DarkAlley: Uncertainty-Aware Goal Selection for Stochastic Sparse-Reward Environments

Olivia Garland, Tim Walter

February 2026

1 Motivation

Exploration in sparse-reward goal-reaching tasks has seen significant recent progress through directed goal selection methods. Approaches such as DISCOVER [2] and MEGA [4] guide exploration by selecting sub-goals from a frontier of achieved states, balancing achievability, novelty, and relevance to the target goal. These methods represent the state of the art on challenging long-horizon tasks, substantially outperforming undirected curiosity-based and count-based alternatives.

However, they have been developed and evaluated almost exclusively in deterministic or near-deterministic environments. Realistic deployment settings generally involve some degree of stochasticity—sensor noise, stochastic contact dynamics, or task-irrelevant dynamic elements in visual observations—that these benchmarks do not capture.

2 Background: Goal-Conditioned Reinforcement Learning

In goal-conditioned reinforcement learning, we consider a Markov decision process extended with a goal space $\mathcal{G} \subseteq \mathcal{S}$. The agent learns a policy $\pi(a|s, g)$ conditioned on both the current state s and a desired goal g. The sparse reward is defined as $r(s, a; g) = -1 \text{ for } s \notin \mathcal{S}_g$, where $\mathcal{S}_g$ is the set of states where goal g is achieved.

The key challenge in this setting is *goal selection*: which intermediate goals should the agent pursue during training to eventually reach a difficult target goal g^*? Recent work has shown that strategic goal selection—choosing goals that are achievable yet novel and relevant to the target—dramatically outperforms random exploration or always pursuing g^* directly.

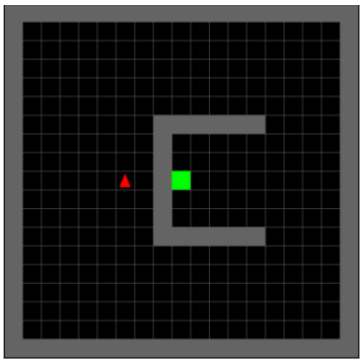

Figure 1: The classical example problem is a maze, the agent in red and goal in green.

3 Problem Statement

Current directed exploration methods typically estimate which sub-goals are worth exploring by measuring *disagreement* across an ensemble of learned value functions—high disagreement suggests the agent is uncertain about a region and should explore it. In deterministic environments, this works well: disagreement reflects genuine knowledge gaps (epistemic uncertainty) that close with more data.

In stochastic environments, however, inherent randomness prevents ensemble members from agreeing, even after many visits. The agent cannot distinguish between:

- **Epistemic uncertainty**: "uncertain because unexplored"
  (novel conditions not yet encountered during training)
- **Aleatoric uncertainty**: "uncertain because unpredictable"
  (inherent stochasticity)

This conflation can cause the agent to waste its exploration budget repeatedly visiting inherently noisy regions where no learning progress is possible. This is structurally analogous to the "noisy TV problem" identified in curiosity-driven exploration [1], but has not been studied in the goal-conditioned setting.

This suggests that effective goal selection in stochastic environments should prioritize sub-goals where:

1. The agent is genuinely uncertain (high epistemic uncertainty)
2. The environment is predictable enough to learn from (low aleatoric uncertainty)
3. Achieving the goal would help reach the target (high relevance)

4 Proposed Work

We propose to first benchmark existing directed exploration methods on stochastic variants of standard goal-reaching environments to characterize how they degrade, and then to augment goal selection with a notion of learnability—an estimate of whether visiting a candidate sub-goal will actually reduce the agent's uncertainty, rather than merely appearing novel. How best to quantify learnability in this setting is itself a core question of the proposed work.

Interesting algorithms would be

- SIERL [5]
- DISCOVER [2]
- MEGA [4]
- SLOPE [3]

5 Desirable Skills

- Strong Python programming skills
- Familiarity with deep reinforcement learning (value functions, policy gradients)
- Experience with JAX or PyTorch
- Interest in exploration strategies and math

References

[1] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by Random Network Distillation, October 2018.

[2] Leander Diaz-Bone, Marco Bagatella, Jonas Hübotter, and Andreas Krause. DISCOVER: Automated Curricula for Sparse-Reward Reinforcement Learning, October 2025.

[3] Yao-Hui Li, Zeyu Wang, Xin Li, Wei Pang, Yingfang Yuan, Zhengkun Chen, Boya Zhang, Riashat Islam, Alex Lamb, and Yonggang Zhang. From Scalar Rewards to Potential Trends: Shaping Potential Landscapes for Model-Based Reinforcement Learning, February 2026.

[4] Silviu Pitis, Harris Chan, Stephen Zhao, Bradly Stadie, and Jimmy Ba. Maximum Entropy Gain Exploration for Long Horizon Multi-goal Reinforcement Learning, July 2020.

[5] Georgios Sotirchos, Zlatan Ajanović, and Jens Kober. Search Inspired Exploration in Reinforcement Learning, January 2026.