DEEP REINFORCEMENT LEARNING THAT MATTERS

Parniyan Malekzadeh

13 May, 2024

CONTENT

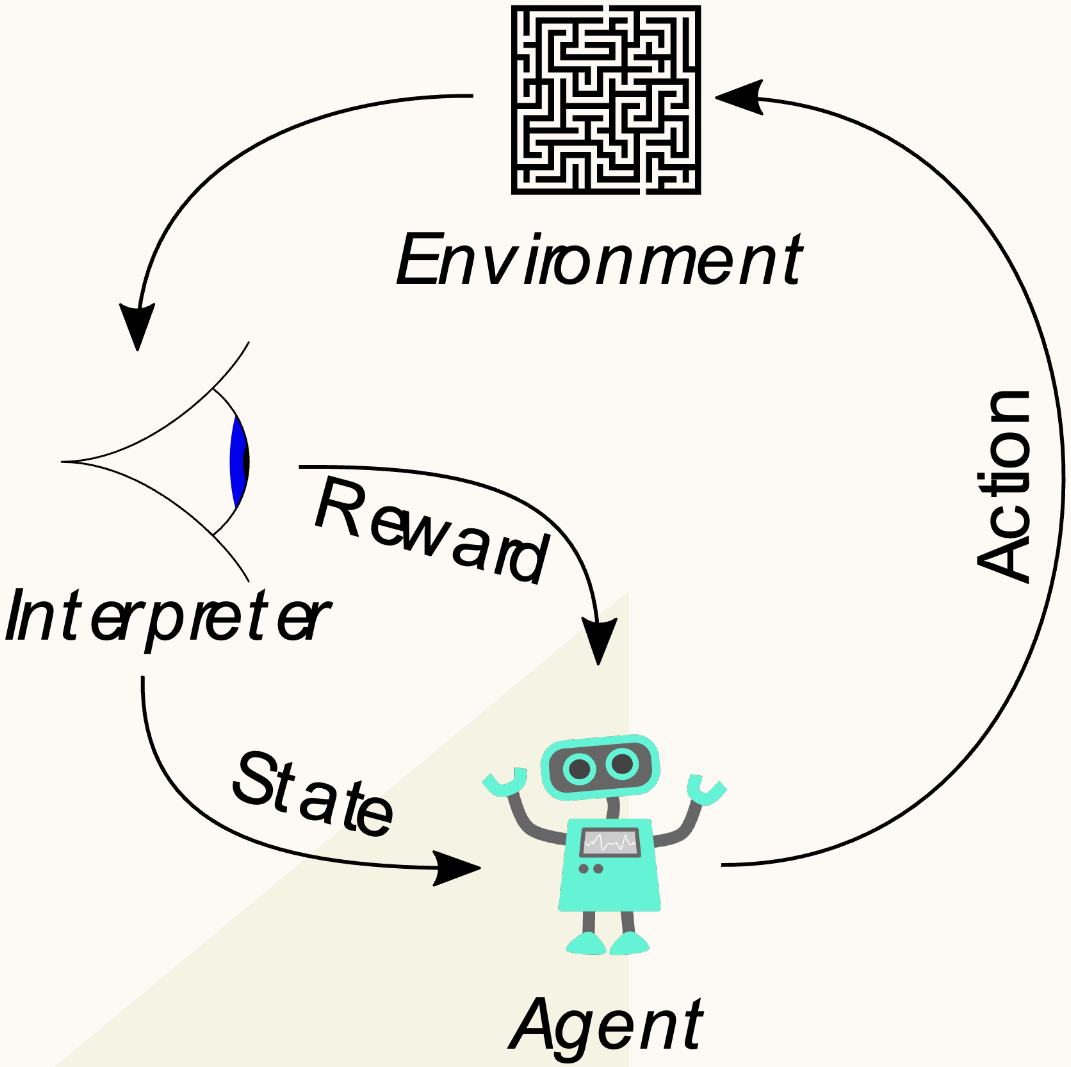
- Introduction
- Environments used for experiments and analyses
- Algorithms used for experiments and analyses
- Factors affecting reproducibility
- Conclusion

CONTENT

- Introduction
- Environments used for experiments and analyses
- Algorithms used for experiments and analyses
- Factors affecting reproducibility
- Conclusion

REINFORCEMENT LEARNING

- Agent
- Environment
- Policy (π)
- Value Function (V(s))
- Reward Function (R(s,a))

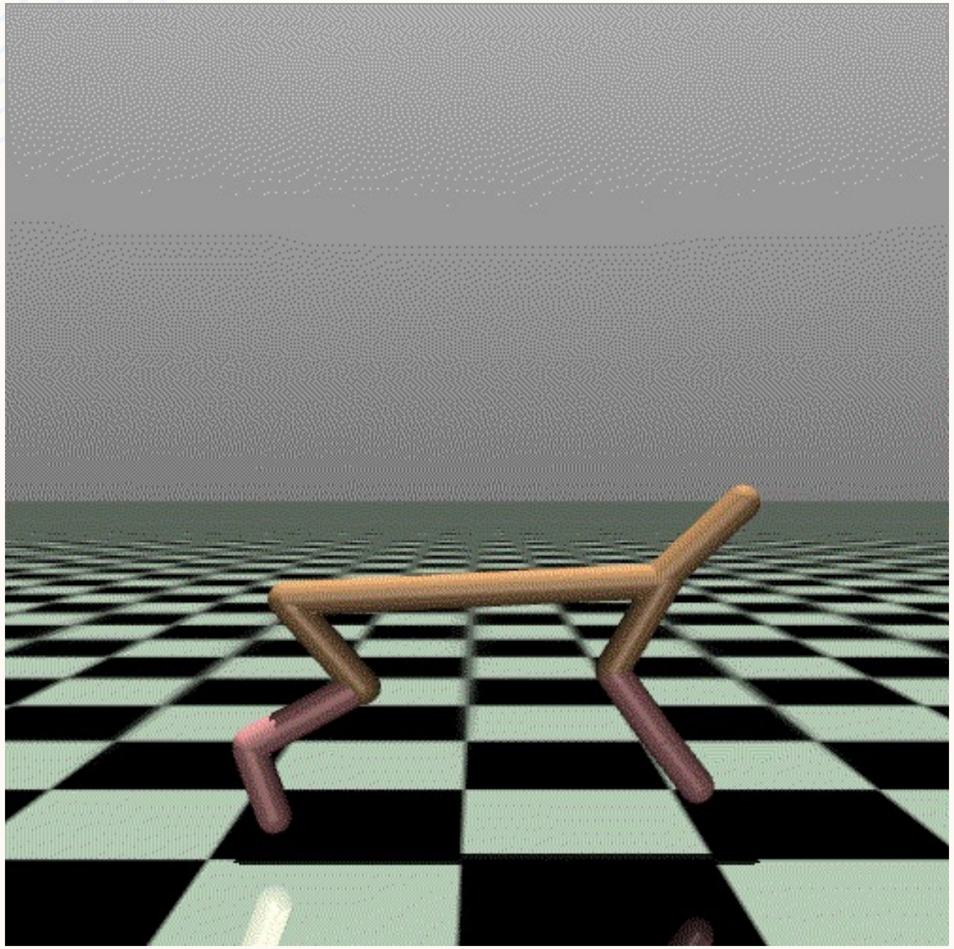


IMPORTANCE OF REPRODUCIBILITY IN RL

- Sustaining Progress: Reproducing existing work and accurately judging the improvements offered by novel methods is vital to sustaining progress in RL research.

CONTENT

- Introduction
- Environments used for experiments and analyses
- Algorithms used for experiments and analyses
- Factors affecting reproducibility
- Conclusion

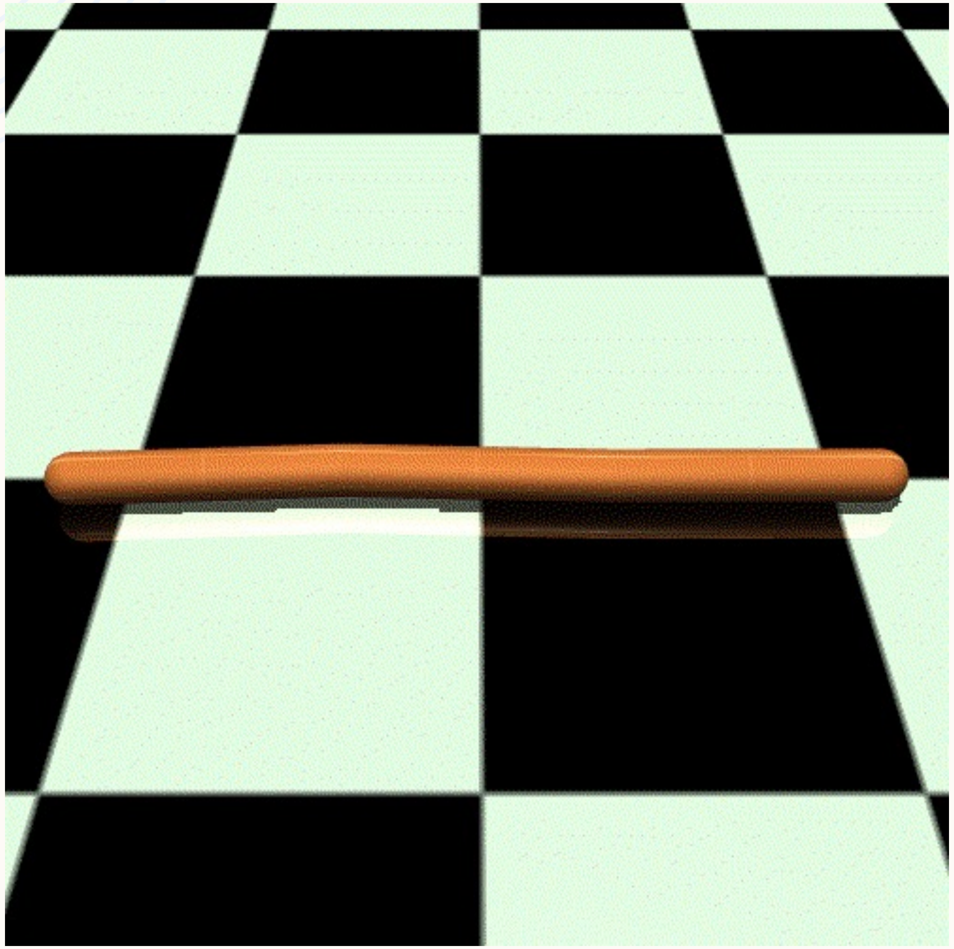


HalfCheetah-v1

Goal: make the cheetah run forward (right) as fast as possible

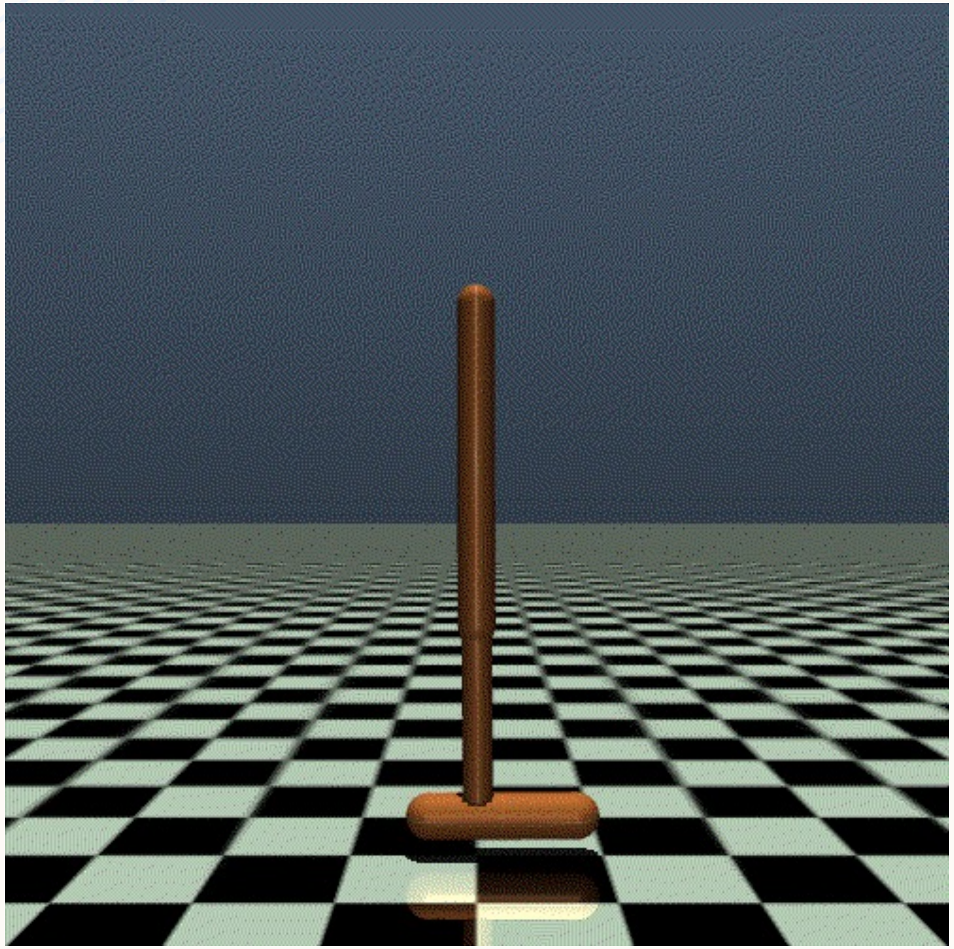
Positive reward: distance moved forward

Negative reward: moving backward



Swimmer

Goal: move as fast as possible toward the right by using the fluids' friction



Hopper

Goal: make hops that move in the forward (right)

CONTENT

- Introduction
- Environments used for experiments and analyses
- Algorithms used for experiments and analyses
- Factors affecting reproducibility
- Conclusion

ALGORITHMS:

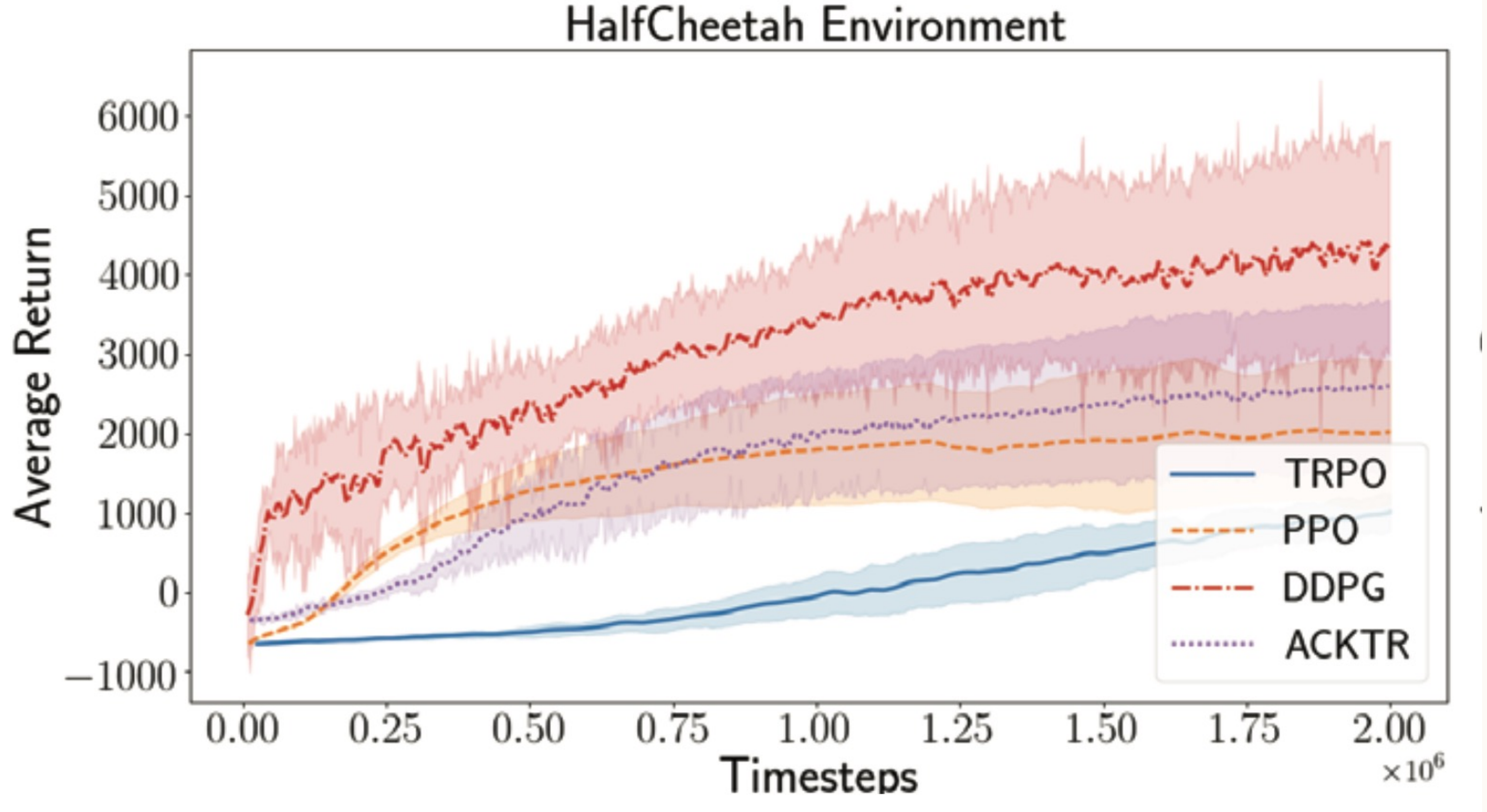
- TRPO (Trust Region Policy Optimization)
- PPO (Proximal Policy Optimization)
- DDPG (Deep Deterministic Policy Gradient)
- ACKTR (Actor-Critic using Kronecker-Factored Trust Region)

CONTENT

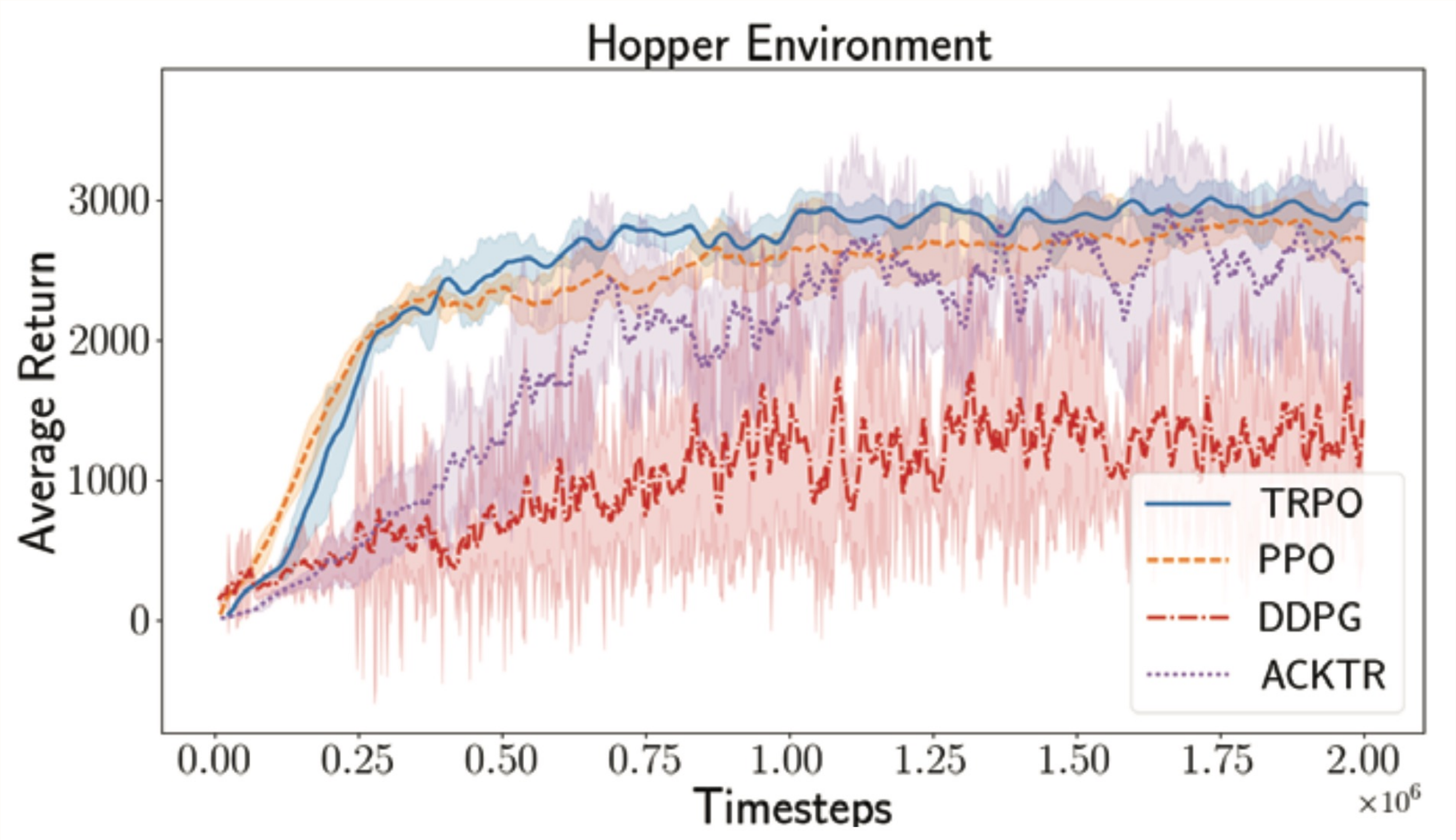
- Introduction
- Environments used for experiments and analyses
- Algorithms used for experiments and analyses
- Factors affecting reproducibility
- Conclusion

ENVIRONMENT PROPERTIES

- random stochasticity
- shortened trajectories
- different dynamic properties



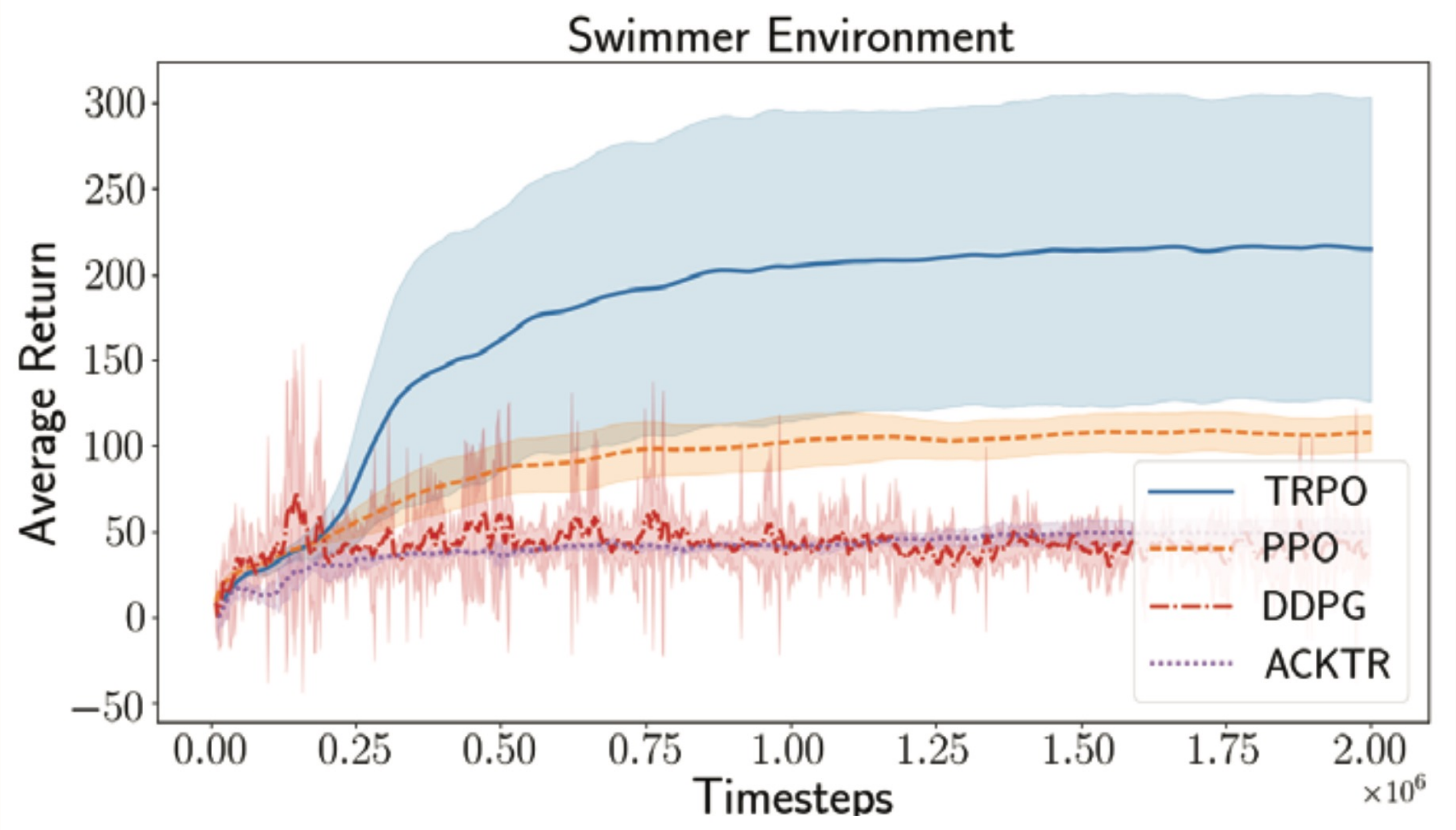
in environments with stable dynamics (e.g. HalfCheetah-v1), DDPG outperforms all other algorithms.



as dynamics become more unstable (e.g. in Hopper-v1)
performance gains rapidly diminish.

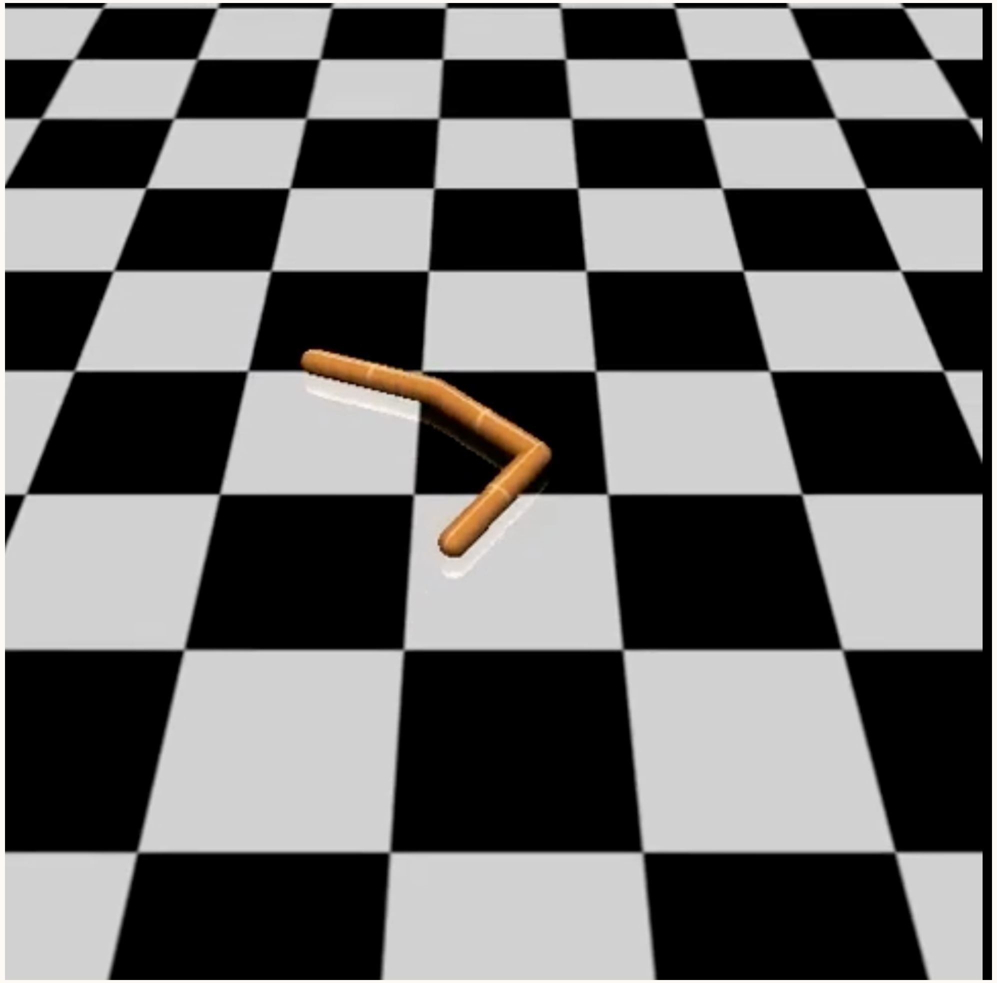
WHY DOES THIS HAPPEN?

- Noise in unstable environments complicates the estimation of an appropriate value function.
- Shortened trajectories characterize failures in such tasks.
- DDPG opts for a survival strategy, seeking a local optimum by enduring until the trajectory's maximum length.
- Reporting only favorable results in stable environments like HalfCheetah would unfairly represent DDPG-based experiments.



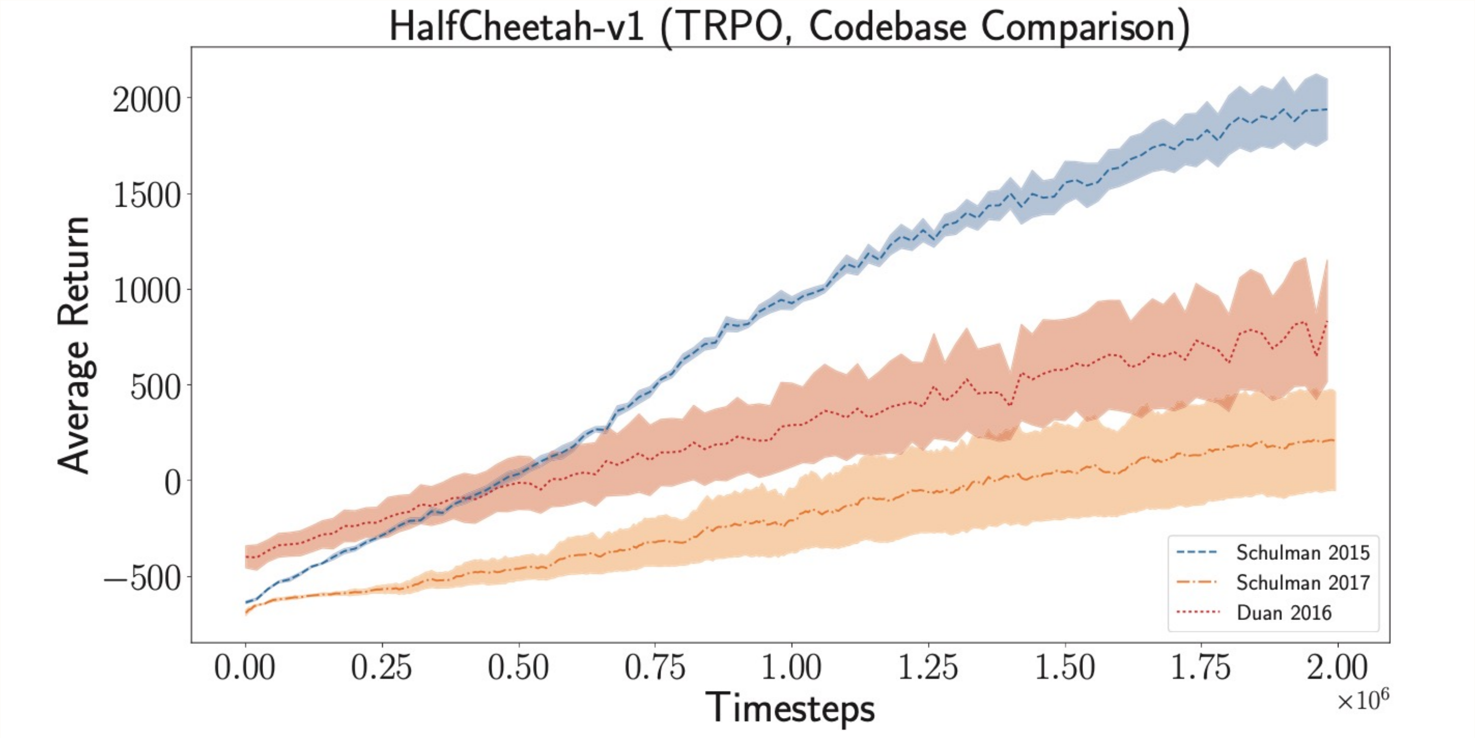
TRPO significantly outperforms all other algorithms. Due to the dynamics of the water-like environment, a local optimum for the system is to curl up and flail without proper swimming.

Swimmer ~ 130

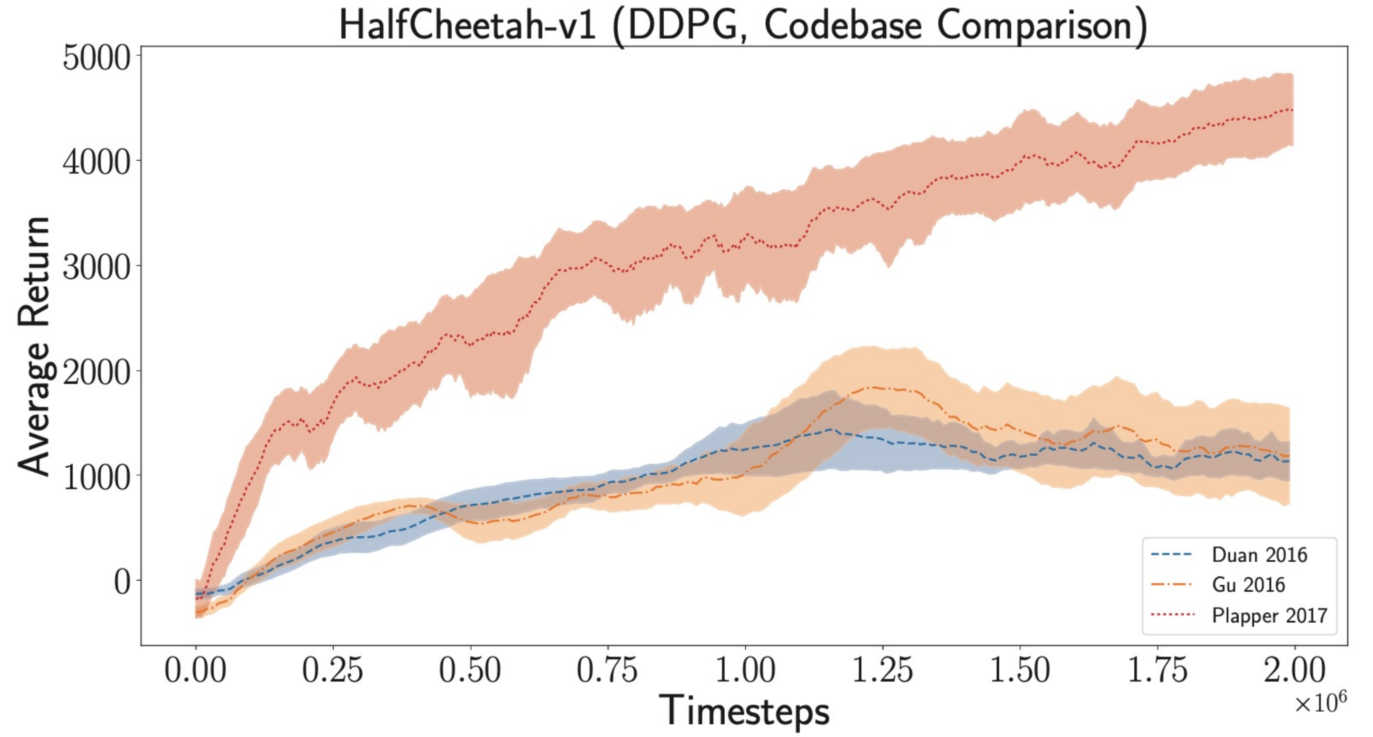


CODEBASES

- authors implement their own versions of baseline algorithms to compare against



This demonstrates the necessity that implementation details be enumerated, codebases packaged with publications, and that performance of baseline experiments in novel works matches the original baseline publication code.


HalfCheetah-v1 (DDPG, Codebase Comparison)
Average Return
5000
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
$\times 10^6$
Timesteps
Duan 2016
Gu 2016
Plapper 2017

CONTENT

- Introduction
- Importance of Reproducibility in RL
- Environments used for experiments and analyses
- Algorithms used for experiments and analyses
- Factors affecting reproducibility
- Conclusion

CONCLUSION

- Reproducibility ensures accurate judgment of novel methods, sustaining progress in RL research.
- Understanding algorithm performance in various environments is crucial for advancing RL
- Factors like codebases and environment properties significantly impact reproducibility.
- Considering the unfairness of selecting top-N rewards from several trials highlights the need for robust experimental practices.