ExpGen: Explore to Generalize

Ev Zisselman, Itai Lavie, Daniel Soudry, Aviv Tamar

NeurIPS 2023

European Research Council
Established by the European Commission

NEURAL INFORMATION PROCESSING SYSTEMS

$(RL)^2$ Reinforcement Learning Research Labs

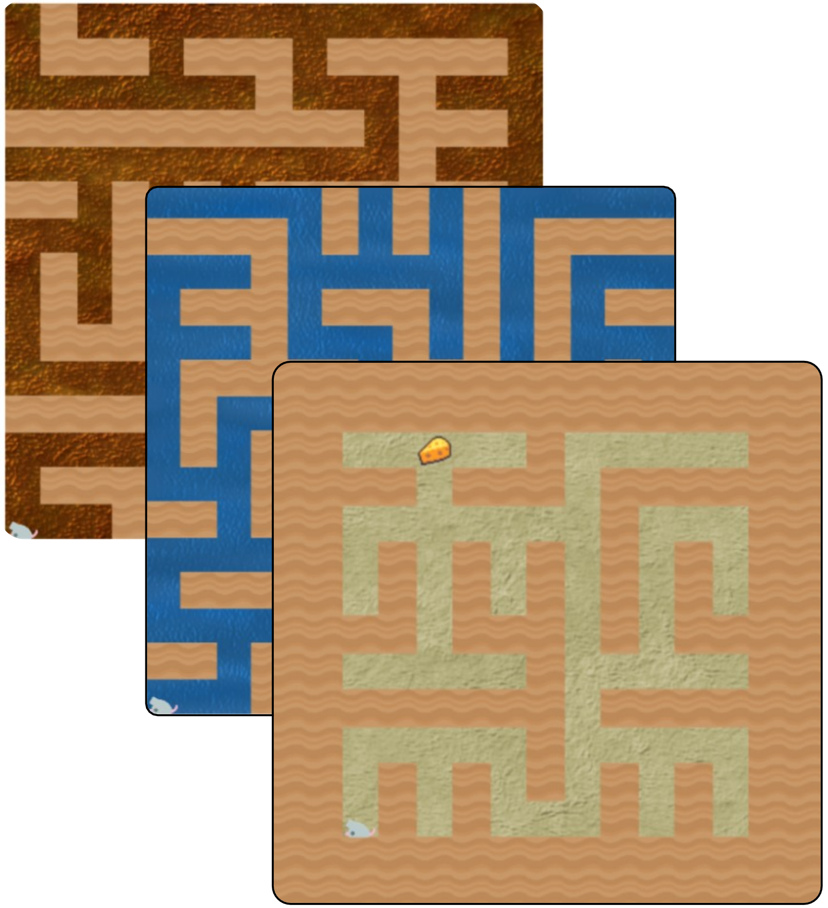
הטכניון
מכון טכנולוגי לישראל
TECHNION
Israel Institute of Technology

The Problem – Zero Shot Generalization in RL

Given N training MDPs
$M_1, M_2, \ldots, M_N \sim P(M)$

$$\mathcal{L}_{emp}(\pi) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\pi;M_i}\left[\sum_{t=0}^{T} C(s_t, a_t)\right]$$

Train

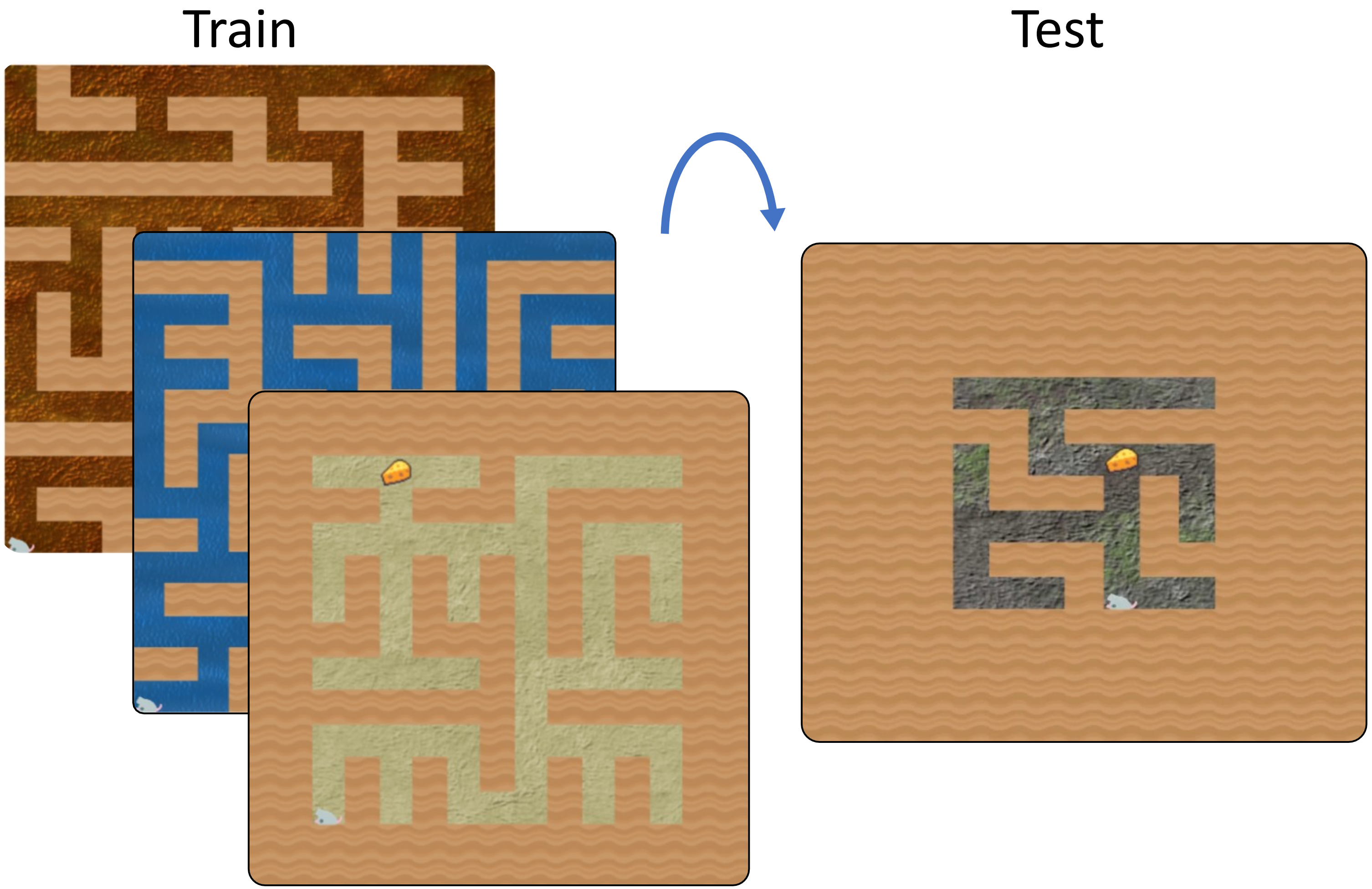


The Problem – Zero Shot Generalization in RL

Generalization is Difficult

In practice – easy to "memorize" training solutions (Overfitting)

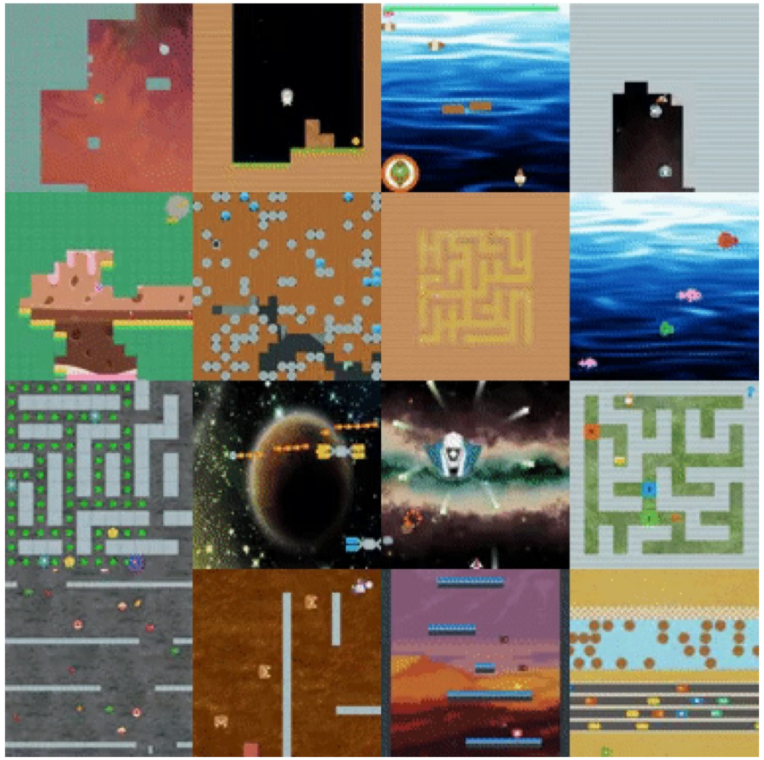
In theory – PAC bounds[1][2] exponential in #DoF of $P(M)$



[1] Tamar, Soudry, Zisselman. Regularization Guarantees Generalization in Bayesian Reinforcement Learning through Algorithmic Stability. AAAI 2022
[2] Rimon, Tamar, Adler. Meta Reinforcement Learning with Finite Training Tasks - a Density Estimation Approach. NeurIPS 2022

ProcGen benchmark

- ProcGen generalization benchmark

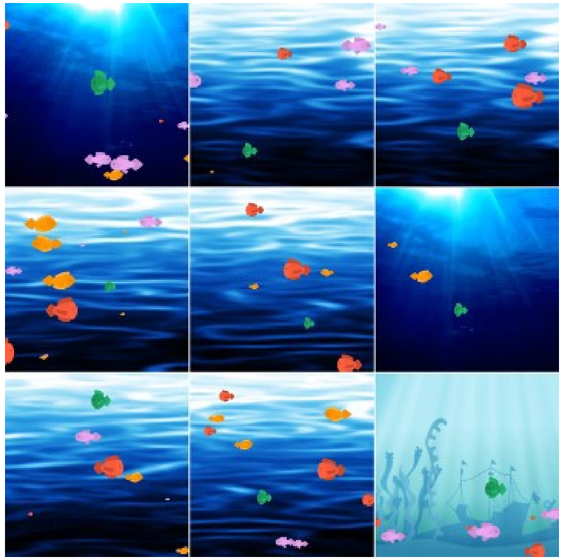


Cobbe et al. Leveraging procedural generation to benchmark reinforcement learning, 2020

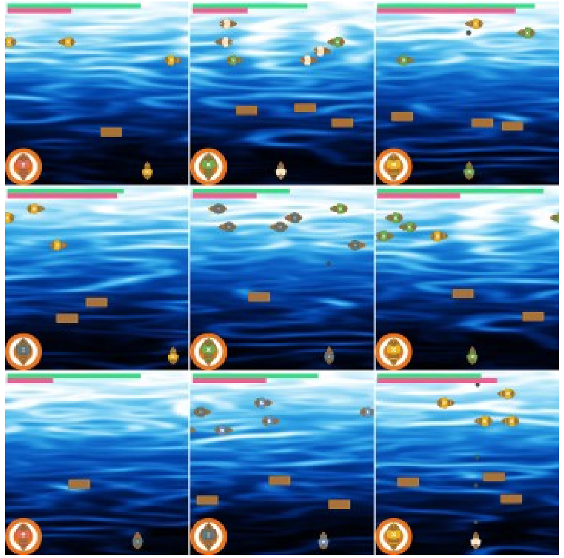
ProcGen benchmark

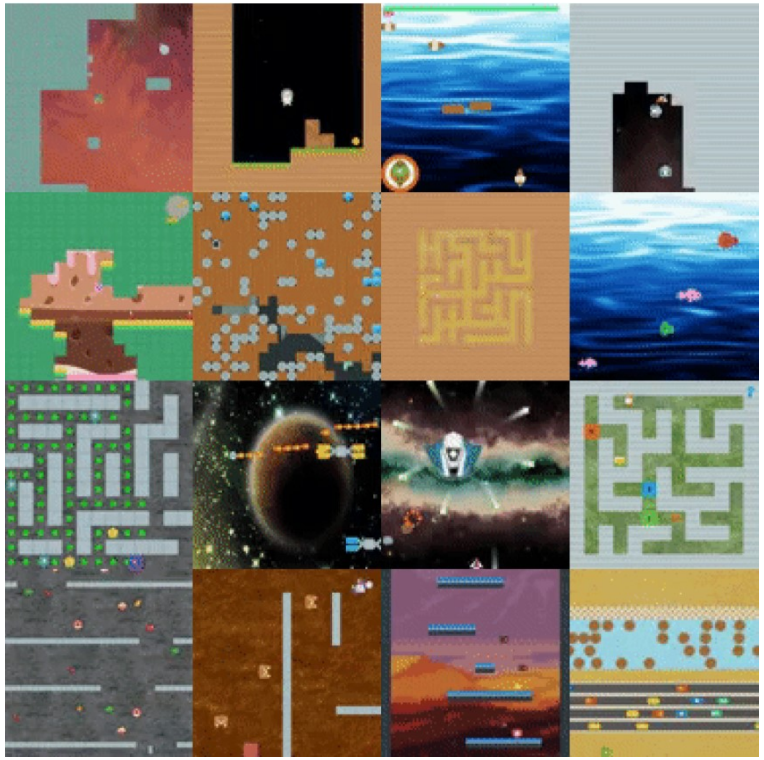
❑ ProcGen generalization benchmark

BigFish



Plunder





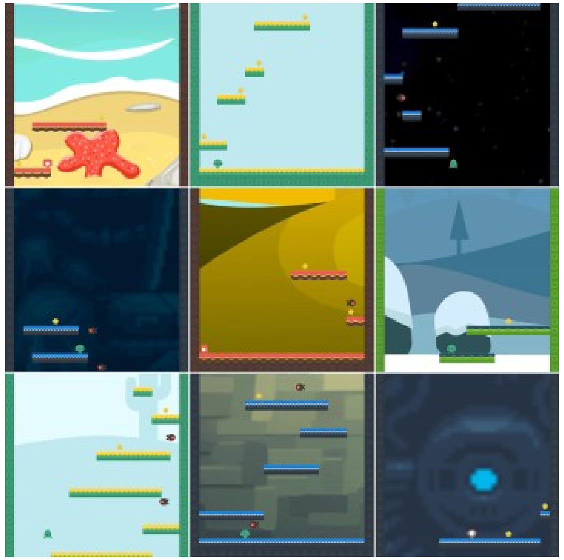
Cobbe et al. Leveraging procedural generation to benchmark reinforcement learning, 2020

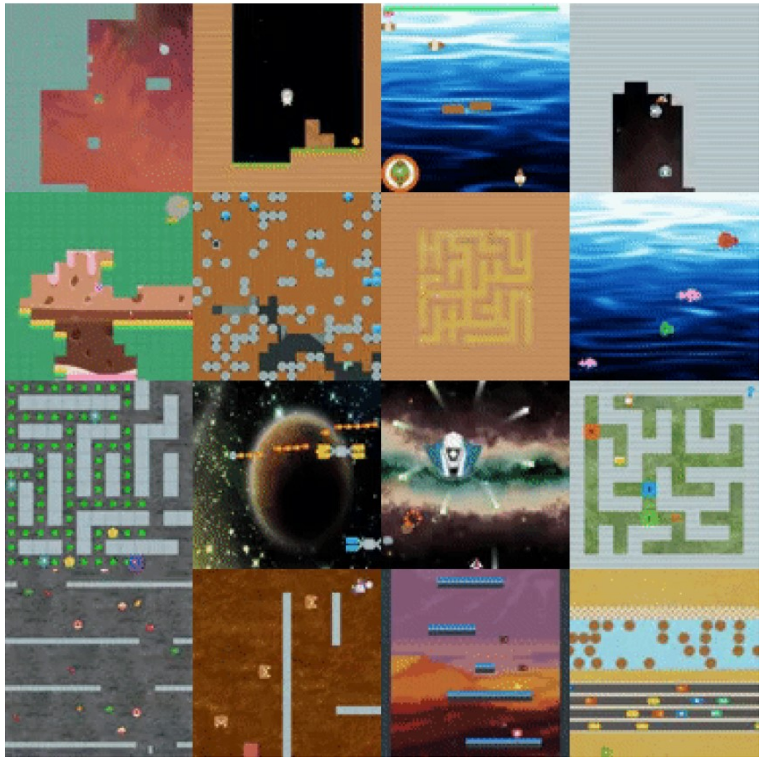
ProcGen benchmark

❑ ProcGen generalization benchmark

Jumper

Climber



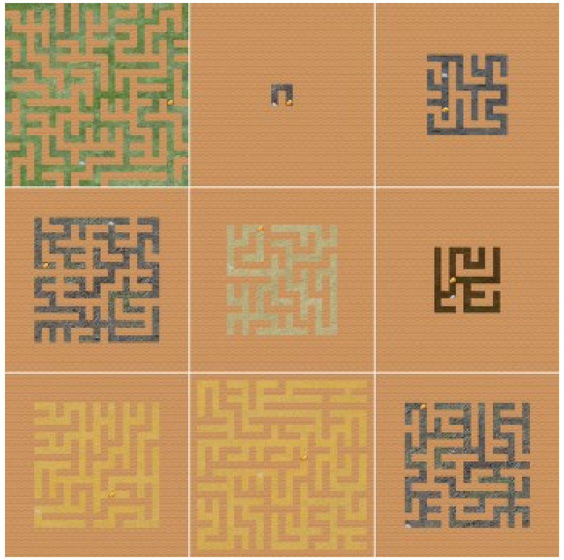


Cobbe et al. Leveraging procedural generation to benchmark reinforcement learning, 2020

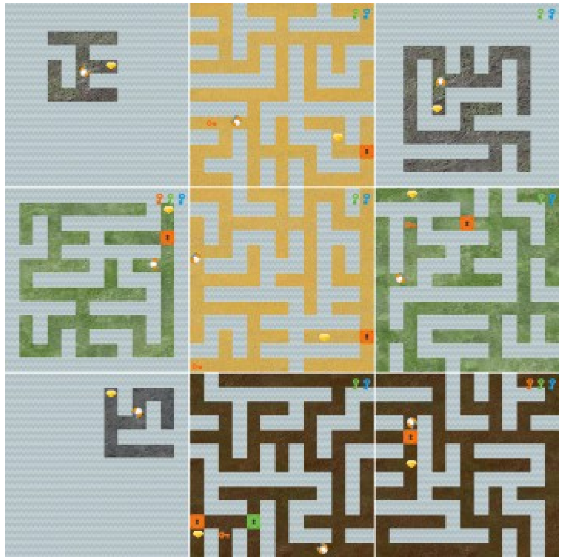
ProcGen benchmark

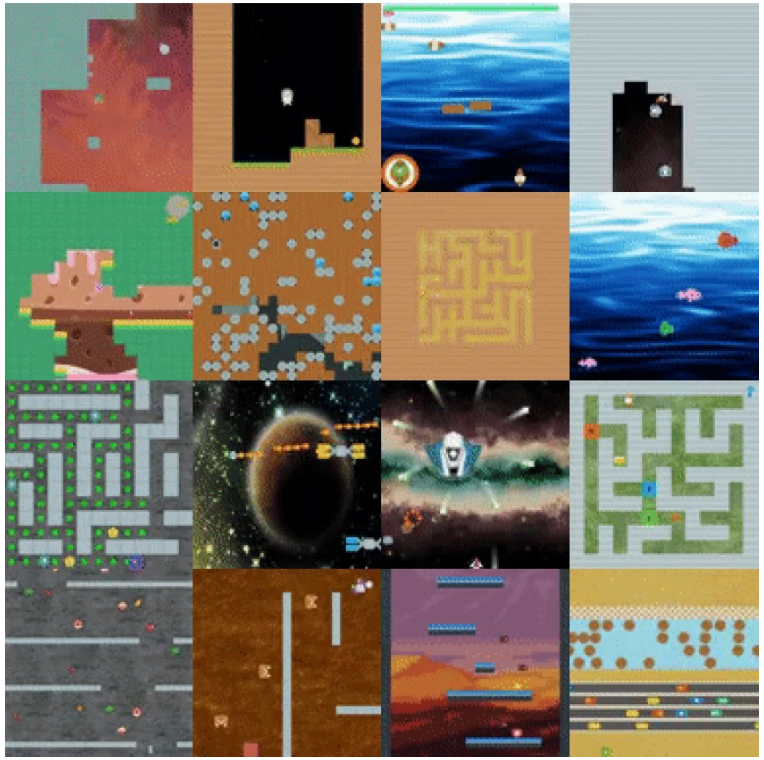
- ProcGen generalization benchmark

Maze



Heist

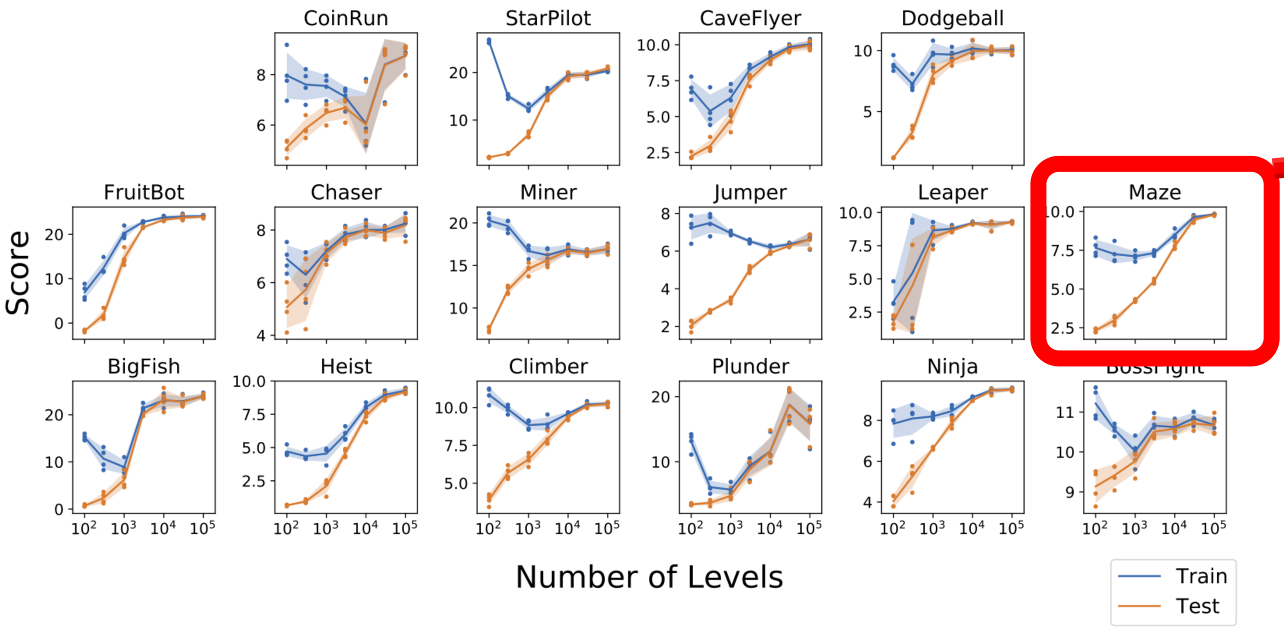




Cobbe et al. Leveraging procedural generation to benchmark reinforcement learning, 2020

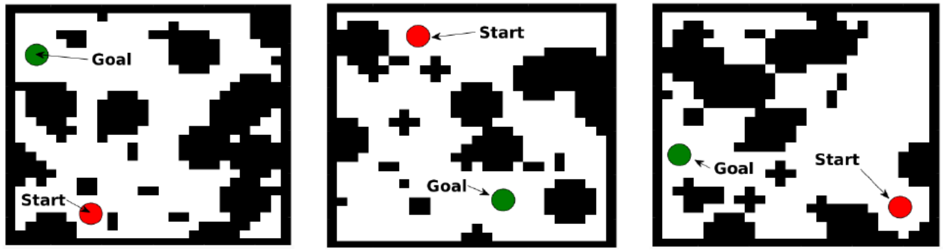
ProcGen benchmark

❑ ProcGen generalization benchmark – still challenging



Cobbe et al. Leveraging procedural generation to benchmark reinforcement learning, 2020

Approaches to Generalization



- ❑ **Dominant Approaches**
  - **Inductive bias for a "planning" policy**
    - E.g., value iteration network[3]
      Issue: Domain-specific

[3] Tamar et al. Value Iteration Networks. NeurIPS 2016

Approaches to Generalization

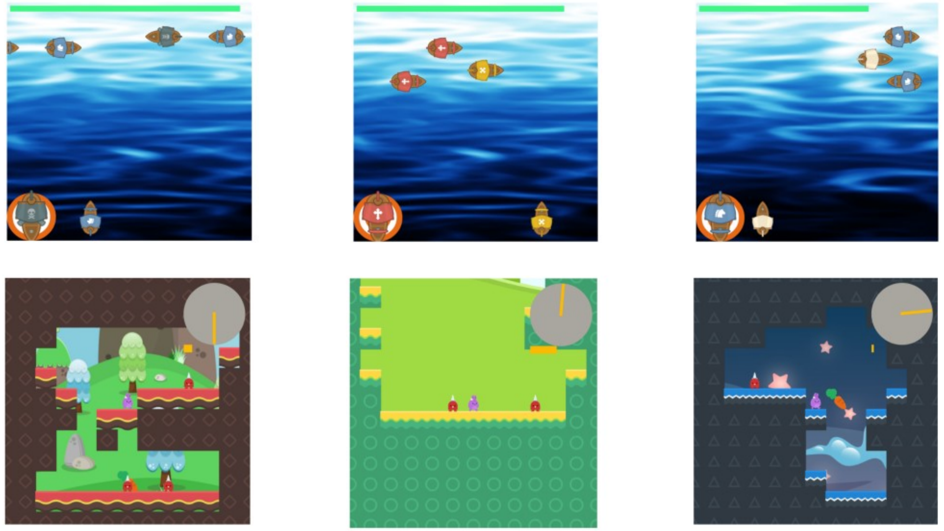
❑ Dominant Approaches

Inductive bias for a “planning” policy

- E.g., value iteration network[3]
  Issue: Domain-specific

Invariant policies

- To appearance (e.g., augmentation UCB-DrAC[4])



Plunder

Jumper

[3] Tamar et al. Value Iteration Networks. NeurIPS 2016
[4] Raileanu et al. Automatic data augmentation for generalization in reinforcement learning. NeurIPS 2021

Approaches to Generalization

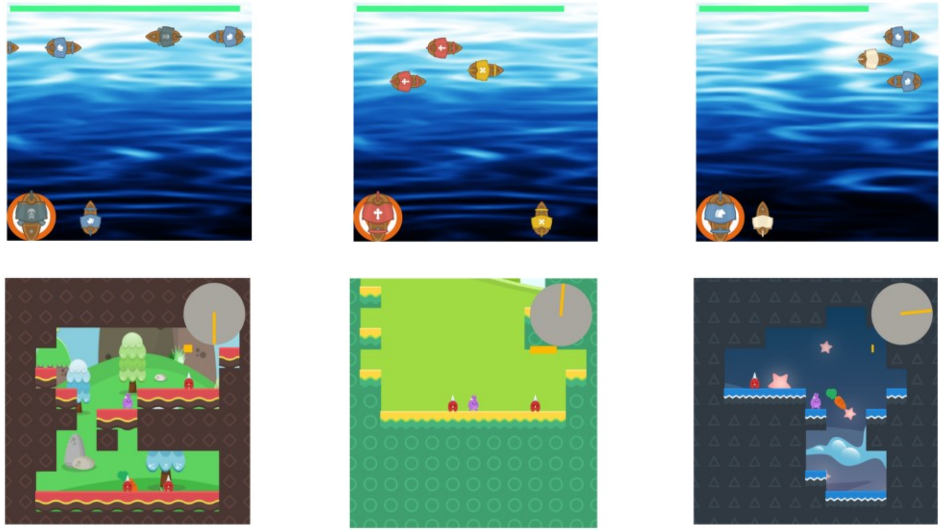
❑ Dominant Approaches

Inductive bias for a “planning” policy

- E.g., value iteration network[3]
  Issue: Domain-specific

Invariant policies

- To appearance (e.g., augmentation UCB-DrAC[4])
- To length of tasks (e.g., decoupling value and policy IDAAC[5])



Plunder

Jumper

Ninja

Level 1 $V_1(s_0) = 7.9,\ \hat{V}_1(s_0) = 8.0$ Level 2 $V_2(s_0) = 6.1,\ \hat{V}_2(s_0) = 6.3$

[3] Tamar et al. Value Iteration Networks. NeurIPS 2016
[4] Raileanu et al. Automatic data augmentation for generalization in reinforcement learning. NeurIPS 2021
[5] Raileanu and Fergus. Decoupling value and policy for generalization in reinforcement learning. ICML 2021

Approaches to Generalization

❑ Dominant Approaches

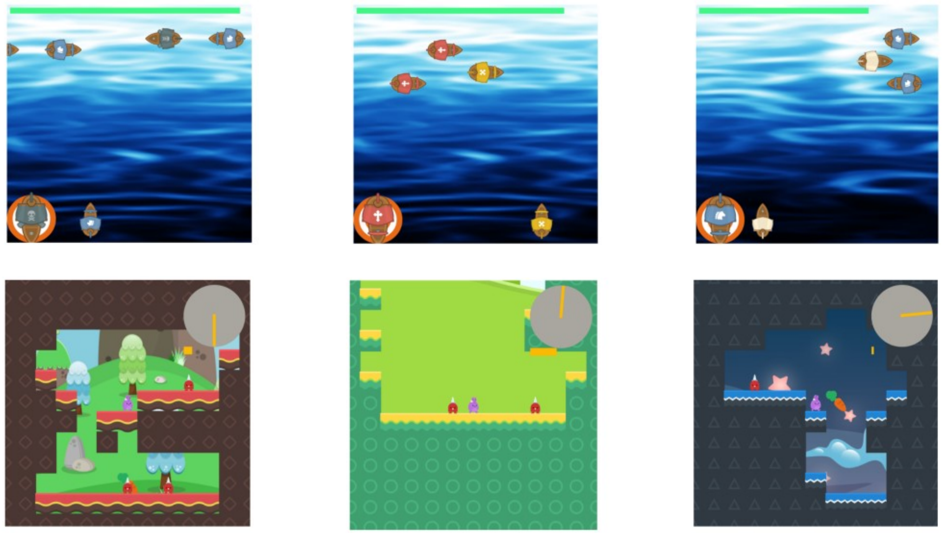
Inductive bias for a "planning" policy

- E.g., value iteration network[3]
  Issue: Domain-specific

Invariant policies

- To appearance (e.g., augmentation UCB-DrAC[4])
- To length of tasks (e.g., decoupling value and policy IDAAC[5])

→ Not enough for <u>All Tasks</u>



Plunder

Jumper

[3] Tamar et al. Value Iteration Networks. NeurIPS 2016
[4] Raileanu et al. Automatic data augmentation for generalization in reinforcement learning. NeurIPS 2021
[5] Raileanu and Fergus. Decoupling value and policy for generalization in reinforcement learning. ICML 2021

Approaches to Generalization

❑ Dominant Approaches

Inductive bias for a "planning" policy

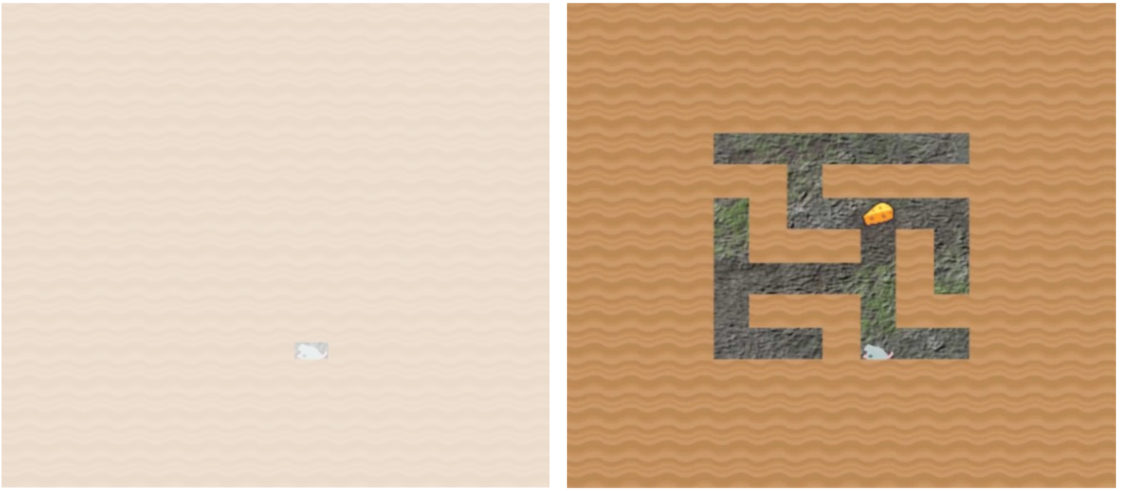
- E.g., value iteration network[3]
  Issue: Domain-specific

Invariant policies

- To appearance (e.g., augmentation UCB-DrAC[4])
- To length of tasks (e.g., decoupling value and policy IDAAC[5])

→ Not enough for <u>All Tasks</u>

Hidden Maze Experiment



[3] Tamar et al. Value Iteration Networks. NeurIPS 2016
[4] Raileanu et al. Automatic data augmentation for generalization in reinforcement learning. NeurIPS 2021
[5] Raileanu and Fergus. Decoupling value and policy for generalization in reinforcement learning. ICML 2021

Approaches to Generalization

❑ Dominant Approaches

Inductive bias for a "planning" policy

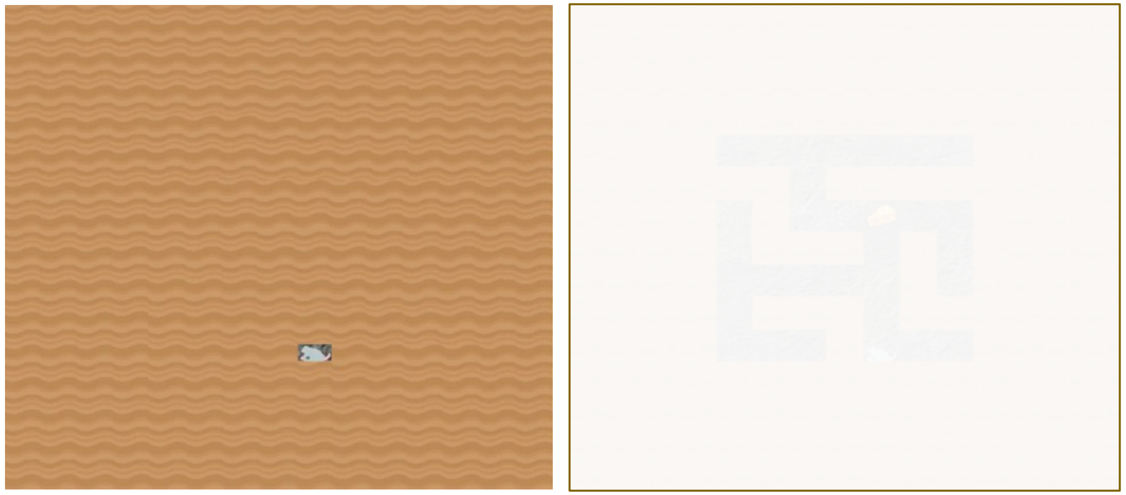
- E.g., value iteration network[3]
  Issue: Domain-specific

Invariant policies

- To appearance (e.g., augmentation UCB-DrAC[4])
- To length of tasks (e.g., decoupling value and policy IDAAC[5])

→ **Not enough for <u>All Tasks</u>**

Hidden Maze Experiment

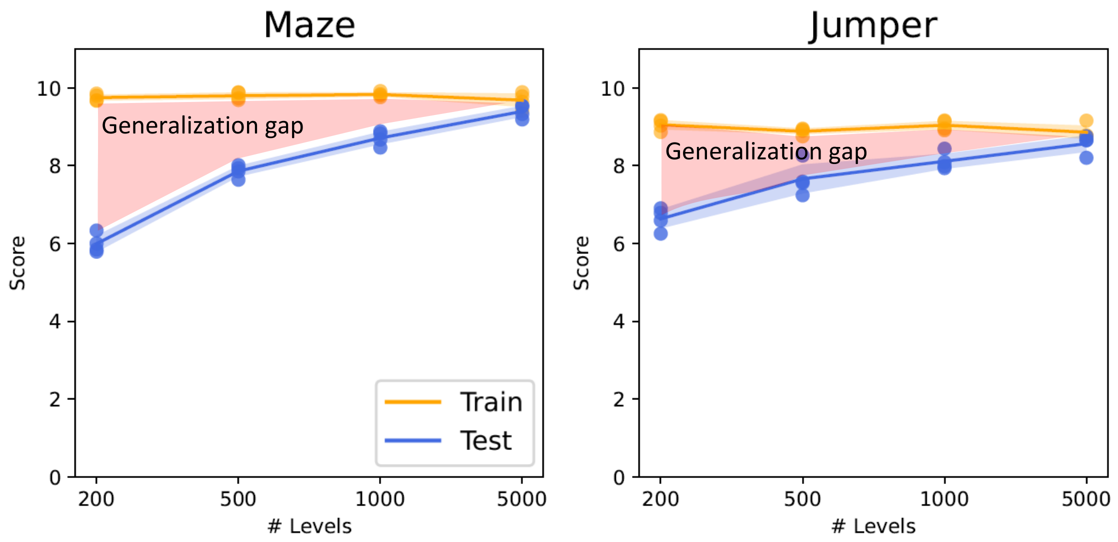


[3] Tamar et al. Value Iteration Networks. NeurIPS 2016
[4] Raileanu et al. Automatic data augmentation for generalization in reinforcement learning. NeurIPS 2021
[5] Raileanu and Fergus. Decoupling value and policy for generalization in reinforcement learning. ICML 2021

Learning to Explore

- ❑ **Observation** – Maximum Entropy[7] exploration **generalizes**[8] !

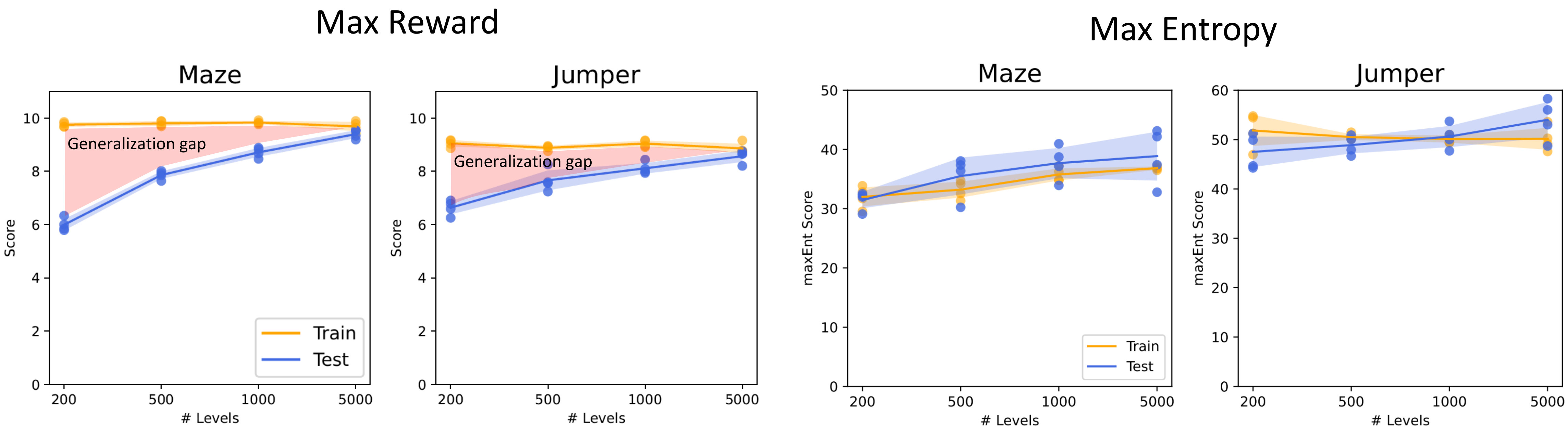
Max Reward



[7] Mutti et al. The importance of non-markovianity in maximum state entropy exploration. ICML 2022
[8] Zisselman, Lavie, Soudry, Tamar. NeurIPS 2023

Learning to Explore

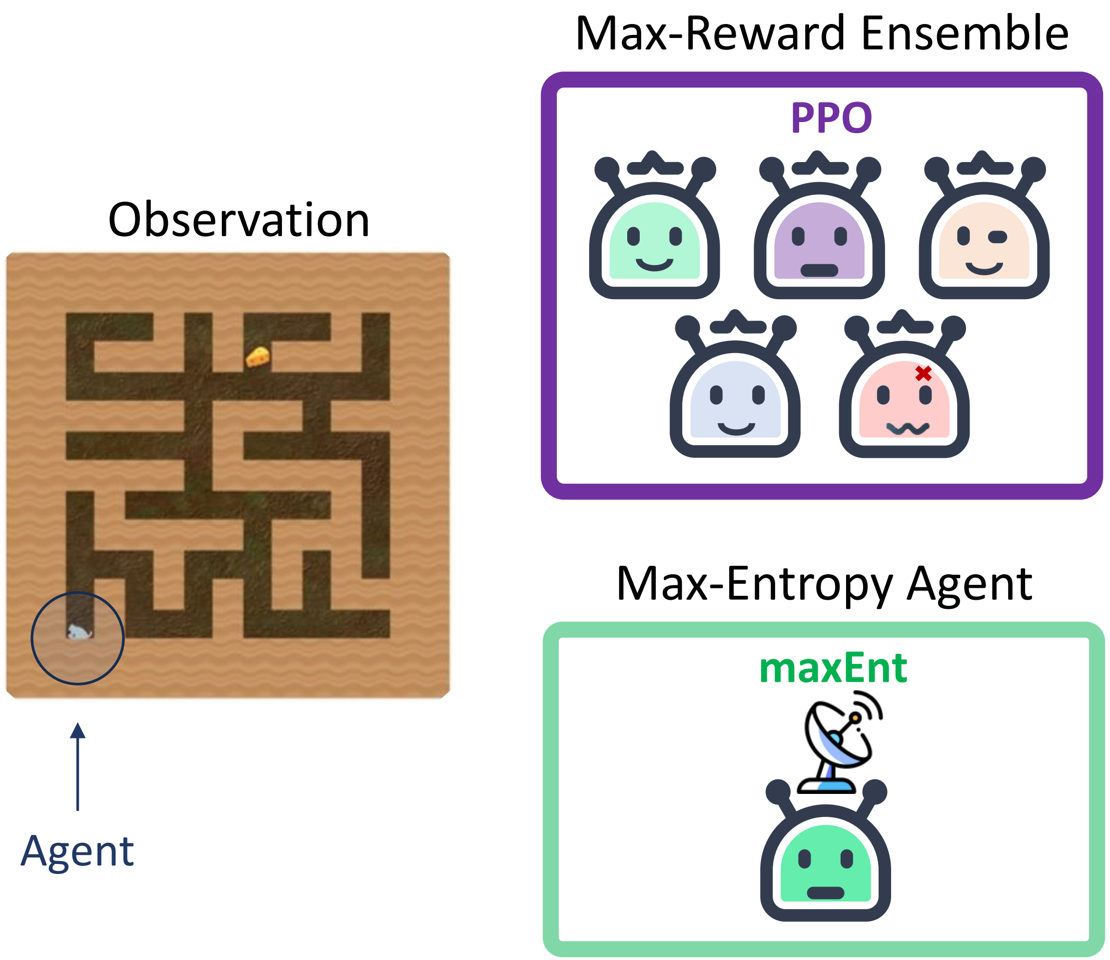
- **Observation** – Maximum Entropy[7] exploration **generalizes**[8] !



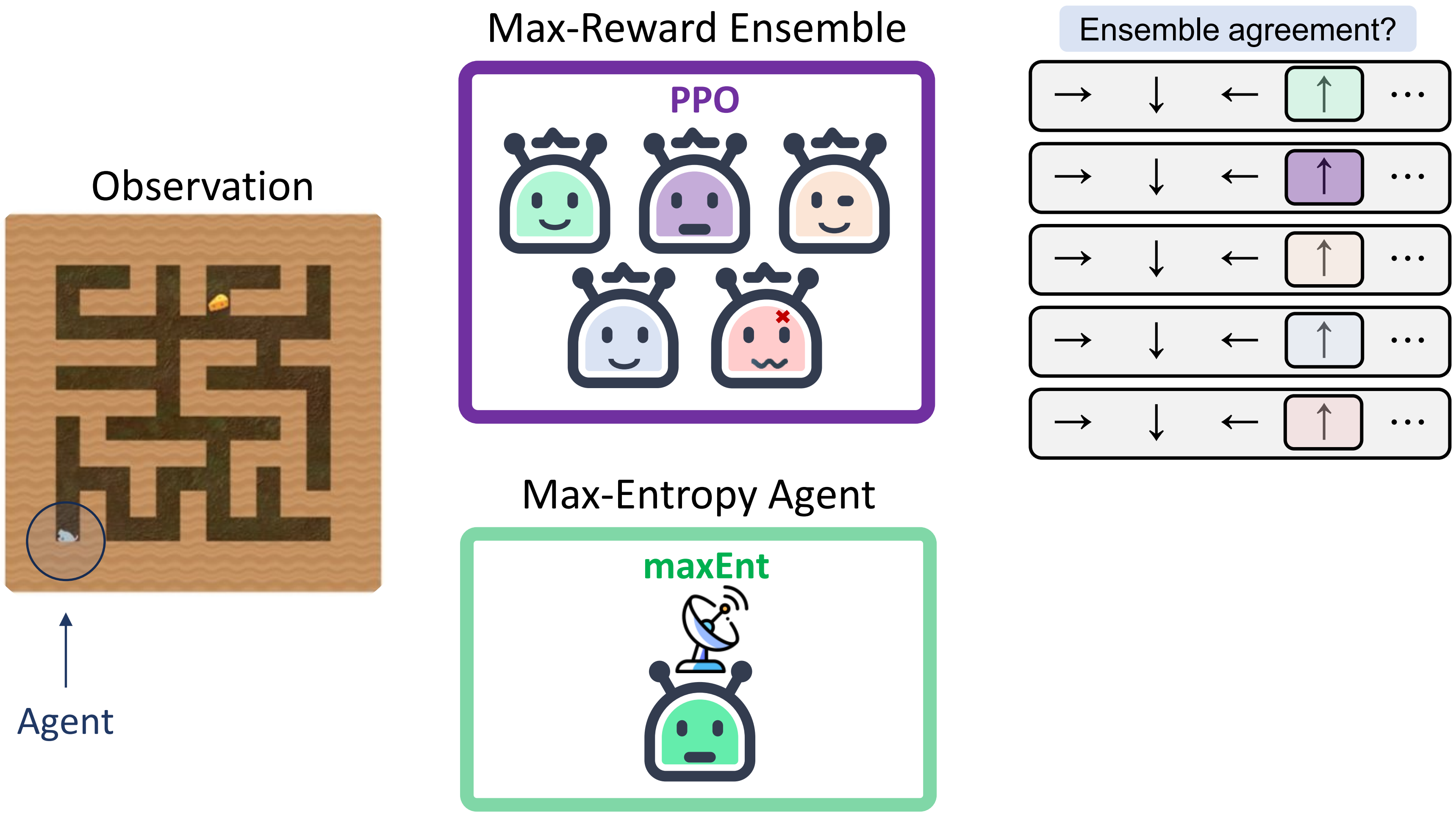
Why? Exploration behavior harder to memorize!

[7] Mutti et al. The importance of non-markovianity in maximum state entropy exploration. ICML 2022
[8] Zisselman, Lavie, Soudry, Tamar. NeurIPS 2023

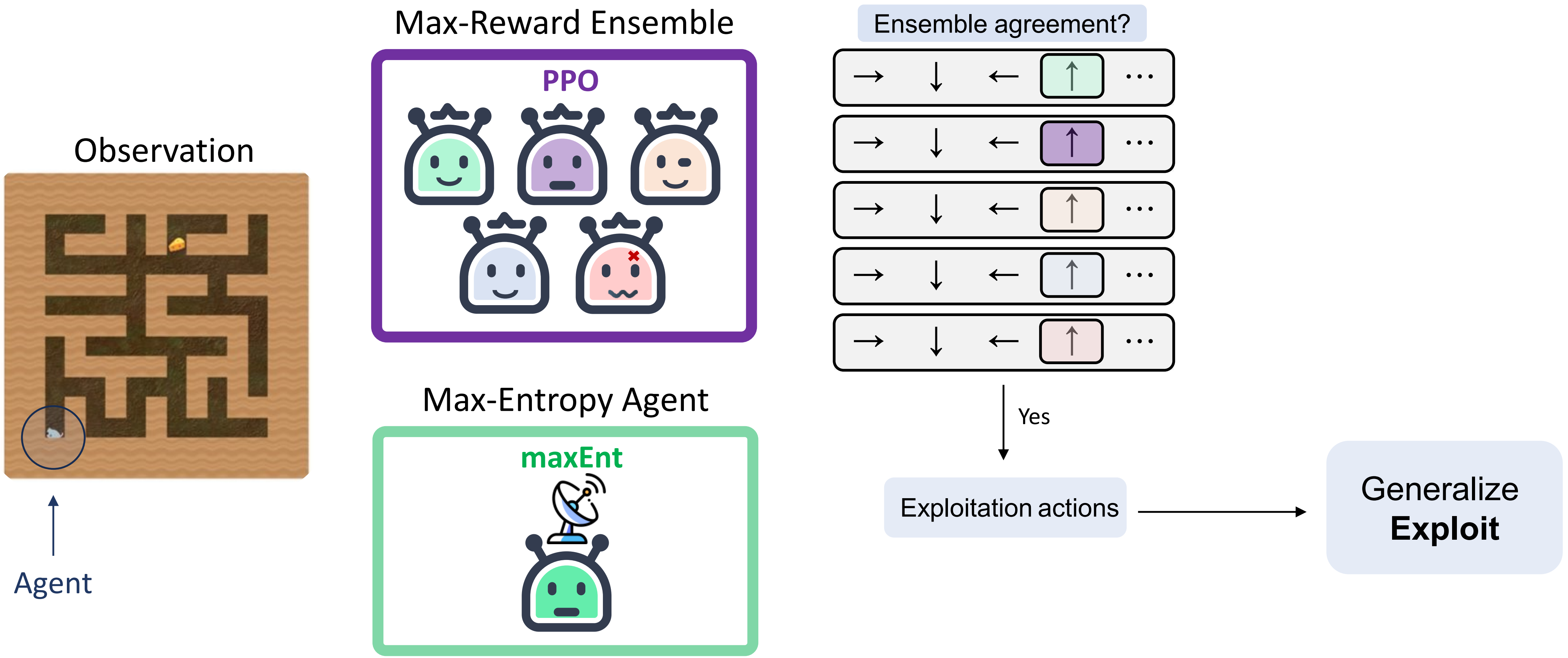
Exp-Gen Algorithm



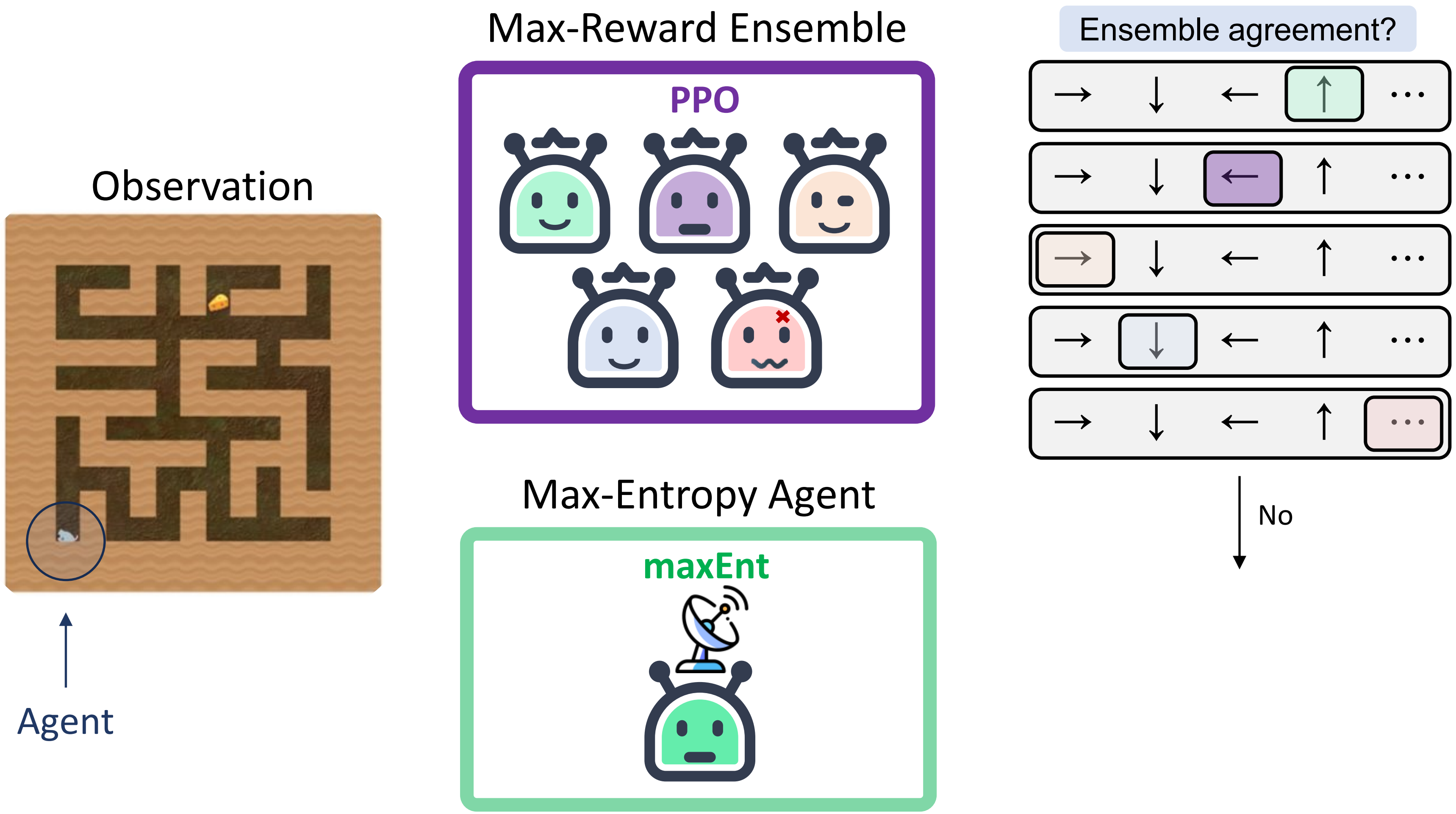
Exp-Gen Algorithm



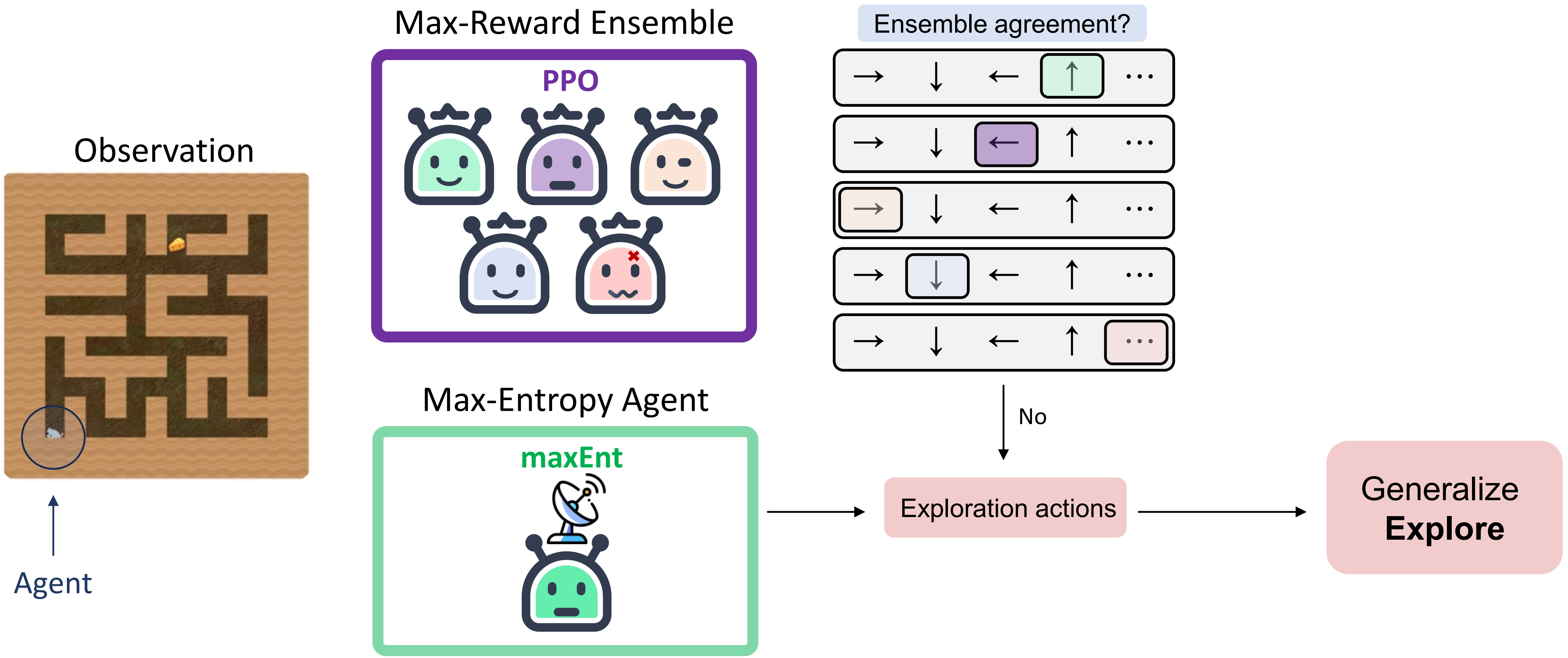
Exp-Gen Algorithm



Exp-Gen Algorithm



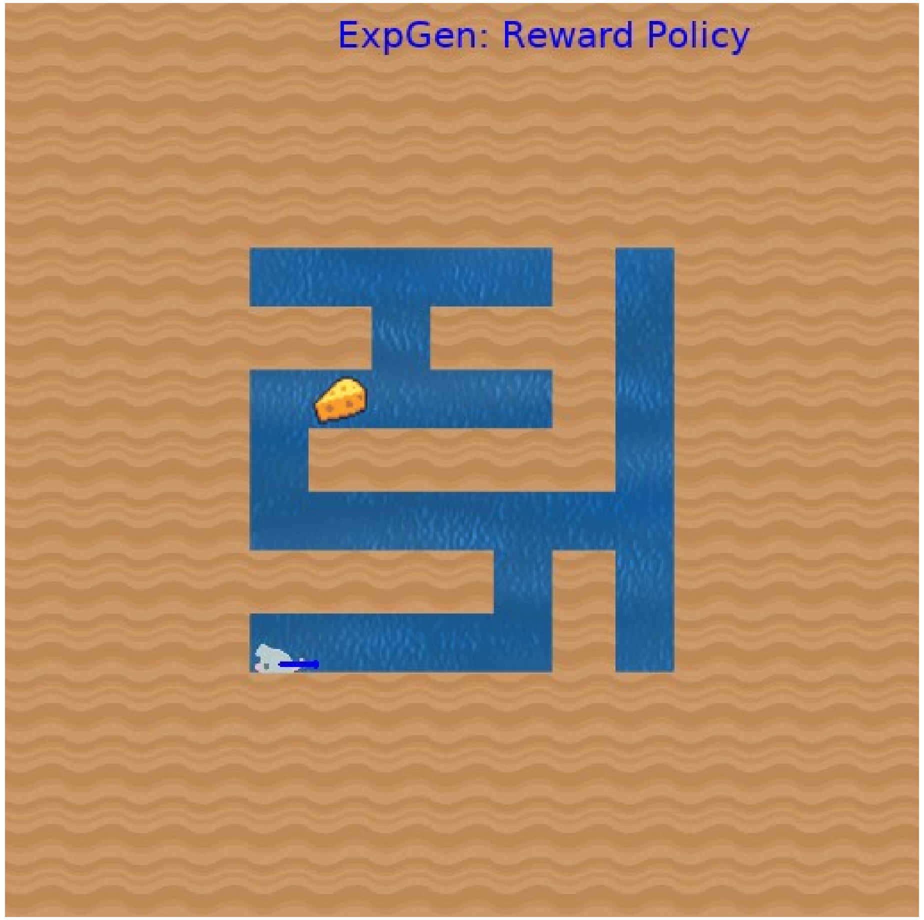
Exp-Gen Algorithm



See it in action

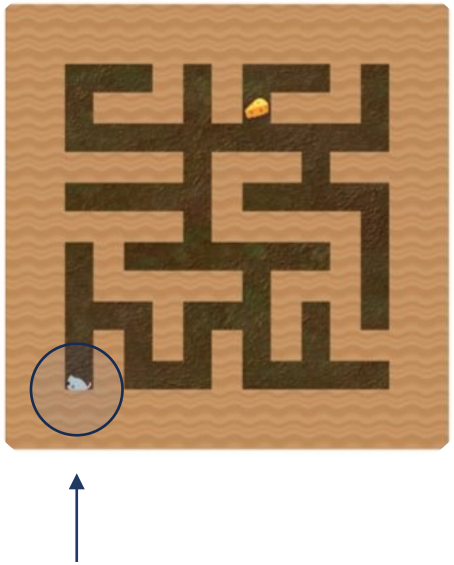
maxEnt starts exploring

Reward ensemble consensus



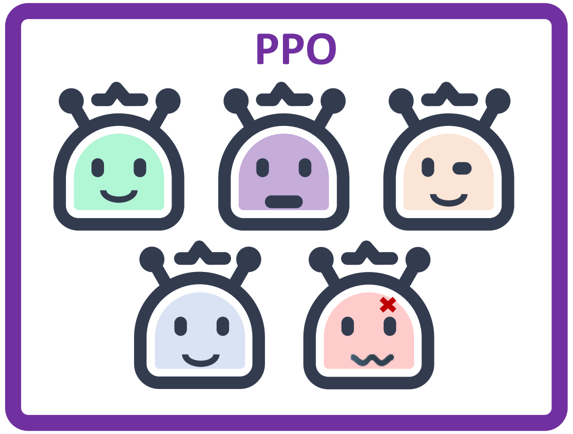
Exp-Gen Algorithm

Observation

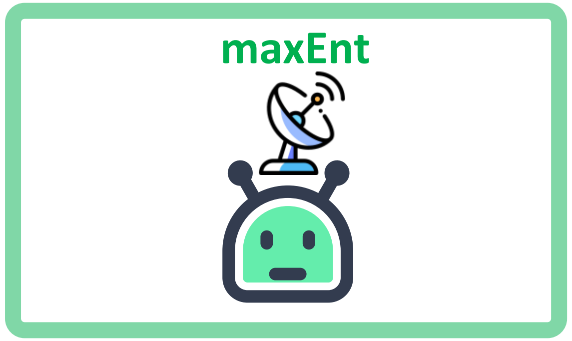


Agent

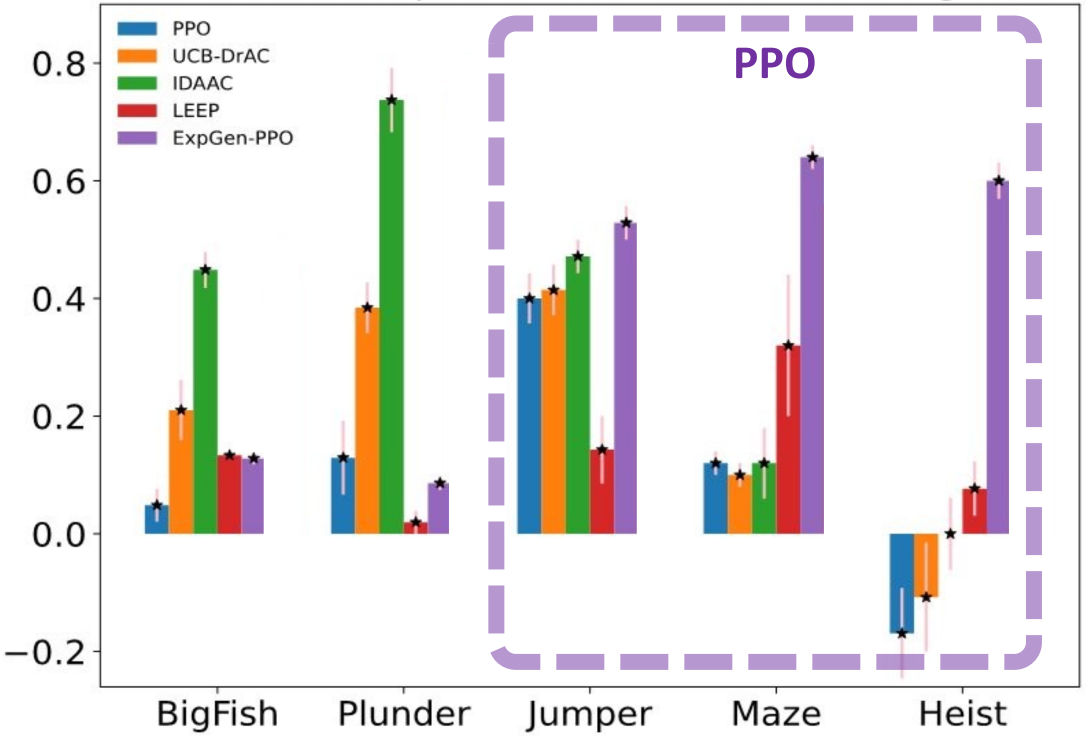
Max-Reward Ensemble



Max-Entropy Agent



Normalized test performance on five ProcGen games

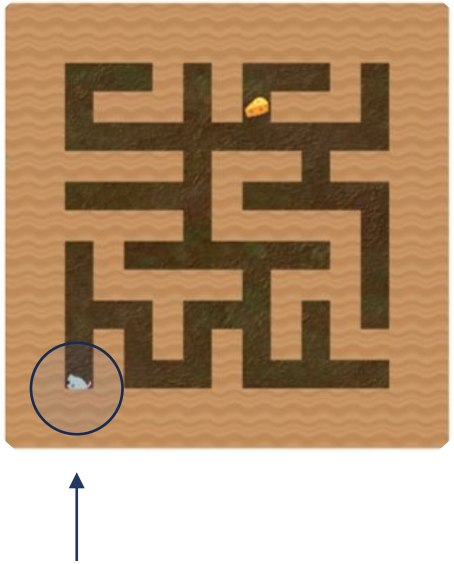


Trained on only 200 levels!

PPO: Schulman et al., 2017
UCB-DrAC: Raileanu et al., 2021
IDAAC: Raileanu et al., 2021
LEEP: Ghosh et al., NeurIPS 2021

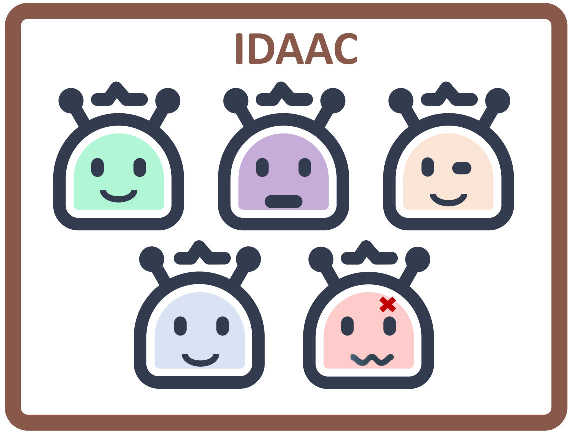
Exp-Gen Algorithm

Observation

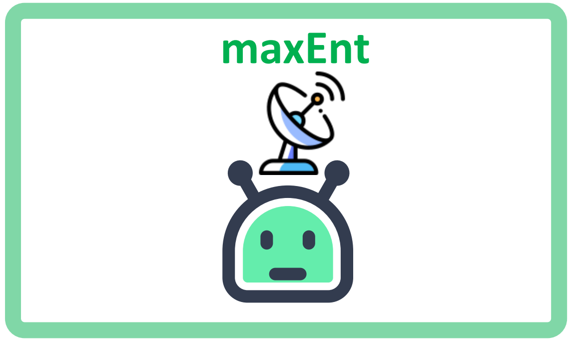


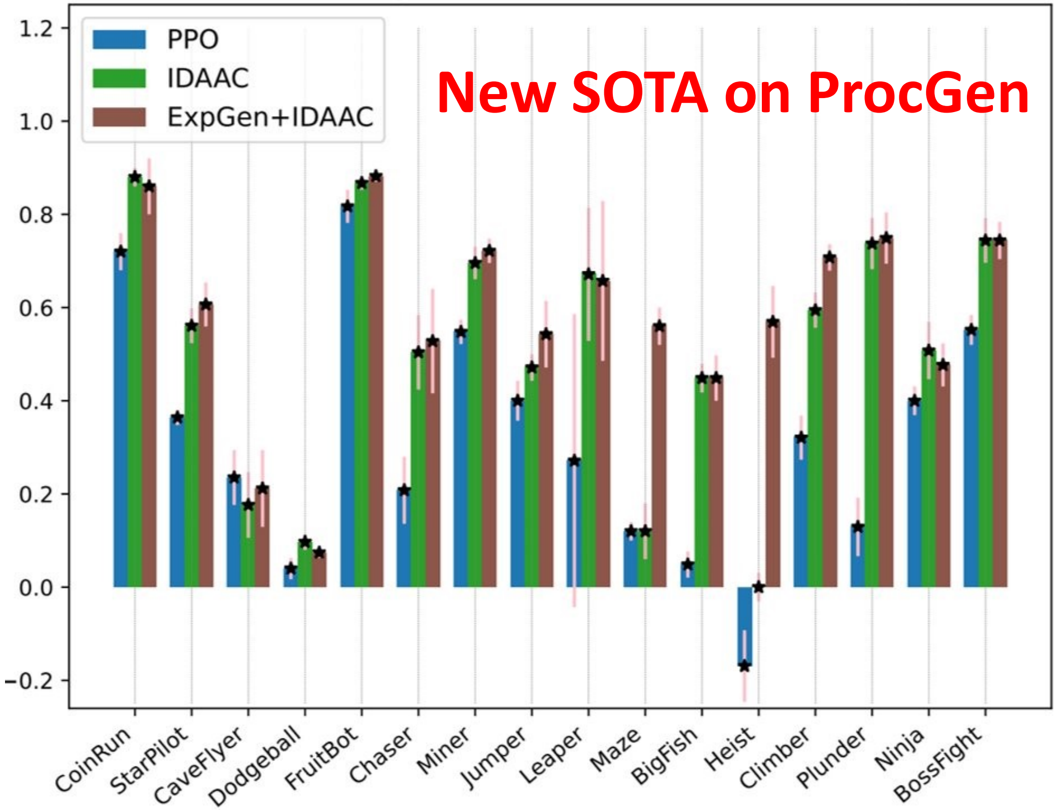
Agent

Max-Reward Ensemble



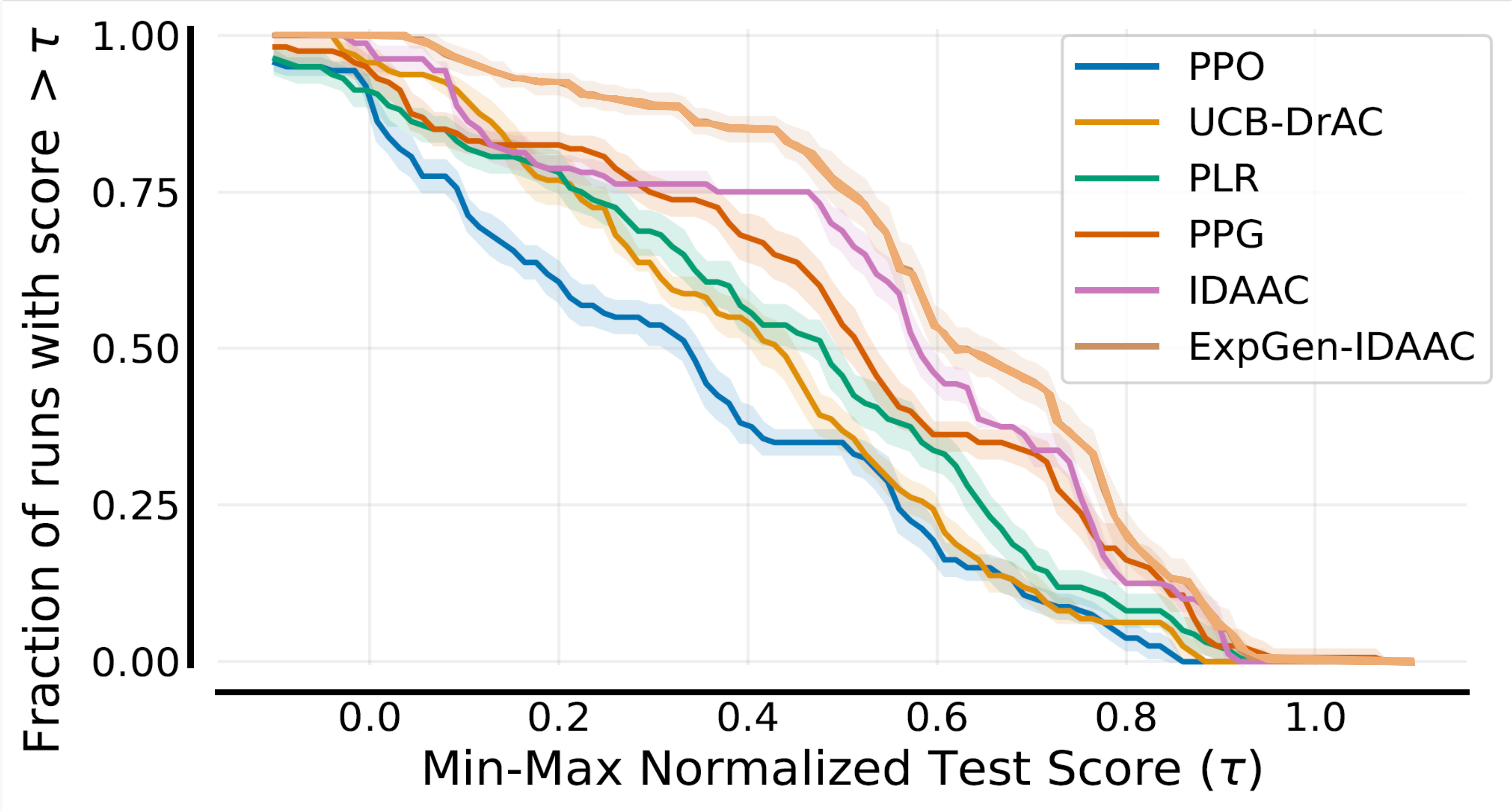
Max-Entropy Agent





PPO: Schulman et al., 2017
IDAAC: Raileanu et al., 2021

Results



Fraction of runs with score > τ
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Min-Max Normalized Test Score (τ)
PPO
UCB-DrAC
PLR
PPG
IDAAC
ExpGen-IDAAC

Summary:

- ❑ Key observation: Maximum-entropy generalizes well
- ❑ Exp-Gen idea
  - Detect uncertainty (ensemble)
  - Explore when uncertain
  - Otherwise exploit
- ❑ Exp-Gen sets a new state-of-the-art on ProcGen.

GitHub

Reference

Ev Zisselman, Itai Lavie, Daniel Soudry, Aviv Tamar

Explore to Generalize in Zero-Shot RL

NeurIPS 2023

Thank you!