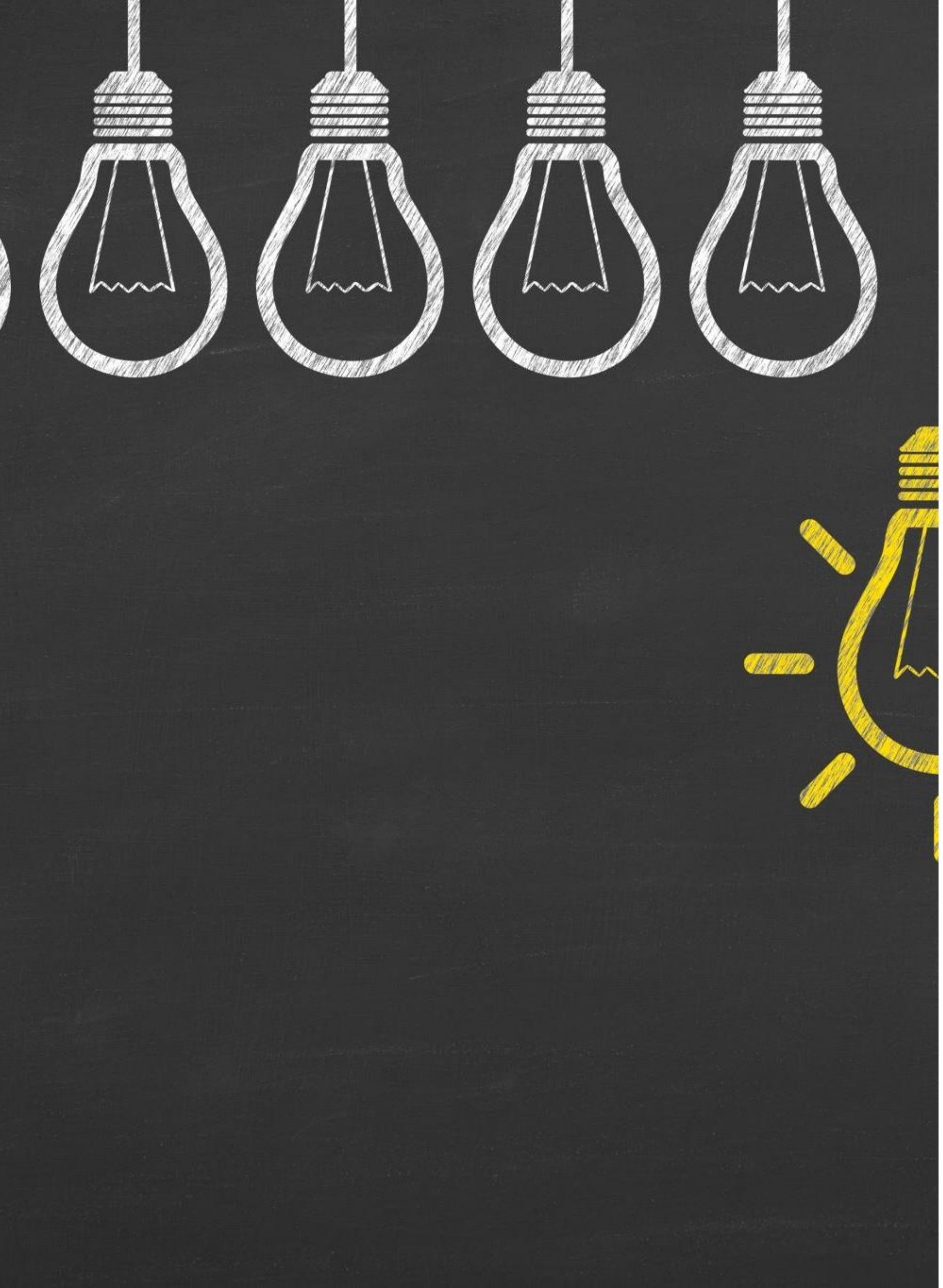


Curiosity-Driven Reinforced Learning of Undesired Actions in Autonomous Intelligent Agents

Christopher Rosser and Khalid Abed

Department of Electrical & Computer Engineering and Computer Science

Jackson State University

SAMI 2021

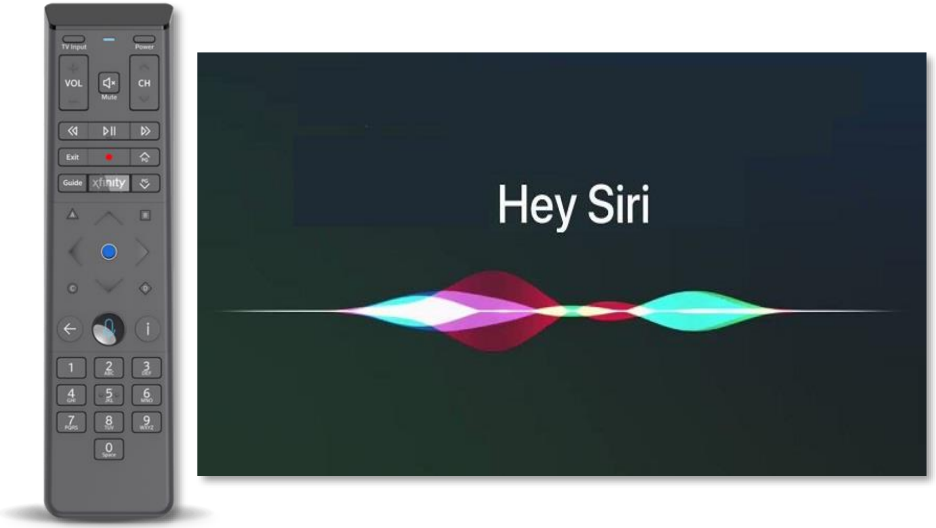
Outline

Introduction

Artificial Intelligence (AI) is everywhere!

Robotics – Warehouse, Cleaning, Delivery

Self-driving cars/Autonomous Underwater Vehicles

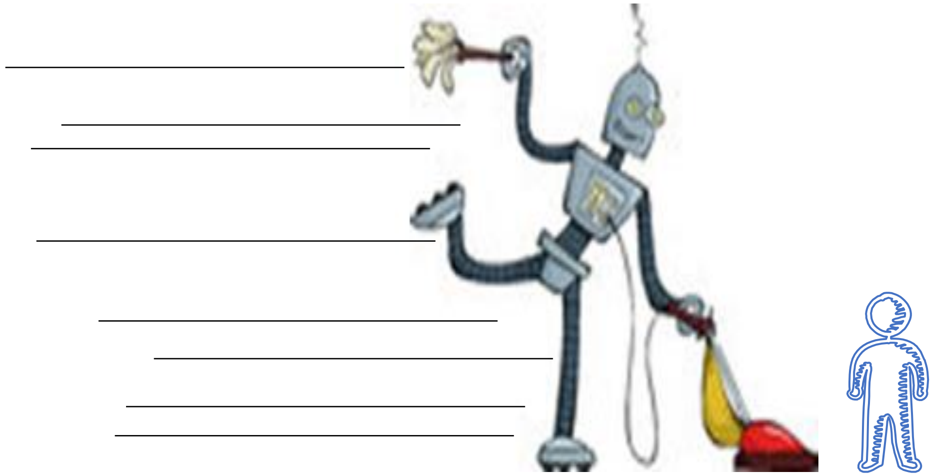


Speech Recognition (Natural Language Processing)

Video Game Playing

Why research autonomously exploring intelligent agents?

- Increase in autonomy of AI systems has risks and can lead to accidents, posing physical threats to human AI users
- AI research is more accessible than ever
- Improvements in safe exploration may have benefits in other domains



Harvard Business Review

TECHNOLOGY

What Changes When AI Is So Accessible That Everyone Can Use It?

by H. James Wilson and Paul R. Daugherty

January 30, 2018

Forbes

19,790 views | Sep 26, 2019, 10:03am EDT

What Happens When Self-Driving Cars Kill People?

Ron Schmelzer Contributor

COGNITIVE WORLD Contributor Group

AI

Why research autonomously exploring intelligent agents?

The following concrete problems in AI were provided in 2016 by researchers from Google Brain, OpenAI, UC Berkley, and Stanford in light of recent incidents

- Avoiding negative side effects
- Avoiding rewards hacking
- Scalable oversight
- Safe exploration
- Robustness to distributional shift

The same researchers assert that AI/ML accidents and risks are attributed to the following three "failures":

- Having the wrong objective function
- Having an objective function that is too expensive to evaluate frequently
- Undesirable behavior during the learning process

Why research autonomously exploring intelligent agents?

Recommended Solutions for Improving Exploration Safety:

- Adversarial Blinding - preventing an agent from understanding how its reward is generated or blinding it to certain variables
- Trip Wires - introduce deliberate vulnerabilities and monitor them so that researchers are alerted and can stop the agents immediately if the vulnerabilities are exploited.
- Simulated Exploration
- Human Oversight

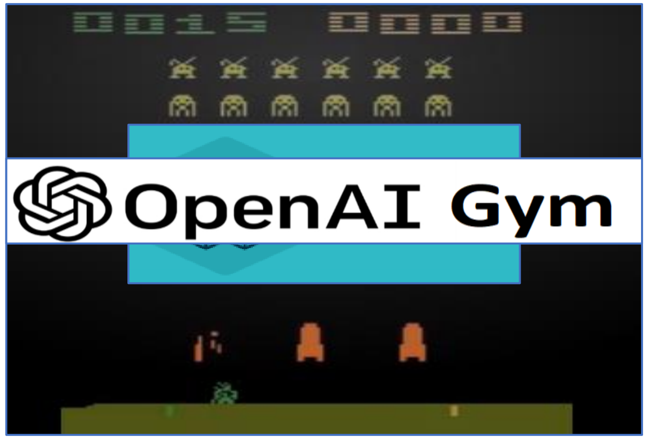
Why research autonomously exploring intelligent agents?

- We want to improve the safety of autonomous exploring AI through simulated exploration and human oversight.
- In this research we use Unity's Machine Learning Agents Toolkit (ML-Agents) to train purposely misbehaving agents to determine when a human should intervene during the agent's learning process.

Related Works

Game Engines/Development Platforms as RL Environments for Investigating Autonomous Exploration

- VizDoom and Mujoco
  - Episodic curiosity-driven exploration
- OpenAI
  - Arcade Learning Environment for general intelligent agents
  - Intrinsic Curiosity Module (ICM) for curiosity-driven exploration
  - Attention-based curiosity-driven exploration



Human-in-the-Loop Reinforcement Learning

- Hard-coded Guidance
  - Defining catastrophes and significant rare events before training
  - Environment-level action blockers
  - Determining where to add actions in RL
- Actual/Learned Human Intervention
  - Improving safety with model-based architectures and human intervention
  - Training agents to imitate human intervention
  - Runtime monitoring framework

Methodology

- Use Unity’s Machine Learning Agents Toolkit (ML-Agents) implementation of Proximal Policy Optimization (PPO) + ICM algorithm to create purposely misbehaving autonomous exploring agents
- Identify PPO+ICM training statistics and custom environment metrics associated with agent misbehavior

Training Agents in ML-Agents

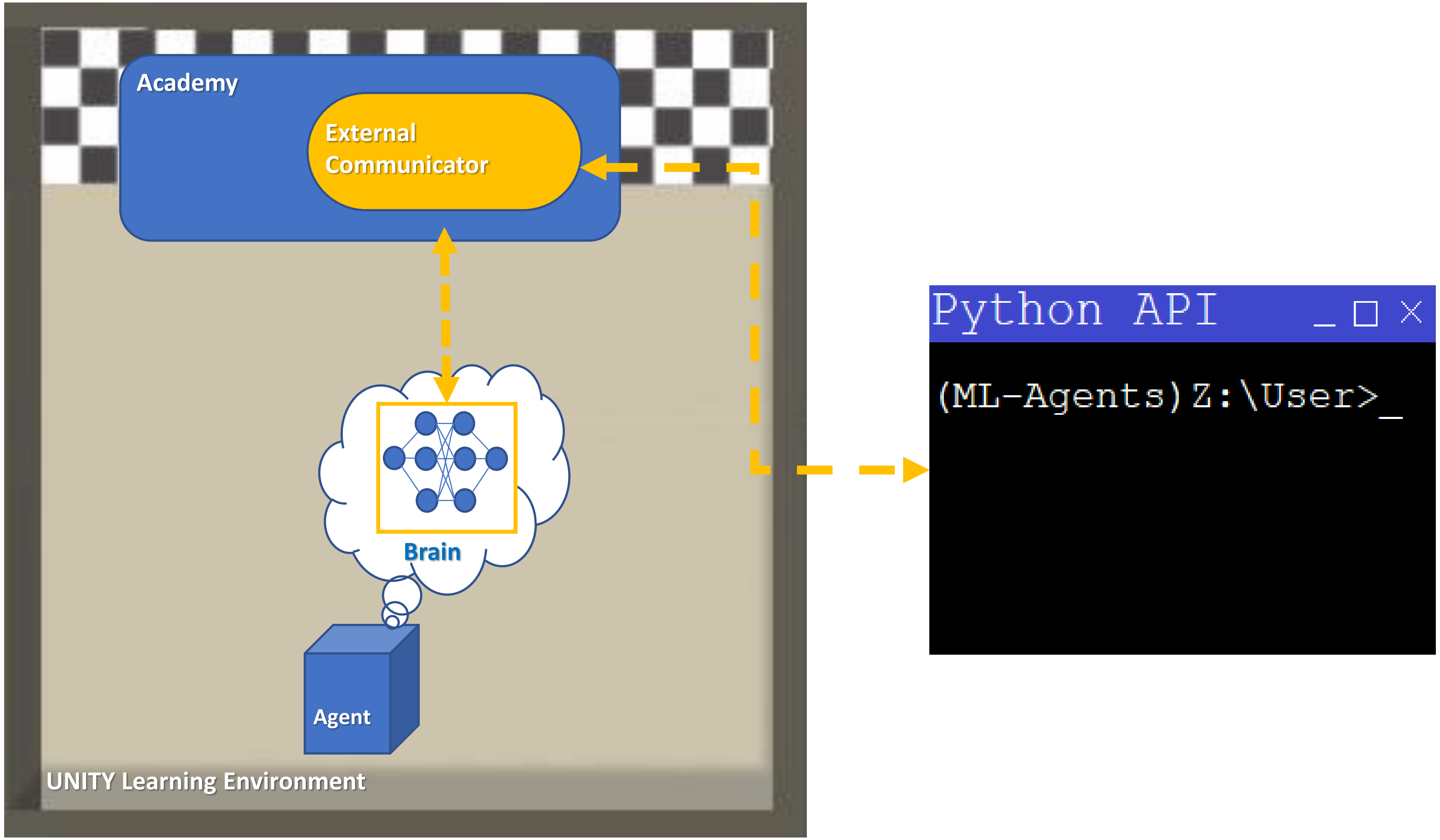
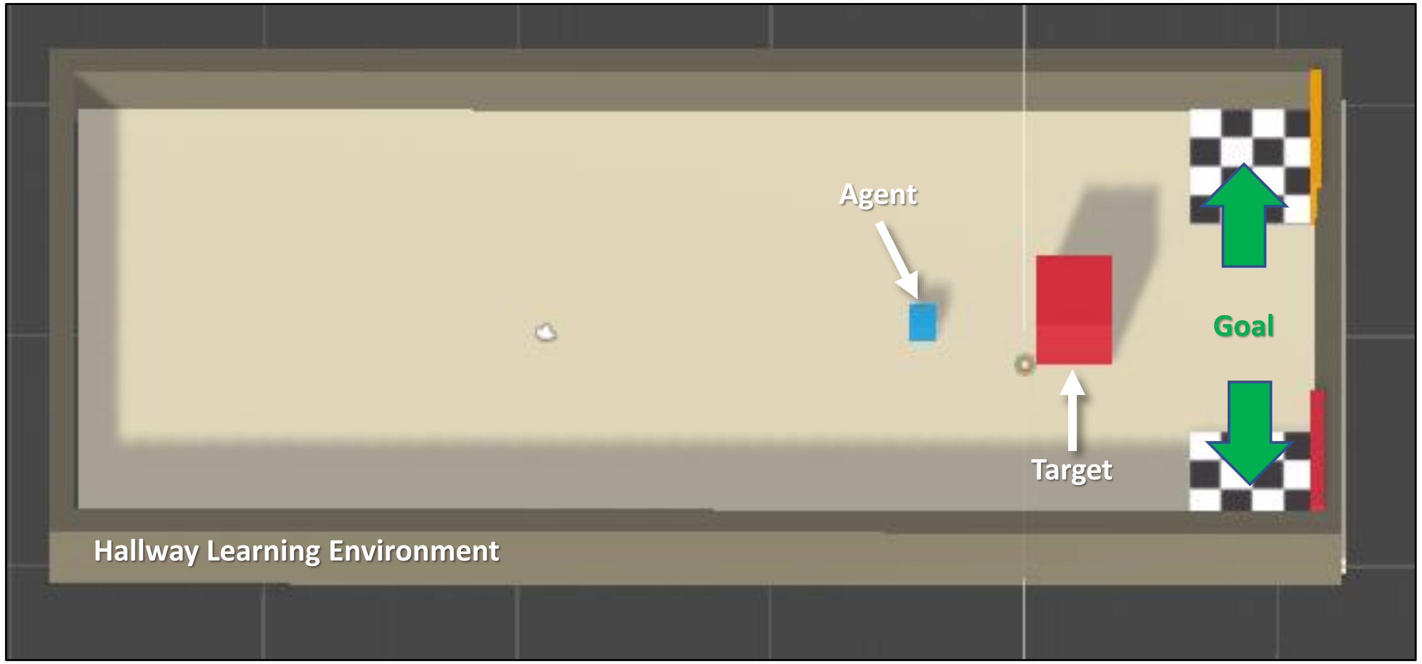


Illustration of externally training neural networks with ML-Agents

PPO + ICM, Learning Environment, and Training Parameters

PPO + ICM Algorithm
for iteration = 1, 2, … **do**
collect set of actions (a) and next states (s+1) with policy (π)
encode current state (s), next state (s+1) → φ(s), φ(s+1)
compute predicted encoded next state $\hat{\varphi}$(s+1)
intrinsic reward ← $\hat{\varphi}$(s+1) – φ(s+1)
optimize π parameters for maximizing extrinsic + intrinsic rewards
update π
end for



PPO + ICM TRAINING PARAMETERS

Parameter	Value
gamma	0.99
lambda	0.95
buffer size	1024
batch size	128
epochs	3
learning rate	0.0003
time horizon	64
max steps	150000
beta	0.01
epsilon	0.2
hidden layers	2
hidden units	128
sequence length	64
memory size	256
curiosity encoding size	128-256
curiosity strength	0.01-10.0

Training Conditions and Experiments

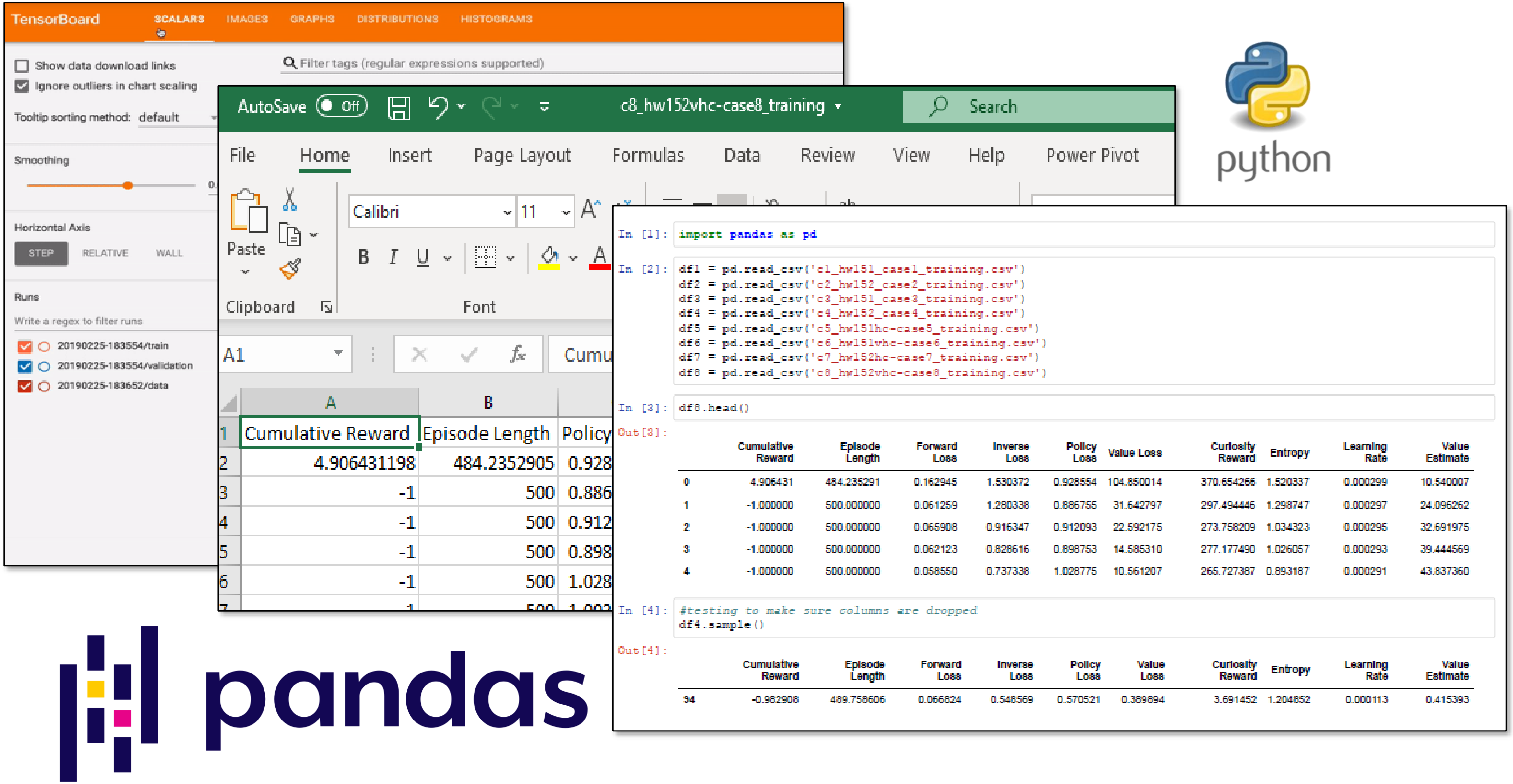
Train autonomous exploring agents with high intrinsic curiosity, large rewards, and large penalties for 150,000 timesteps under the following six conditions:

1. Low/Zero Curiosity Strength, Small Rewards, Large Penalty
2. Low/Zero Curiosity Strength, Large Rewards, Small Penalty
3. Max Recommended Curiosity Strength, Small Rewards, Large Penalty
4. Max Recommended Curiosity Strength, Large Rewards, Small Penalty
5. Very High Curiosity Strength, Small Rewards, Large Penalty
6. Very High Curiosity Strength, Large Rewards, Small Penalty

CURIOSITY STRENGTHS AND REWARDS FOR TEST CASES 1 - 8

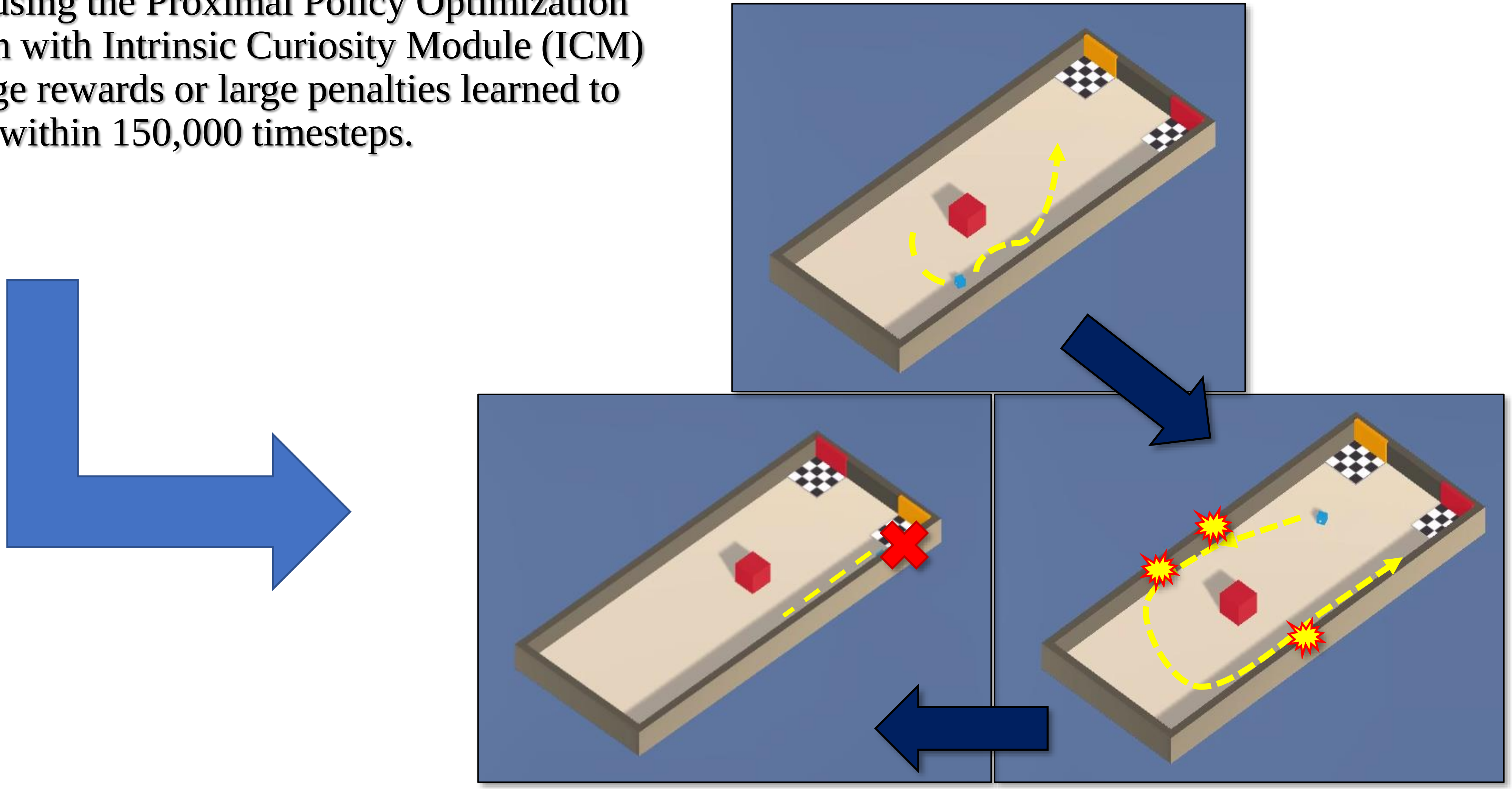
Case	Curiosity Strength	Rewards	Penalties
1	Disabled	1	-100, -0.1
2	Disabled	100	-1, -0.1
3	0.1	1	-100, -0.1
4	0.1	100	-1, -0.1
5	1.0	100	-1, -0.1
6	1.0	1	-100, -0.1
7	10.0	1	-100, -0.1
8	10.0	100	-1, -0.1

Analysis of ML-Agents Data



Results and Discussion

Agents trained using the Proximal Policy Optimization (PPO) algorithm with Intrinsic Curiosity Module (ICM) enabled and large rewards or large penalties learned to act undesirably within 150,000 timesteps.



- Implemented a collision counter and goal-to-collision ratio to identify three test cases for further analysis

AVERAGE VALUES FOR AGENT WITH DEFAULT NEURAL NETWORK

Goals	Collisions	GC-Ratio
52	2.56	0.9566

AVERAGE VALUES FOR AGENT WITH CASES 1-8 NEURAL NETWORKS ATTACHED

Case	Goals	Collisions	GC-Ratio
1	0	12.66	0.001
2	1.333	183.33	0.008
3	0.333	140.67	0.003
4	0	418.33	0.001
5	0.333	268.67	0.006
6	0	362.33	0.001
7	0	436.33	0.001
8	0	170.67	0.001

Case 6

Two strongest negative correlations

(1) Episode Length v. Cumulative Reward, (2) Value Loss v. Value Estimate

Third strongest

(3) Value Estima

Two strongest p

(1) Entropy v. Va

Third strongest

(3) Learning Rat

```
In [18]: df_norm6.corr()
```

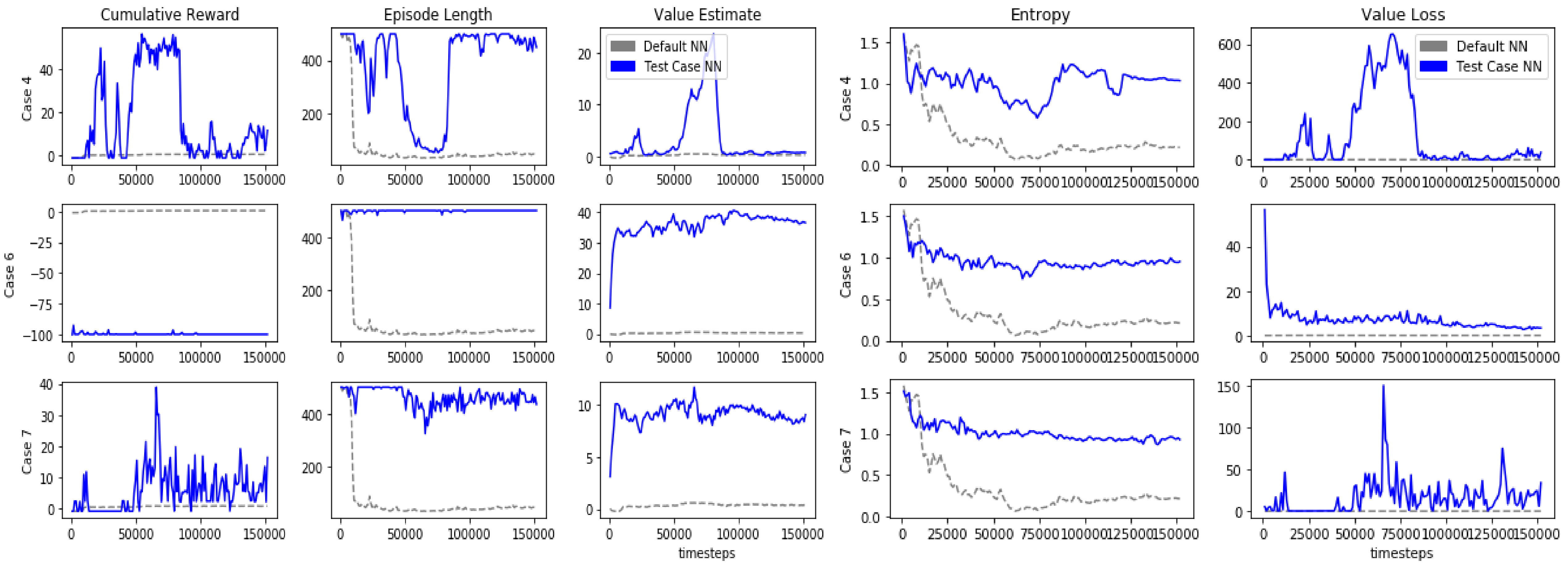
Out[18]:

	Cumulative Reward	Episode Length	Policy Loss	Value Loss	Entropy	Learning Rate	Value Estimate
Cumulative Reward	1.000000	-0.999761	0.228822	0.331685	0.434574	0.274277	-0.274521
Episode Length	-0.999761	1.000000	-0.232948	-0.346523	-0.442640	-0.276359	0.285040
Policy Loss	0.228822	-0.232948	1.000000	0.351071	0.554644	0.159015	-0.320367
Value Loss	0.331685	-0.346523	0.351071	1.000000	0.643581	0.529072	-0.801056
Entropy	0.434574	-0.442640	0.554644	0.643581	1.000000	0.470622	-0.665674
Learning Rate	0.274277	-0.276359	0.159015	0.529072	0.470622	1.000000	-0.593451
Value Estimate	-0.274521	0.285040	-0.320367	-0.801056	-0.665674	-0.593451	1.000000

Average Cumulative Reward vs. Episode Length Correlation of

-0.96!

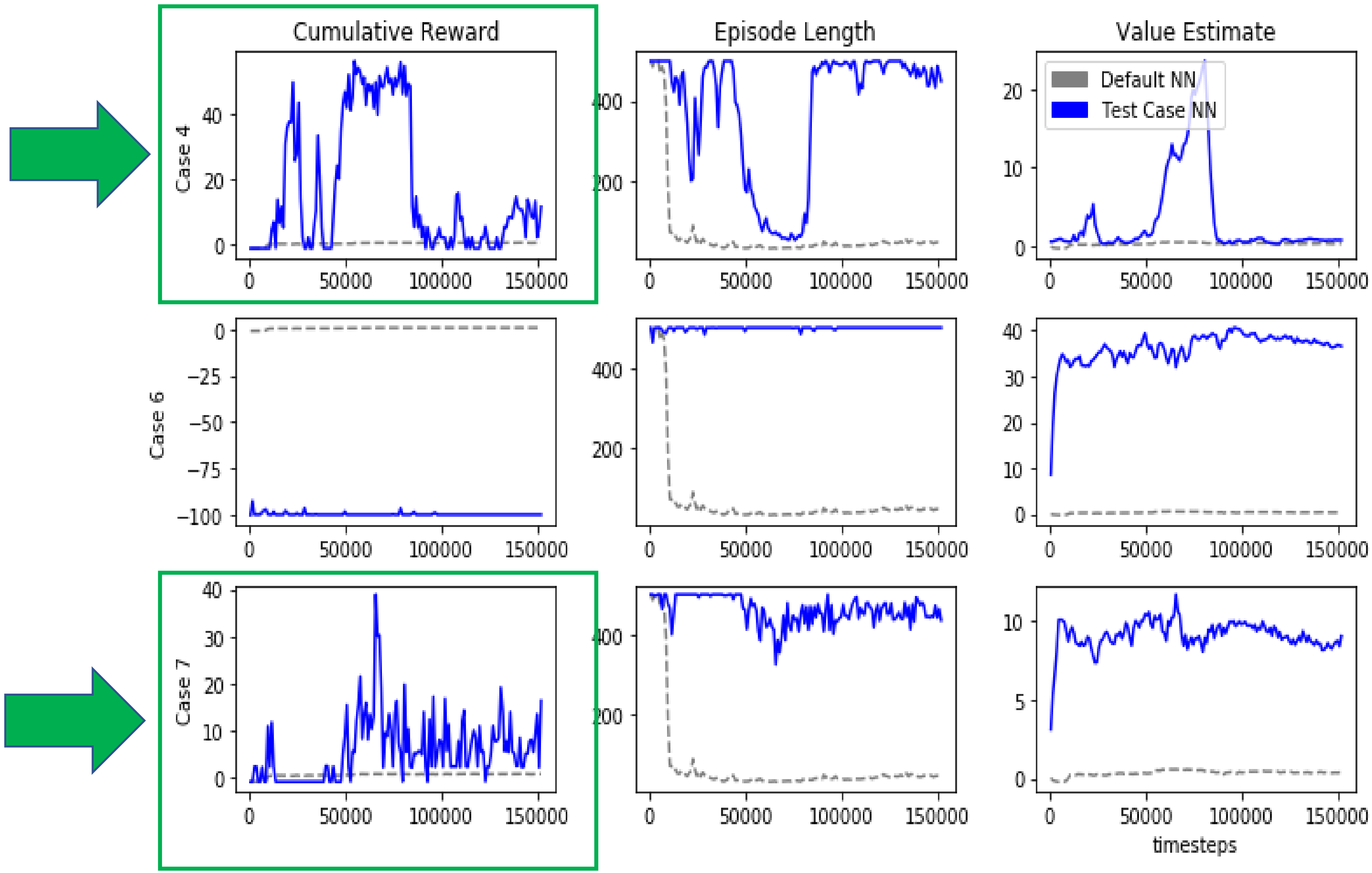
PPO+ICM for Hallway Cases 4, 6, and 7



Conclusion and Future Work

Conclusion

High-frequency oscillation of cumulative rewards indicate agent misbehavior



Future Work

- Expand and conduct experiments in additional learning environments
- Investigate Research Questions (RQs) and Research Objectives (ROs) 1 & 2
- RQ1: How can we modify RL algorithms to detect anomalies in training statistics?
  - RO1: Incorporate findings from experiments discussed in this research into a modified PPO + ICM algorithm.
- RQ2: How can we accommodate a distracted human’s efforts to intervene during the agent’s learning process in human-in-the-loop RL?
  - RO2: Design and implement a scheme for alerting and receiving input from the human during RL through an external smart device.

The End

Thank you!

References

Images used in this presentation were obtained from the following websites:

https://www.militaryaerospace.com/unmanned/article/14069139/swarming-uavs-counter
https://www.wired.com/story/waymo-google-arizona-phoenix-driverless-self-driving-cars/
https://www.xfinity.com/tips/xfinity-x1-remote-tips-and-tricks
https://www.republicworld.com/technology-news/mobile/how-to-make-siri-say-things-in-ios-14-learn-in-simple-steps-here.html
https://en.wikipedia.org/wiki/Pac-Man
https://gym.openai.com/
http://www.mujoco.org/
https://github.com/Unity-Technologies/ml-agents
https://en.wikipedia.org/wiki/Pandas_(software)
https://www.libraries.rutgers.edu/node?page=23
https://www.slidescarnival.com/
https://www.fotosearch.com/CSP992/k13227884/