When should agents explore? A DeepMind research paper

Niklas Forsstroem
forsstroemniklas@gmail.com

Introduction

- ❑ Exploration is about the balance between taking the familiar choice that is known to be rewarding and learning about unfamiliar options which could be more valuable than the familiar options.

- ❑ Kembro et al. showed that bees modulate their foraging excursions by considering trade-off dynamics between
  - ❑ Visiting and exploiting flowers close to the nest,
  - ❑ Searching for new routes and resources
  - ❑ Exploiting learned flower visitation sequences.

- ❑ Experienced bees combine these behavioural strategies even after they find an optimal route that minimizes travel distances between flowers

- ❑ This paper focuses on when to explore, expanding the class of exploratory behaviours beyond the commonly used monolithic ones (where modes are merged homogeneously in time).

Exploration periods

Varying exploration length

An exploration period is an uninterrupted sequence of steps in explore mode

01 **Step-level:** Decision to explore is taken independently at each step, affecting one action. The canonical example is ε-greedy (with decaying ε)

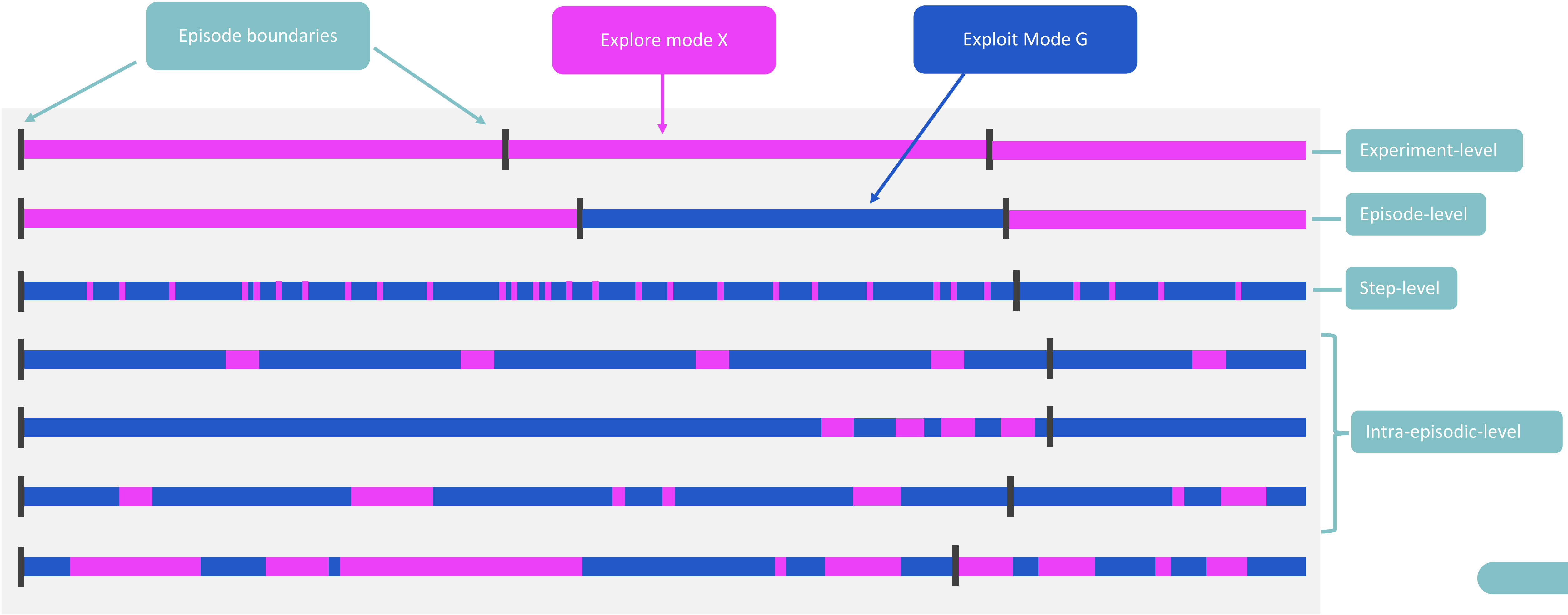
02 **Experiment-level:** All behaviour in training comes from explore mode, and learning is off-policy (greedy policy is only used for evaluation).

03 **Episode-level:** Mode is fixed for an entire episode at a time (e.g., training games versus tournament matches in a sport)

04 **Intra-episodic:** Falls in-between step- and episode-level exploration, where exploration periods last for multiple steps, but less than a full episode.

Exploration periods

Illustrations



Switching mechanisms

Transitions between exploration/exploitation

01

Blind switching: Does not take any state into account

- Only concerned with switching at desired time resolution
- Can be deterministic or probabilistic

02

Informed switching: Informed by agent’s internal state.

- A scalar trigger signal (which proxies uncertainty) at each step
- Binary switching decision is taken based on the trigger signal

03

Starting mode: Whether the agent explores early or later in an episode

- Starting with exploit mode can be beneficial since early states have often been visited many times
- In other domains, early actions may disproportionately determine the available future paths

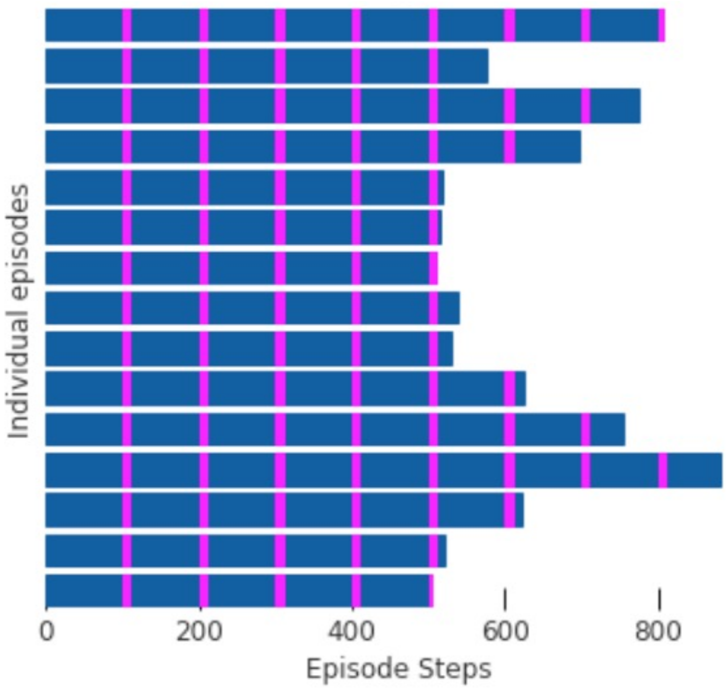
Exploration triggers

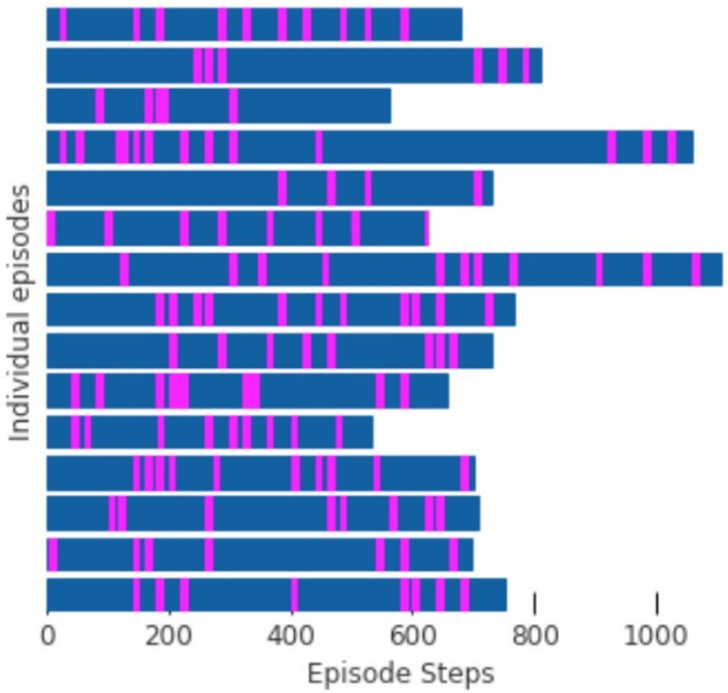
When to start exploring

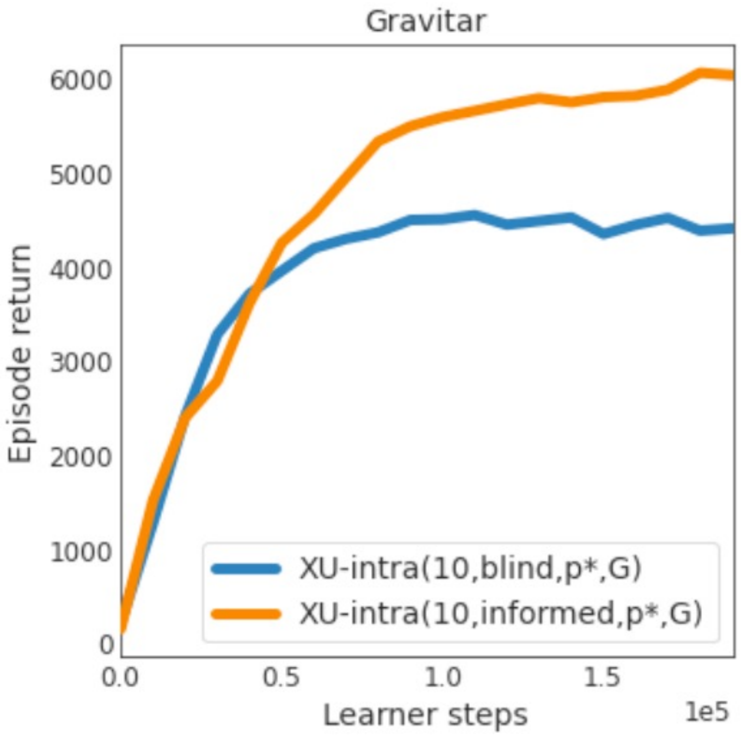
blind, step-based trigger

Informed trigger signal

Superior performance for informed exploration trigger







Two modes of exploration
Uniform and intrinsic reward

In addition to deciding when to explore, we must also agree on how to select our actions whilst exploring

01

Uniform explore mode $\mathcal{X}_U$: The naive uniform random policy

- Selects action uniformly from viable alternatives
- Works well when environment has dense rewards that are easy to find by taking random action sequences
- Tends to fail when the rewards are sparse and hard to find (e.g. Montezuma's Revenge)

02

RND intrinsic reward explore mode $\mathcal{X}_I$: Favors unfamiliar states

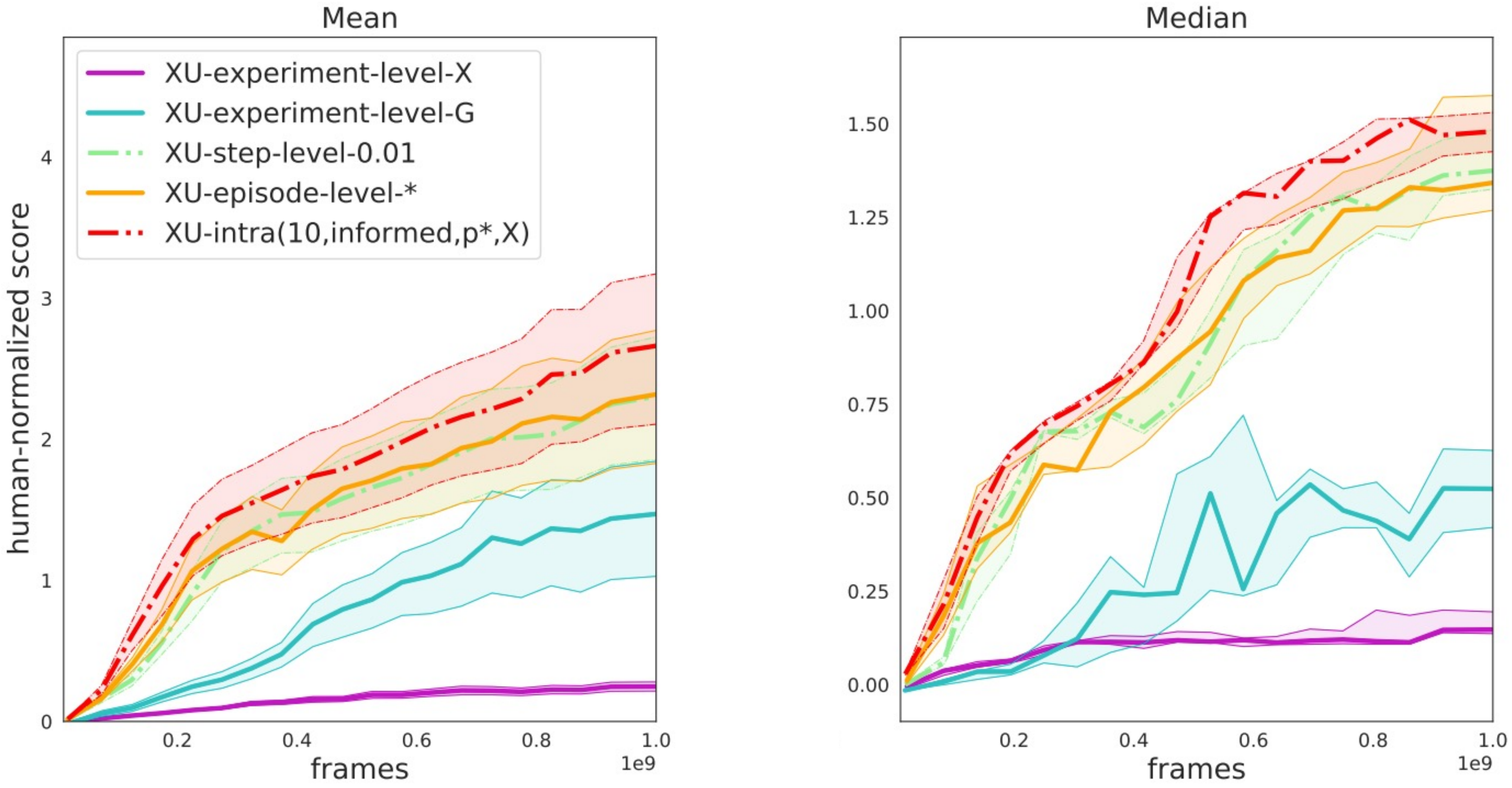
- Measures how hard it is to predict the output of a fixed random neural network on visited states.
- Unfamiliar state outputs are hard to predict and get favored. [1]

03

Softer explore-exploit modes: Alternative to standard uniform exploration, e.g:

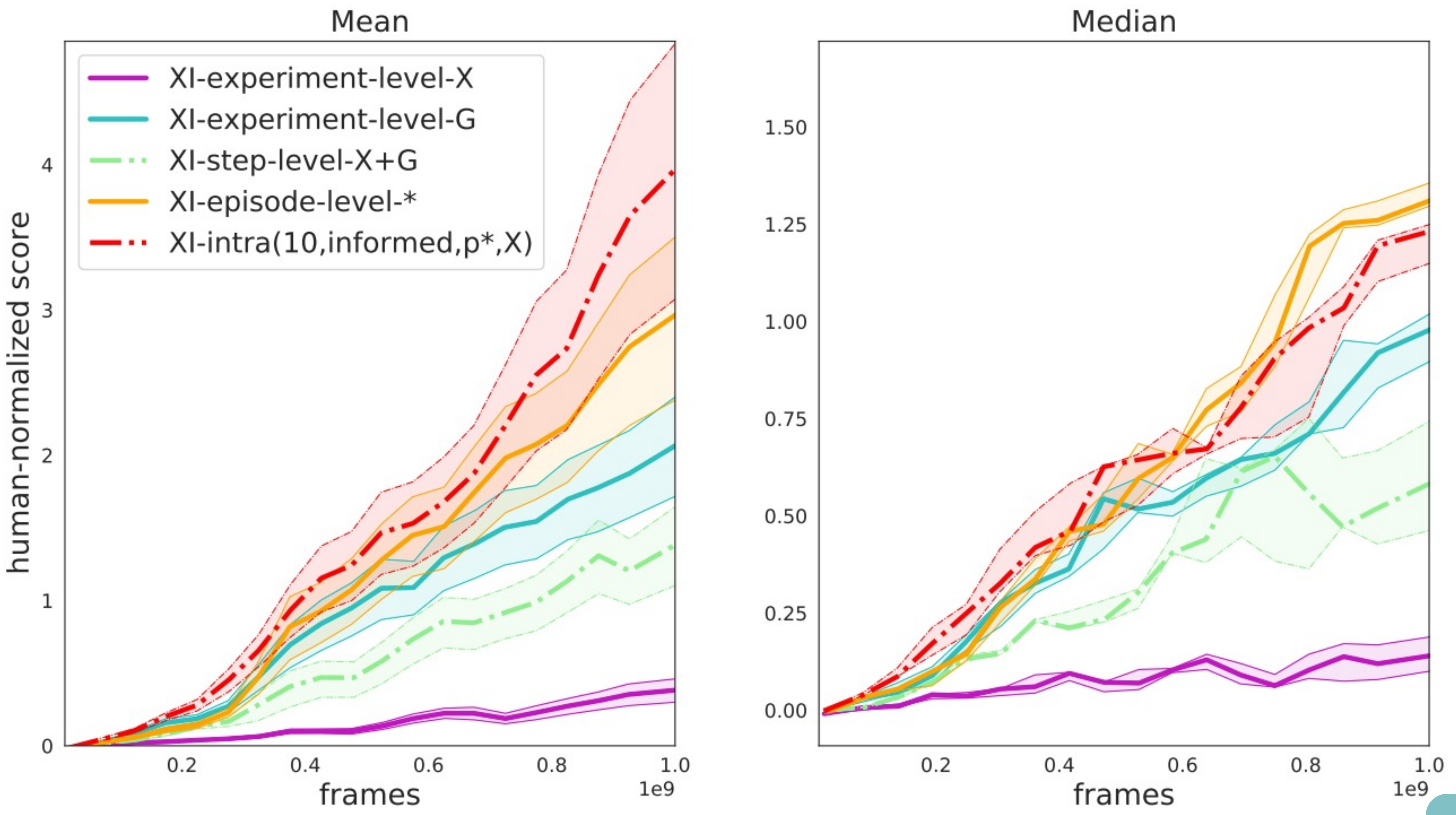
- ε-greedy explore mode with ε = 0.4
- ε-greedy exploit mode with ε = 0.1

Uniform exploration



Intrinsic reward exploration

Intra-episodic exploration improves over both step-level and episode-level baselines



Self-regulation
Flexibility without additional burden

01

Bandit adaptation

Allows for dynamic adaptation of when to enter/exit explore mode

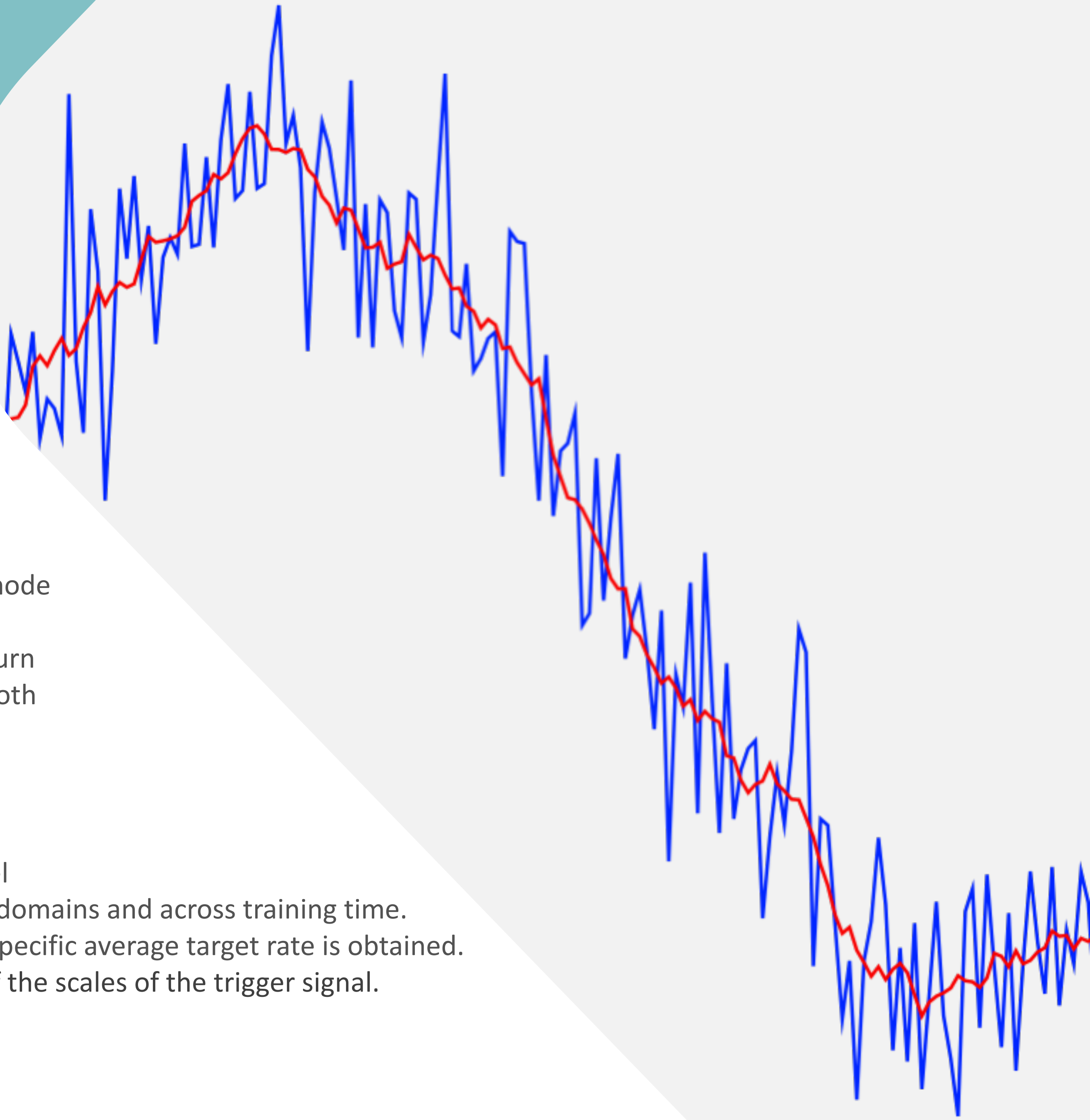
- Adaptation by a meta-controller, implemented as a non-stationary multi-armed bandit that maximizes episodic return
- Allows the “when” of exploration to become adaptive to both the task, and the stage of learning

02

Homeostasis

Makes algorithm robust against variability in trigger signal level

- The informed trigger signals may vary substantially across domains and across training time.
- Homeostasis adapts the threshold for switching so that a specific average target rate is obtained.
- The target rate of switching is configured independently of the scales of the trigger signal.



Starting mode effect

Uniform exploration

Overall exploratory proportion: $p_{\mathcal{X}}$
length of an exploratory period: $n_{\mathcal{X}}$
Episode length: L

Relative summary statistic of exploration:

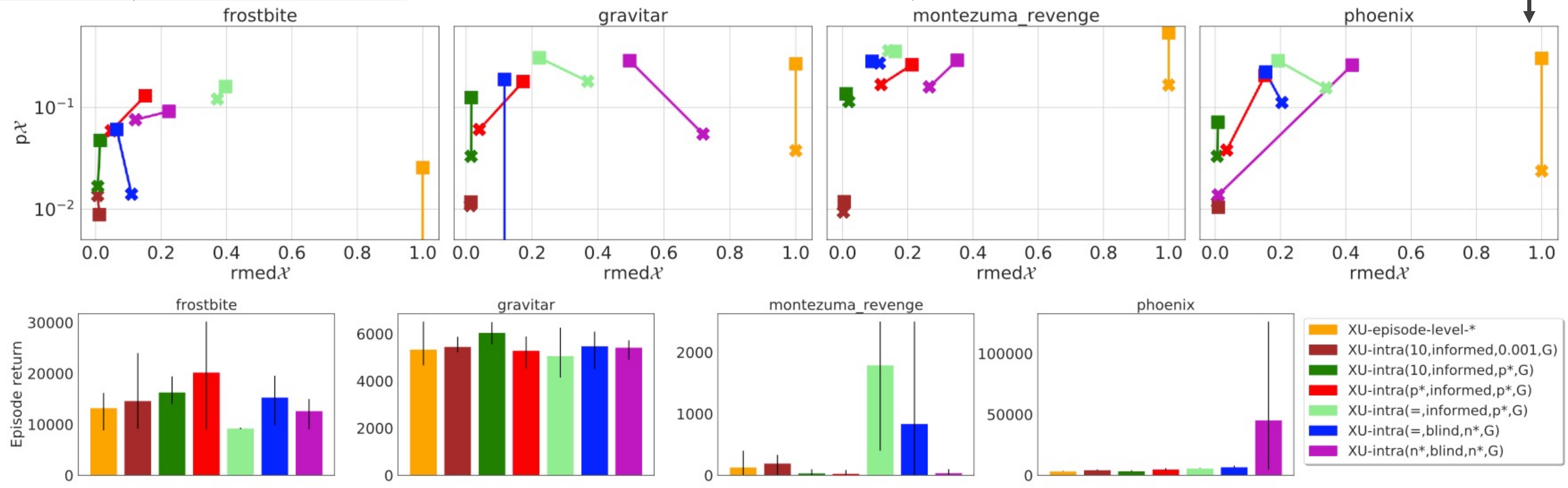
$$rmed_{\mathcal{X}} := median(n_{\mathcal{X}}/L)$$

Only game where **informed trigger** clearly outperforms its **blind equivalent**

Best results on Montezuma's revenge come from **symmetric, informed trigger variant**, which is forced to retain a high $p_{\mathcal{X}}$.

Only the **blind, doubly adaptive model** escapes the local optimum of phoenix.

The bandits radically shift exploration statistics over the course of training.



Starting mode effect

Intrinsic reward exploration

Overall exploratory proportion: $p_\mathcal{X}$
length of an exploratory period: $n_\mathcal{X}$
Episode length: L

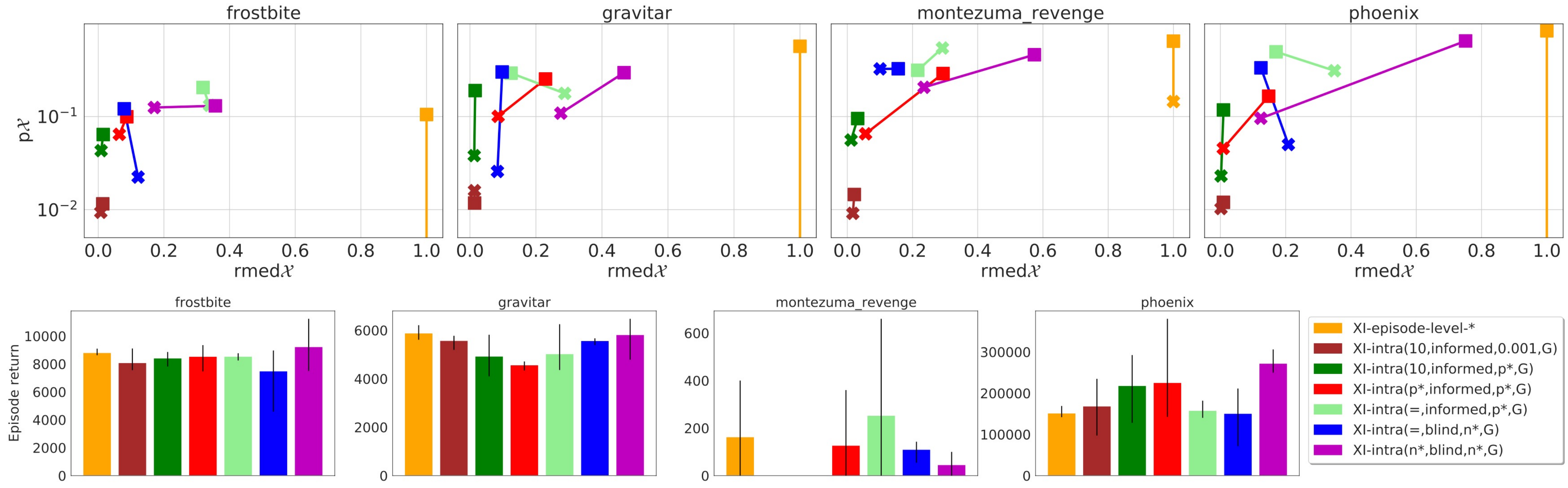
Relative summary statistic of exploration:

$$rmed_\mathcal{X} := median(n_\mathcal{X}/L)$$

The bandit adaptation shifts the statistics into different directions for different games.

a common pattern is that reducing $p\_\mathcal{X}$ far below 0.5 is needed for high performance.

The raw amount of exploration $p_\mathcal{X}$ is not a sufficient predictor of performance, implying that the temporal structure matters.



Starting mode effect

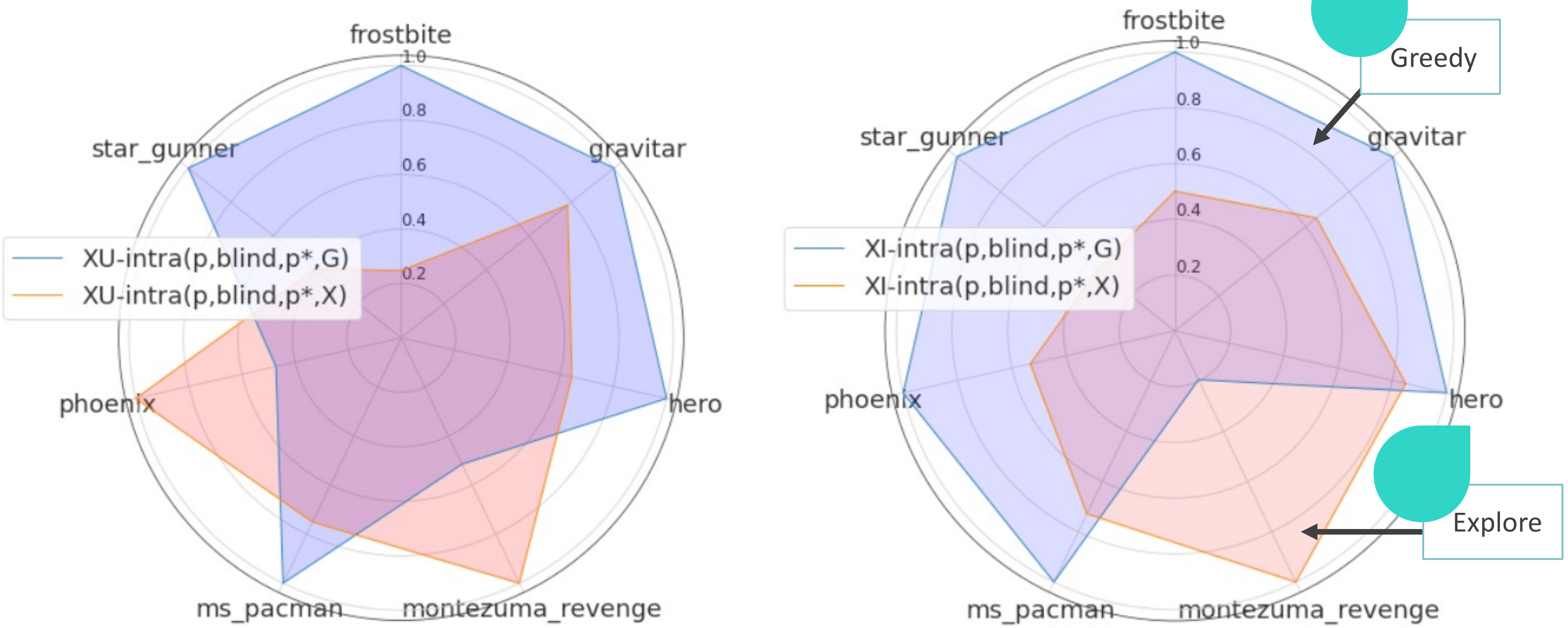
Greedy vs explore mode

Starting mode: Whether the agent explores early or later in an episode. Especially important for games with long periods

- ❑ Starting with exploit mode can be beneficial since early states have often been visited many times
- ❑ In other domains, early actions may disproportionately determine the available future paths

Starting in explore mode can be consistently beneficial in some games, and consistently harmful in others

When starting in exploit mode, agent often prefers long initial exploit periods (up to 10’000 steps)



Greedy

Explore

Conclusions

- ❑ For exploration, there seems to be an optimum in terms of temporal granularity, and intra-episodic exploration is the right step towards finding it.
- ❑ The raw amount of exploration p_X is not a sufficient predictor of performance, implying that the temporal structure matters.
- ❑ Reducing p_X far below 0.5 seems necessary for high performance.
    - ❑ Appears common between XU and XI, despite very different explore modes (and differing performance)
- ❑ Prolonged intrinsic exploration seems more useful than prolonged randomness

Thank You