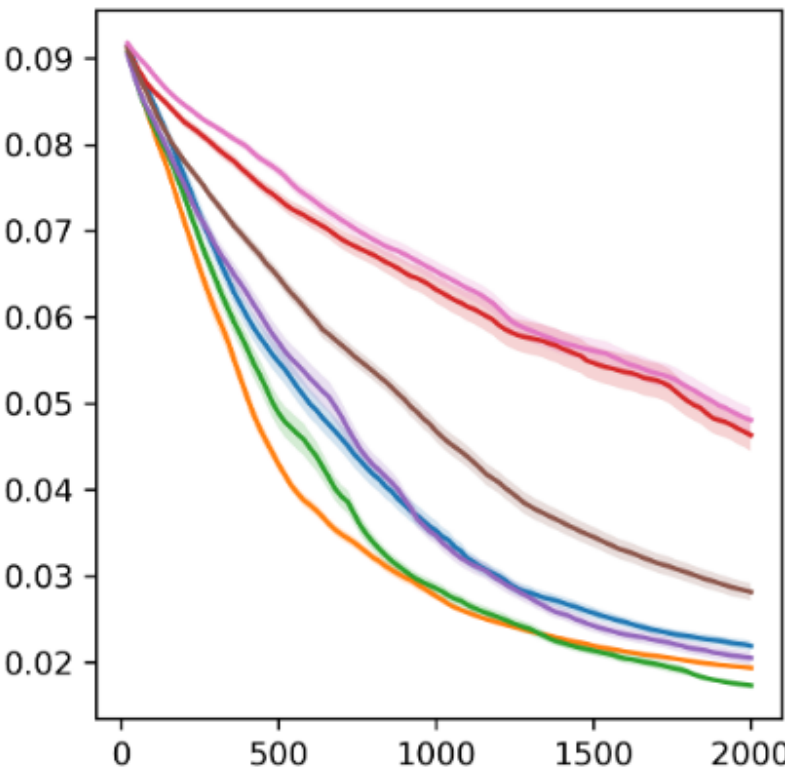
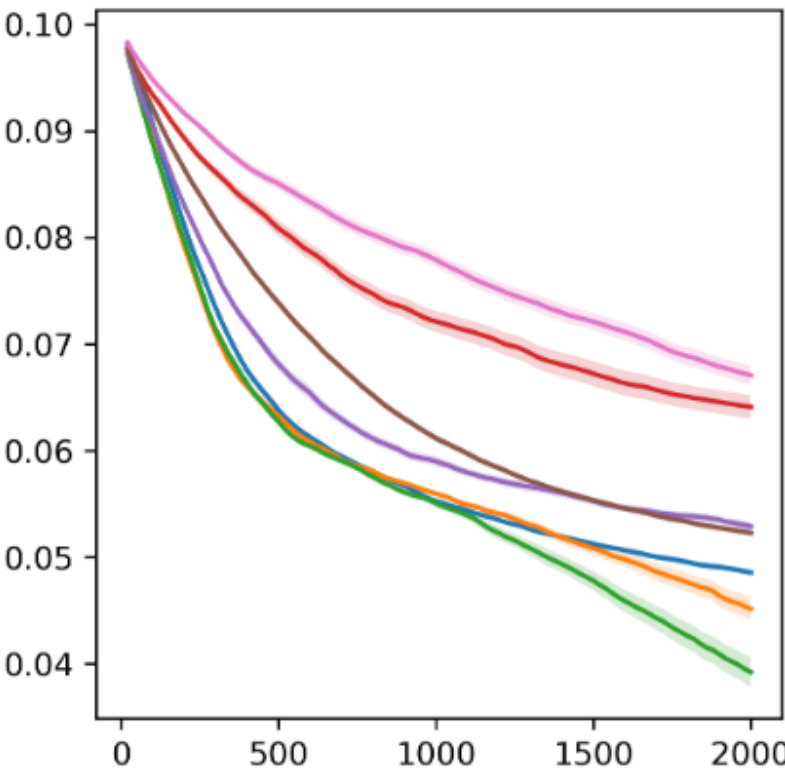
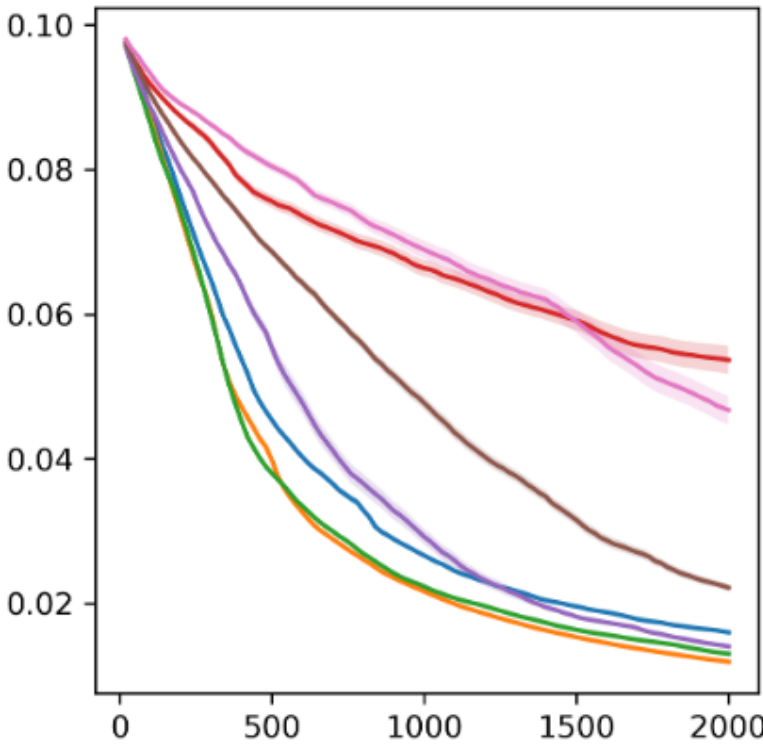
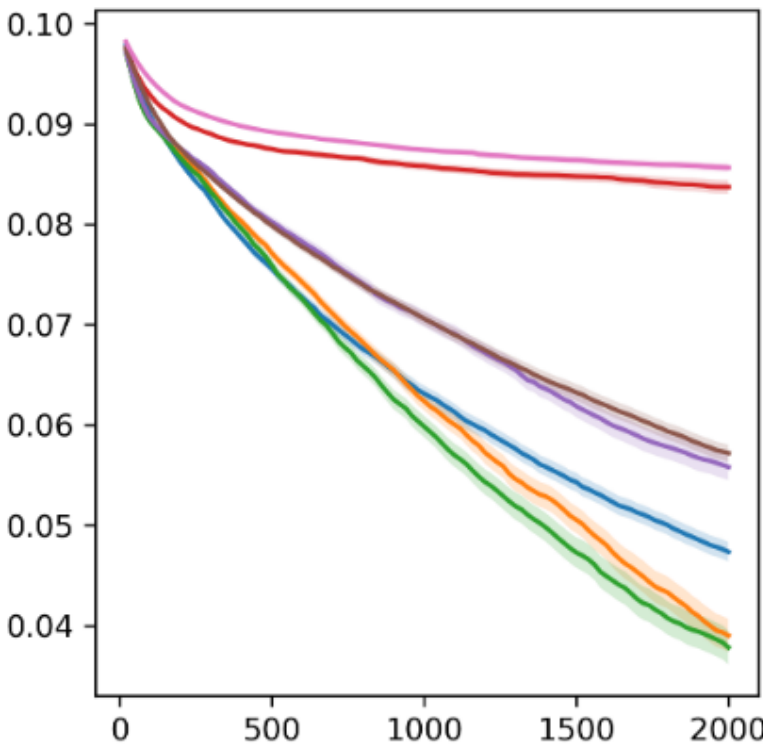
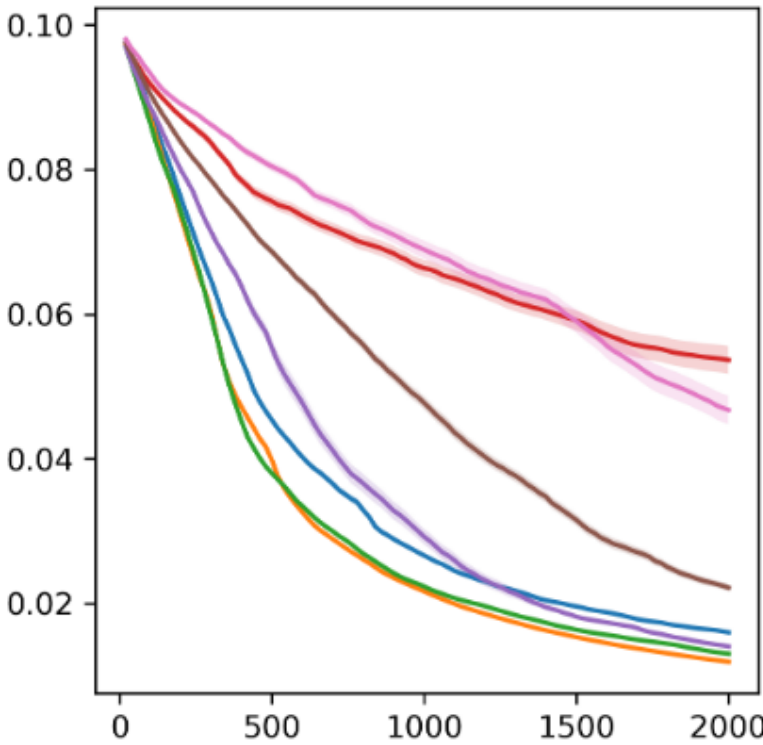
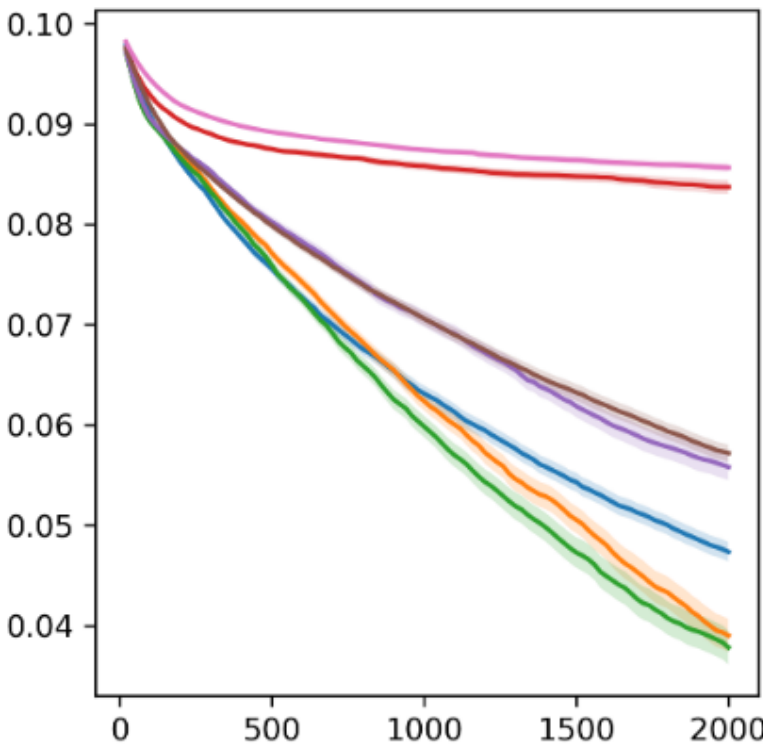
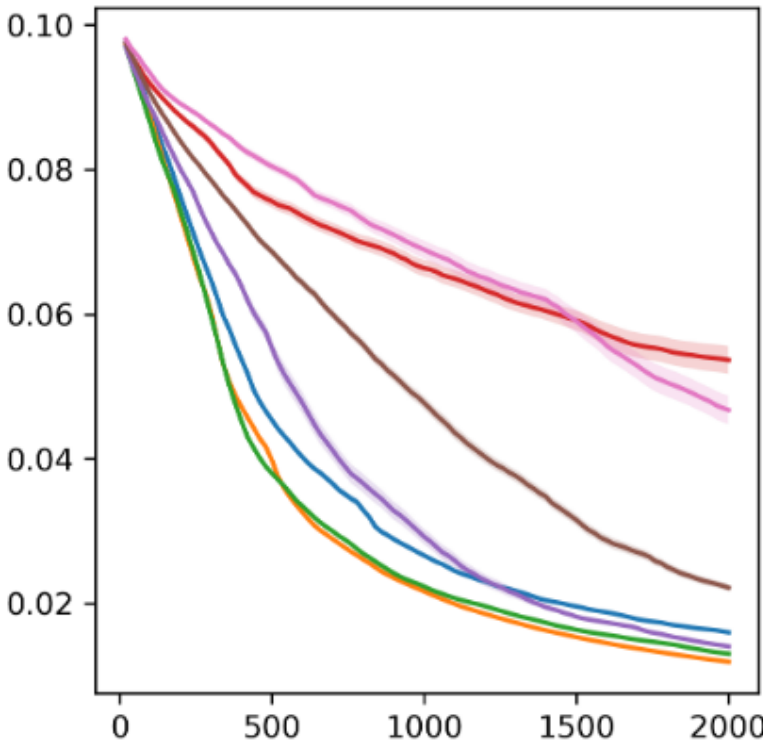
Neutral

Sink

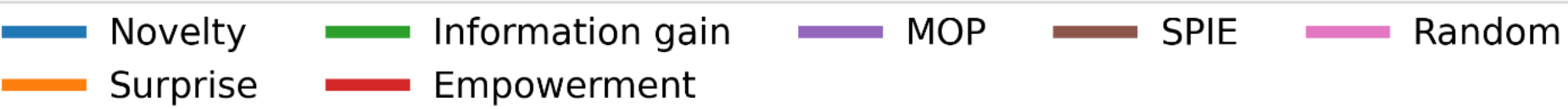
Source

Stochastic

Mixed



Lower is better



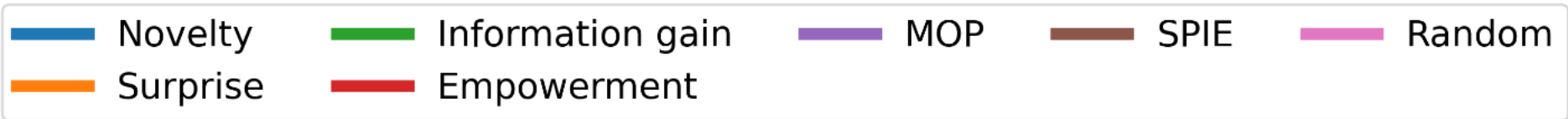


# Uniform state visitation

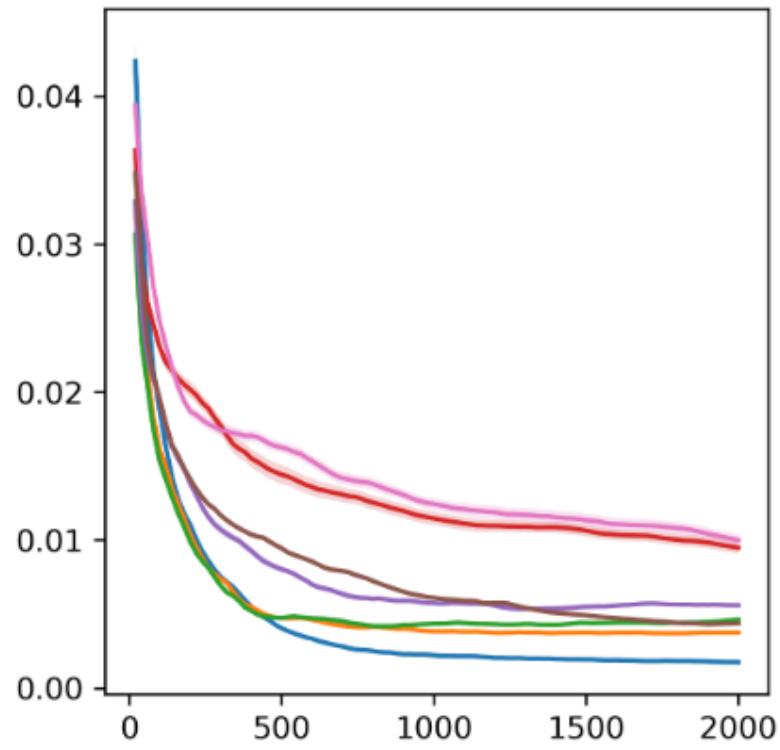
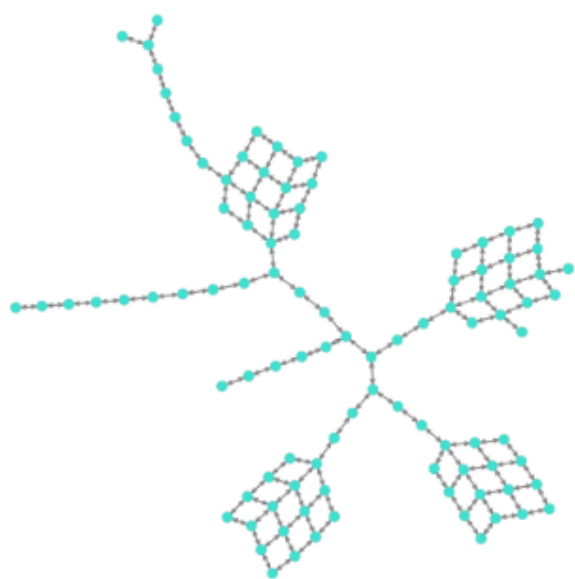
$$\text{RMSE} \left( \hat{P}(S), \text{Unif}(S) \right)$$

- Novelty
- SPIE
- Information Gain
- Surprise
- MOP
- Empowerment
- Uniform state visitation
- Random policy
- Stochastic policy/area
- Diverse area

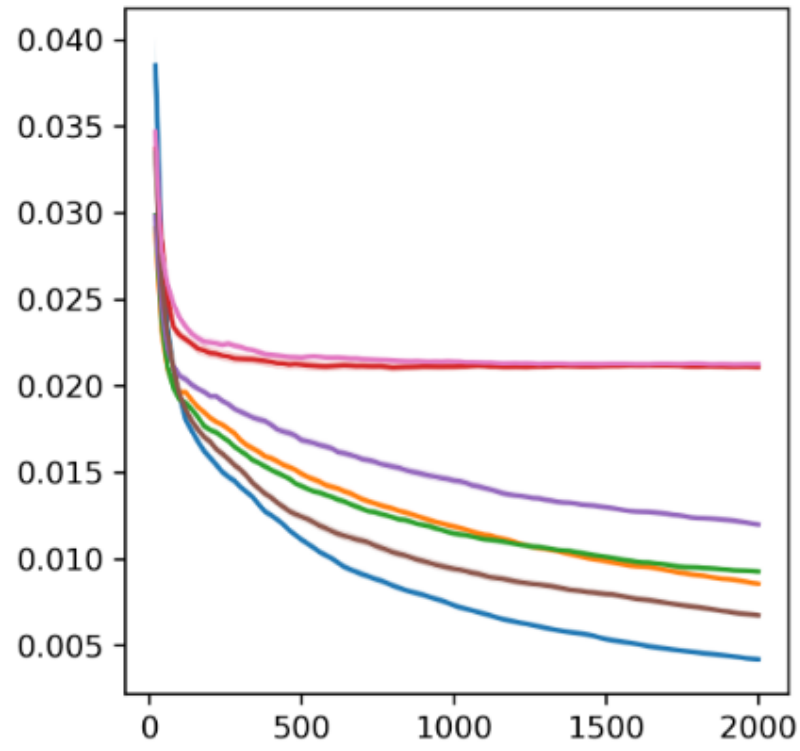
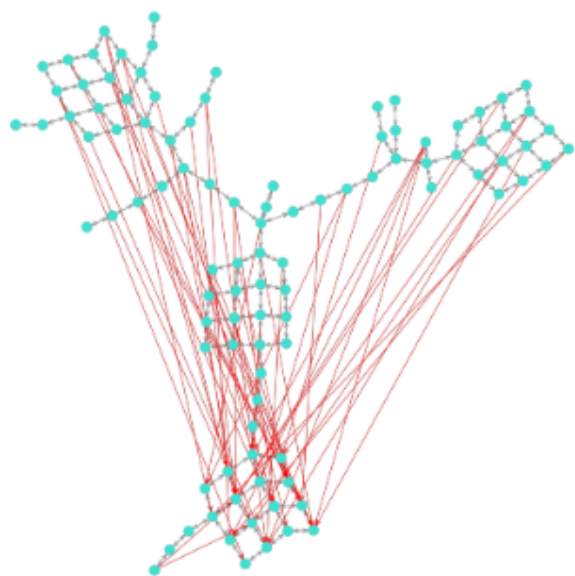
Lower is better



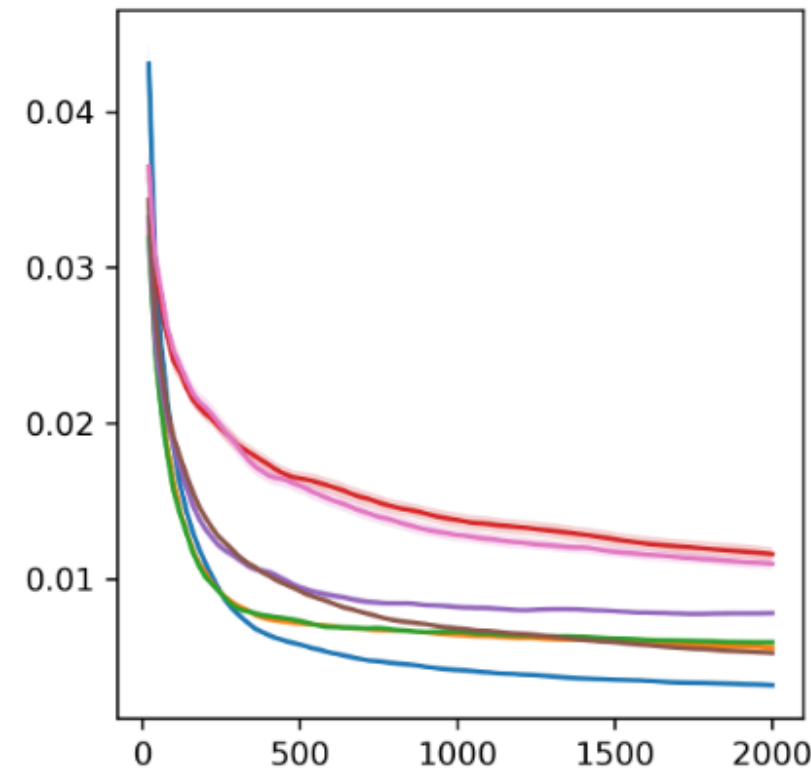
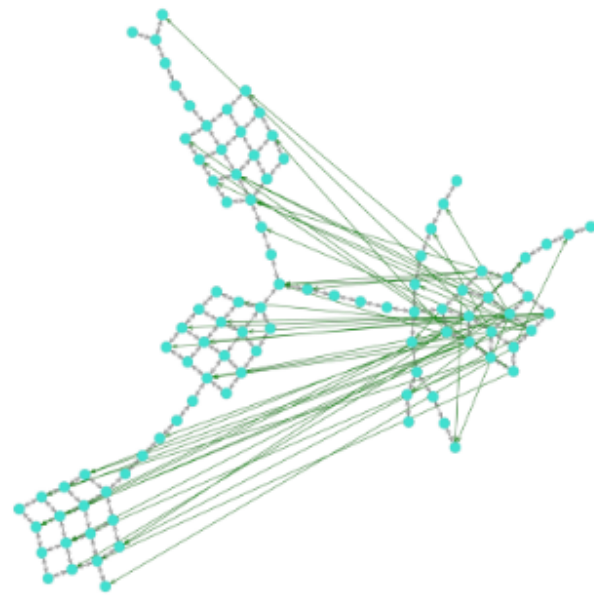
Neutral



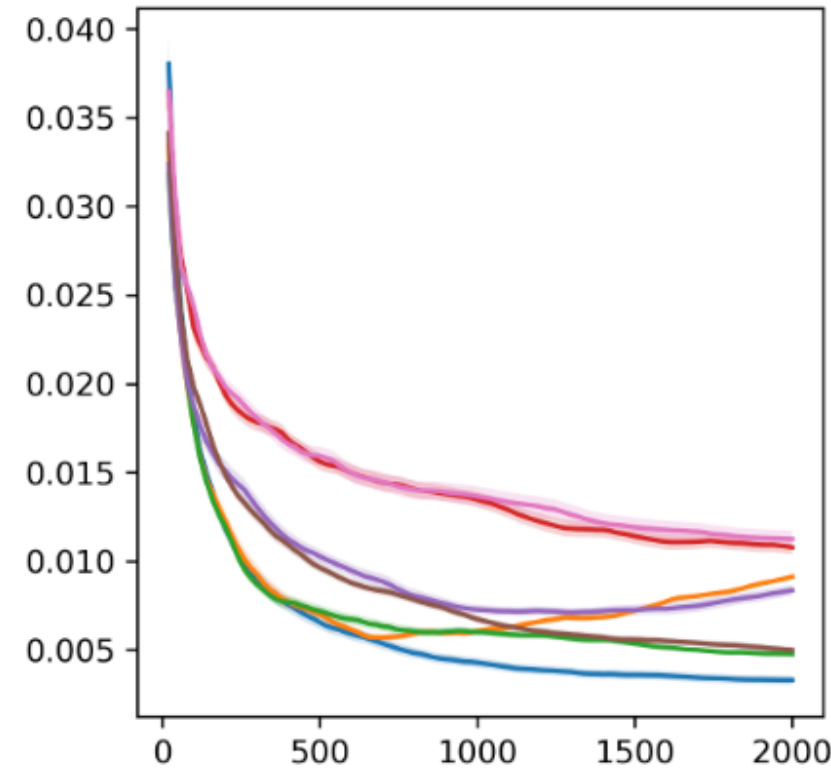
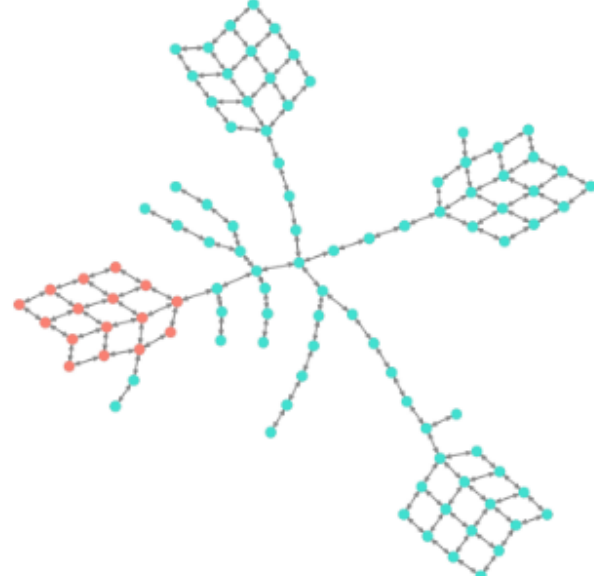
Sink



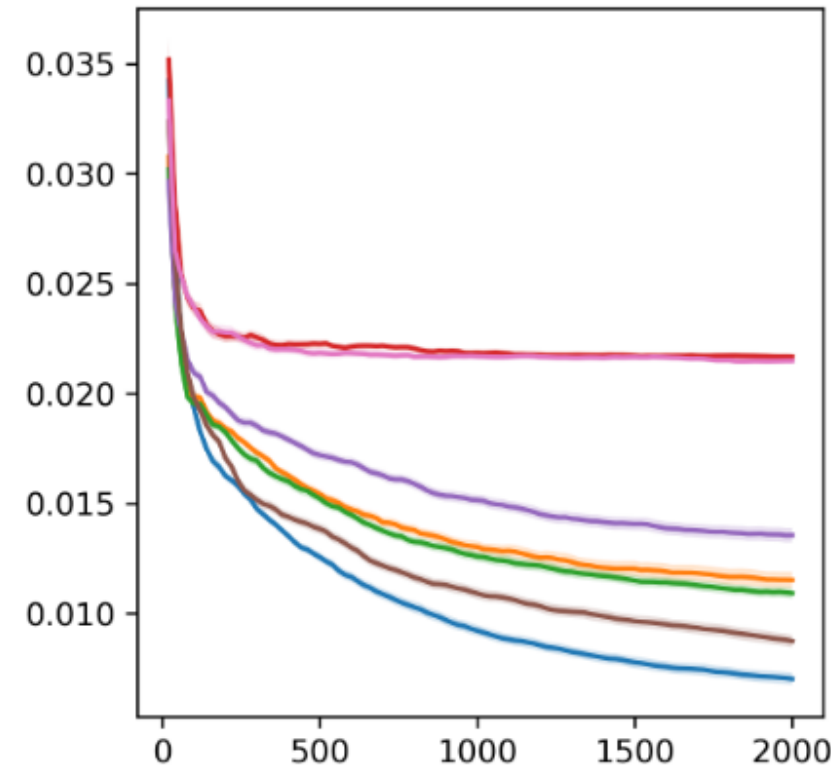
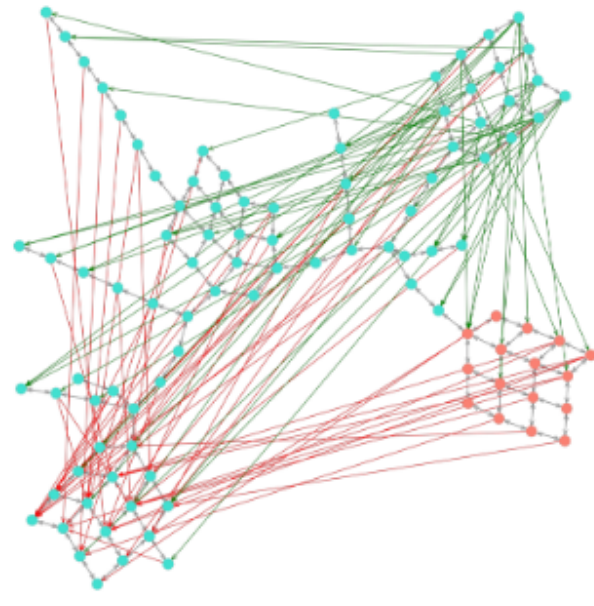
Source



Stochastic



Mixed



# Performance analysis

## Among intrinsic rewards

- Novelty-seeking consistently have the best performance according to Measure 3 (uniform state visitation)
- Surprise or Information Gain excel on Measures 1 and 2 but perform consistently worse than novelty-seeking agents for Measure 3
- Empowerment perform poorly across all scenarios; this is essentially because they avoid unknown regions, which are perceived as non-empowering due to uncertainty.
- MOP and SPIE, perform worse than agents seeking surprise, information-gain, or even novelty on Measures 1 and 2.



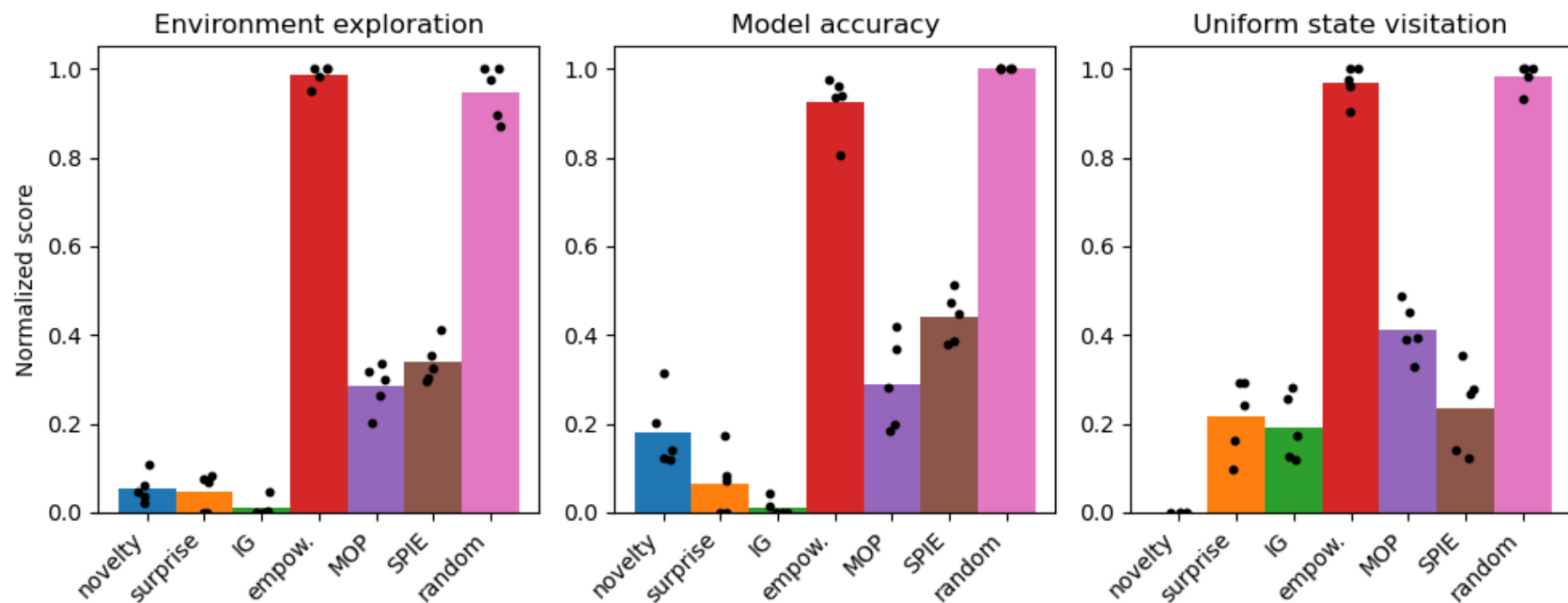
# Performance analysis

## Among environments

- Different environment types affect performance in distinct ways:
  - Neutral environments offer a good reference point.
  - Sink environments can more vividly show the differences in the performance of different agents
- Source environments benefit Surprise and Information Gain, is specifically detrimental for Novelty
- Stochastic environments Surprise and MOP tend to stay in the stochastic room after learning sufficiently about the environment, resulting in poor performance on Measure

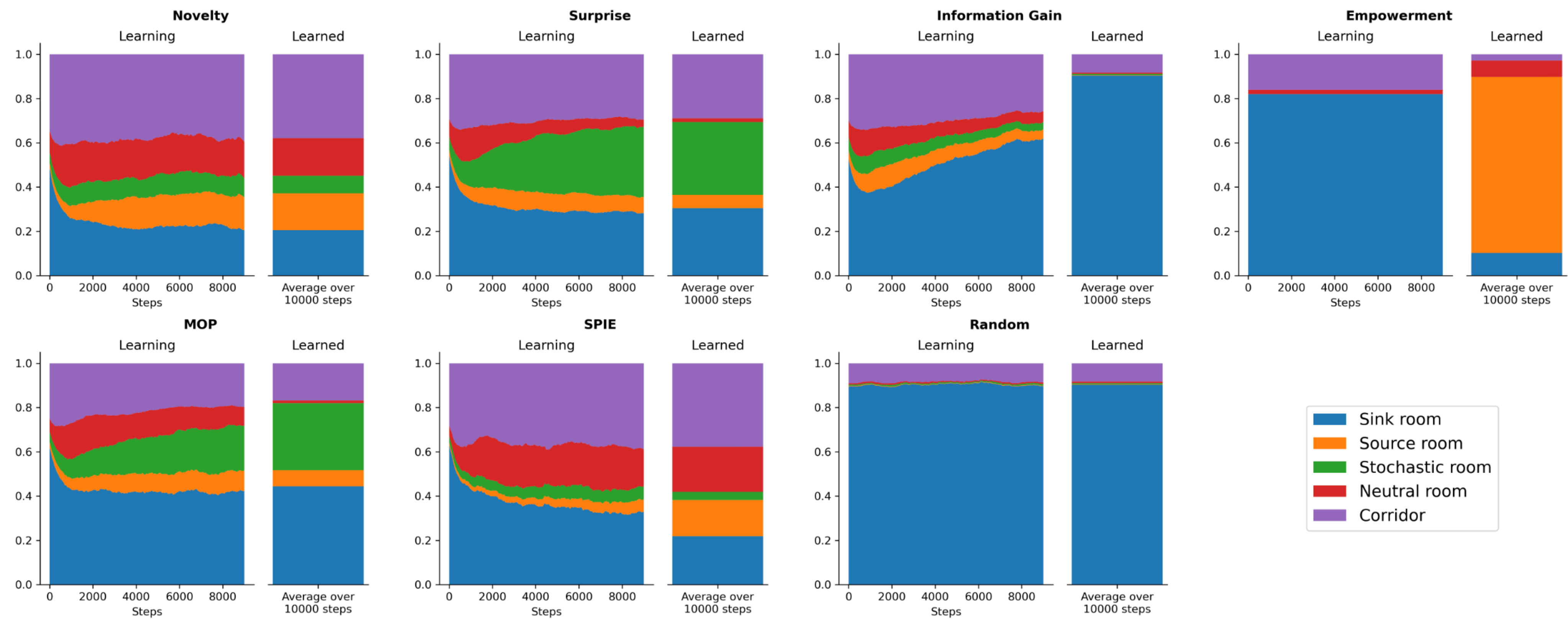
# Combination of intrinsic rewards

## Novelty + information gain



# Performance analysis

## Time spent in each room



# Overview

- The performance of curiosity-driven agents depends highly on the structure of their environment
- Different intrinsic motivations produce different exploratory patterns
- Information gain and novelty are the two most effective drivers of curiosity; information gain helps with exploring and understanding environments better  
-> NOTE that this is also dependent on the objective functions

**Applying to knowledge  
acquisition**

# Environment / knowledge network

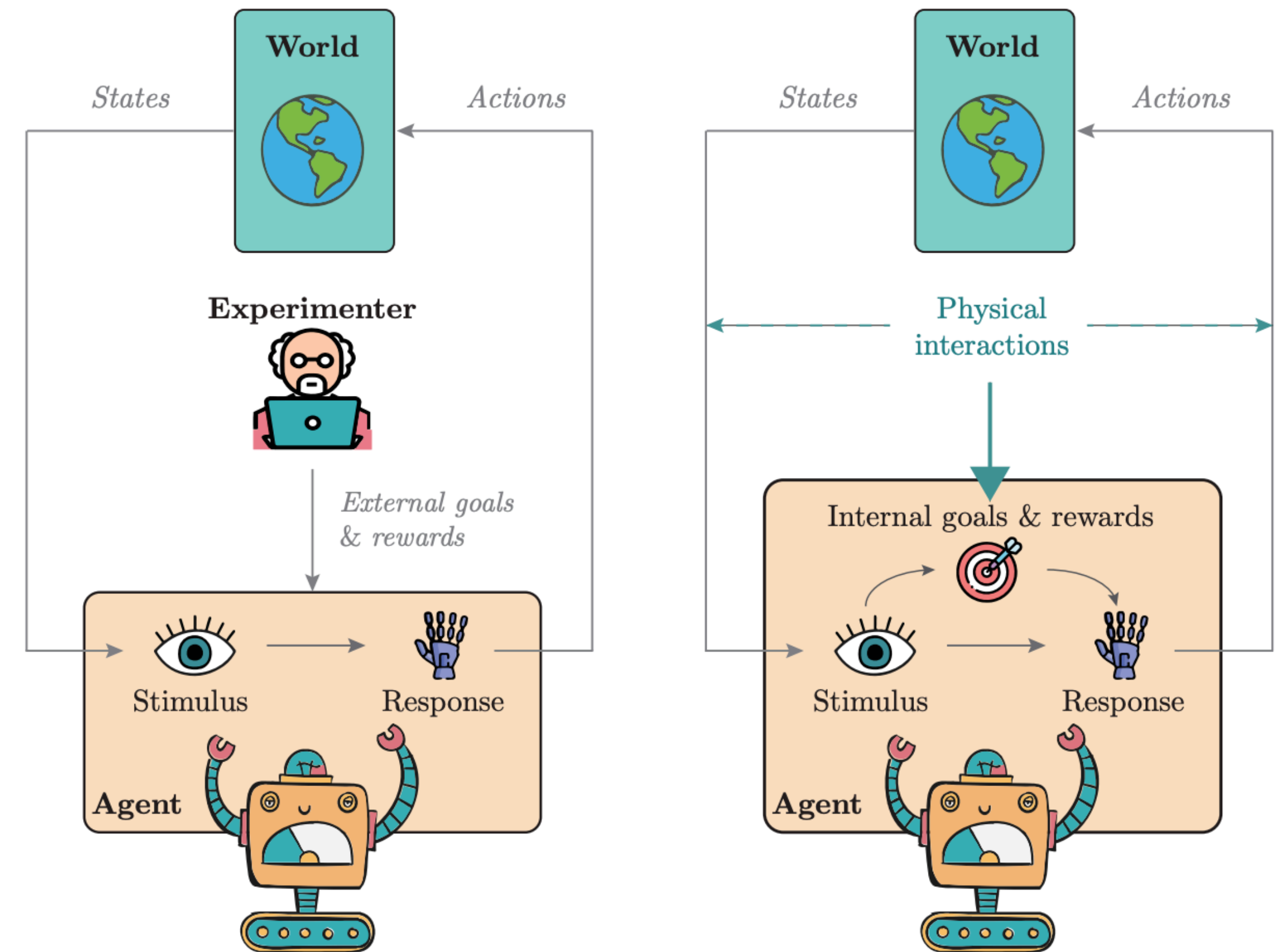
- **Semantic network structure**
- Steyvers & Tenenbaum, 2005 analyze a large free-association database involving more than 6000 participants was collected by Nelson et al. (1999). Over 5000 words served as cues (e.g. “cat”) for which participants had to write down the first word that came to mind (e.g. “dog”)
  - Small-world and scale-free properties
- Christianson et al., 2020 explore the structure of semantic networks in linear algebra textbooks
  - Core-periphery, community/modularity structure
- Lynn & Bassett, 2021 find hierarchical organization with tight clustering and heterogeneous degrees—increases compressibility
- Budel et al., 2023 study the properties of semantic networks from ConceptNet, defined by 7 semantic relations from 11 different languages
  - they are sparse, highly clustered, and exhibit power-law degree distributions. Our findings show that the majority of the considered networks are scale-free.

# Environment / knowledge network

- Knowledge structure
- **Semantic knowledge value**
  - Existing intrinsic motivations do not distinguish states by semantics - instead, they only care about frequency
  - Similarity, prerequisite, difficulty

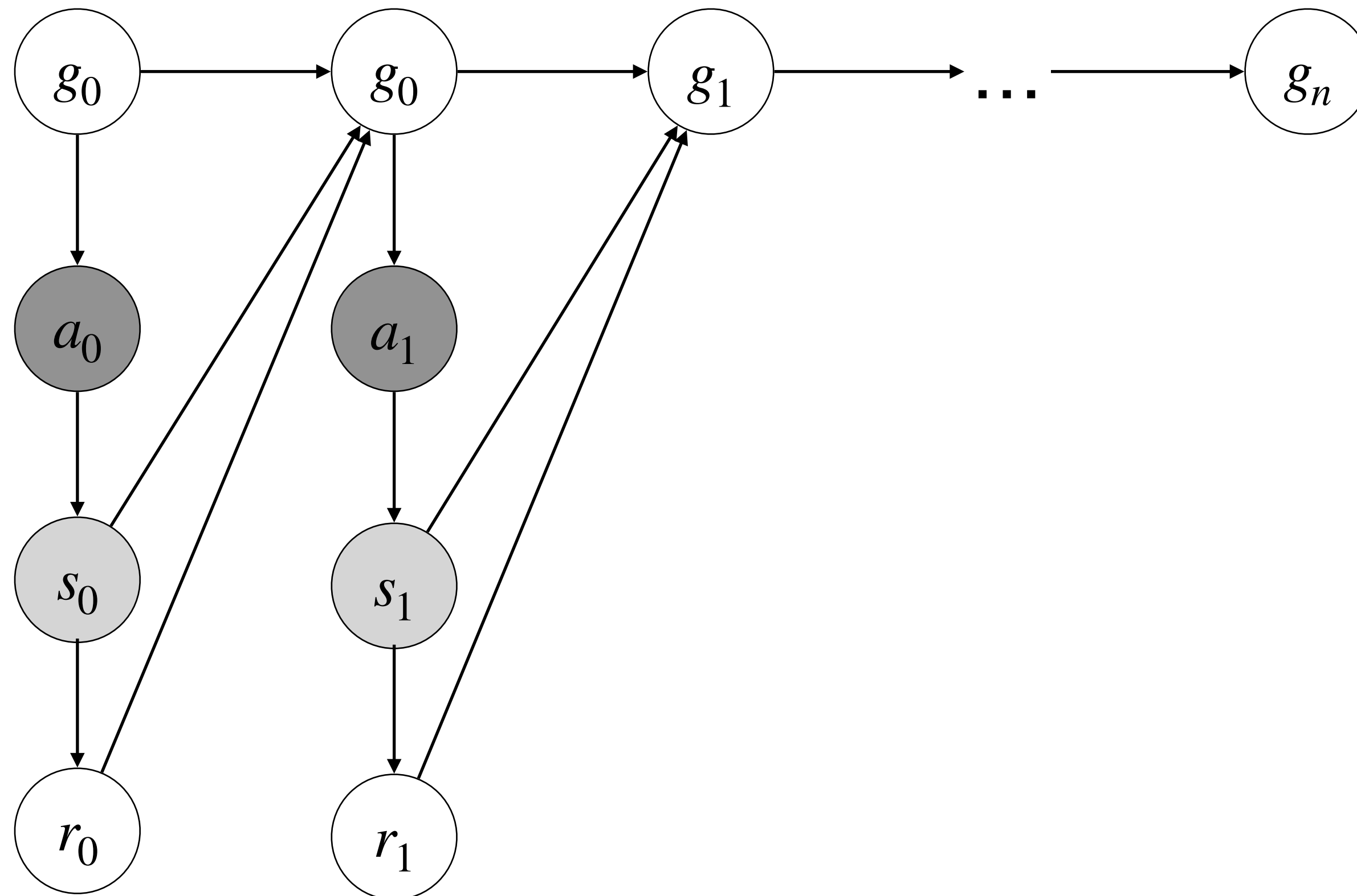
# Agent: autotelic

- self-generation,
- self-selection,
- self-ordering and,
- self-experimentation of learning goals.
- How the dynamics of goals relate to the performance?





# Agent: autotelic



# Autotelic agent aiming for flexible networks

- Goal generation from a *pre-defined goal space*
  - $g_t \sim p_{\pi_g}(g \mid s_t)$
  - Reward  $r_g$ , action policy  $\pi_a$
- Objective
  - Local & global feature of the learned network
- What is the optimal trajectory of goals?

