Decoupled Reinforcement Learning to Stabilise Intrinsically-Motivated Exploration

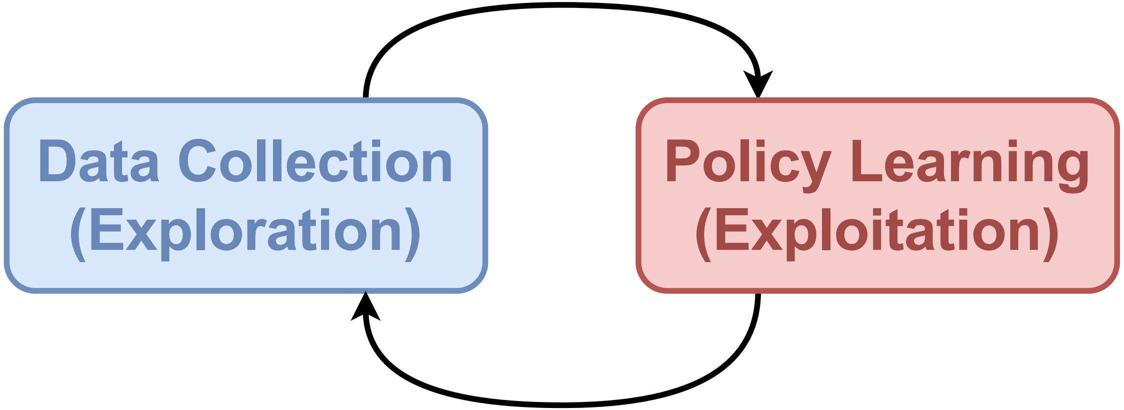
Lukas Schäfer, Filippos Christianos, Josiah P. Hanna, Stefano V. Albrecht

International Conference on Autonomous Agents and Multi-Agent Systems 2022

THE UNIVERSITY of EDINBURGH
informatics

Autonomous Agents
Research Group

RL: Exploration and Exploitation



Intrinsically-Motivated Exploration

Optimise for combined reward signal $r = \boxed{r^e} + \boxed{\lambda\, r^i}$

extrinsic reward (task objective)

intrinsic reward (exploration objective)

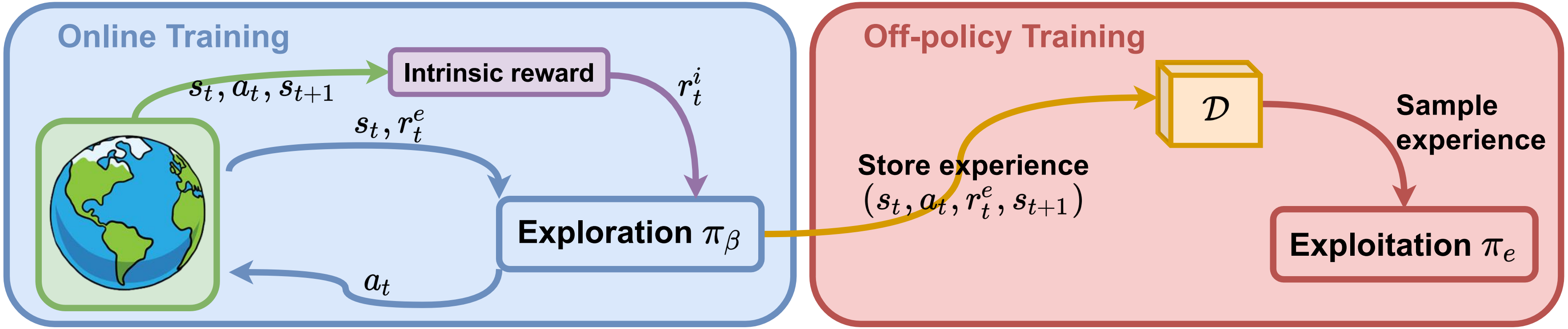
Challenges

1. Non–stationary reward shaping with r^i
2. Sensitivity to scaling factor λ
3. Sensitivity to rate of decay of r^i

Task-specific challenges and sensitivity require extensive hyperparameter search

→ Already biased exploration!

Decoupled Reinforcement Learning (DeRL)



Decoupled Reinforcement Learning (DeRL)

Algorithm Decoupled Reinforcement Learning

Initialise: parameters ϕ, θ and π_β
$\mathcal{D} \leftarrow \emptyset$
$i \leftarrow 0$
for ep $= 0, \ldots, N_{ep}$ **do**
  for $t = 0, \ldots, T$ **do**
    $a_t \sim \pi_\beta(s_t)$
    $s_{t+1}, r_t^e \leftarrow$ environment step with a_t
    Update π_β using RL with intrinsic rewards
    $\mathcal{D} \leftarrow \mathcal{D} \cup (s_t, a_t, r_t^e, s_{t+1})$
    $i \leftarrow i + 1$
    if $i \mod T_{Dec} = 0$ **then**
      Update π_e using off-policy RL on $\mathcal{D}$
    end if
  end for
end for

1. Follow exploration policy π_β
2. Update exploration policy
3. Store experience with extrinsic rewards in $\mathcal{D}$
4. Update exploitation policy π_e from **off-policy** experience in $\mathcal{D}$ generated by π_β

Evaluation - Algorithms

RL Baselines (w/o and w/ intrinsic rewards)

- Advantage Actor-Critic (A2C)
- Proximal Policy Optimisation (PPO)

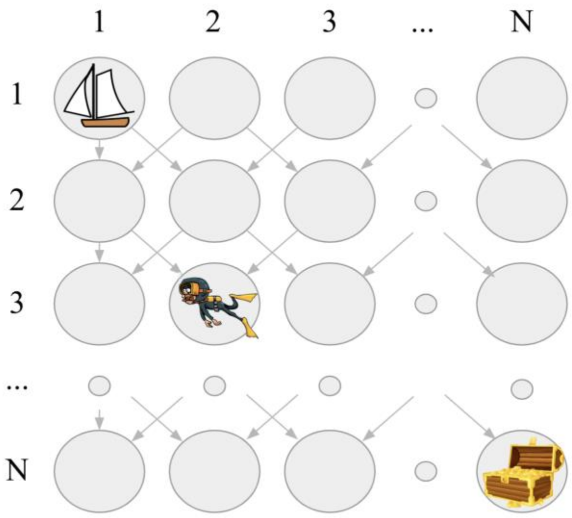
DeRL (π_β trained with A2C and intrinsic rewards)

- DeA2C: π_e trained with A2C
- DePPO: π_e trained with PPO
- DeDQN: π_e trained with DQN

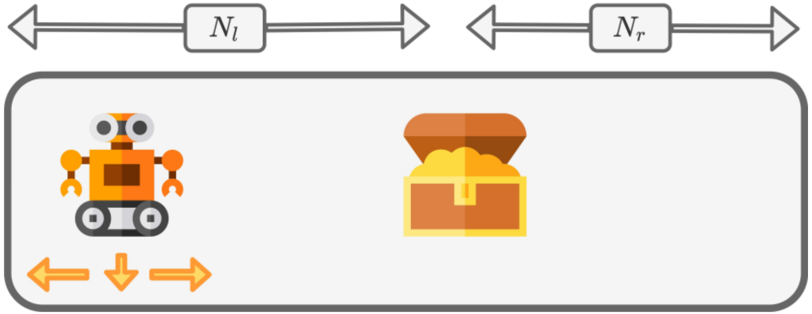
Intrinsic rewards

- **Count**: state counts
- **Hash-Count:** count of state hashes
- **ICM:** Intrinsic Curiosity Module
- **RND:** Random Network Distillation
- **RIDE:** Rewarding Impact-Driven Exploration

Evaluation - Environments

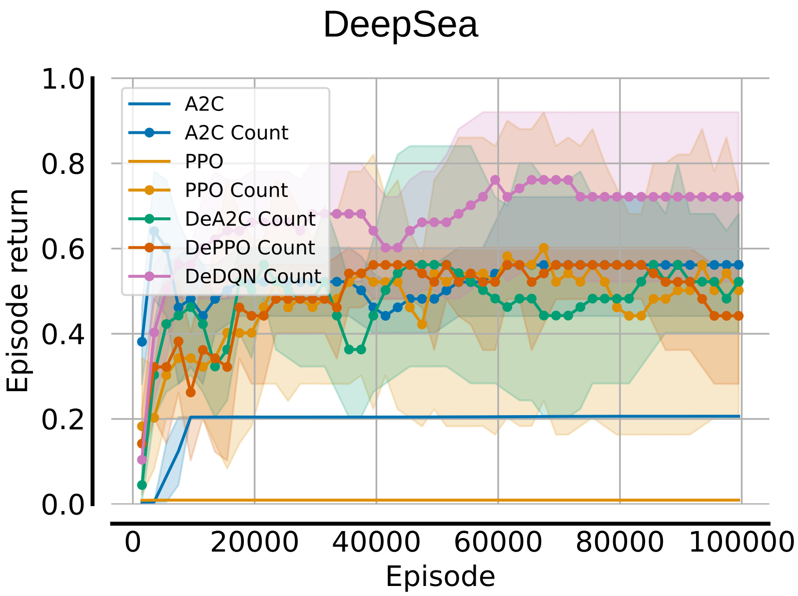


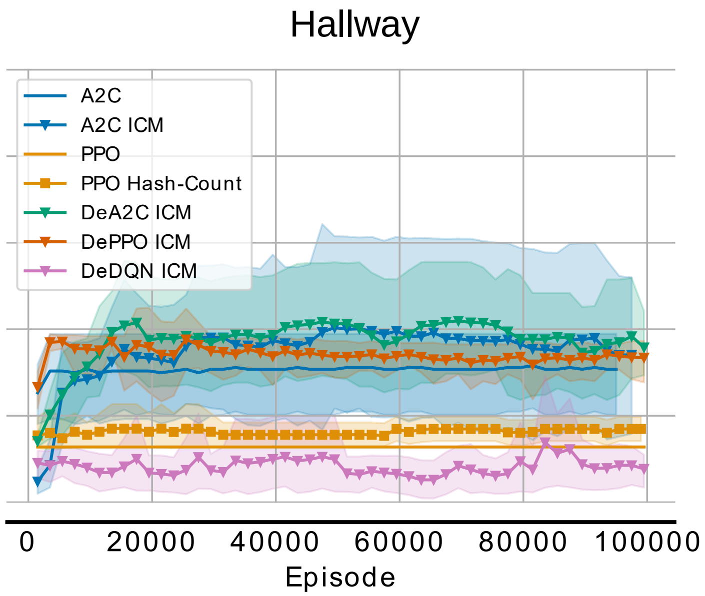
DeepSea environment



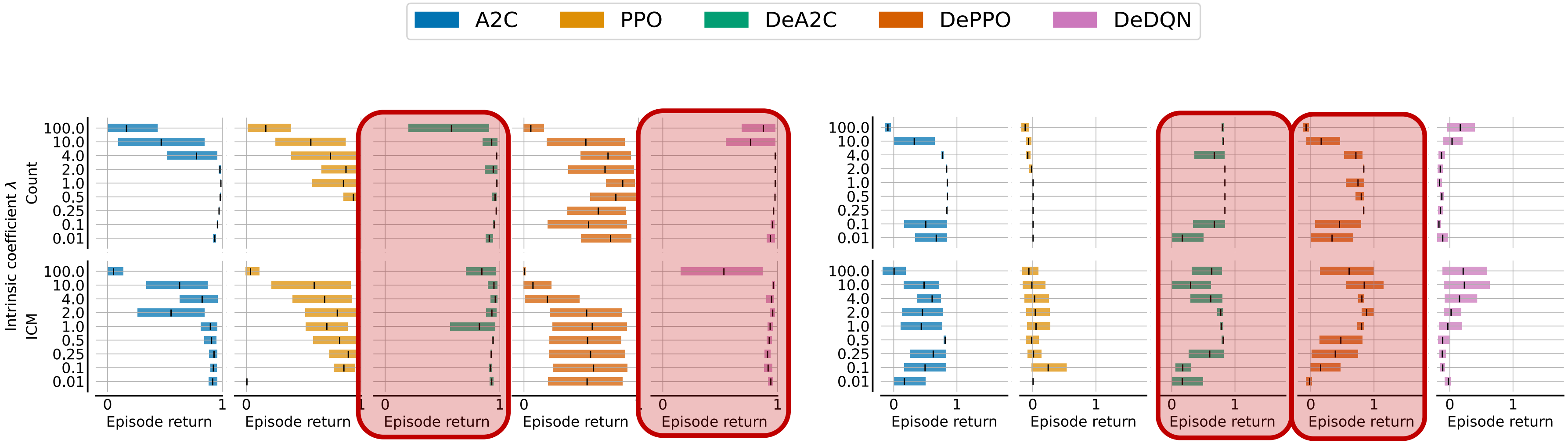
Hallway environment

Evaluation – Normalised Returns





Evaluation – Sensitivity to Intrinsic Reward Scale



DeepSea 10

Hallway $N_l = N_r = 10$

Decoupled Reinforcement Learning to Stabilise Intrinsically-Motivated Exploration

Arxiv: https://arxiv.org/abs/2107.08966
Code: https://github.com/uoe-agents/derl

Contributions:

1. We demonstrate the sensitivity of intrinsically-motivated exploration to hyperparameters.
2. We propose to train decoupled policies for exploration and exploitation to stabilise returns.

See us during slots **1A5-3 (Day 1)** and **3C1-2 (Day 3)**

THE UNIVERSITY of EDINBURGH informatics
Autonomous Agents Research Group