Intervening in Co-evolution

Stefan Niculae, Alejandro Marin Parra, Daniel Paul Pena, Allen Kim, Abel John
Team *ASADA*, TA: Karl Pertsch

Problem Statement

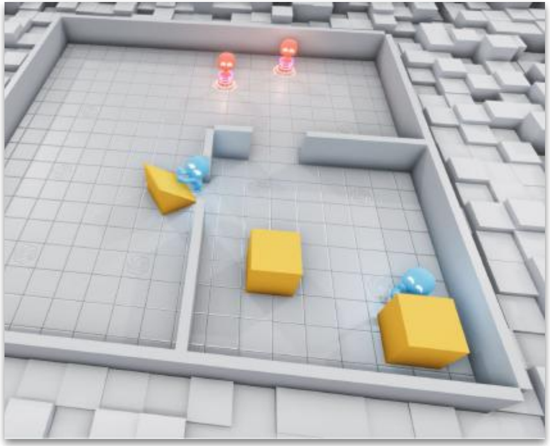
Both sides stagnate when one of them dominates the other

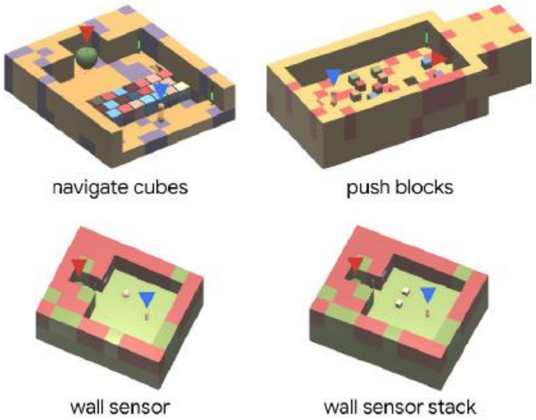
Both sides co-evolve by playing against a strong opponent

- Context: zero-sum game, two sides
- **Hypothesis:** Balanced learning evolves ultimately better teams

- Intervene to balance learning: help loser/hinder winner
- Test in: Multi-Agent Reinforcement Learning (MARL)

Motivation

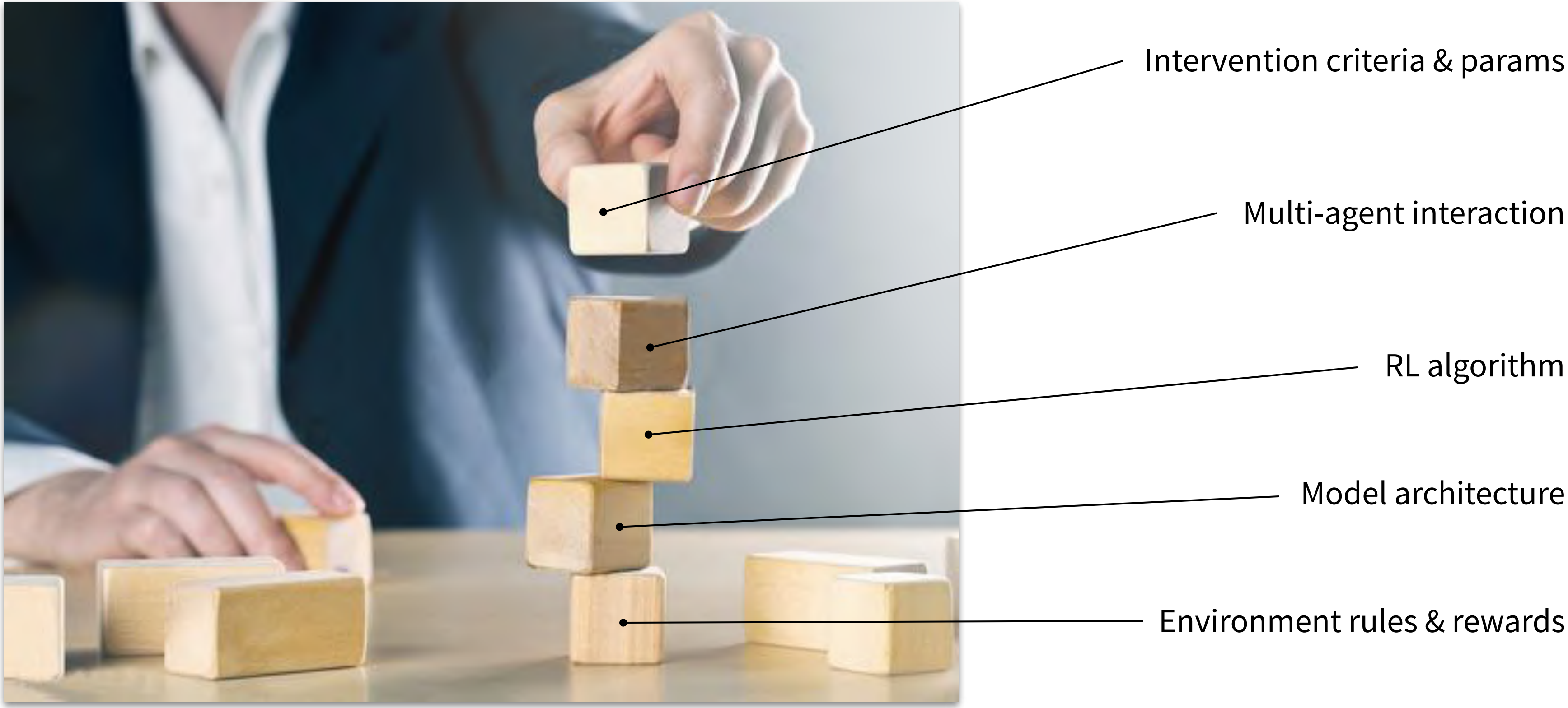




- Abstract concepts can be learned through MARL [1]
- Human guidance can greatly improve RL performance [2]
- MARL is starting to becoming more and more relevant

Setup

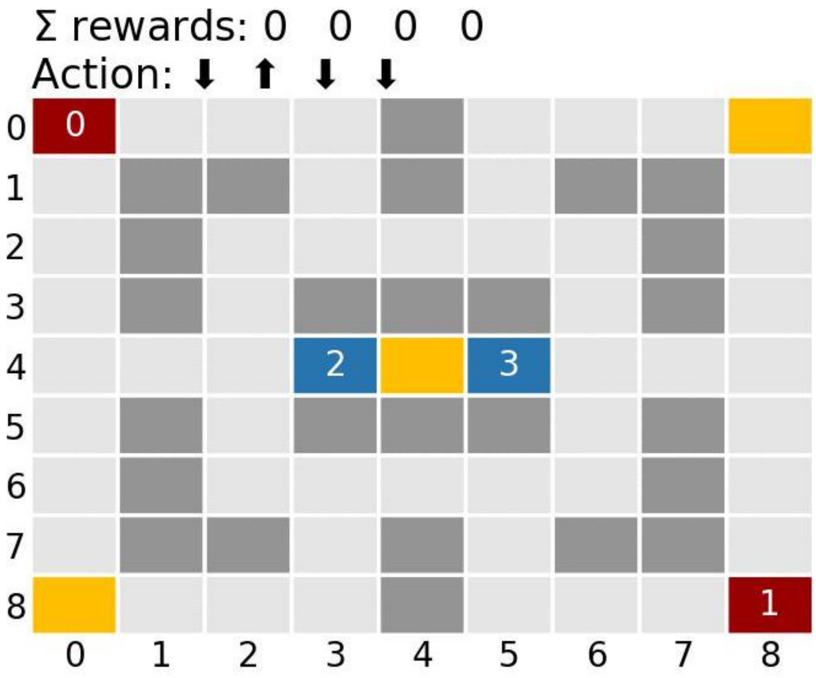
Components


Intervention criteria & params
Multi-agent interaction
RL algorithm
Model architecture
Environment rules & rewards

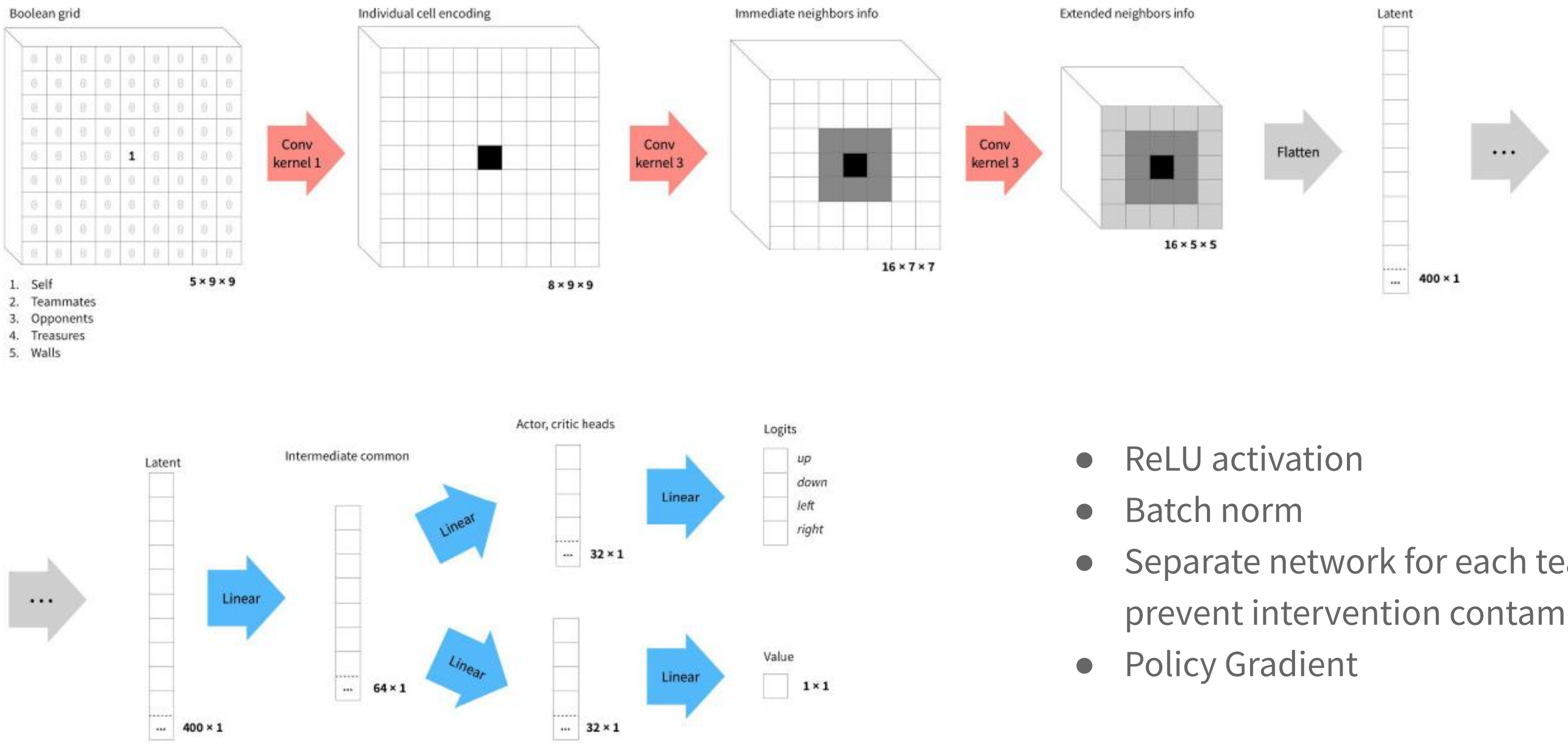
Environment

- Requirements:
  - Computationally cheap
  - Asymmetrical objectives

- Thieves win by collecting two treasures
- Guardians win by catching all thieves
- +1 reward per collect/catch



Model



- ReLU activation
- Batch norm
- Separate network for each team, to prevent intervention contamination
- Policy Gradient

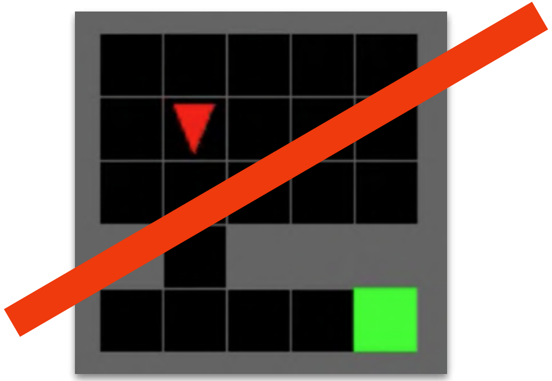
Intervening

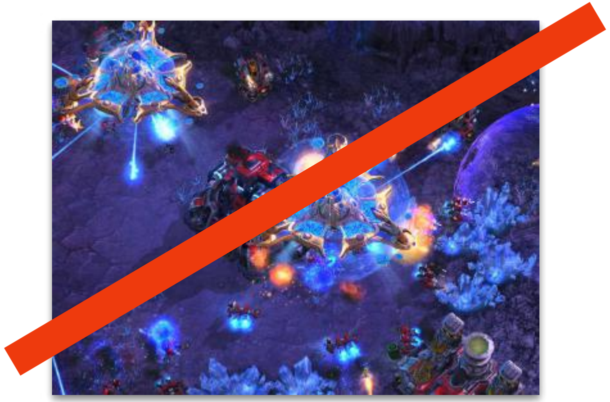
- Freeze learning rate
- Constrain *MI(input, latent)*
- Add noise to policy

Goal: balanced skills

- Guide exploration

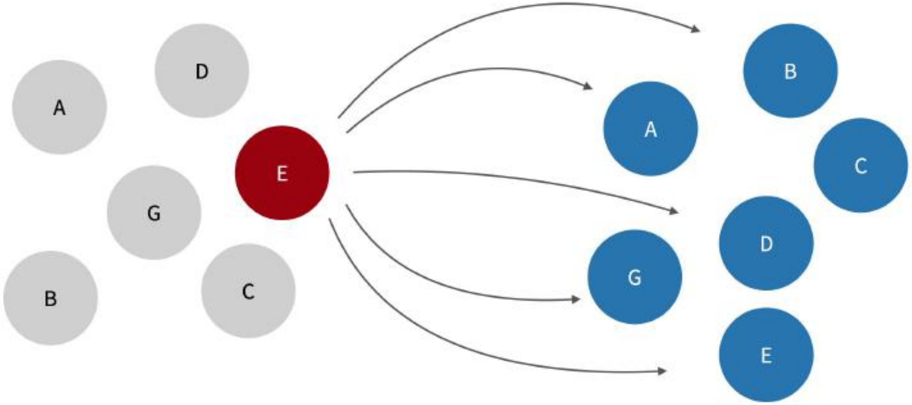
Measuring Performance





- High reward/winrate can come from a weak opponent
- Can't pit Thief A against Thief B directly
- Thief A > Thief B on Guardian X
- Thief B < Thief A on Guardian Y

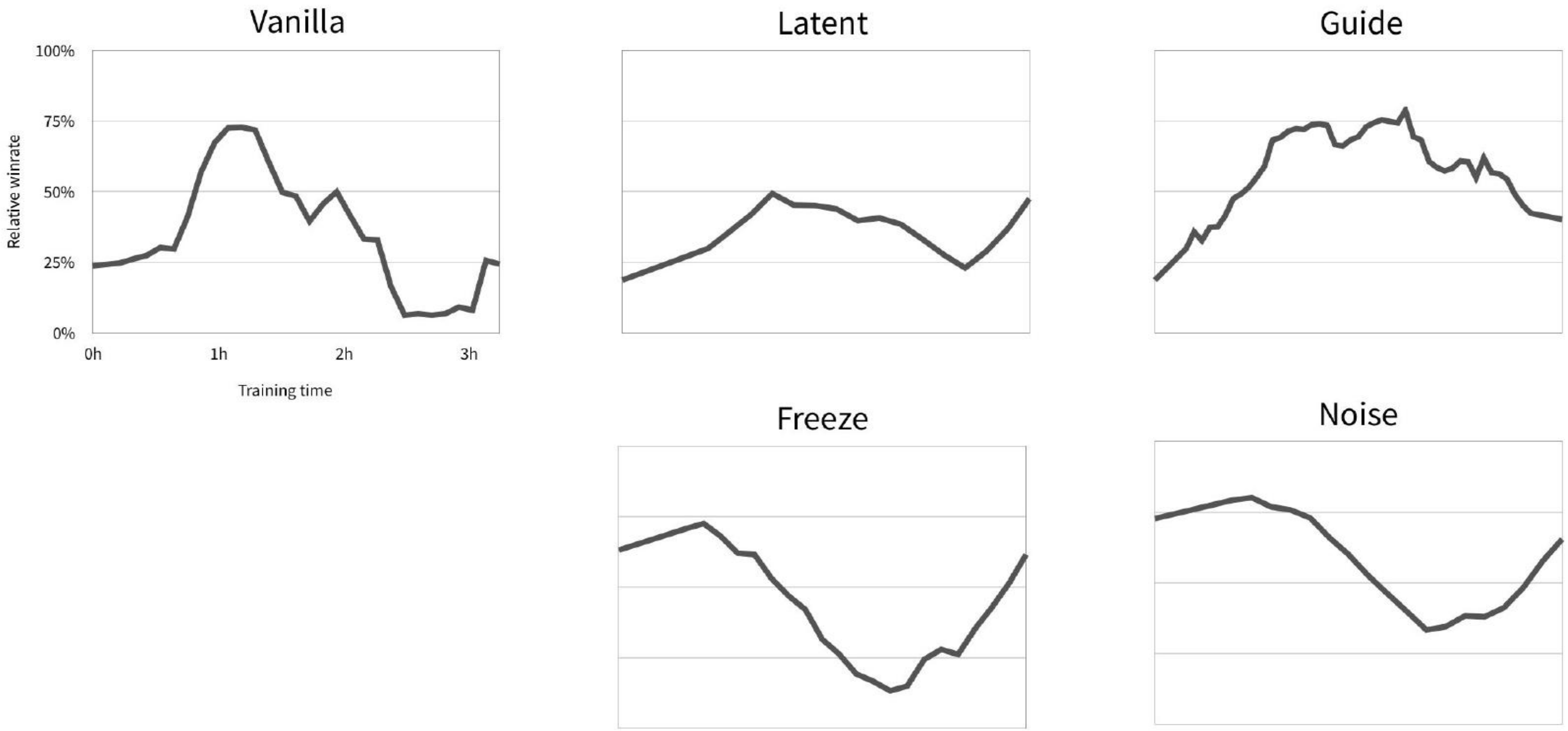
League



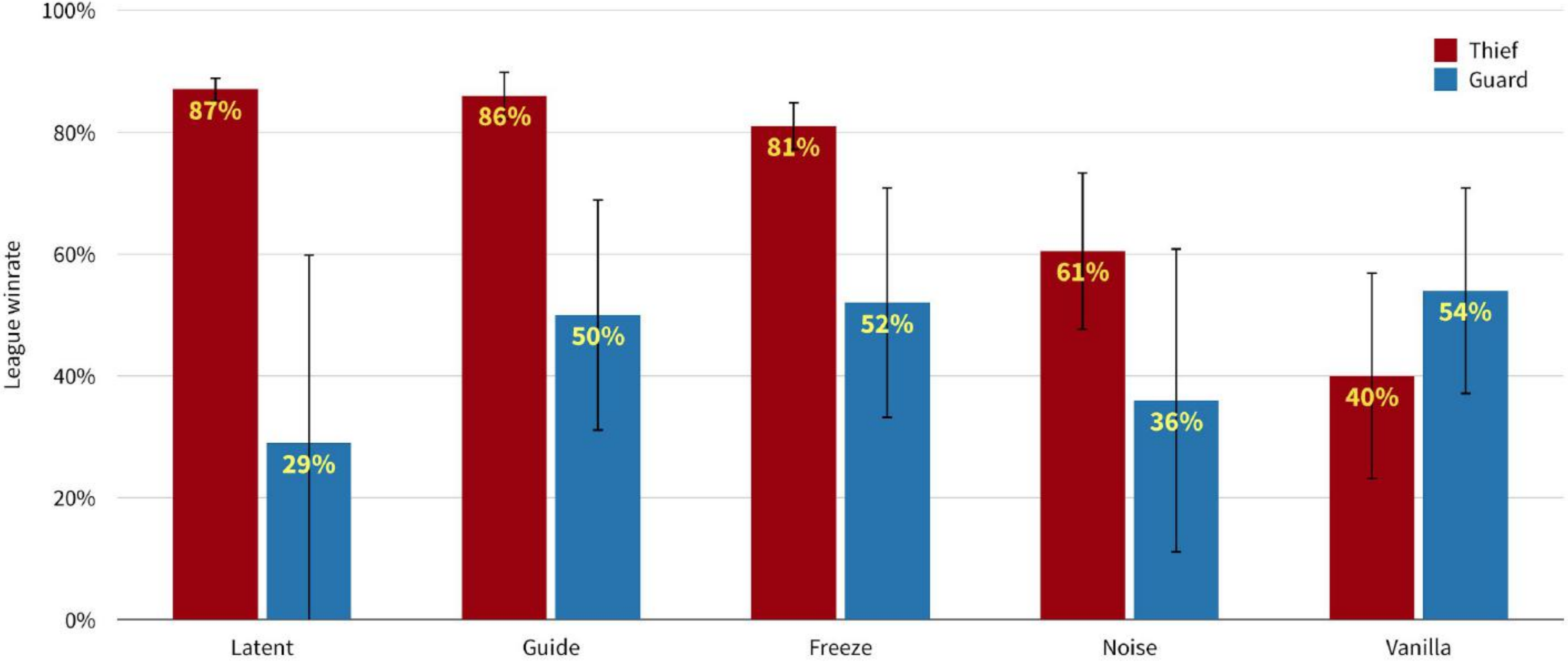
- Performance against many opponents is a proxy for absolute skill
- All Thieves play against all Guardians

Results

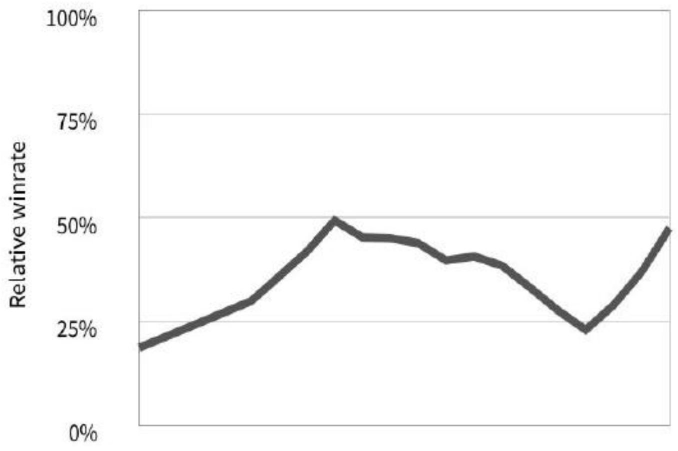
Training Balance

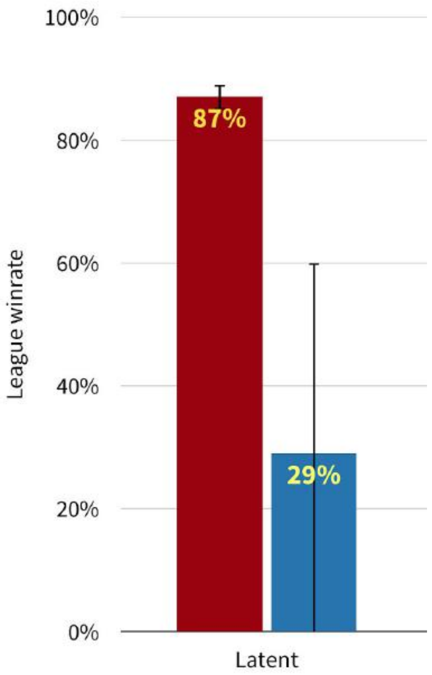


League Performance (best teams)



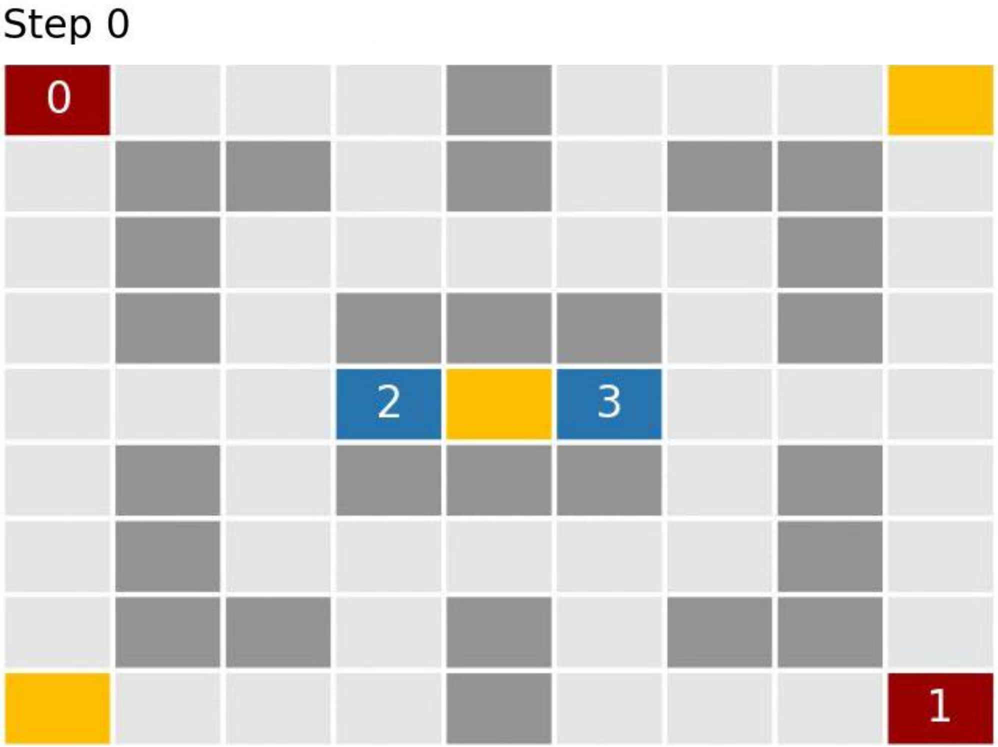
Zoom: Latent





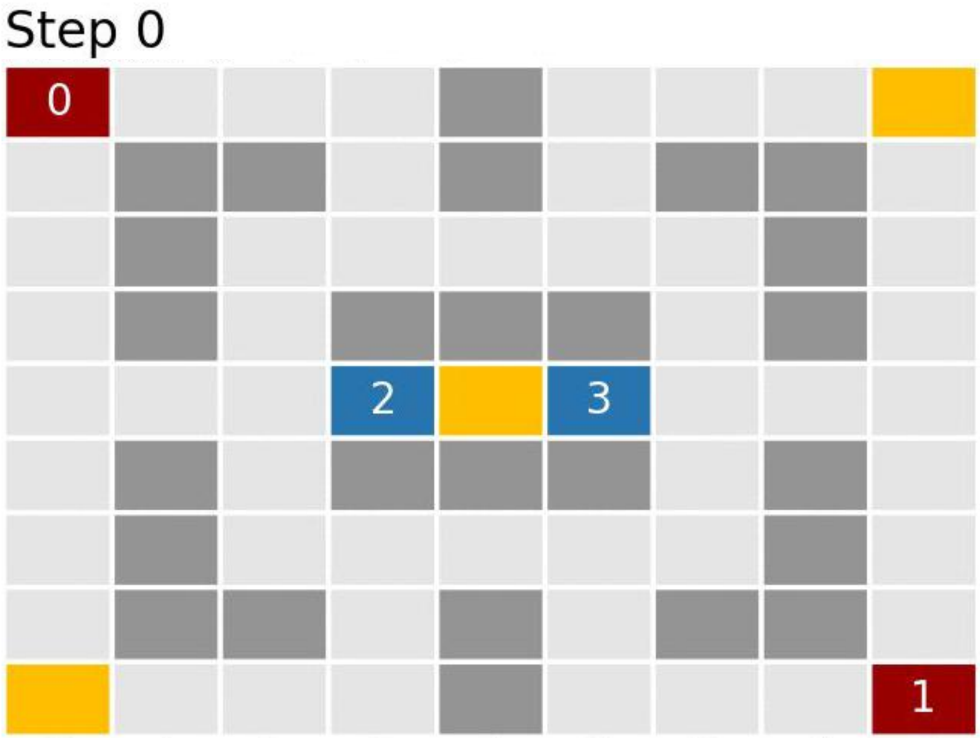
- Thieves are weaker than their training opponent… but win against others
- Guardians beat their training opponent… but don't generalize against others

Vanilla vs Vanilla



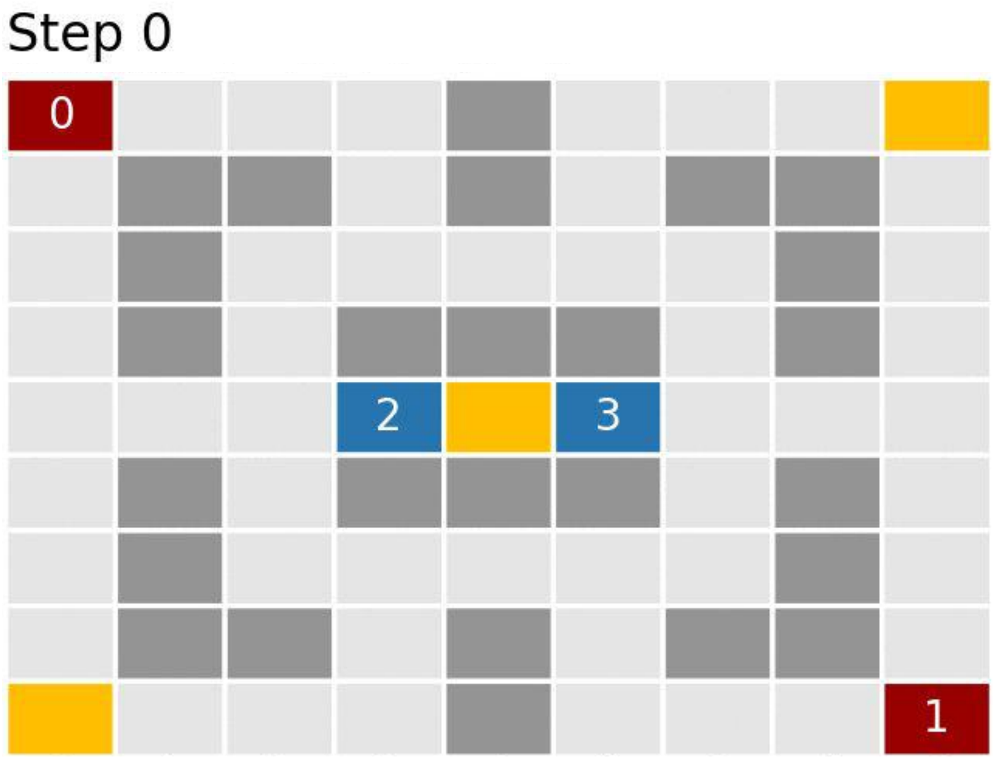
- Step 3: G2 blocks access then lures T0
- Step 36: T1 is cornered

Best Thief v Best Guard



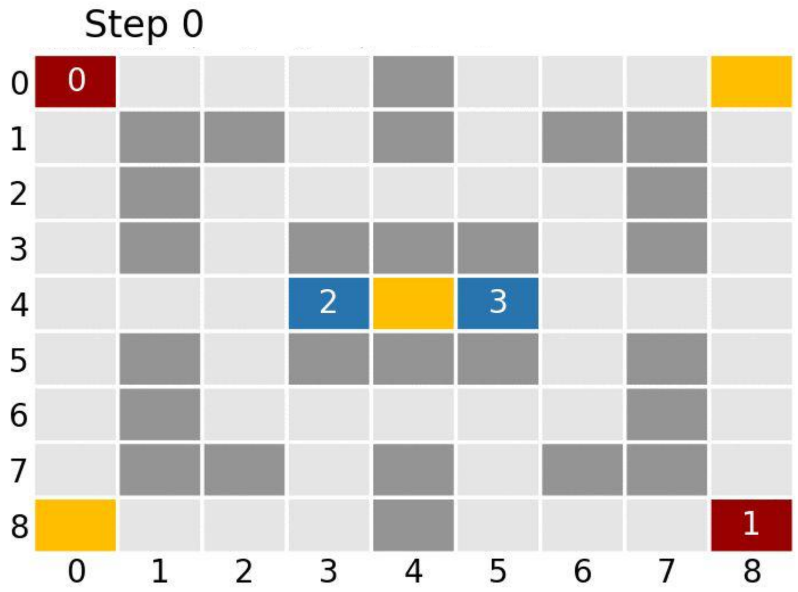
- Step 1: G3 guesses T1 trajectory wrongly
- Step 28: T0 gets the to the treasure before guardians catch it

Specialized Strategy

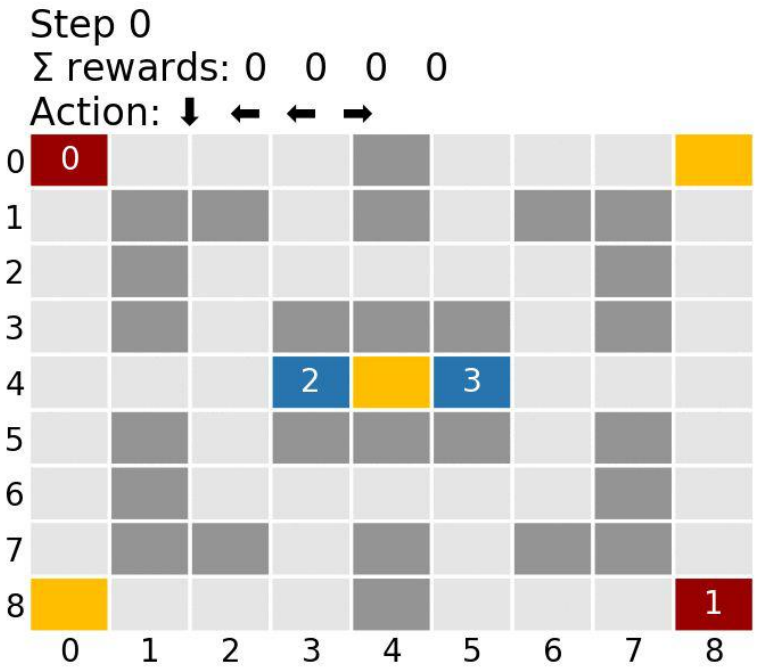


- Guardians learned how to beat their training opponents
- But their strategy falls short on unseen opponents
- Step 9: guardians block paths but thieves never come there

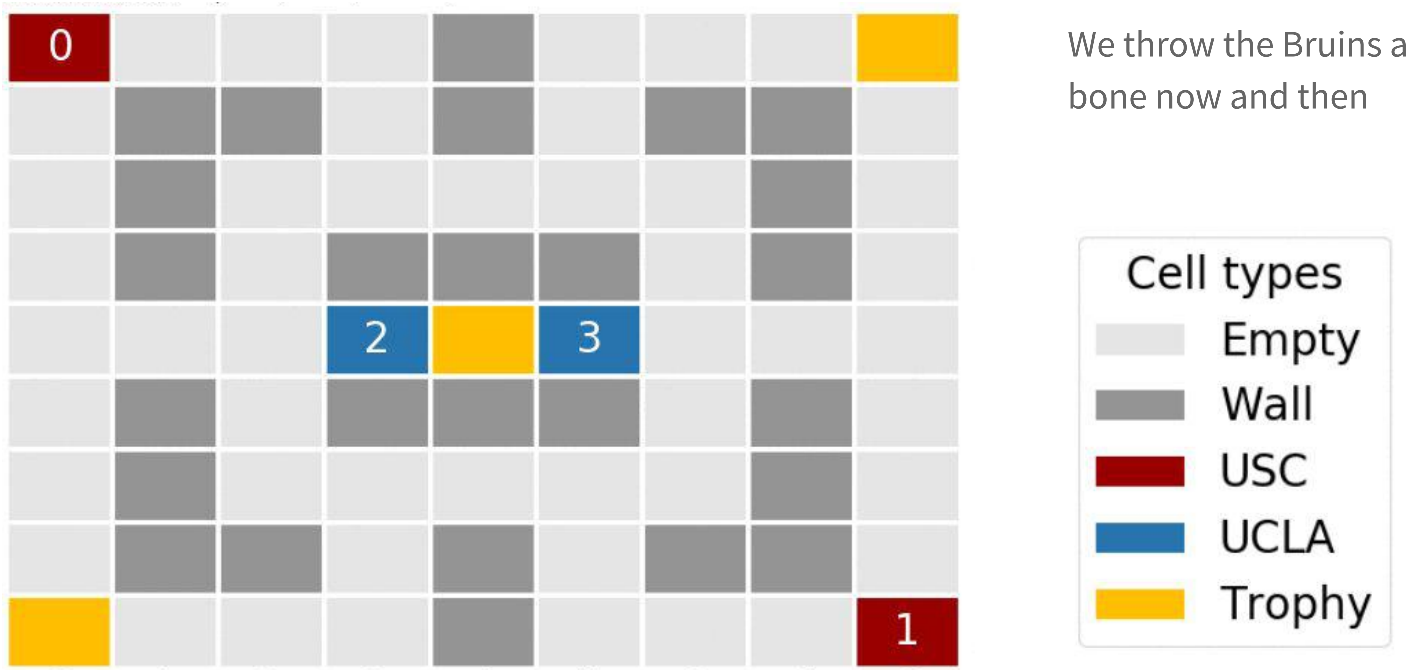
#2 Thief v Best Guard



Best Thief v Best Guard



Trojans are Generous



We throw the Bruins a bone now and then

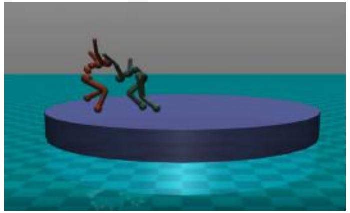
Future work

Intervening

- Combine intervention tactics
- Other intervention criteria
  - Relative gradients magnitude
  - Relative league performance
- Intervention proportional to criteria deviance
- Better assimilation of expert demonstrations [4]
- More sophisticated MI constraint [3]
- Revert winner to previous checkpoint
- Trap: opponents can learn to rely on exploiting this artificial weakness

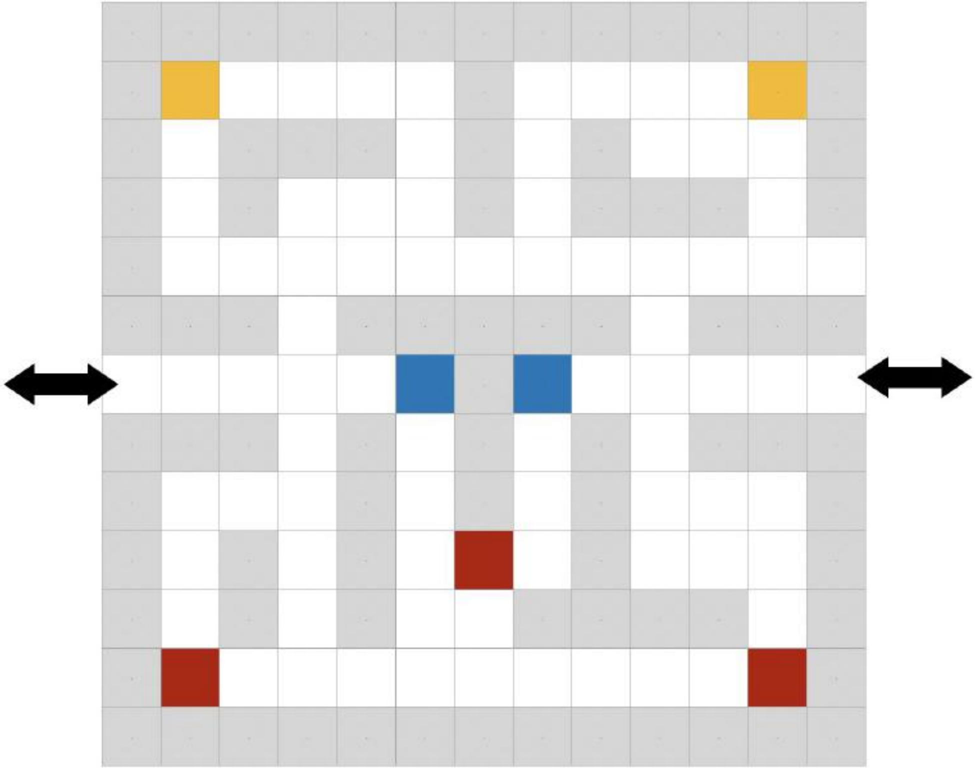
Environment

- Scenarios that enable more complex strategies
  - Larger board size
  - Diagonal movement
- Non-euclidean topologies
  - Screen wrap-around (Pacman-style)
  - Portals
- External environments:



Sumo

Soccer



Intervening impact on learned representations

- Transformer encoder: generalize to variable length inputs
- Scaling: train on small, evaluate on large
- Composability: train on independent tasks, evaluate together
- Curriculum: best sequence of Guardian teams to play against in order to train the ultimate Thief team
- When samples are shown as pre-training
- Variability across random seeds/throughout training

Key takeaways

- Asymmetric MARL evaluation is not straightforward
- Intervening had a positive impact on final performance
  - **BUT**: Very brittle wrt the huge amount of params
- More investigation needed

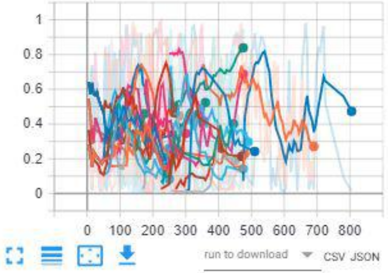
Thank you!

Q&A

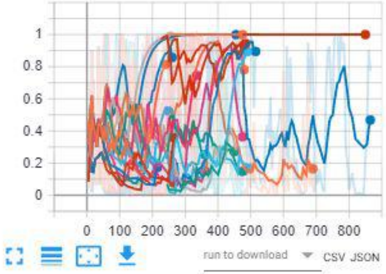
thieves_model	avg_rewards	winrate
#5 - seed=1; threshold=0.7; scripted=0.5 -- checkpoint-300	0.83	0.68
#0 - seed=1 -- checkpoint-400	0.83	0
#1 - seed=1; threshold=0.6; scripted=0.5 -- checkpoint-300	0.76	0.66
#1 - seed=1; threshold=0.6; scripted=0.5 -- checkpoint-400	0.73	0.6
#3 - seed=1; threshold=0.66; uniform=0.33 -- checkpoint-300	0.71	0.62
#7 - seed=1; threshold=0.8; lr=0 -- checkpoint-300	0.63	0.48
#2 - seed=1; threshold=0.6; lr=0 -- checkpoint-300	0.54	0.24
#4 - seed=1; threshold=0.66; mi=0.2 -- checkpoint-300	0.51	0.26
#3 - seed=1; threshold=0.66; uniform=0.33 -- checkpoint-400	0.47	0.11
#10 - seed=2; threshold=0.6; lr=0 -- checkpoint-300	0.46	0.24
#14 - seed=2; threshold=0.7; lr=0 -- checkpoint-400	0.45	0.023
#9 - seed=2; threshold=0.6; scripted=0.5 -- checkpoint-400	0.45	0.23
#9 - seed=2; threshold=0.6; scripted=0.5 -- checkpoint-300	0.45	0.21
#10 - seed=2; threshold=0.6; lr=0 -- checkpoint-400	0.44	0.21
#2 - seed=1; threshold=0.6; lr=0 -- checkpoint-400	0.43	0.046
#13 - seed=2; threshold=0.7; scripted=0.5 -- checkpoint-400	0.4	0.24
#11 - seed=2; threshold=0.66; uniform=0.33 -- checkpoint-300	0.38	0.21
#11 - seed=2; threshold=0.66; uniform=0.33 -- checkpoint-400	0.37	0.2
#8 - seed=2 -- checkpoint-300	0.37	0.14
#0 - seed=1 -- checkpoint-300	0.3	0.046
#7 - seed=1; threshold=0.8; lr=0 -- checkpoint-400	0.29	0.17
#8 - seed=2 -- checkpoint-400	0.29	0.16
#6 - seed=1; threshold=0.7; lr=0 -- checkpoint-300	0.25	0.14
#6 - seed=1; threshold=0.7; lr=0 -- checkpoint-400	0.25	0.13
#14 - seed=2; threshold=0.7; lr=0 -- checkpoint-300	0.12	0.046
scripted	0.086	0.034
#15 - seed=2; threshold=0.8; lr=0 -- checkpoint-300	0.052	0.034
#15 - seed=2; threshold=0.8; lr=0 -- checkpoint-400	0.046	0.023
#13 - seed=2; threshold=0.7; scripted=0.5 -- checkpoint-300	0.034	0.034

end-reason-per

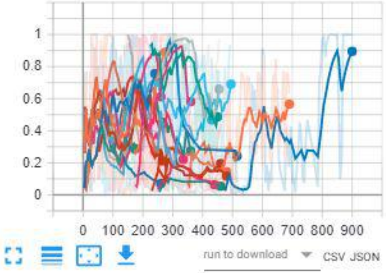
All_thieves_caught
tag: end-reason-per/All_thieves_caught



Out_of_time
tag: end-reason-per/Out_of_time

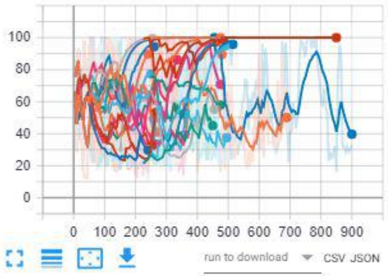


Treasure_s__collected
tag: end-reason-per/Treasure_s__collected

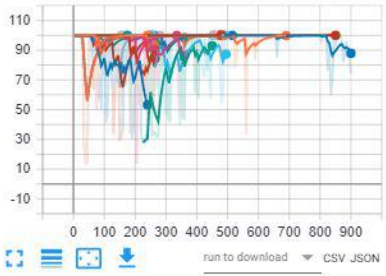


episode-steps-alive

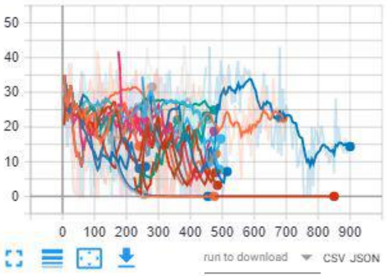
Guardian-2/avg
tag: episode-steps-alive/Guardian-2/avg



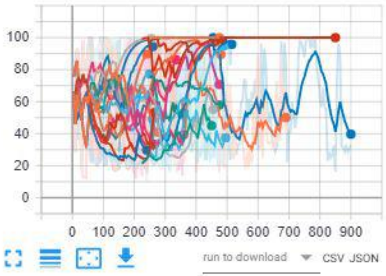
Guardian-2/max
tag: episode-steps-alive/Guardian-2/max



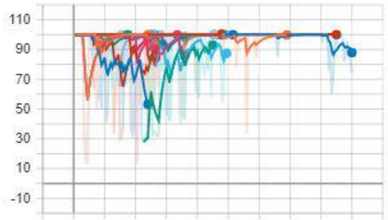
Guardian-2/std
tag: episode-steps-alive/Guardian-2/std



Guardian-3/avg
tag: episode-steps-alive/Guardian-3/avg

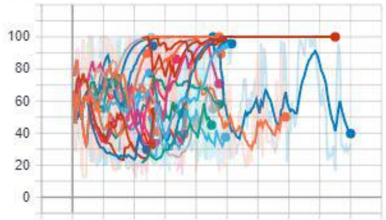


Guardian-3/max
tag: episode-steps-alive/Guardian-3/max

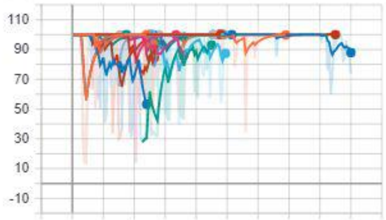


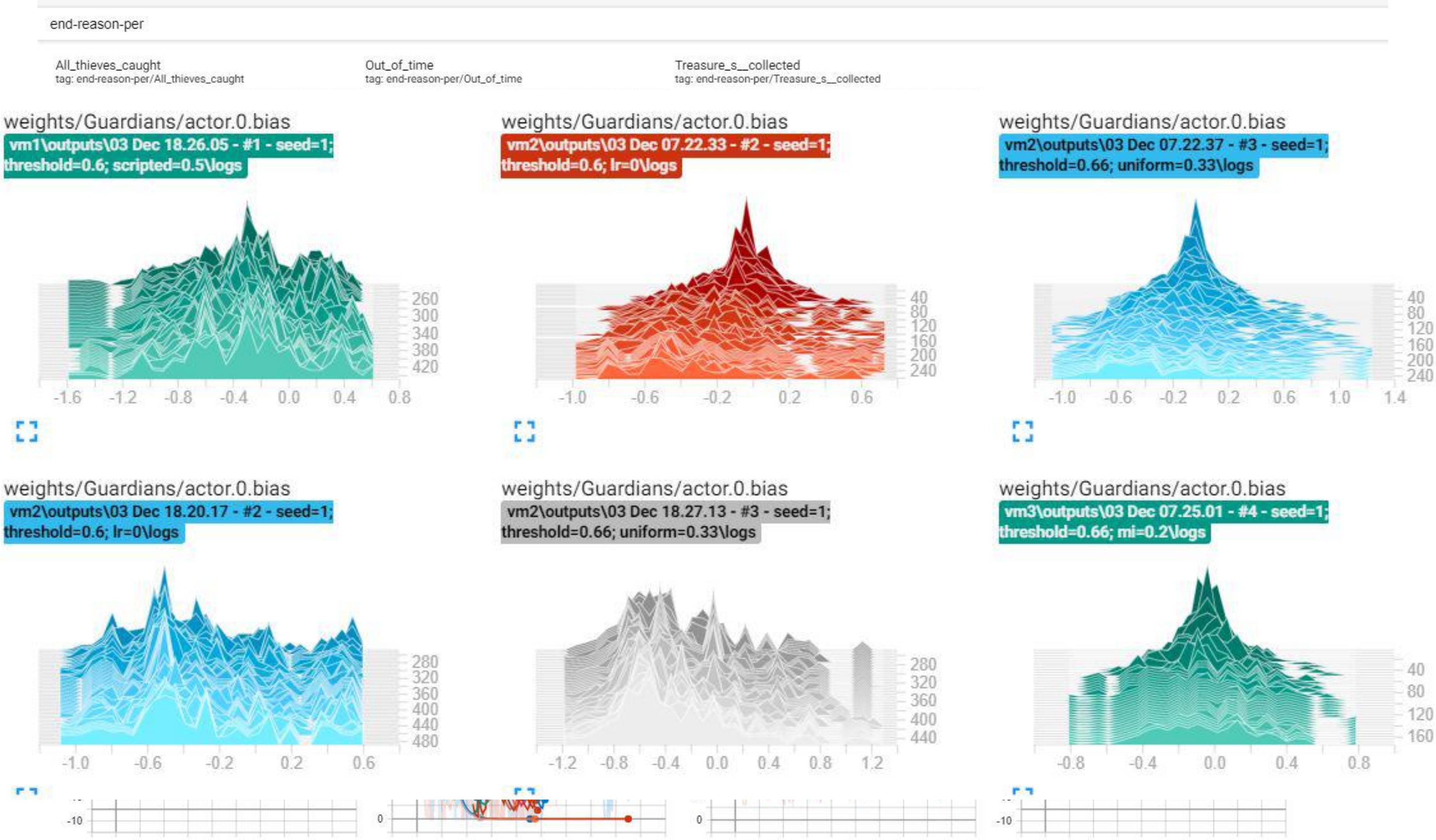
Guardian-3/std
tag: episode-steps-alive/Guardian-3/std

Guardians-average/avg
tag: episode-steps-alive/Guardians-average/avg



Guardians-average/max
tag: episode-steps-alive/Guardians-average/max




end-reason-per
All_thieves_caught
tag: end-reason-per/All_thieves_caught
Out_of_time
tag: end-reason-per/Out_of_time
Treasure_s__collected
tag: end-reason-per/Treasure_s__collected
weights/Guardians/actor.0.bias
vm1\outputs\03 Dec 18.26.05 - #1 - seed=1; threshold=0.6; scripted=0.5\logs
weights/Guardians/actor.0.bias
vm2\outputs\03 Dec 07.22.33 - #2 - seed=1; threshold=0.6; lr=0\logs
weights/Guardians/actor.0.bias
vm2\outputs\03 Dec 07.22.37 - #3 - seed=1; threshold=0.66; uniform=0.33\logs
weights/Guardians/actor.0.bias
vm2\outputs\03 Dec 18.20.17 - #2 - seed=1; threshold=0.6; lr=0\logs
weights/Guardians/actor.0.bias
vm2\outputs\03 Dec 18.27.13 - #3 - seed=1; threshold=0.66; uniform=0.33\logs
weights/Guardians/actor.0.bias
vm3\outputs\03 Dec 07.25.01 - #4 - seed=1; threshold=0.66; mi=0.2\logs

References

1. Baker et al — Emergent Tool Use From Multi-agent Autocurricula (2019)
2. Paine et al — Making Efficient Use of Demonstrations to Solve Hard Exploration Problems (2019)
3. Hjelm et al — Learning Deep Representations By Mutual Information Estimation And Maximization (2018)
4. Ho & Ermon — Generative Adversarial Imitation Learning (2016)
5. Sukhbaatar et al — Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play (2018)
6. Peng et al — Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse Rl, And Gans By Constraining Information Flow (2018)
7. Goyal et al — InfoBot: Transfer And Exploration Via The Information Bottleneck (2019)
8. Kim et al — EMI: Exploration with Mutual Information (2019)
9. Grau-Moya et al — Soft Q-learning With Mutual-information Regularization (2019)
10. Haarnoja et al — Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor (2018)
11. Haarnoja et al — Soft Actor-Critic Algorithms and Applications (2019)
12. Christodoulou — Soft Actor-Critic For Discrete Action Settings (2019)
13. Jaques et al — Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning (2019)