White-Box Adversarial Policies in Deep Reinforcement Learning

Stephen Casper, Taylor Killian, Dylan Hadfield-Menell, Gabriel Kreiman

Algorithmic Alignment

CENTER FOR Brains Minds+ Machines

Boston Children's Hospital

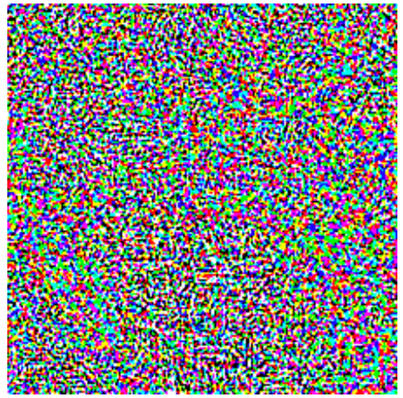
Stephen Casper, scasper@mit.edu
SafeAI, February 2023

Adversaries in Supervised Learning

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

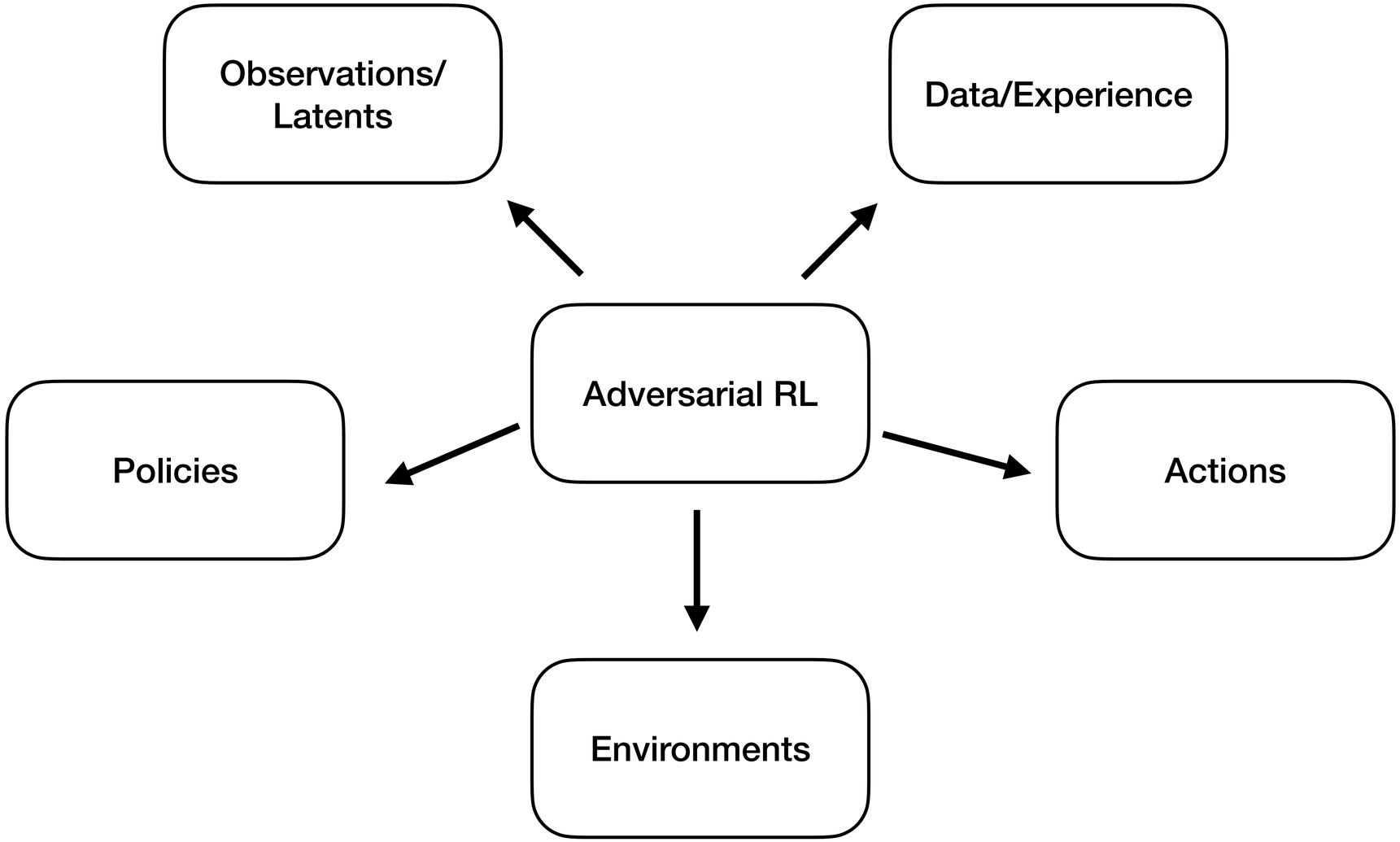
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

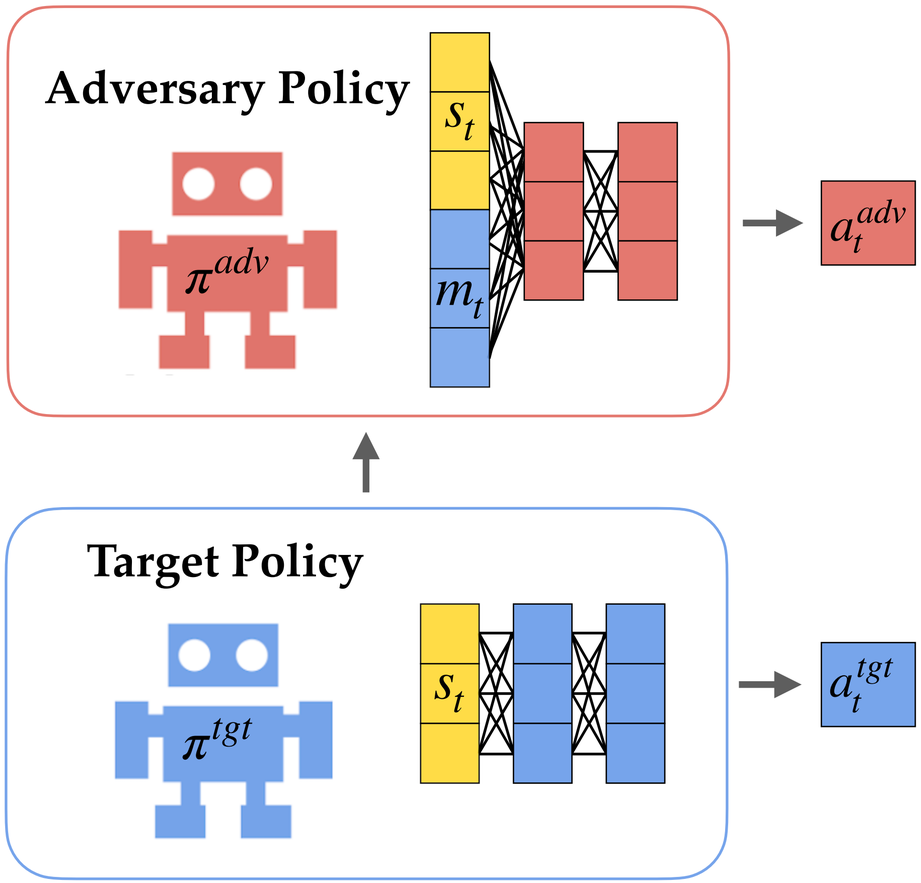
"gibbon"
99.3 % confidence

Goodfellow et al. (2014)

Adversaries are interesting in RL



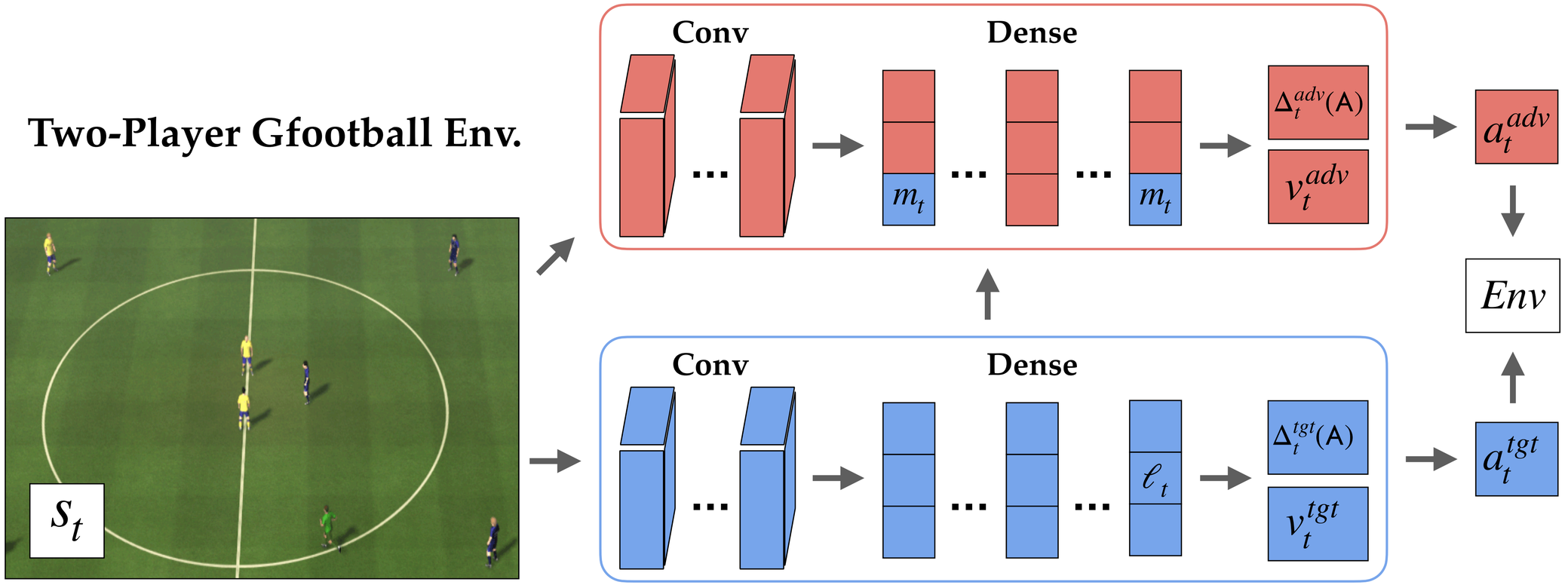
White box adversaries: “reading the target agent’s mind”



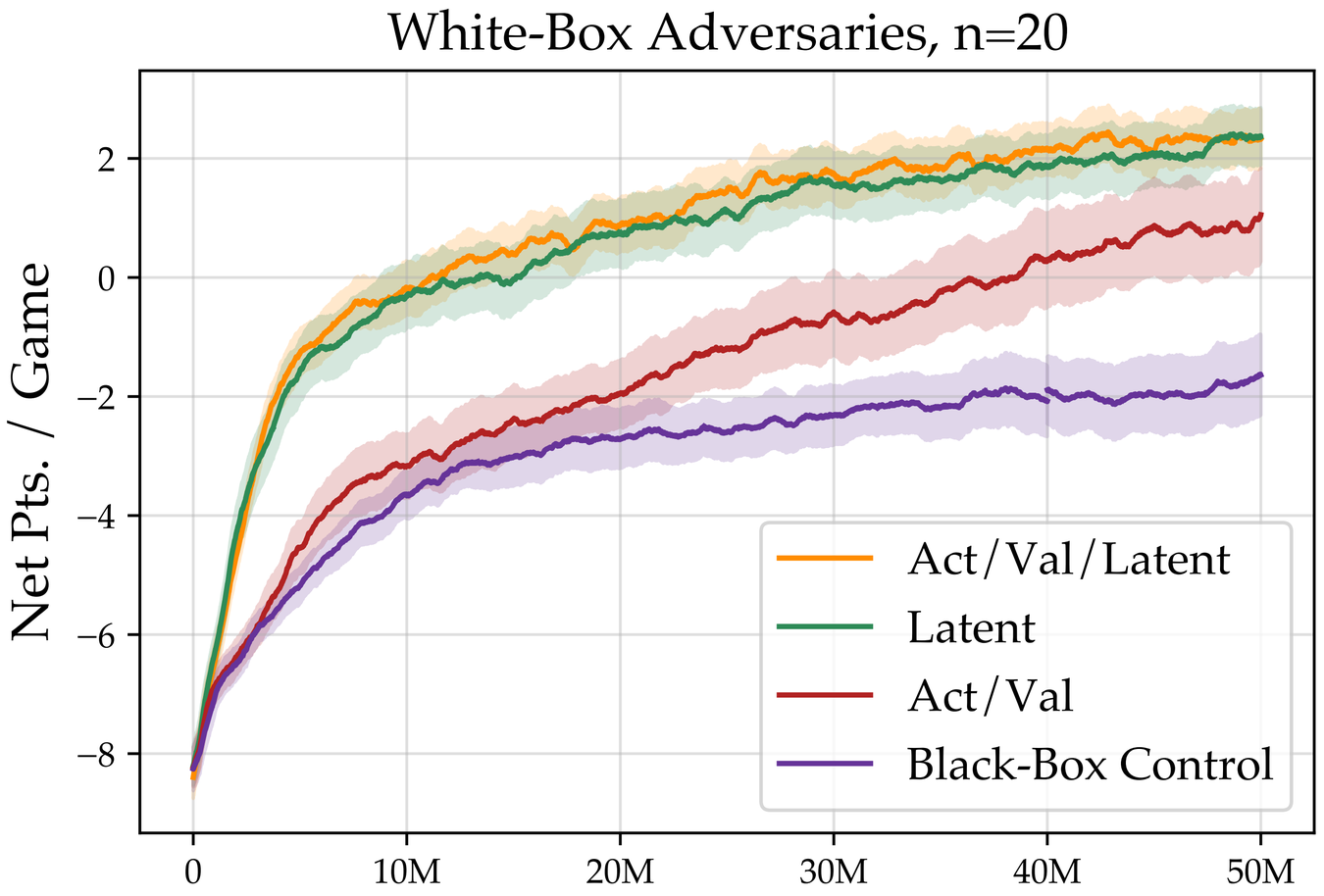
Why?

- Identifying weaknesses with attacks
- Improving robustness with robust adversarial reinforcement learning

Attacks

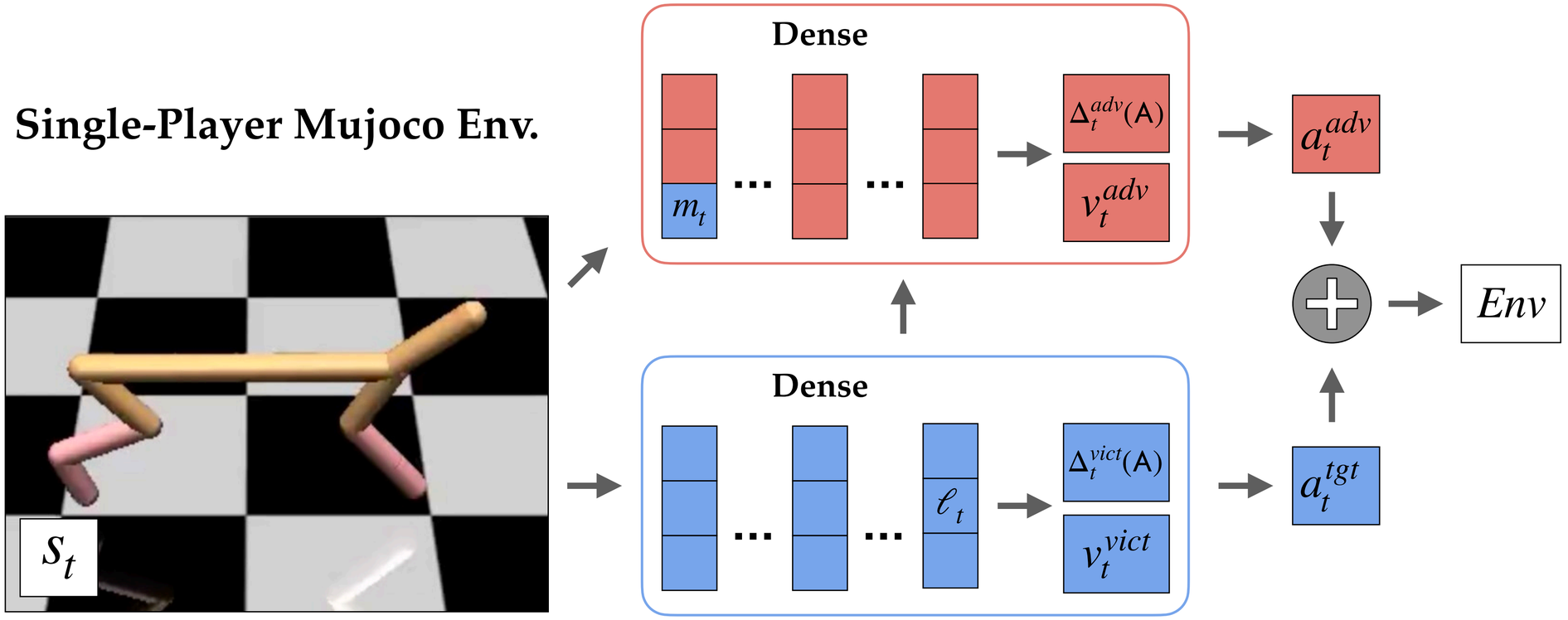

Two-Player Gfootball Env.
s_t
Conv
Dense
m_t
m_t
$\Delta_t^{adv}(A)$
v_t^{adv}
a_t^{adv}
Env
Conv
Dense
ℓ_t
$\Delta_t^{tgt}(A)$
v_t^{tgt}
a_t^{tgt}

Attacks: better initial and asymptotic performance

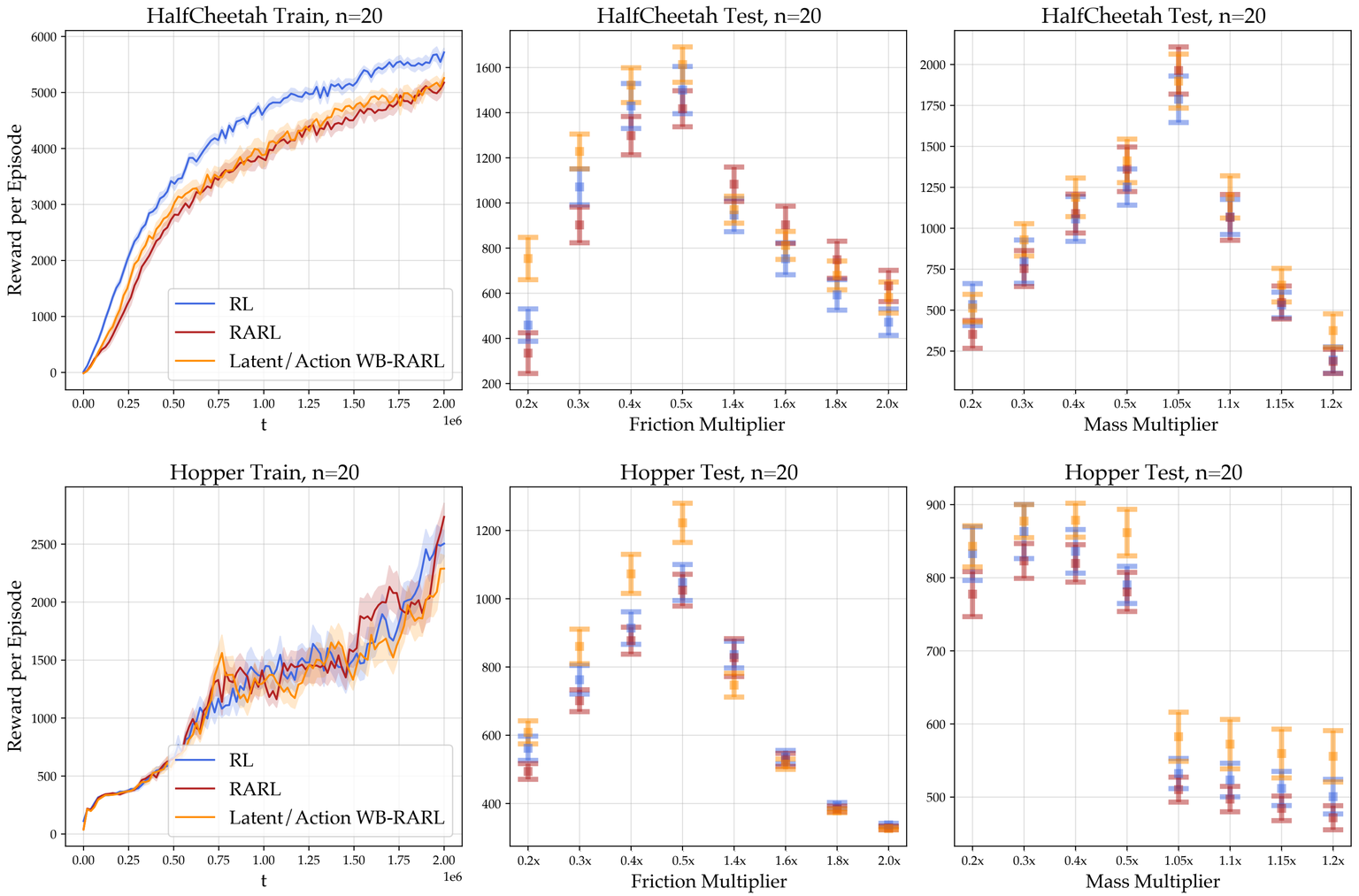


Act/Val/Latent v. Ctrl p: 2e-05
Latent v. Ctrl p: 1e-05
Act/Val v. Ctrl p: 0.00638

Robust Adversarial Reinforcement Learning



Robust Adversarial Reinforcement Learning



So what?

1. White box adversarial attacks are more effective.
2. Leveraging model internals for diagnostic/debugging tasks is useful.

Thanks!

Taylor Killian

Dylan Hadfield-Menell

Gabriel Kreiman

Correspondence:
scasper@mit.edu