

No Representation, No Trust

Connecting Representation, Collapse,
and Trust Issues in PPO



Skander Moalla¹ Andrea Miele¹ Daniil Pyatko¹ Razvan Pascanu² Caglar Gulcehre¹

¹. EPFL

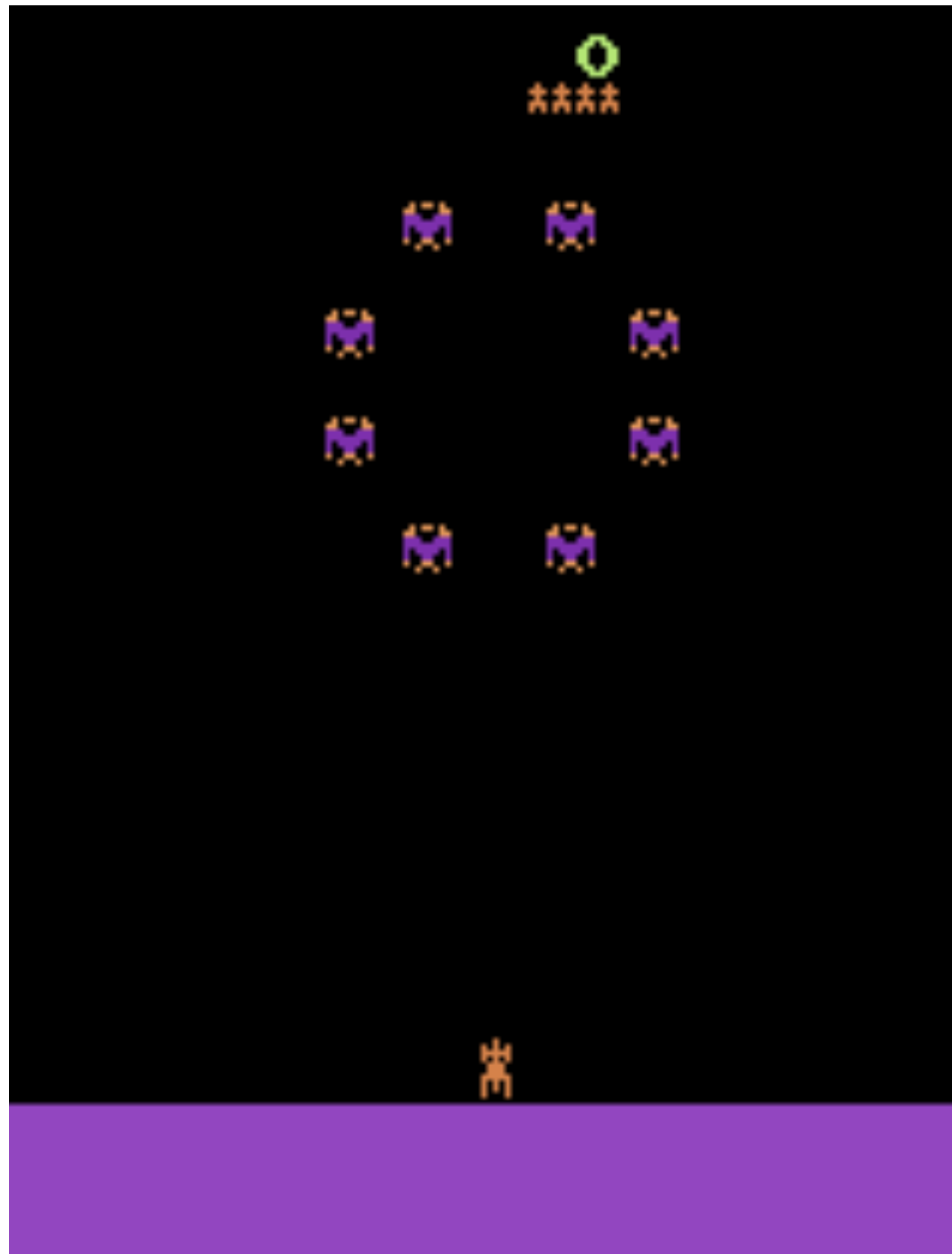


¹. CLAIRE

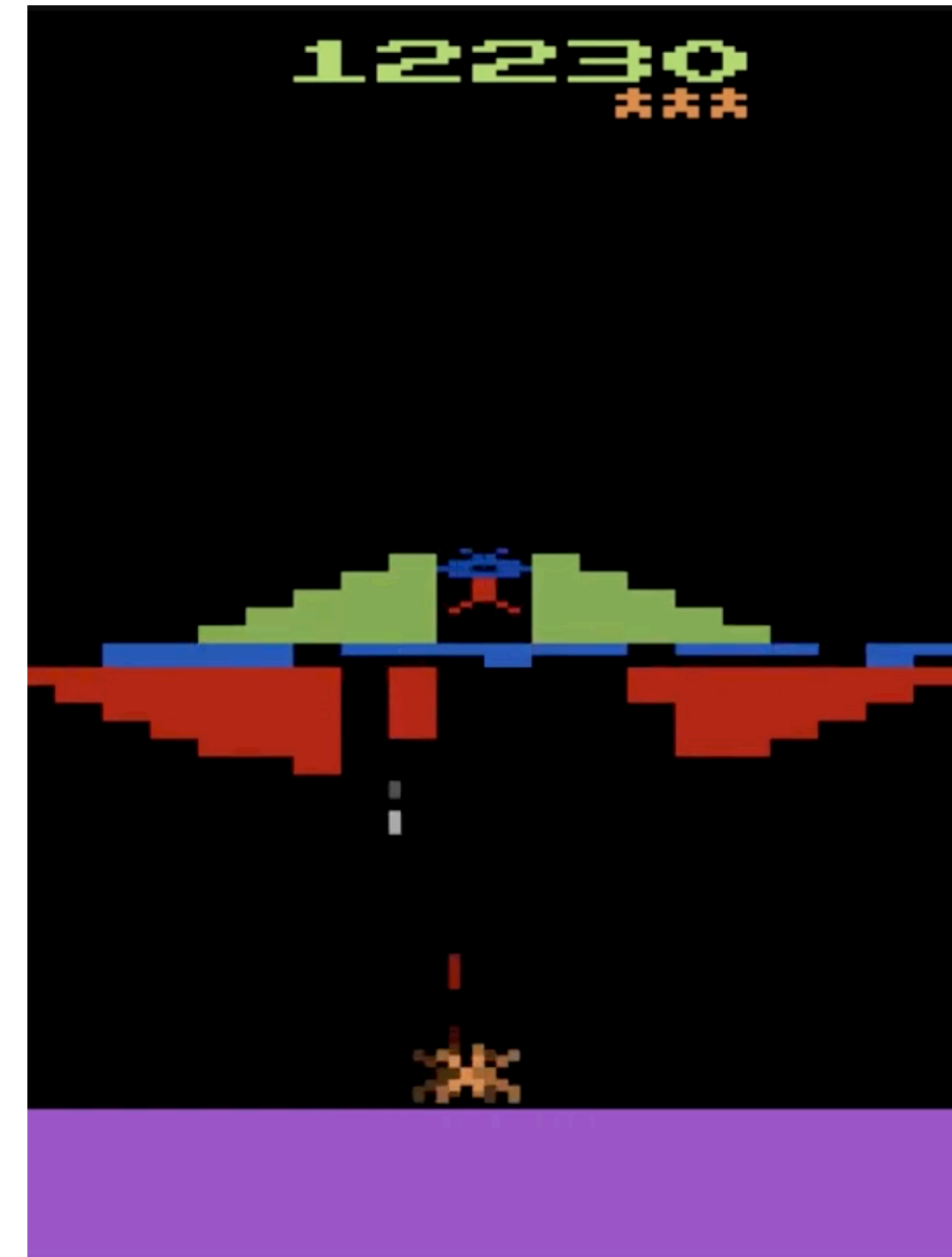
². Google DeepMind

Non-stationarity

A core feature of reinforcement learning



https://ale.farama.org/_images/phoenix.gif



<https://www.youtube.com/watch?v=3ILULkcBRG0>

Non-stationarity

Neural networks in deep RL need to adapt to changing distributions

Policy network π_θ and value network \hat{v}_w

$$L(w) = \mathbb{E}_{\pi_{\theta_{curr}}} \left[\sum_t (\hat{v}_w(S_t) - G_t)^2 \right]$$

$$\tilde{J}(\theta) = \mathbb{E}_{\pi_{\theta_{curr}}} \left[\sum_t G_t \log \pi_\theta(A_t | S_t) \right]$$

$\mathbb{E}_{\pi_{\theta_{curr}}}$: expectation over trajectories from the current policy

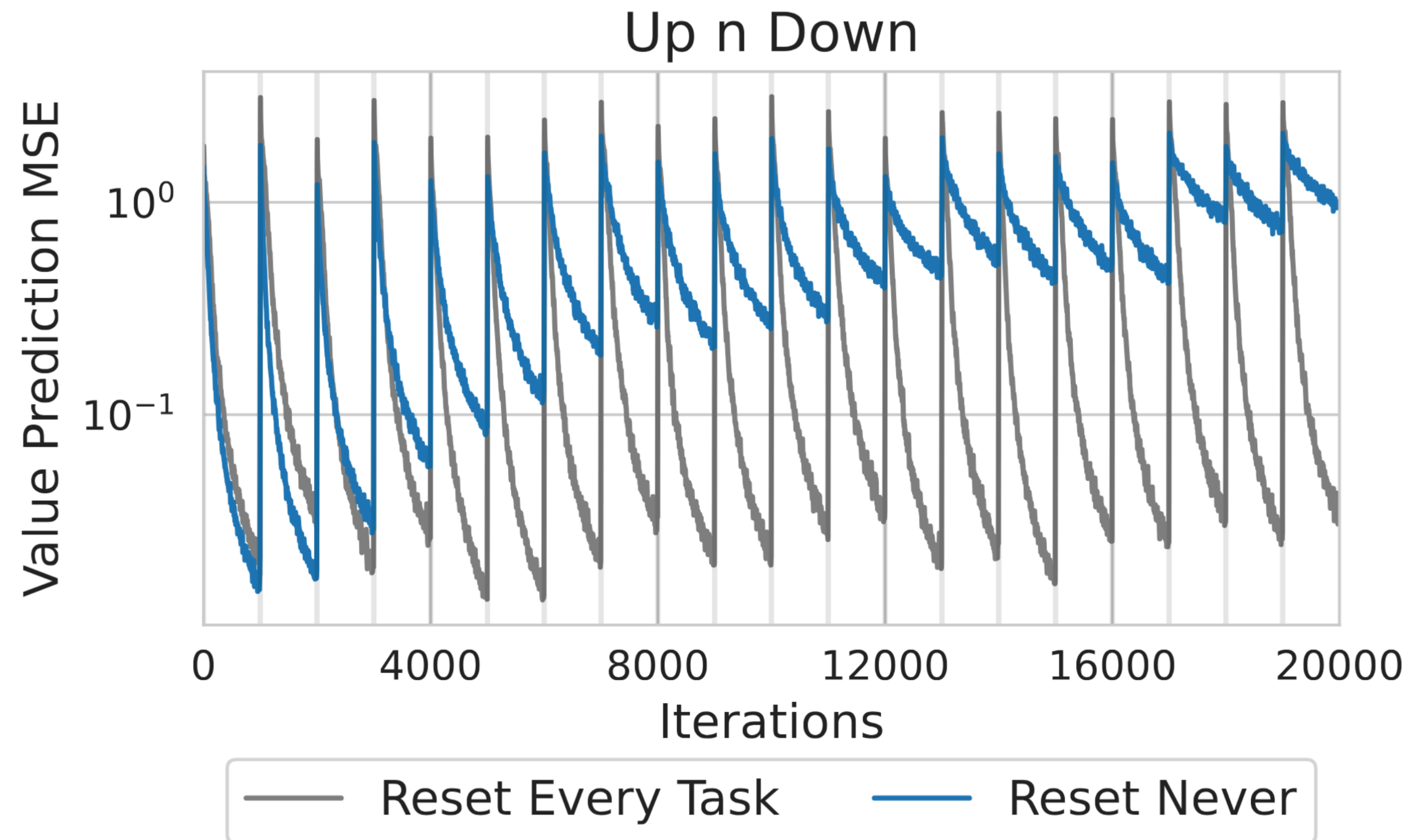
S_t : state at timestep t

A_t : action at timestep t

G_t : return (sum of rewards) after timestep t

Plasticity/Capacity loss

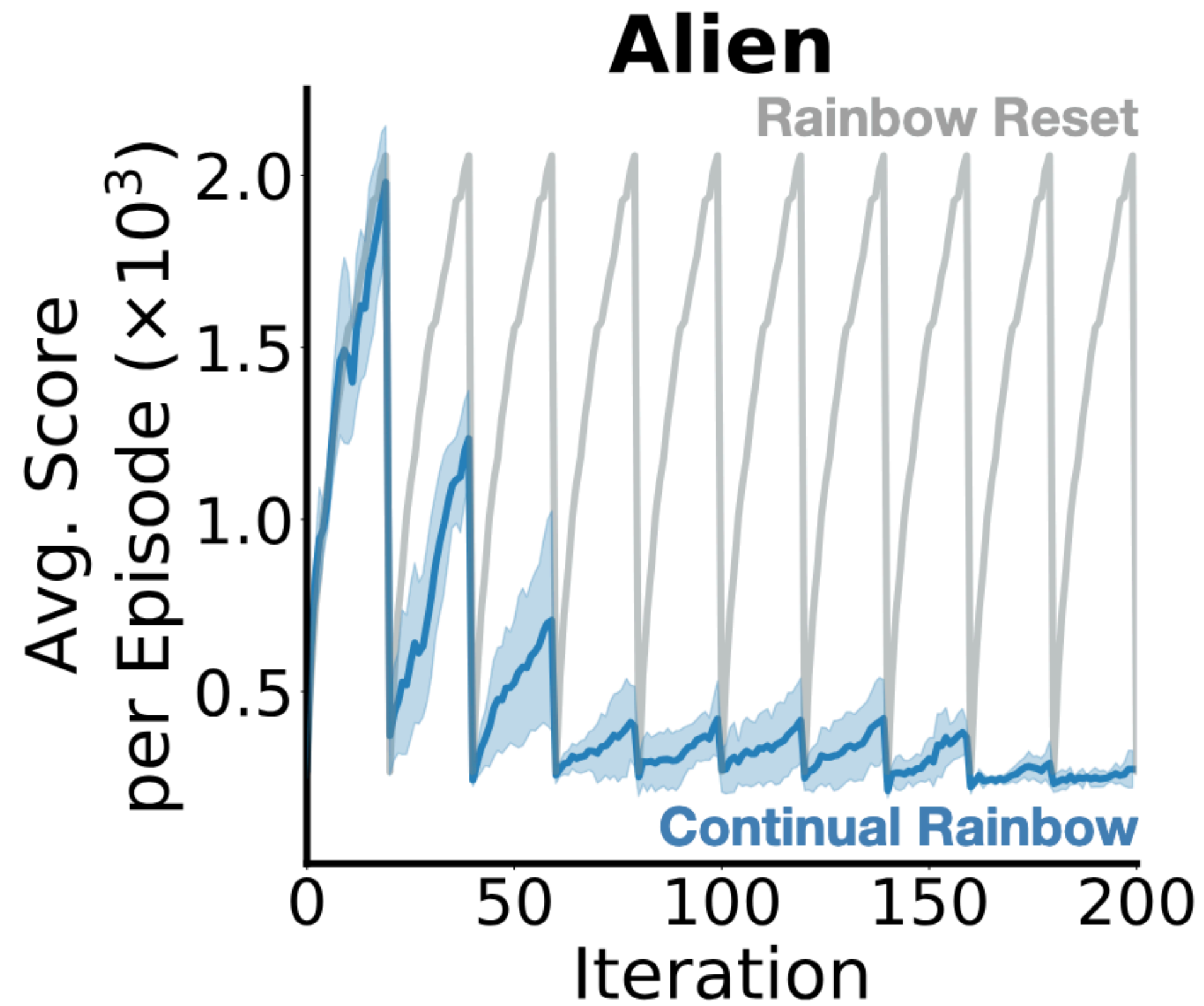
Same network is less able to fit a sequence of targets than a re-initialized network



Nikishin, Evgenii, et al. "Deep reinforcement learning with plasticity injection." *Advances in Neural Information Processing Systems* 36 (2024).

Plasticity loss

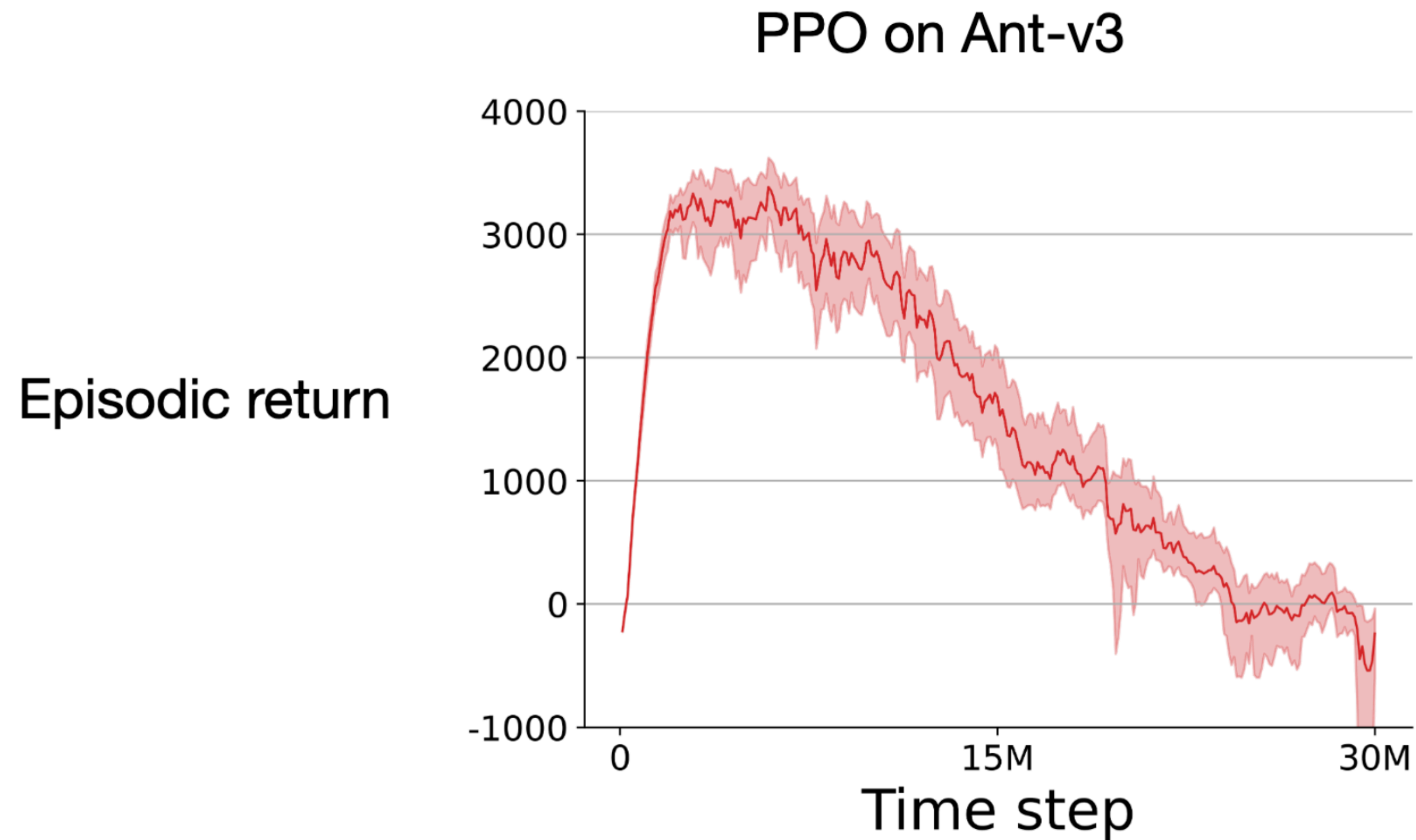
Same network is less able to learn than a re-initialized network



Abbas, Zaheer, et al. "Loss of plasticity in continual deep reinforcement learning." Conference on Lifelong Learning Agents. PMLR, 2023.

Plasticity loss

Inability to continue learning with non-stationarity



Dohare, Shibhansh, et al. "Maintaining plasticity in deep continual learning." arXiv preprint arXiv:2306.13812 (2023).

Rank collapse

Feature layer's rank decreases rapidly

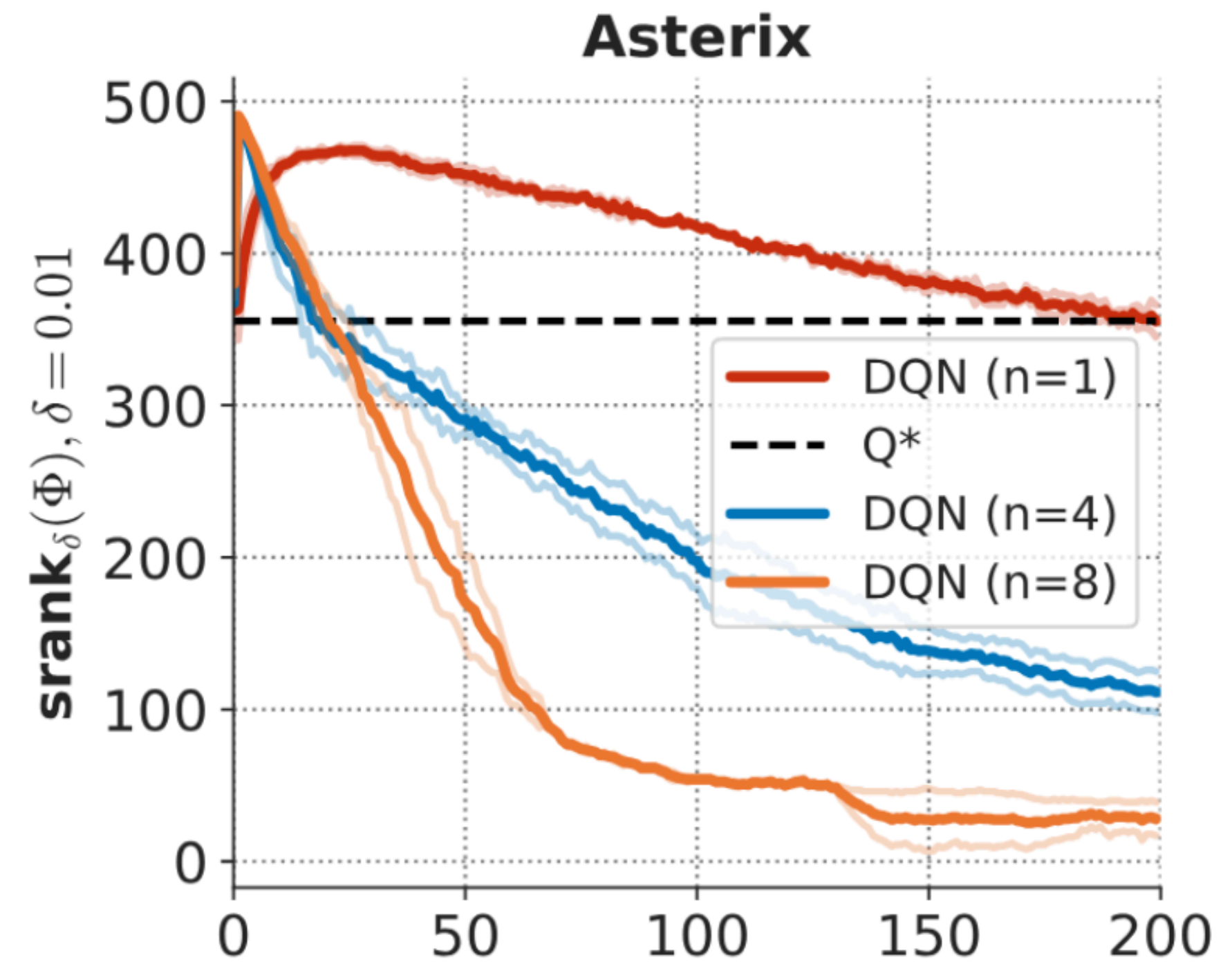
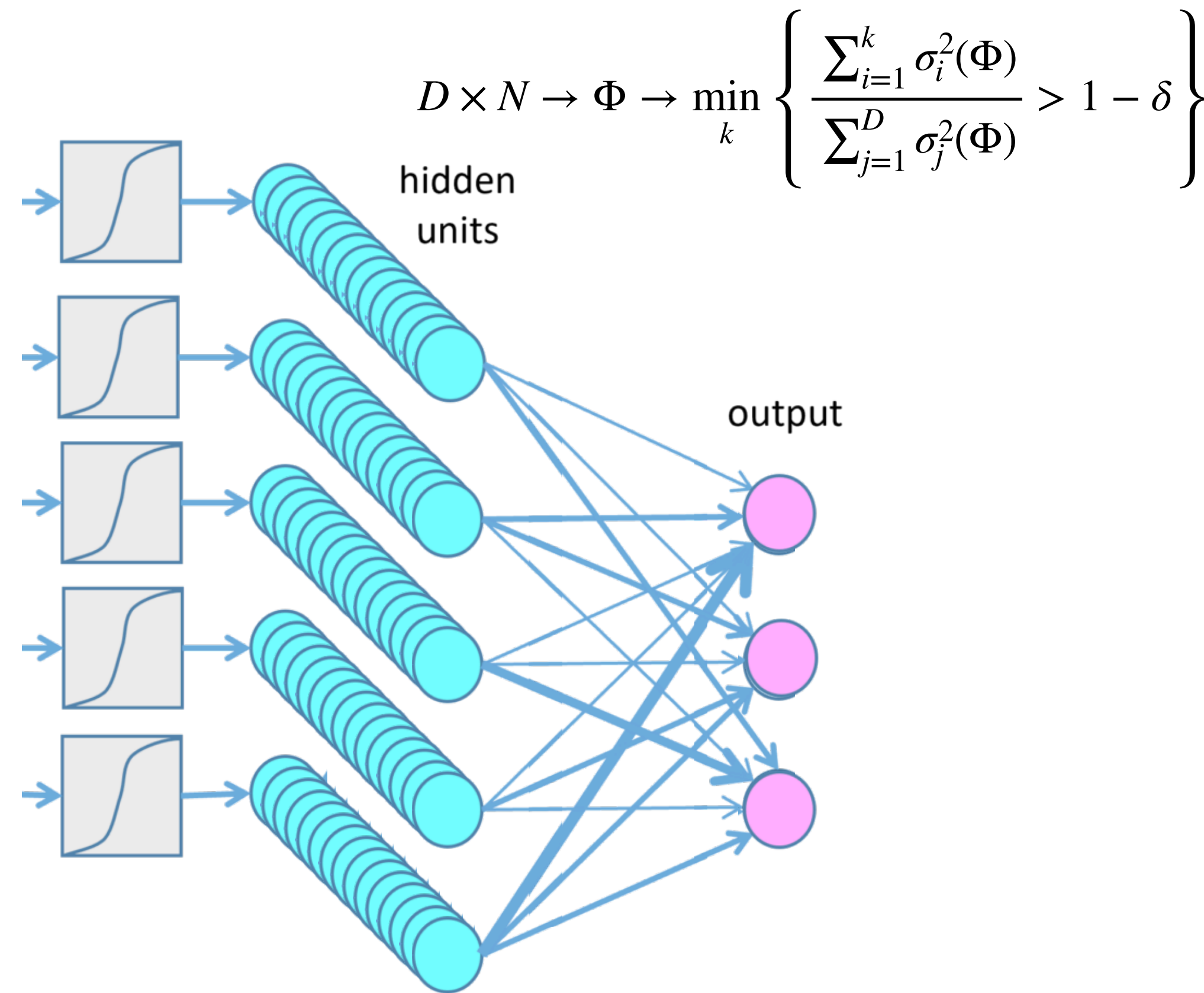


Diagram adapted from <https://lamarr-institute.org/blog/deep-neural-networks/>
Kumar, Aviral, et al. "Implicit under-parameterization inhibits data-efficient deep reinforcement learning." arXiv preprint arXiv:2010.14498 (2020).

Previous work

Non-stationarity is detrimental to deep learning

	Non-stationary supervised learning	RL: Value-based methods (<i>DQN</i>)	RL: Policy optimization (<i>PPO</i>)
Performance Performance collapse <i>Plasticity loss</i>	✓	✓	✓
Representation <i>Rank decrease</i> <i>Capacity loss</i>	✓	✓	?

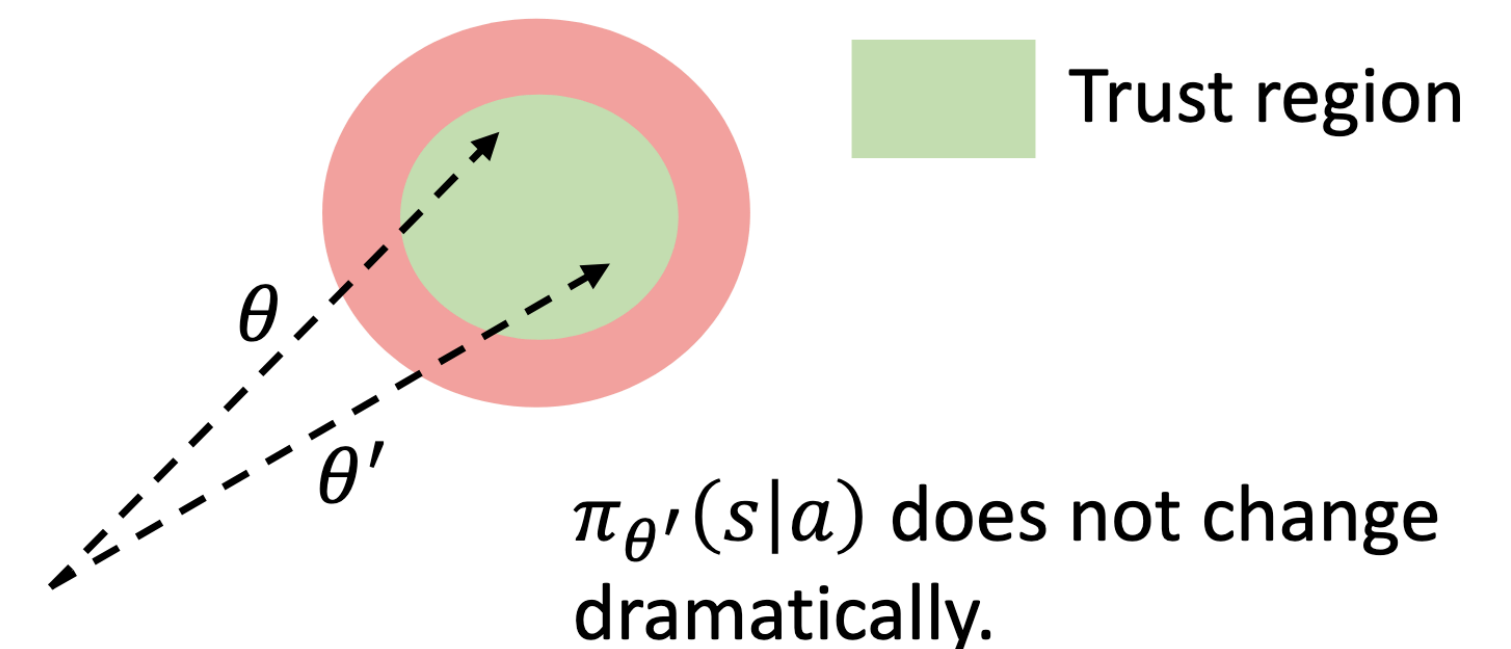
Open questions

Much more to understand about Proximal Policy Optimization (PPO)

- Are representations affected by non-stationarity?
- How does multi-epoch optimization play with non-stationarity?










$$L_{\pi_{\text{old}}}^{CLIP}(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\text{old}}} \left[\sum_{t=0}^{t_{\text{max}}-1} \min \left(\frac{\pi_{\boldsymbol{\theta}}(A_t|S_t)}{\pi_{\text{old}}(A_t|S_t)} \Psi_t, \text{clip} \left(\frac{\pi_{\boldsymbol{\theta}}(A_t|S_t)}{\pi_{\text{old}}(A_t|S_t)}, 1 + \epsilon, 1 - \epsilon \right) \Psi_t \right) \right]$$

- How can PPO collapse despite its trust region?



Our contributions

PPO suffers from a deteriorating representation that breaks its trust region

	Findings	Implications
Representation Multi-epoch optimization	 Collapse  Faster collapse	 More perspective on plasticity loss
Trust region	 Fails with poor representations	 Better understanding of trust-region failure
Interventions	 Better representation -> better trust region	 Representations should be monitored
Proximal Feature Optimization	 Extending trust region to features helps	 Design more interventions

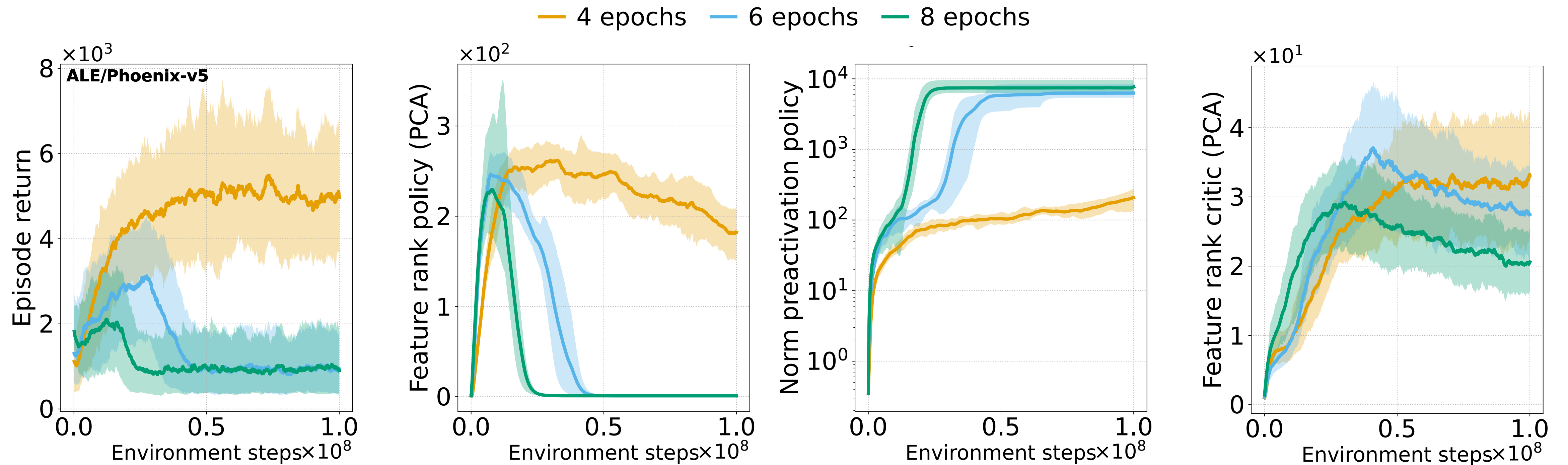


Fully reproducible and replicable!

All runs available on W&B and with raw logs and checkpoints!

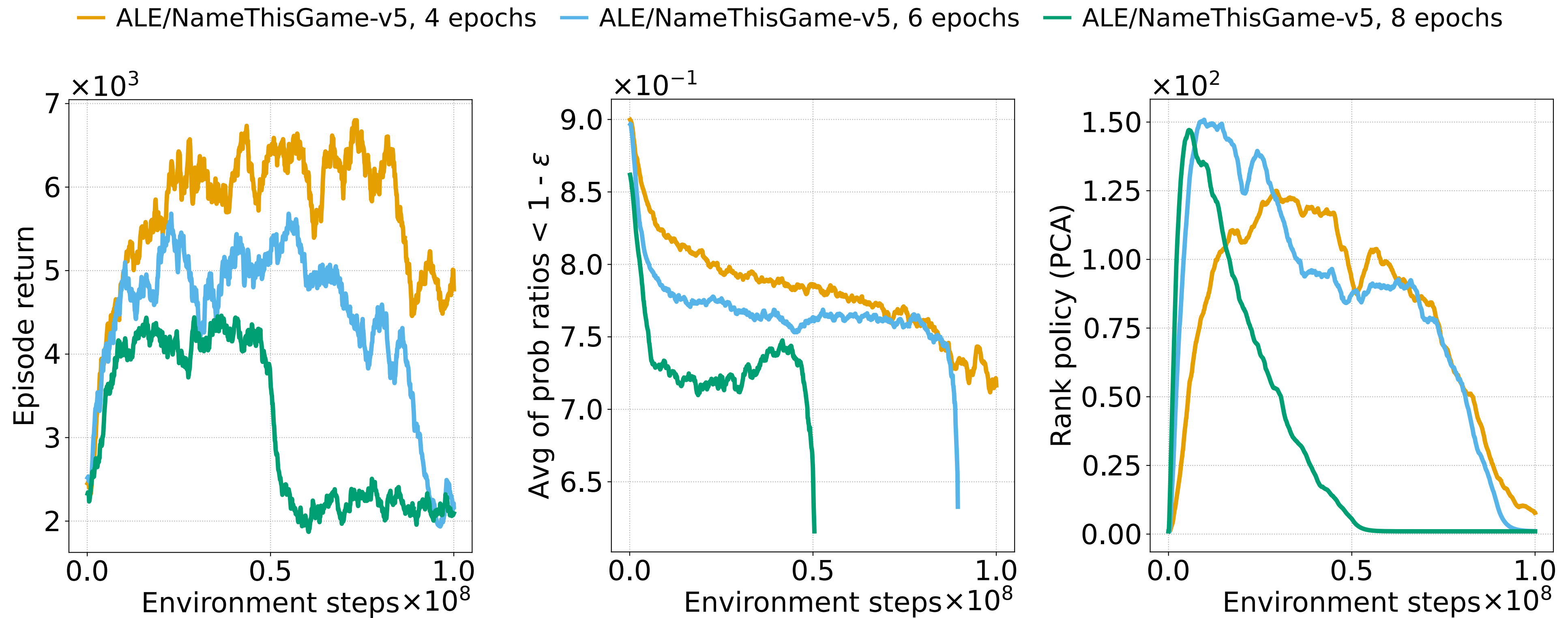
PPO suffers from deteriorating representations

The collapse is faster with stronger non-stationarity, achieved with more epochs



The trust region fails

It cannot prevent the catastrophic change; it breaks down with a poor representation



Why does the trust region fail?

It's not possible to maintain the trust region with a collapsed representation

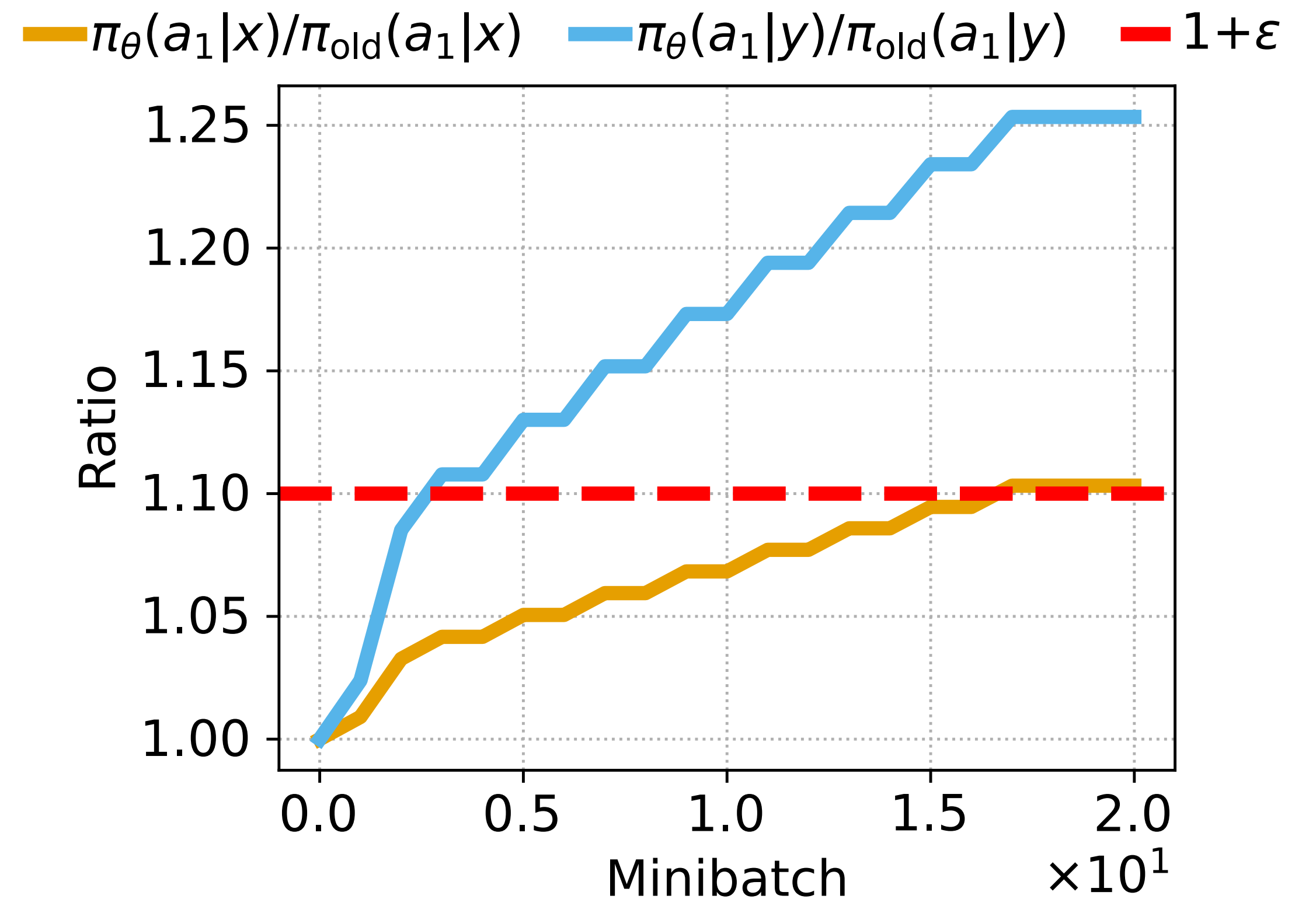
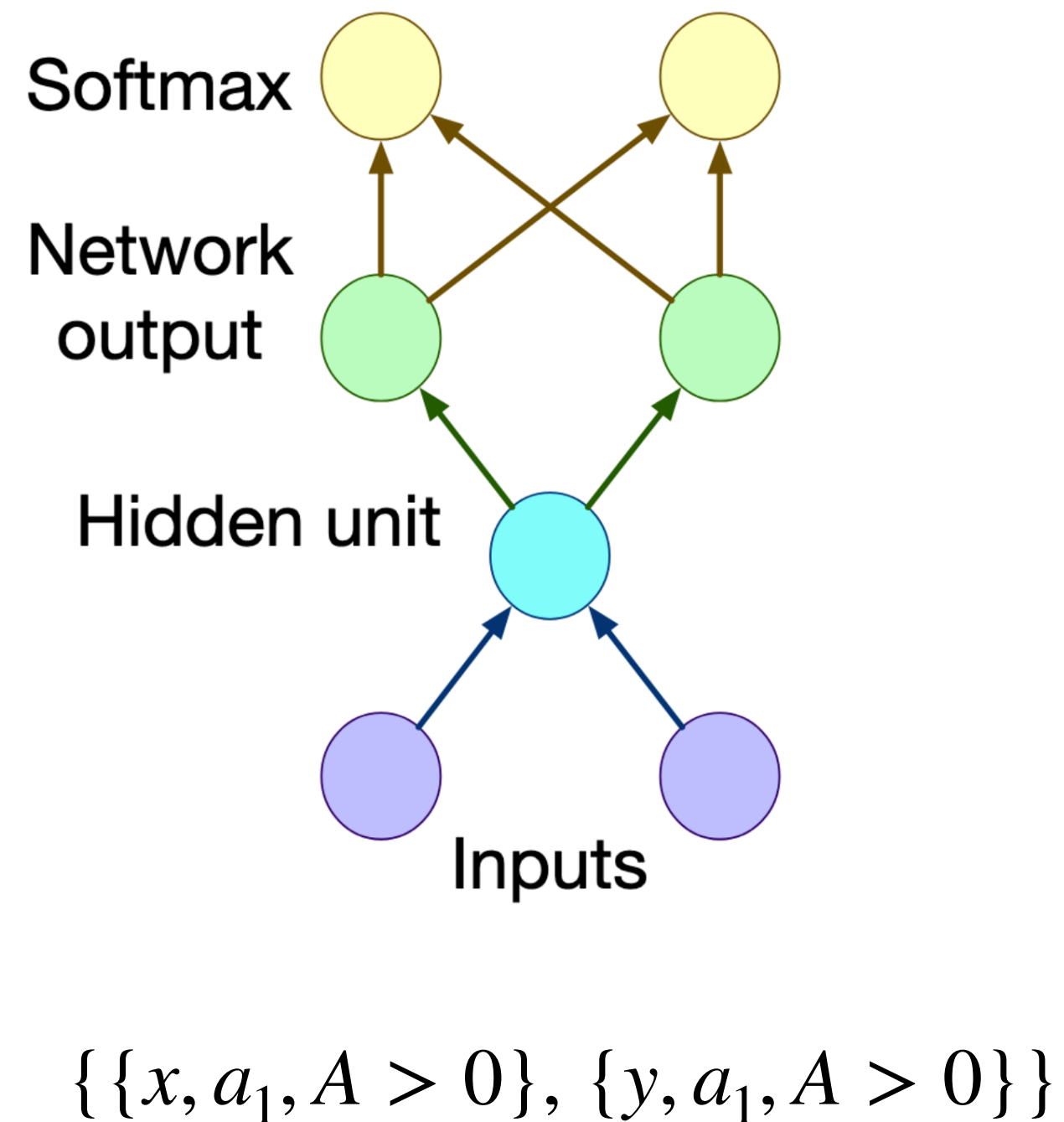
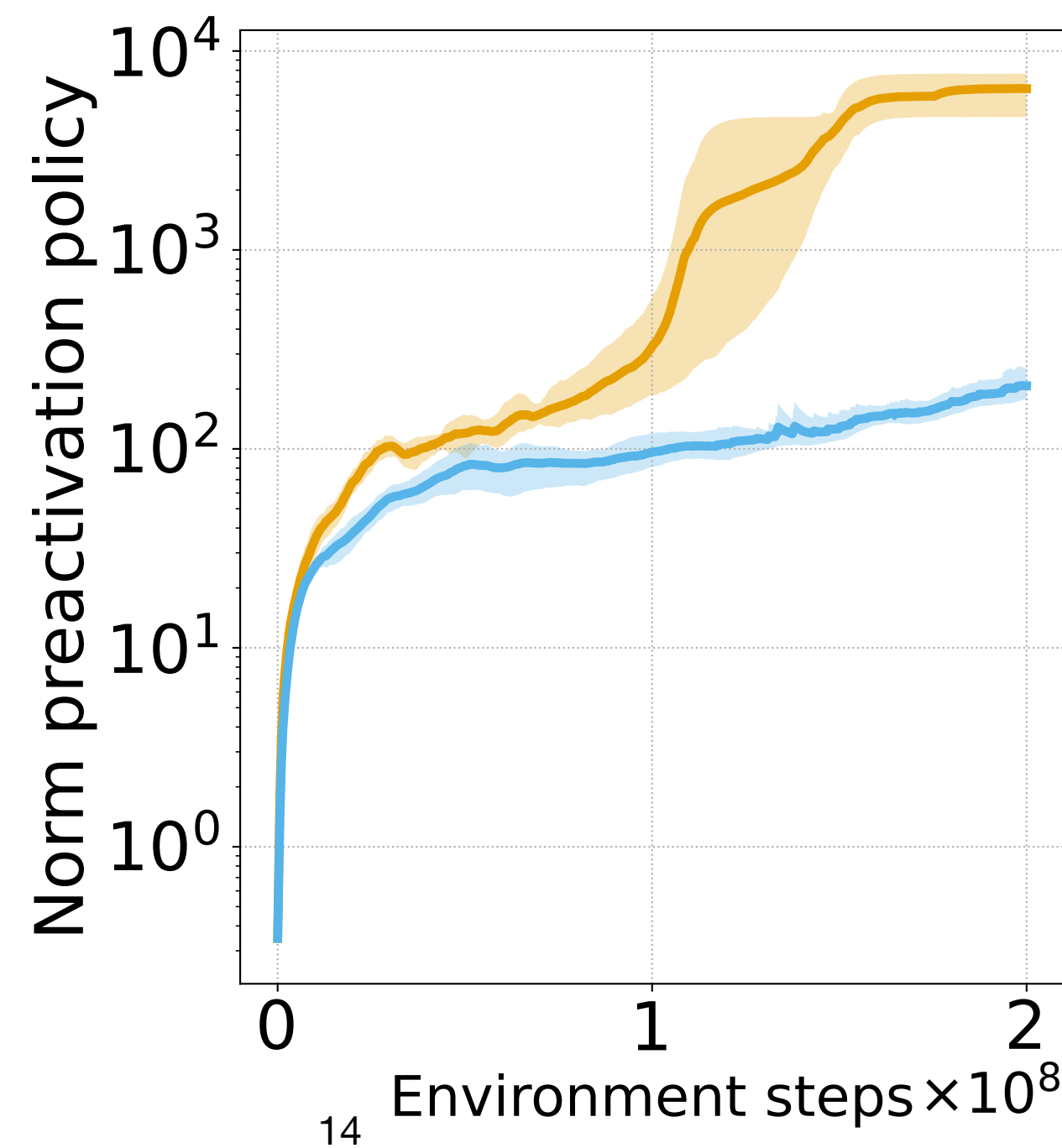
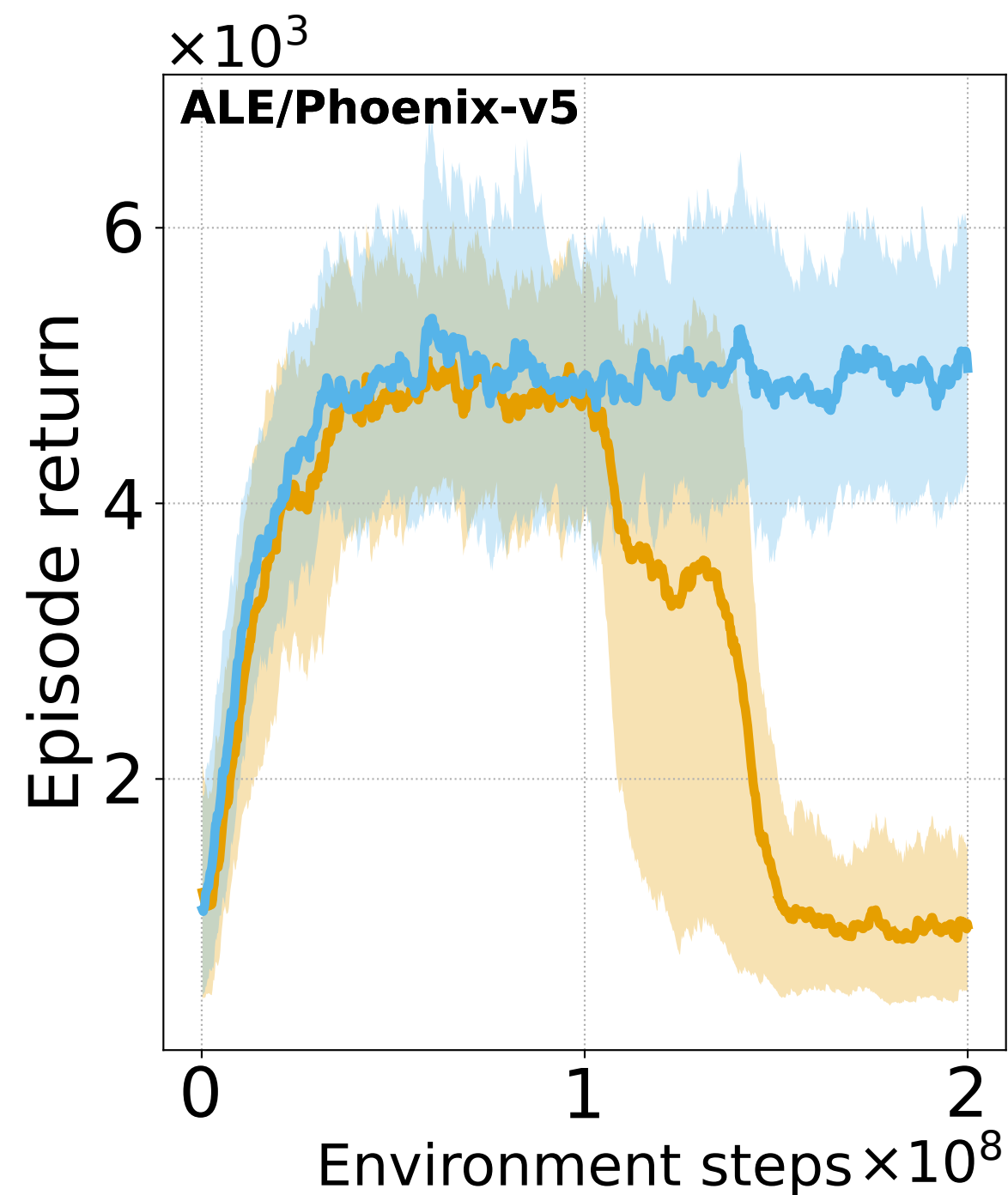


Diagram from Dohare et al (2023).

Proximal Feature Optimization (PFO)

Extending trust region to features helps

$$L_{\pi_{old}}^{PFO}(\theta) = \mathbb{E}_{\pi_{old}} \left[\sum_t \|\phi_{\theta}(S_t) - \phi_{\pi_{old}}(S_t)\|_2^2 \right]$$

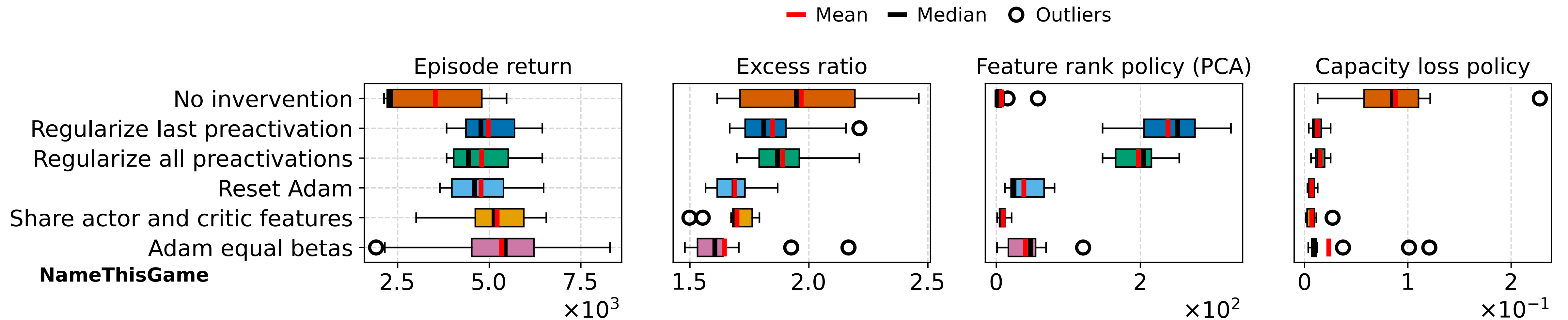


— PFO
— Baseline

$\mathbb{E}_{\pi_{curr}}$: expectation over trajectories from the current policy
 S_t : state at timestep t
 ϕ_{θ} : feature layer

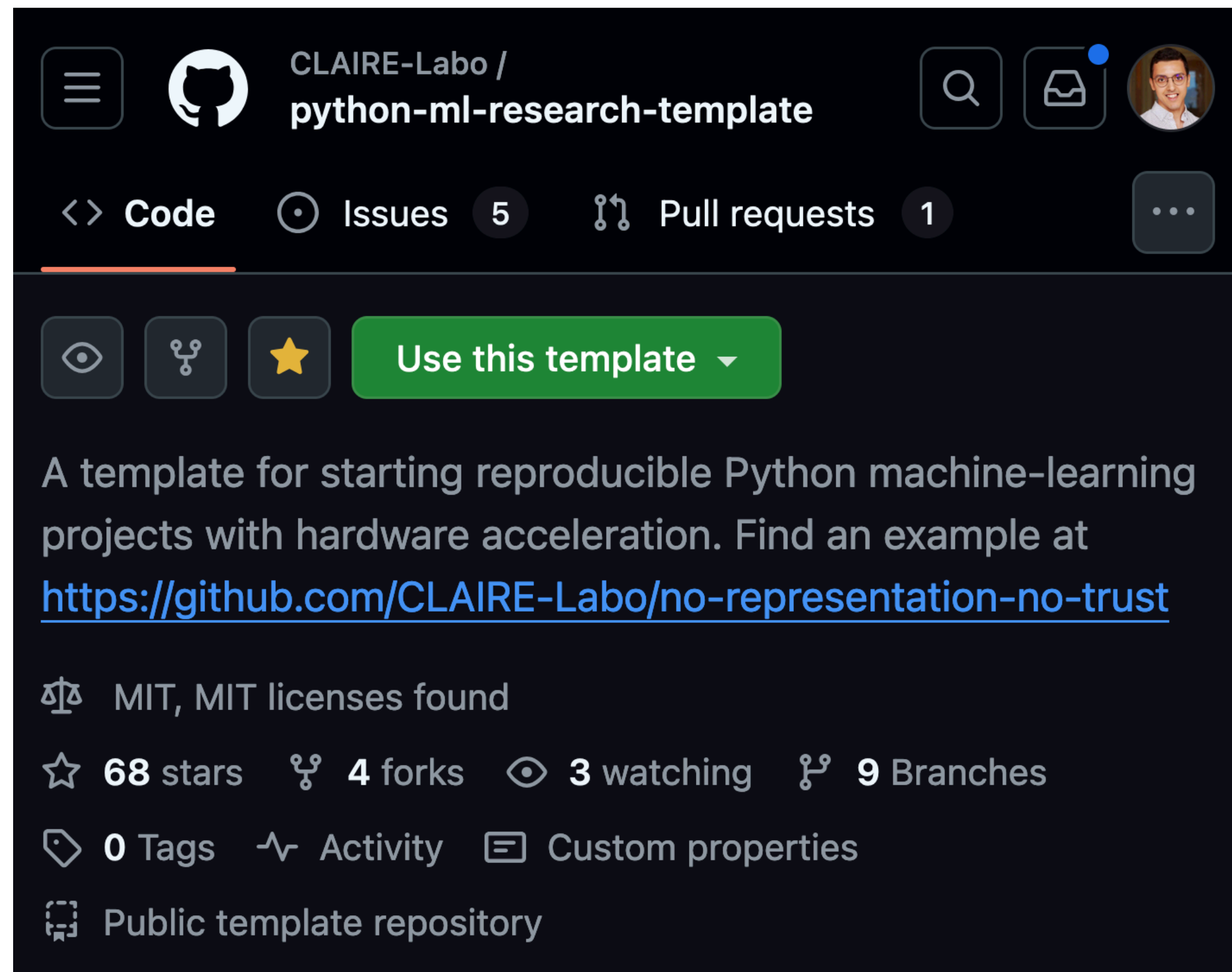
Interventions to corroborate the connection

Better representation, better trust region, mitigated collapse



Tooling by-product

Template for deploying ML projects on all clusters



- Easily switch between IC HaaS/CaaS, RCP CaaS, SCITAS, CSCS clusters
- Reproduce outside EPFL, Facilitate collaboration
- ★ for support! ★



No Representation, No Trust

Connecting Representation, Collapse,
and Trust Issues in PPO



Skander Moalla¹ Andrea Miele¹ Daniil Pyatko¹ Razvan Pascanu² Caglar Gulcehre¹



Code

Fully reproducible
and replicable!

All runs available on
W&B and with raw
logs and
checkpoints!



Template

★ for support! ★

Easily switch between
clusters
Reproduce outside
EPFL, facilitate
collaboration