Degrees of Guidance

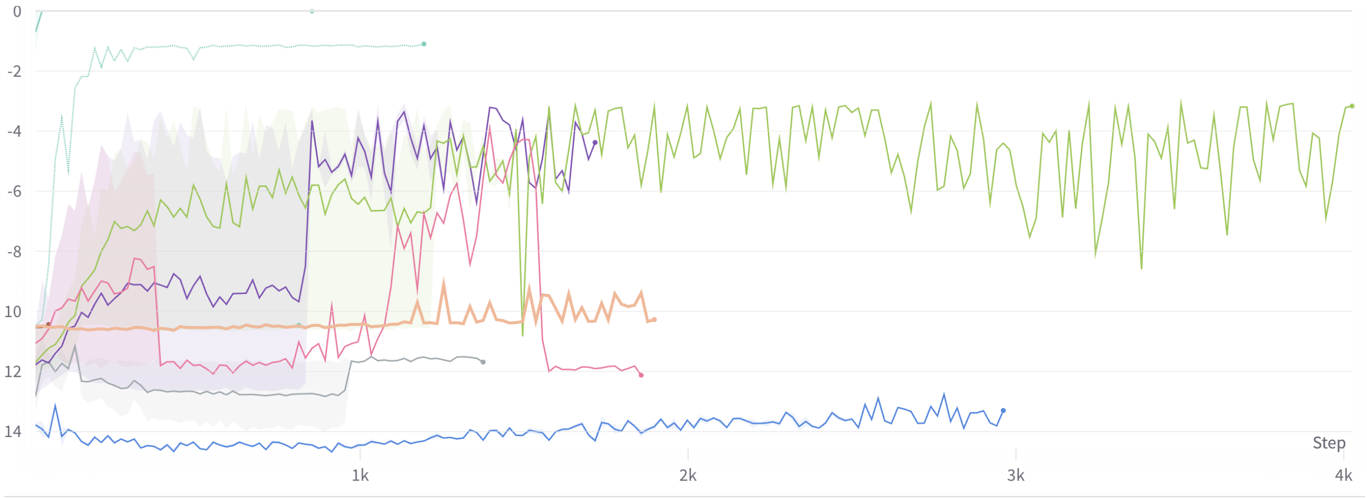
European Master Team Project – AI BugPlus

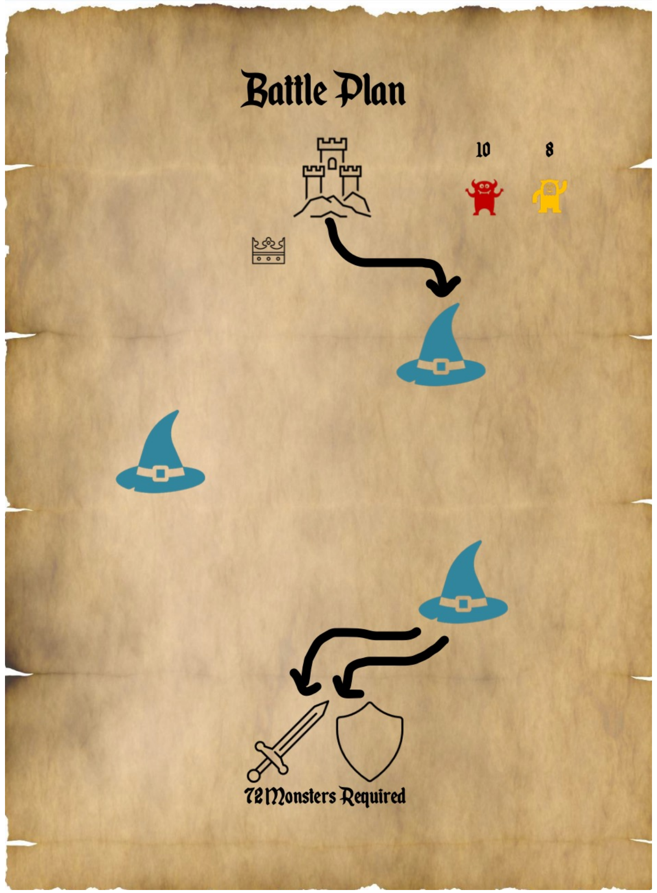
Aaron Steiner, Mae Turner, Mayte Dächer, Radu Tarean, Rares Matisan

18 April 2023

Goals of the Project

- Explore RL using BugPlus
- Create a platform that helps in understanding BugPlus and makes it easier to interact with it





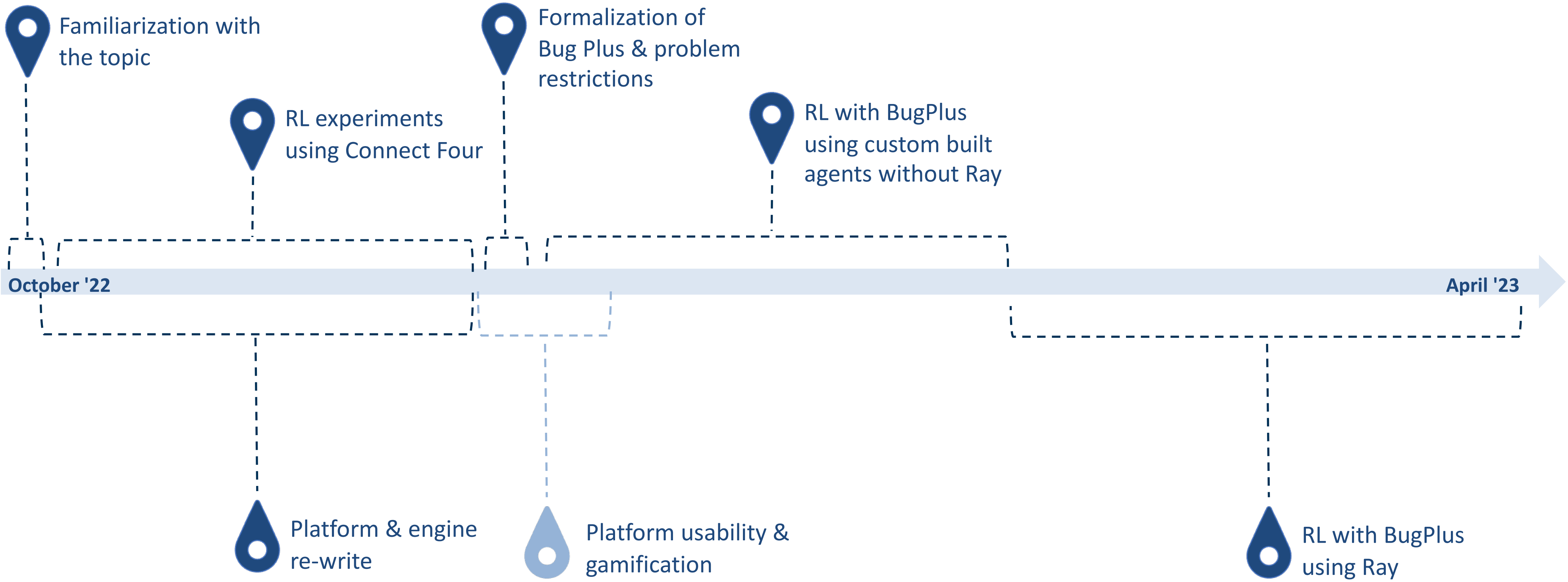
Teamwork

Team Members

Mode of Operation:

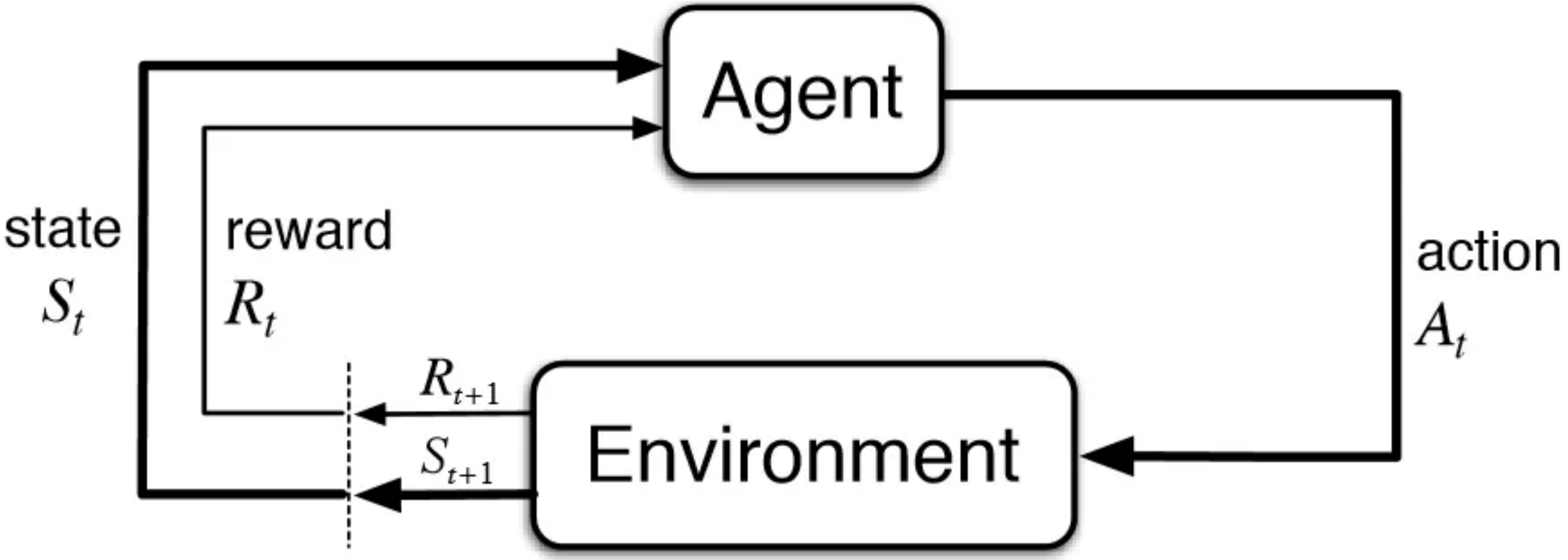
- Sprint duration of 2 weeks → Sprint Review and planning together with advisors on a bi-weekly basis
- Weekly internal team meetings for alignment on progress and problem discussions

Project Phases


Familiarization with the topic
RL experiments using Connect Four
Formalization of Bug Plus & problem restrictions
RL with BugPlus using custom built agents without Ray
October '22
April '23
Platform & engine re-write
Platform usability & gamification
RL with BugPlus using Ray

Reinforcement Learning

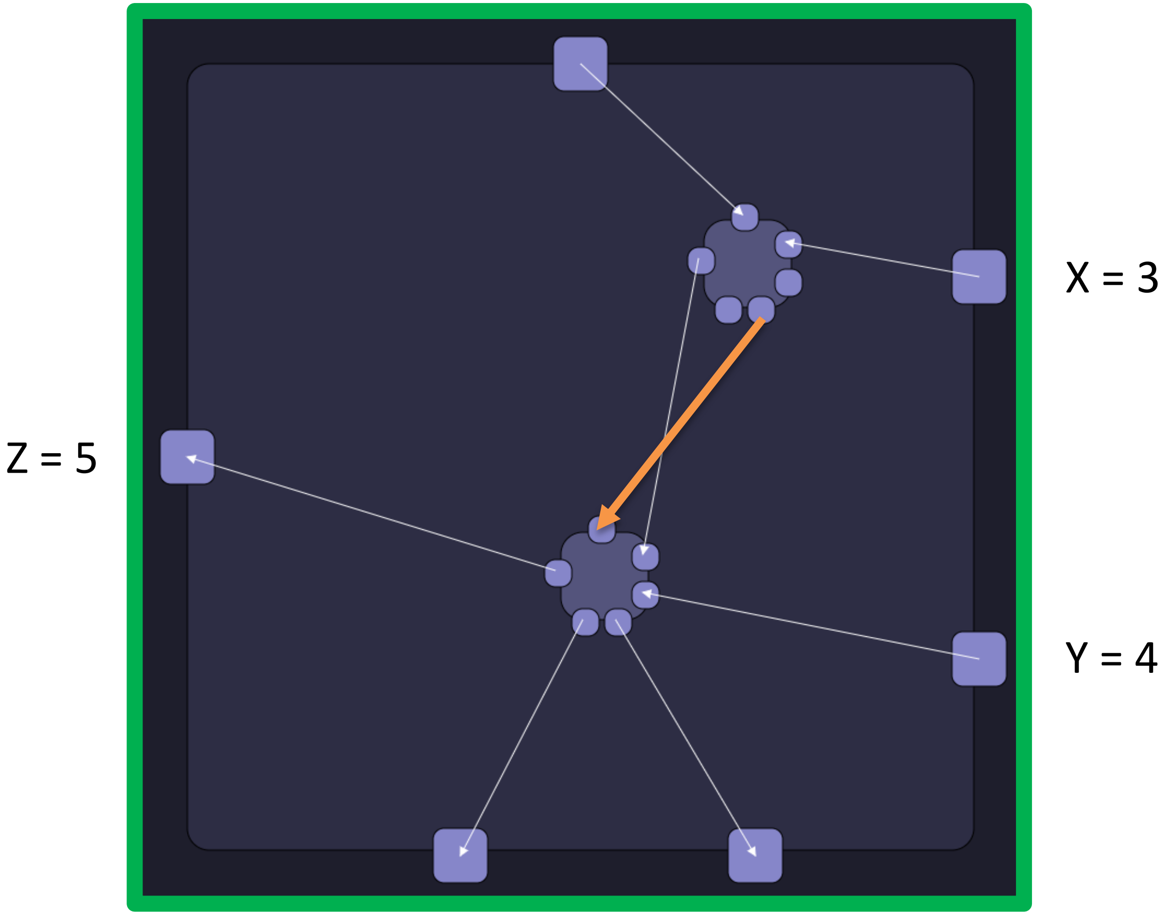
Intro



Reinforcement Learning

Terms

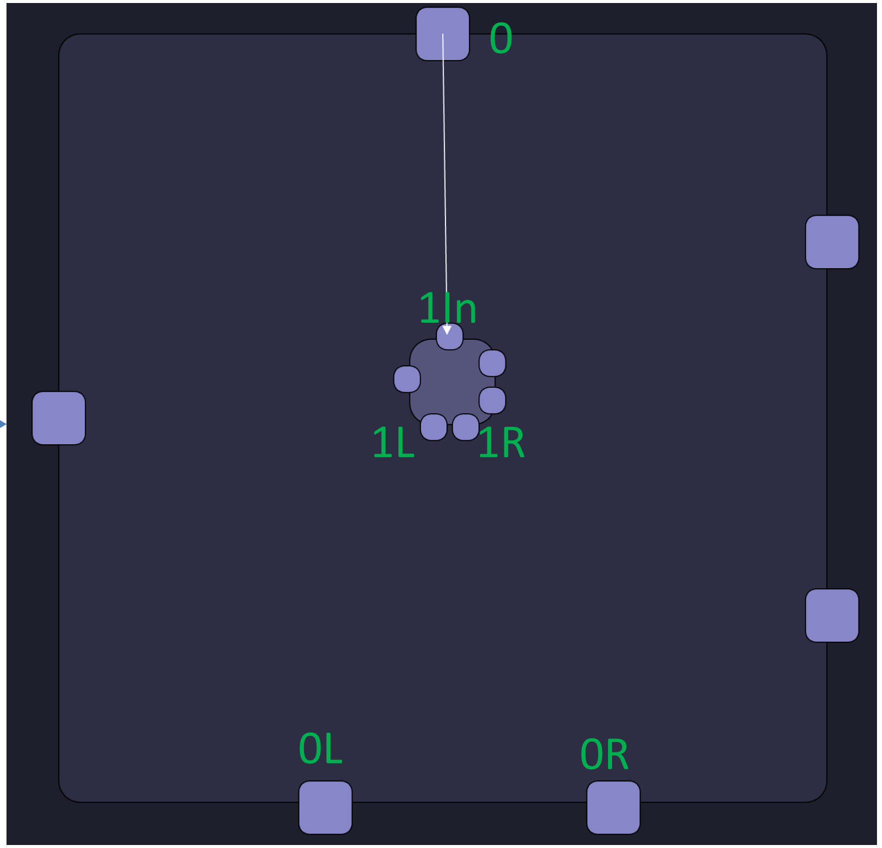
- Problem (y + 1)
- Config
- Edge
- Step = Agent acts (place one edge), gets reward
- Episode = Sequence of steps until termination



Reinforcement Learning

Environment

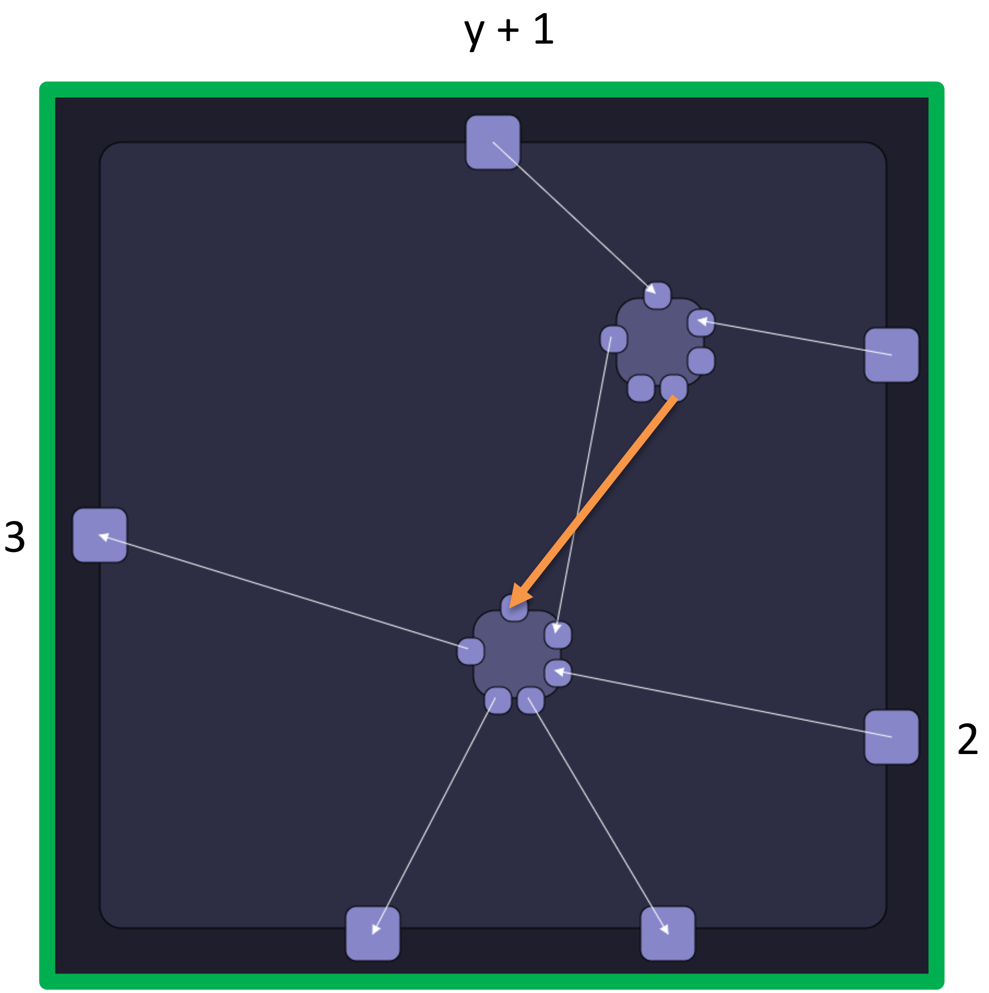
to\from	0	1 L	1 R
0L			
0R			
1 in	1		



Reinforcement Learning

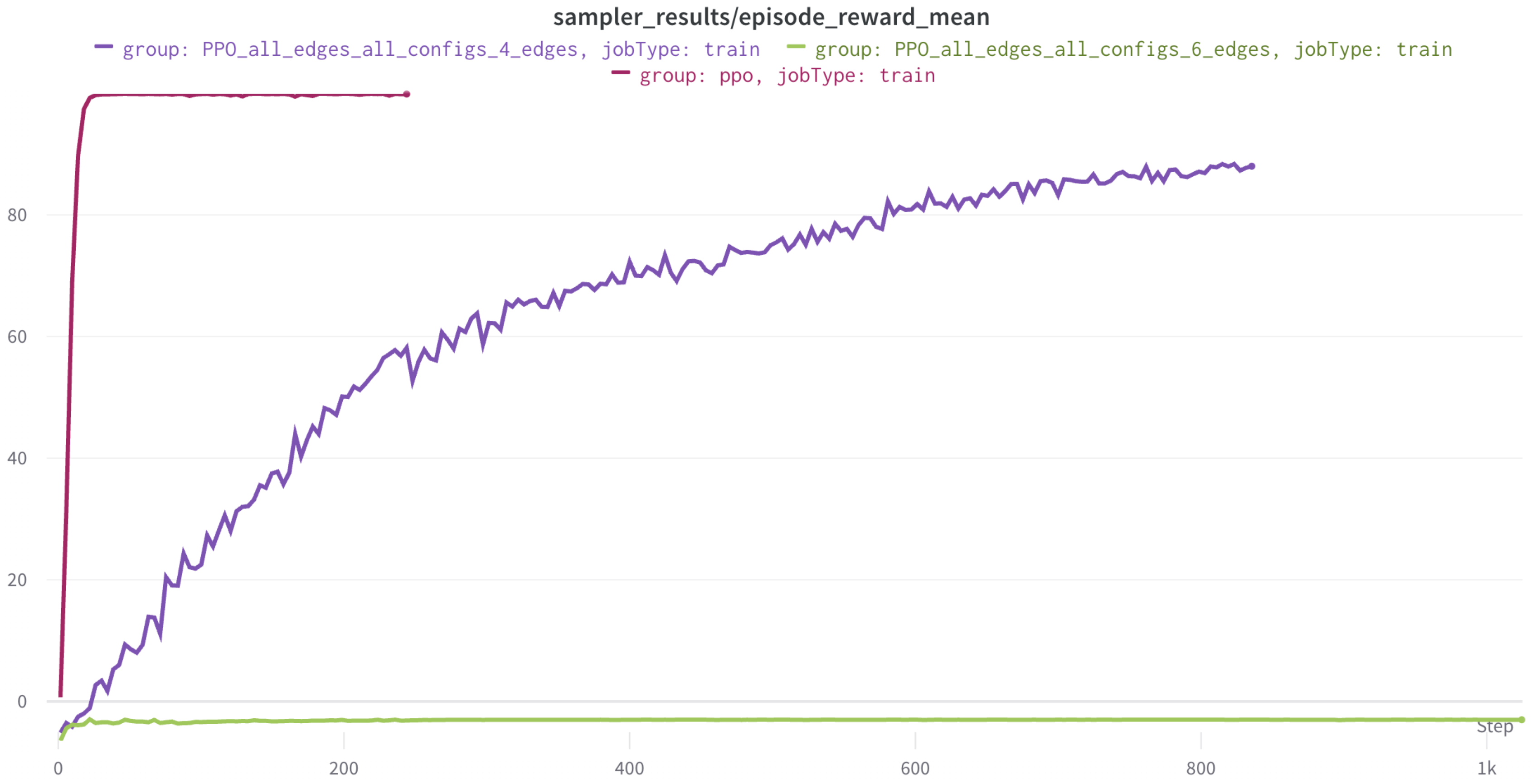
Config- Generation

- We build all (52) problems that are possible to create with our restrictions
  - No loops
  - 3 bugs
- Delete n edges
- Create input + output values



Degrees of Guidance

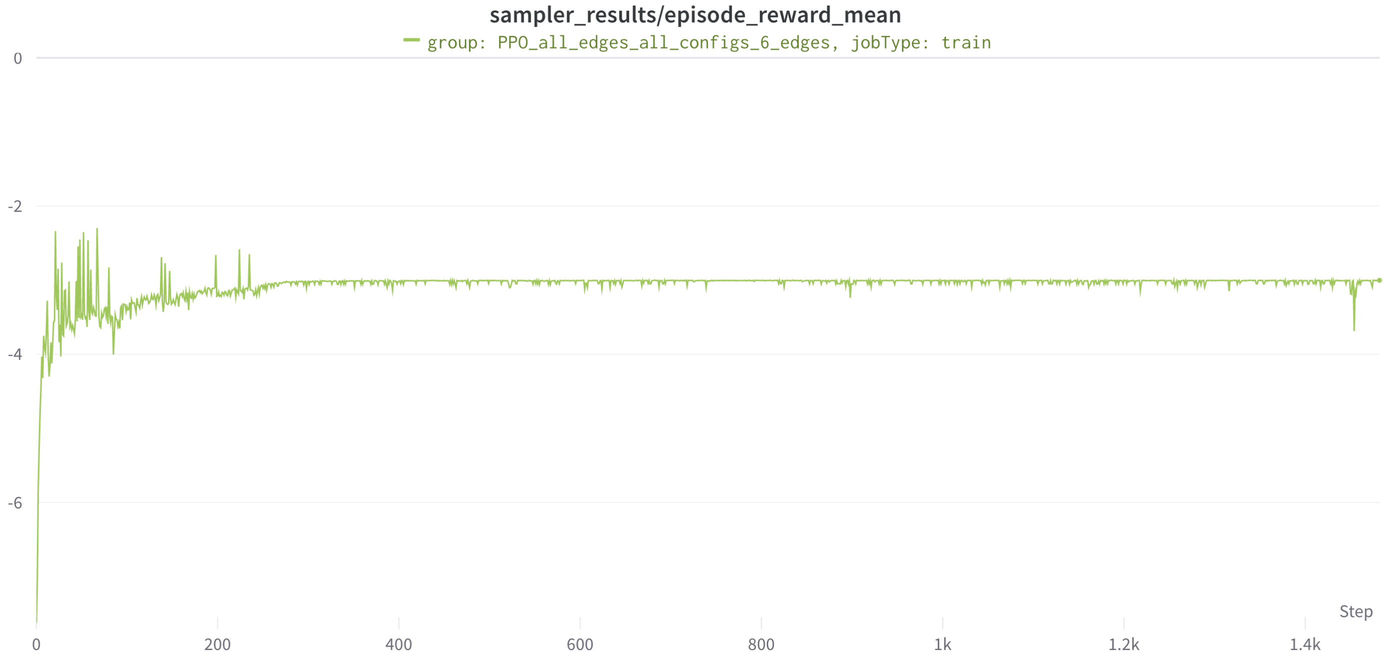
Baseline Agents



Degrees of Guidance

Reward System

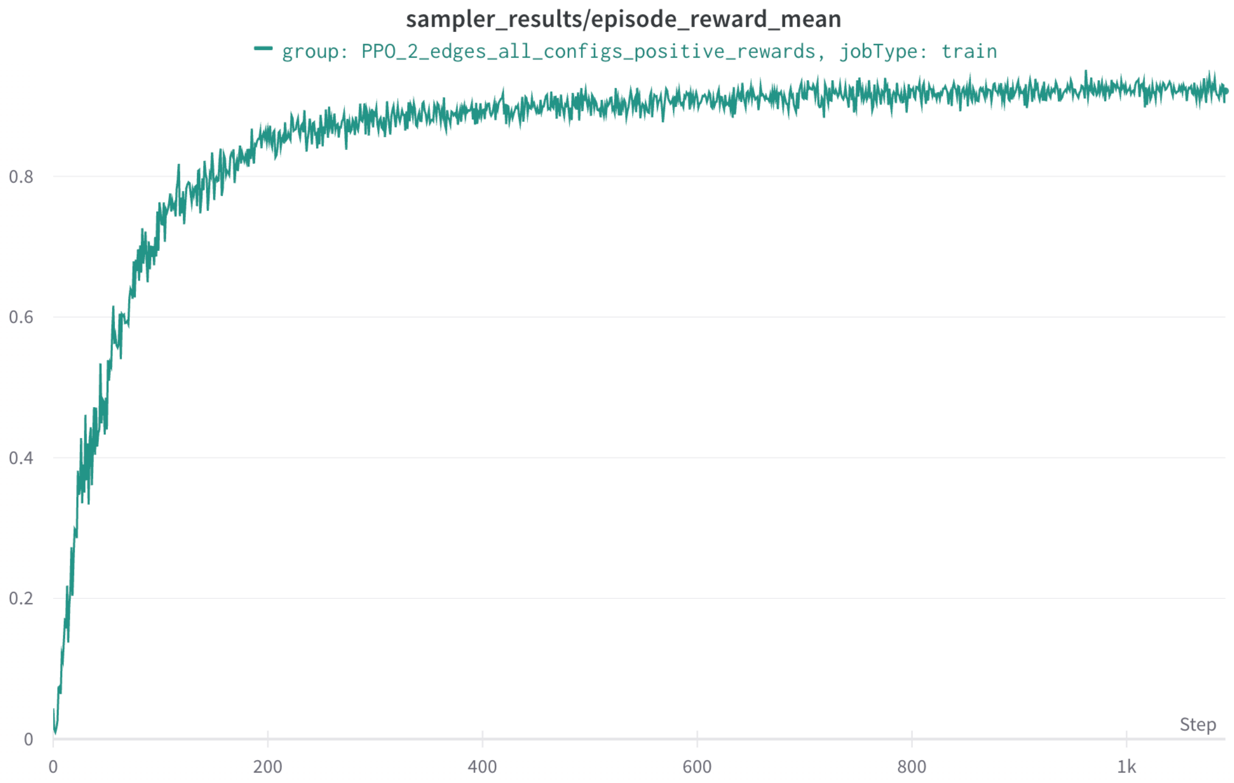
→ Idea: appeal to agent's "intrinsic motivation" & encourage exploration

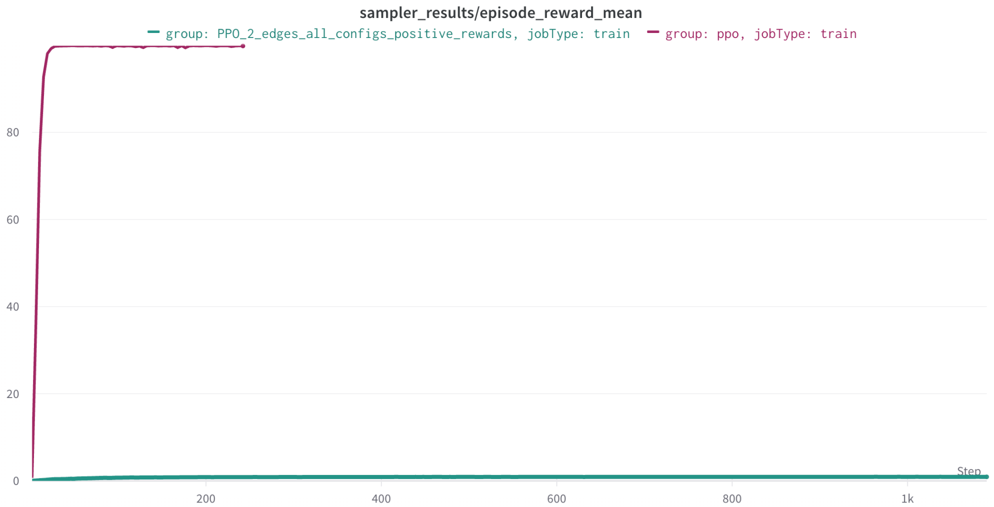


Burda, Y., Edwards, H., Pathak, D., Storkey, A. J., Darrell, T., & Efros, A. A. (2019). Large- scale study of curiosity-driven learning. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Degrees of Guidance

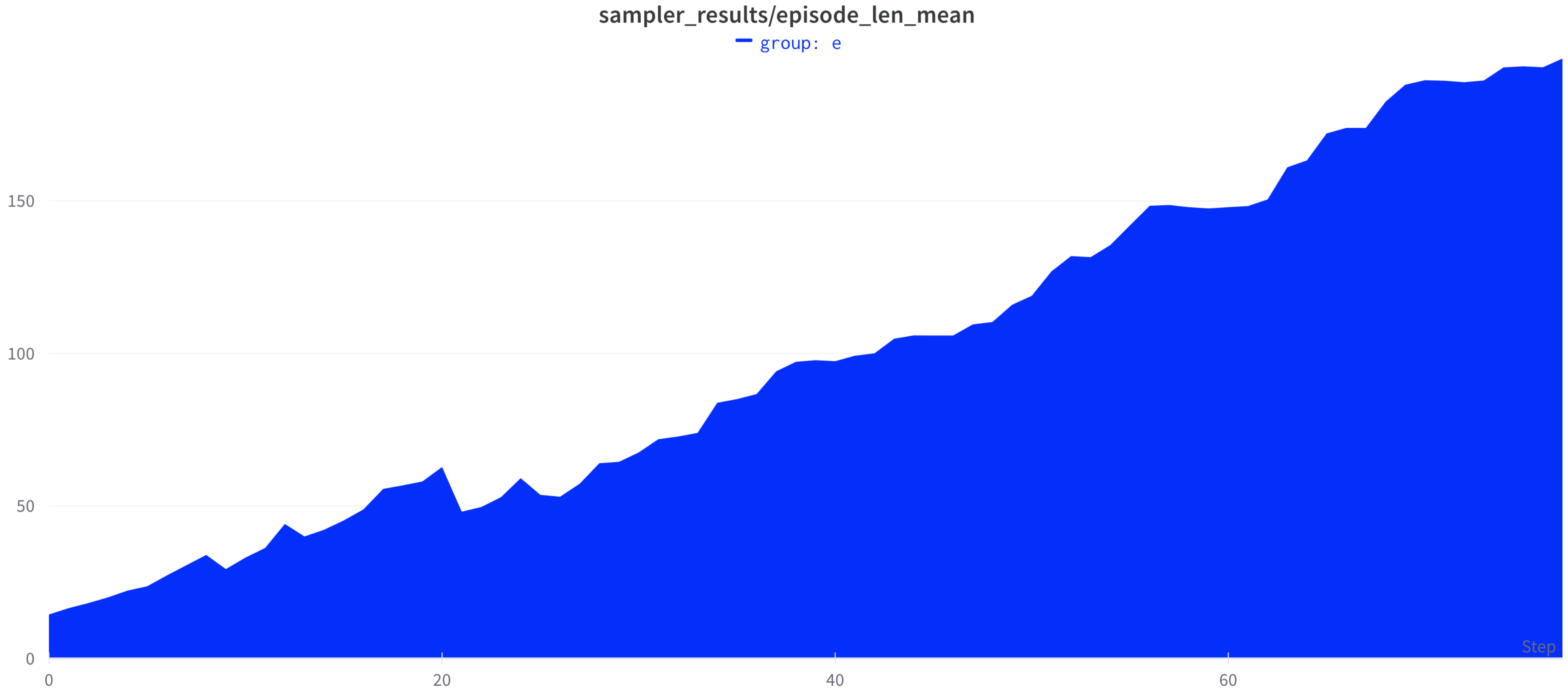
Reward System: Sparse Rewards





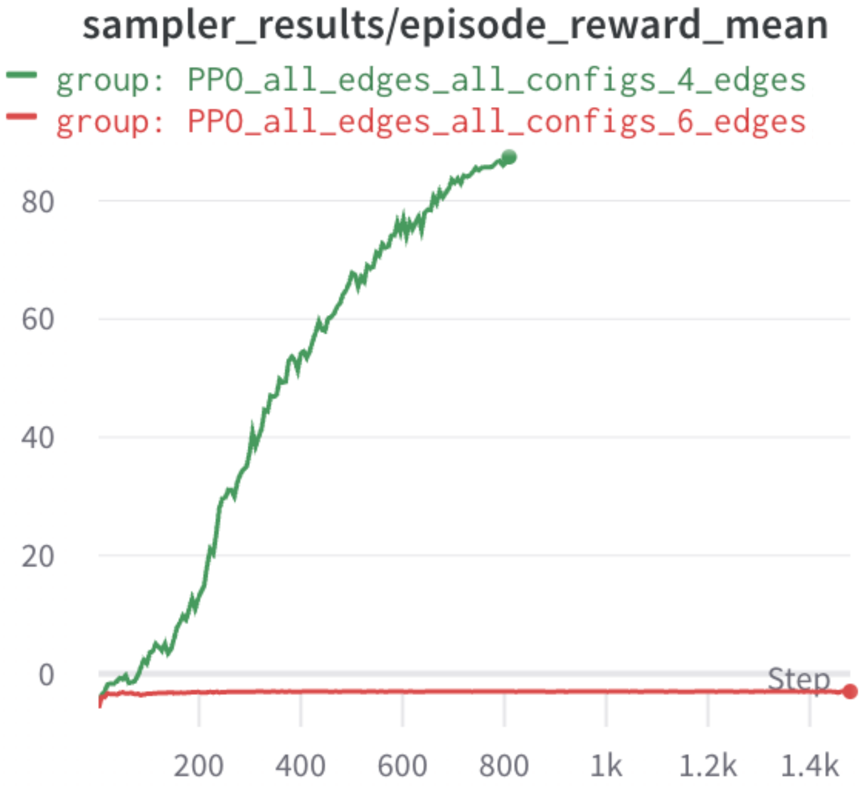
Reinforcement Learning

Episode length

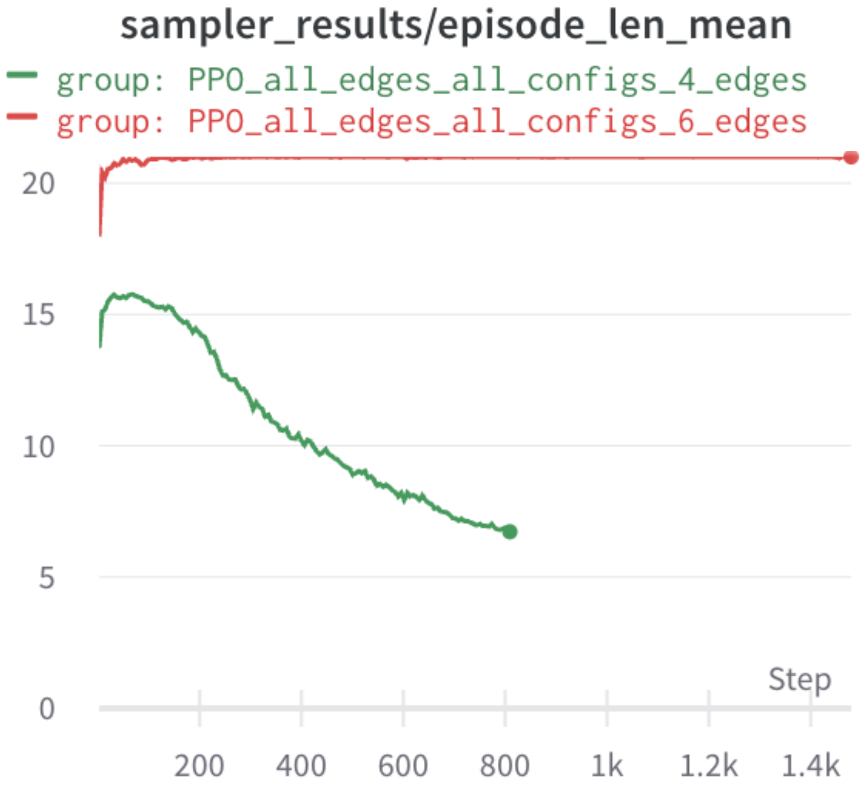


Reinforcement Learning

Where are we at?



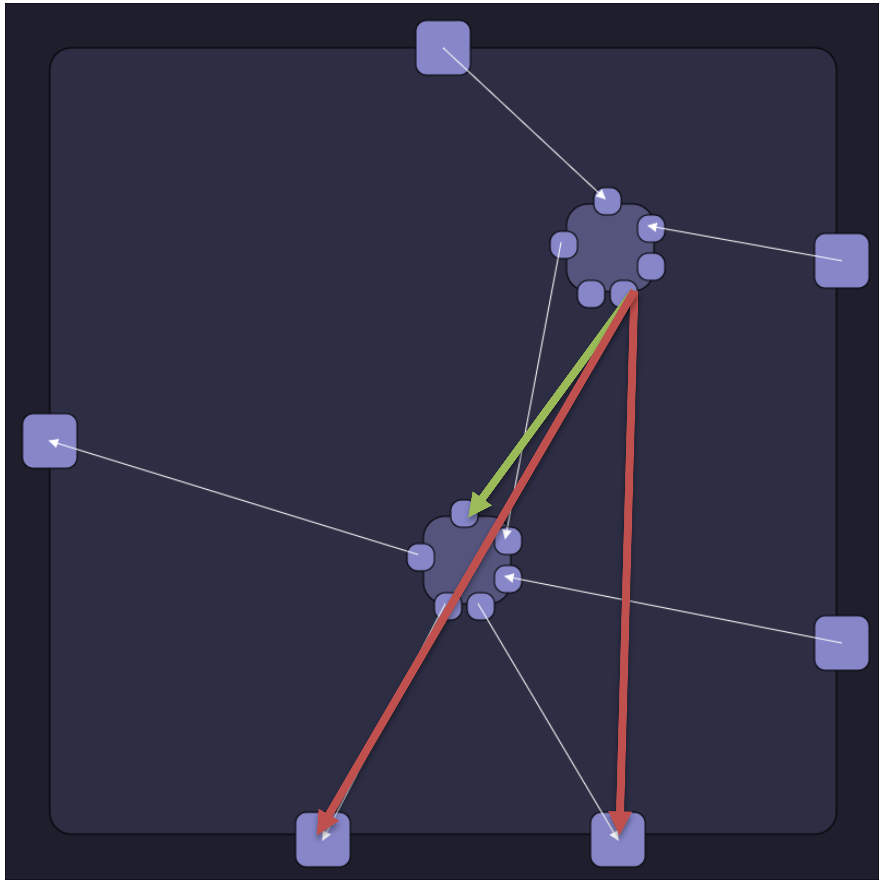
sampler_results/episode_reward_mean



sampler_results/episode_len_mean

Degrees of Guidance

Feedback from Engine

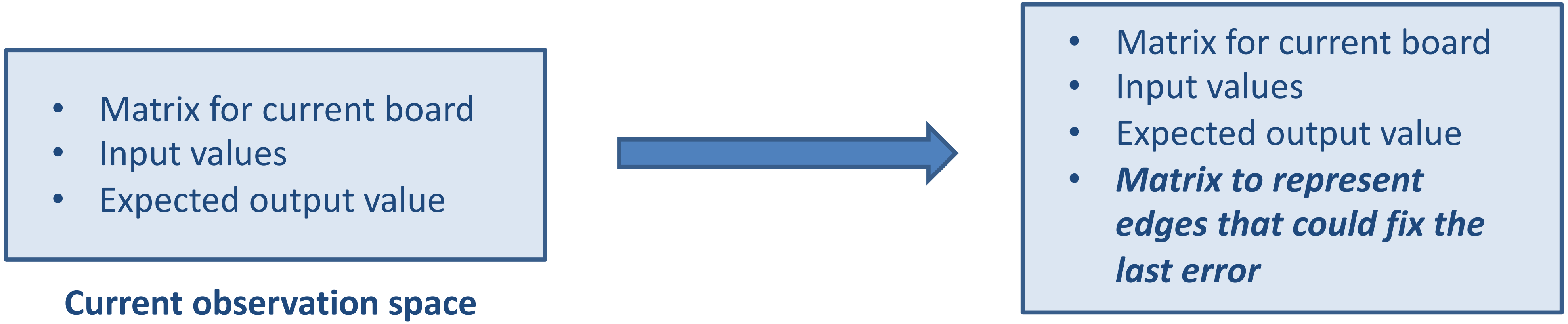


→ **ValueError:** Port Right of bug 1 is not connected to anything

- This information would likely be very helpful to the agent, how do we present it to the agent?

Degrees of Guidance

Feedback from Engine: Observation Space



- Matrix for current board
- Input values
- Expected output value

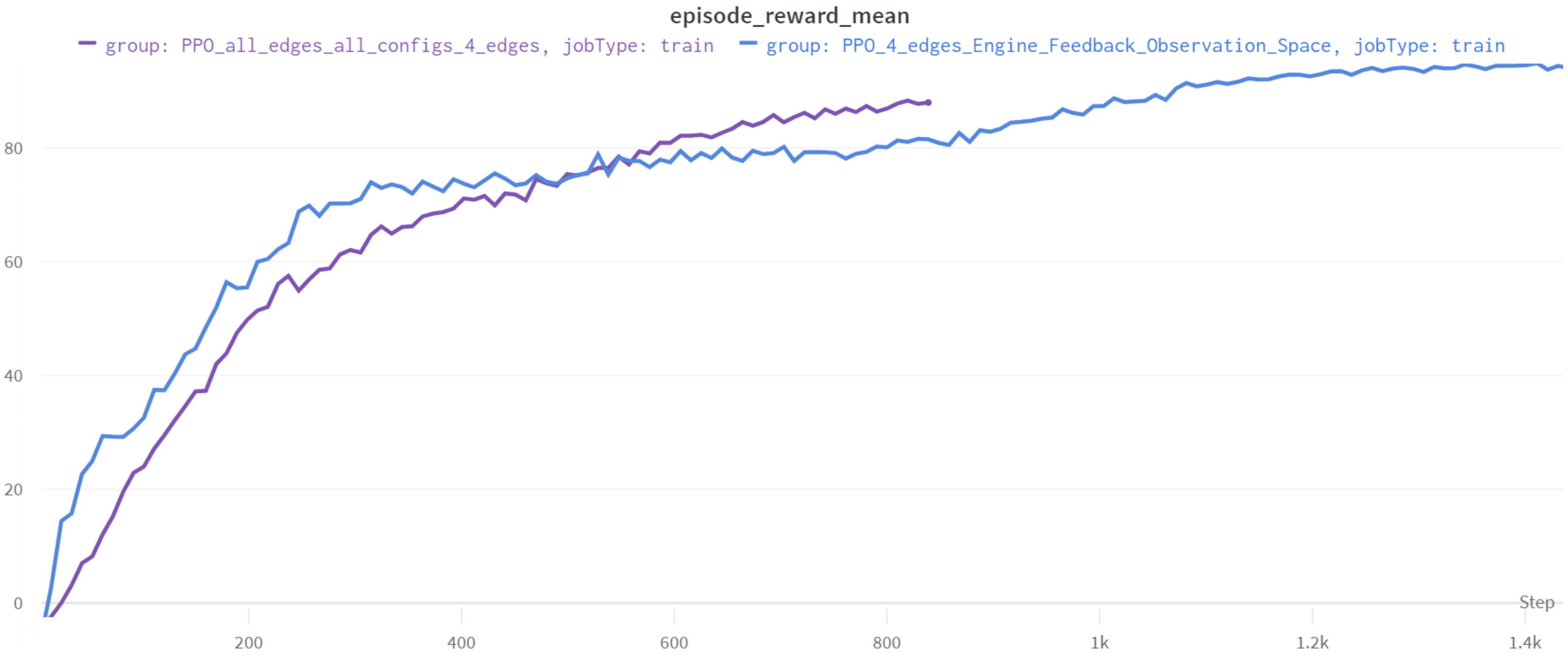
Current observation space

- Matrix for current board
- Input values
- Expected output value
- ***Matrix to represent edges that could fix the last error***

New observation space

Degrees of Guidance

Feedback from Engine: Observation Space



Degrees of Guidance

Feedback from Engine: Utilizing Rewards

- Matrix for current board
- Input values
- Expected output value
- ***Matrix to represent edges that could fix the last error***

New observation space

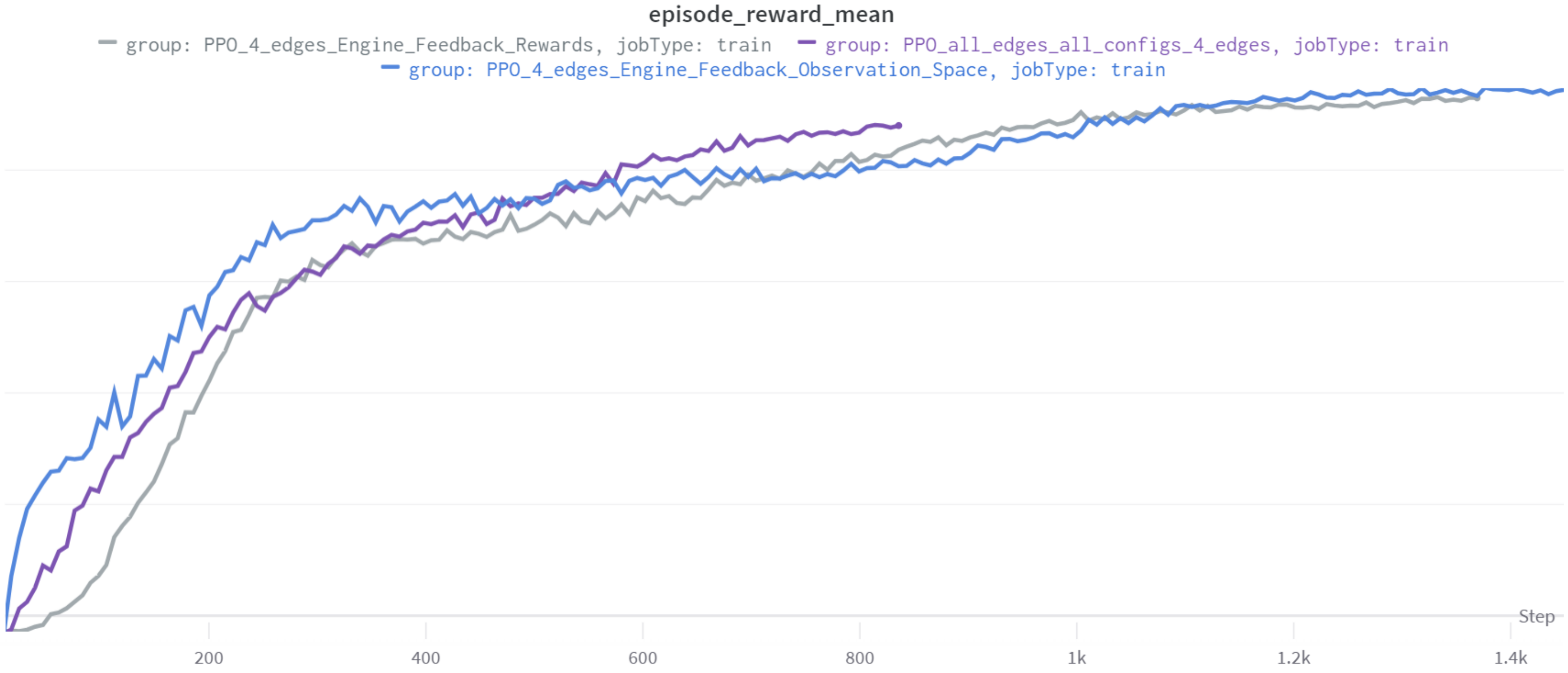
+

Action is chosen that fixes the previous error **AND** still does not solve the config:

Positive reward

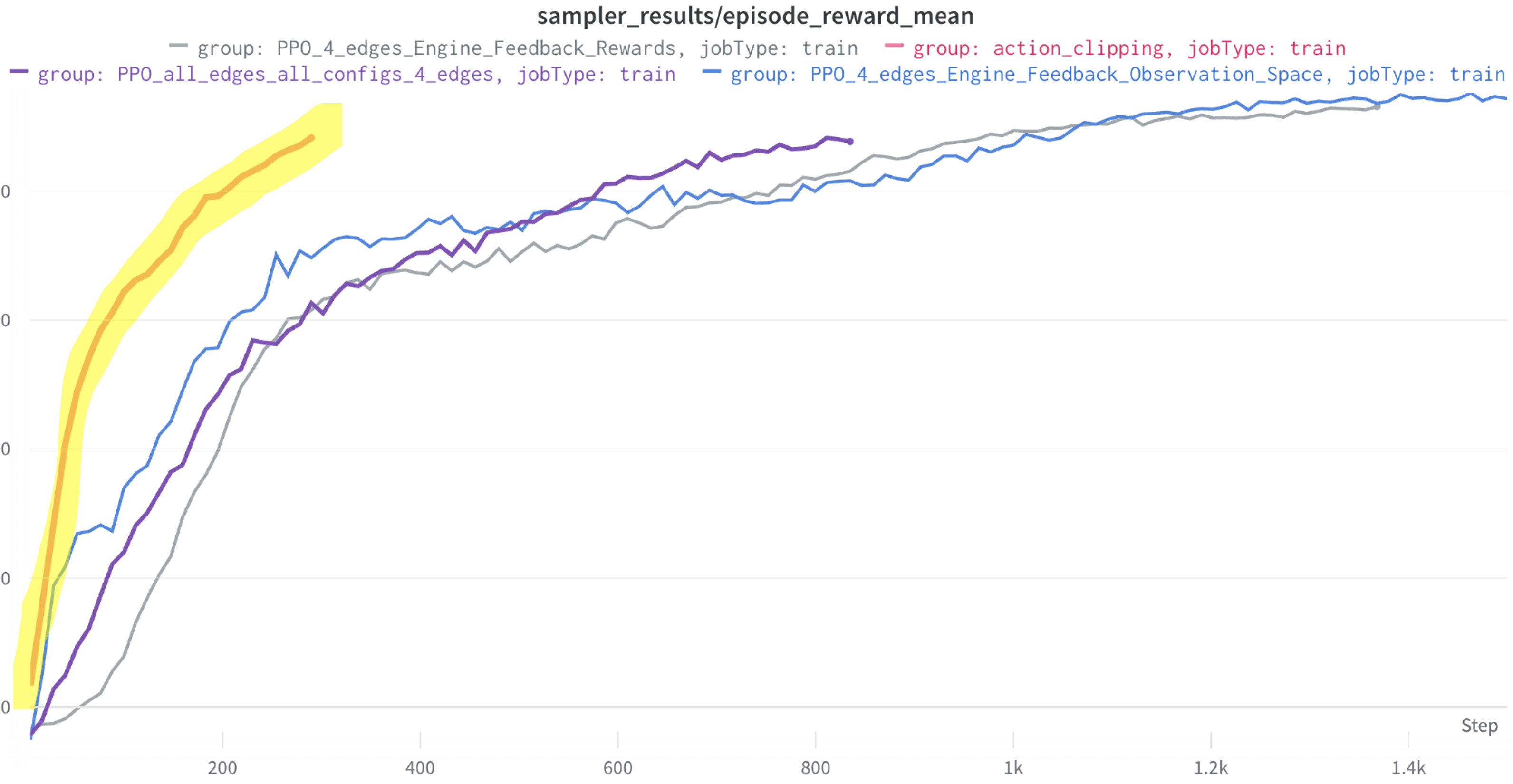
Degrees of Guidance

Feedback from Engine: Utilizing Rewards



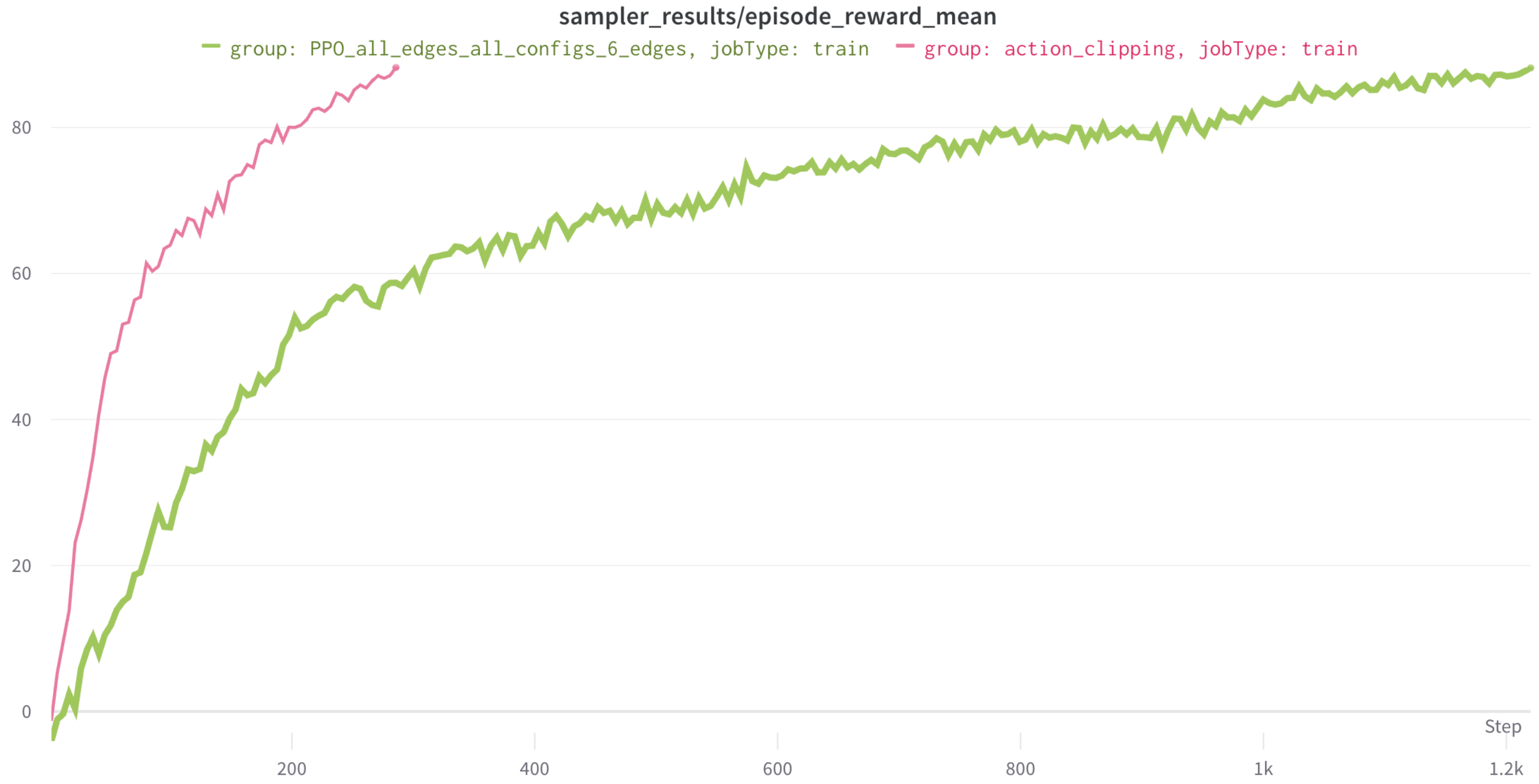
Degrees of Guidance

Feedback from Engine: Action Restriction



Degrees of Guidance

Feedback from Engine: Action Restriction



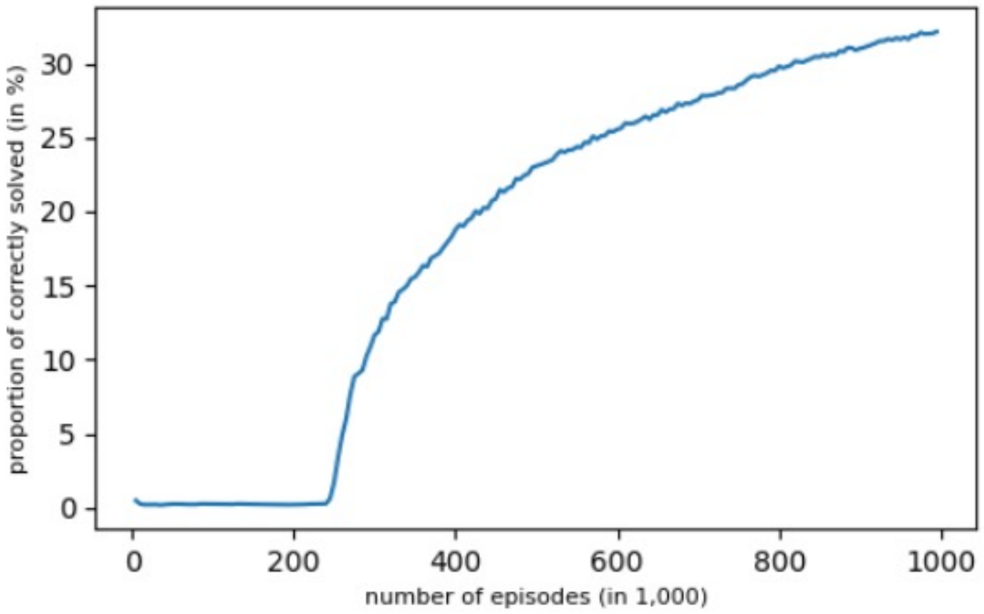
Project Review

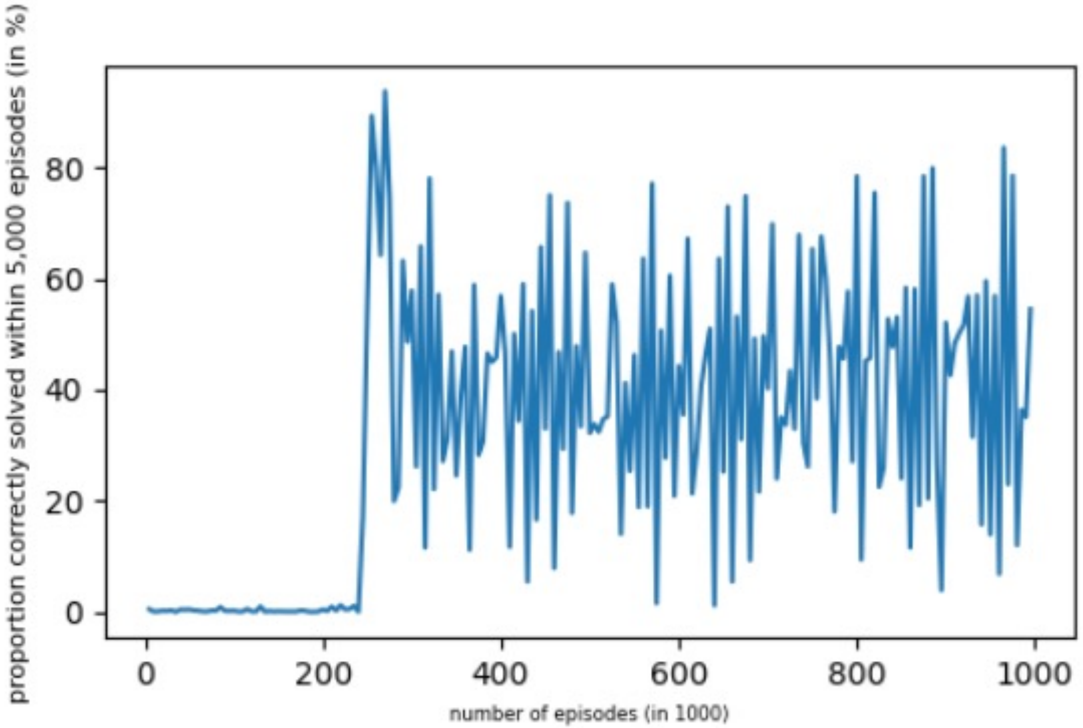
- Setup for RL using our BugPlus implementation and custom environment in Python
  - Own evaluation engine
  - Config/training data generator
  - Exploration of different performance factors

    → Agent can solve configs with 6 missing edges
- Editor to visually work with BugPlus
- Importance of Exploration

Project Review

Our Learning ~~Oscillation~~ Curve





Let’s talk

Questions, comments?

Possible Problems for 3 Bugs

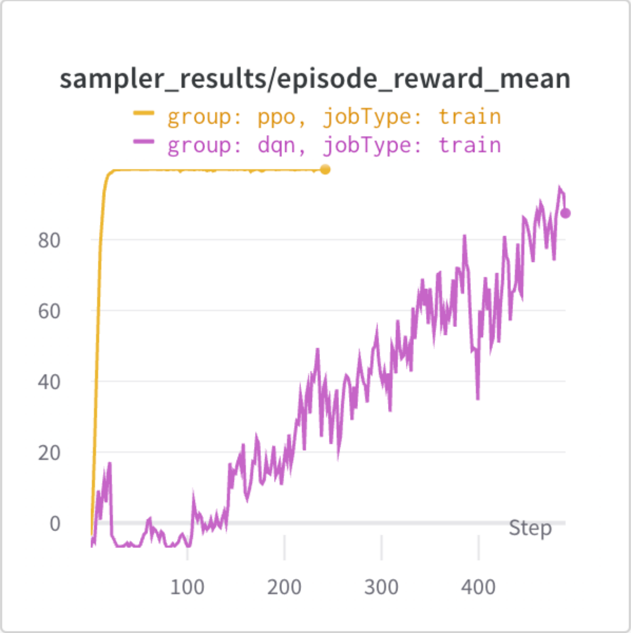
52 unique problems are possible with 3 bugs:

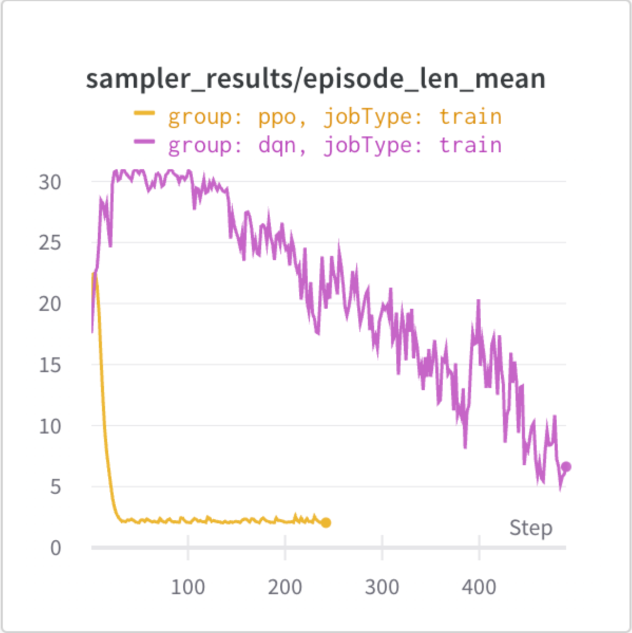
-4	2x	2y	4x	8x	x+y
-3	2x-1	2y-1	4x+2y	8y	x+y-1
-2	2x-2	2y-2	4x+4y	x-1	x+y+1
-1	2x+1	2y+1	4x+y	x-2	y-1
0	2x+2	2y+2	4y	x+1	y-2
1	2x+2y	3x	5x	x+2	y+1
2	2x+3y	3x+2y	5y	x+2y	y+2
3	2x+4y	3x+y	6x	x+3y	
4	2x+y	3y	6y	x+4y	

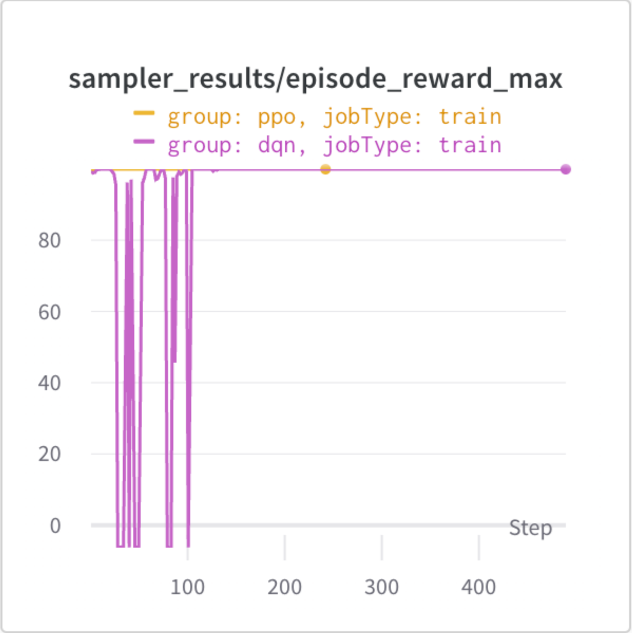
DQN vs PPO

Hypothesis: DQN quicker learner, but less stable in comparison to PPO

Result: PPO quicker and more stable

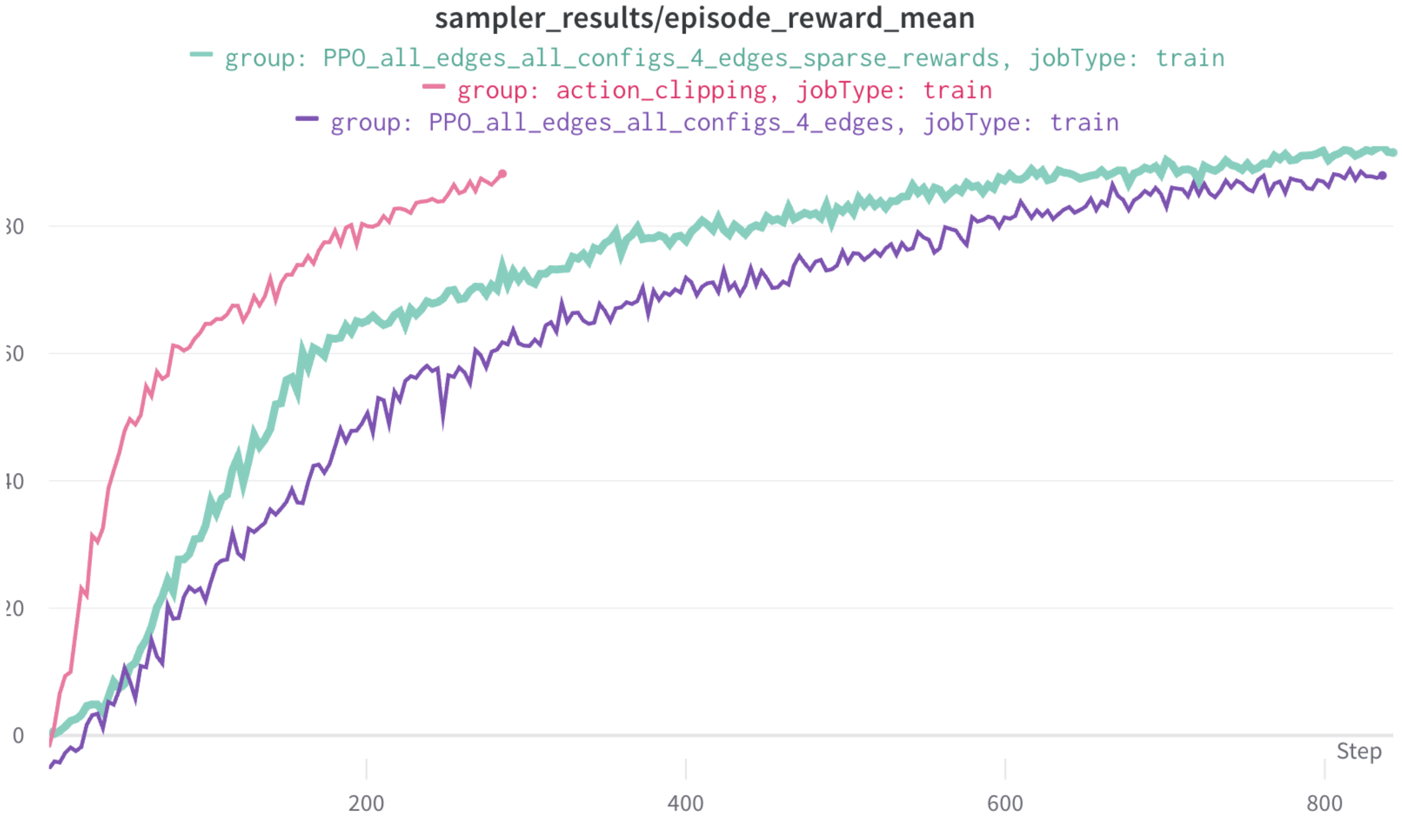






Degrees of Guidance

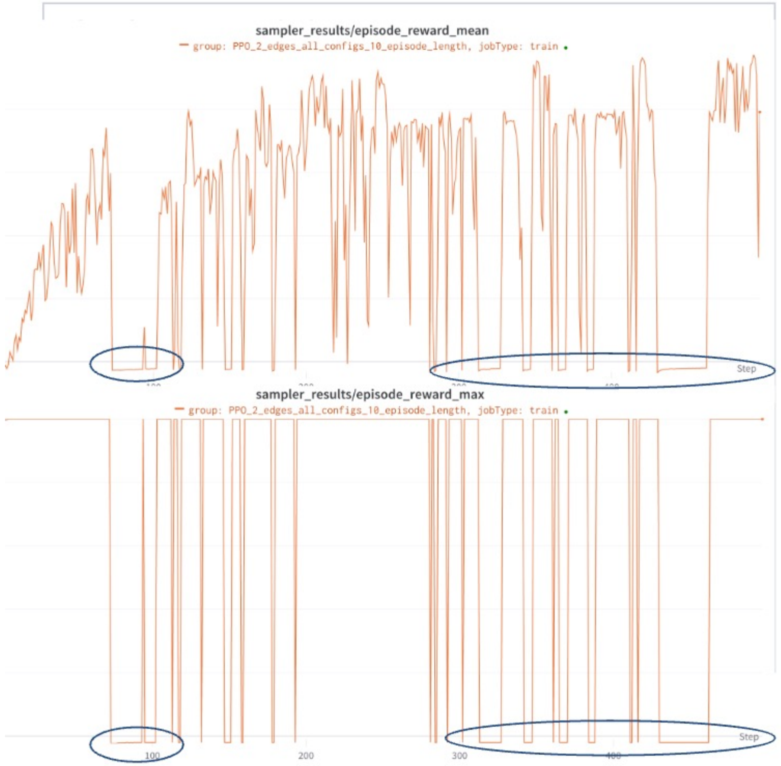
Reward System: Dependency on Difficulty



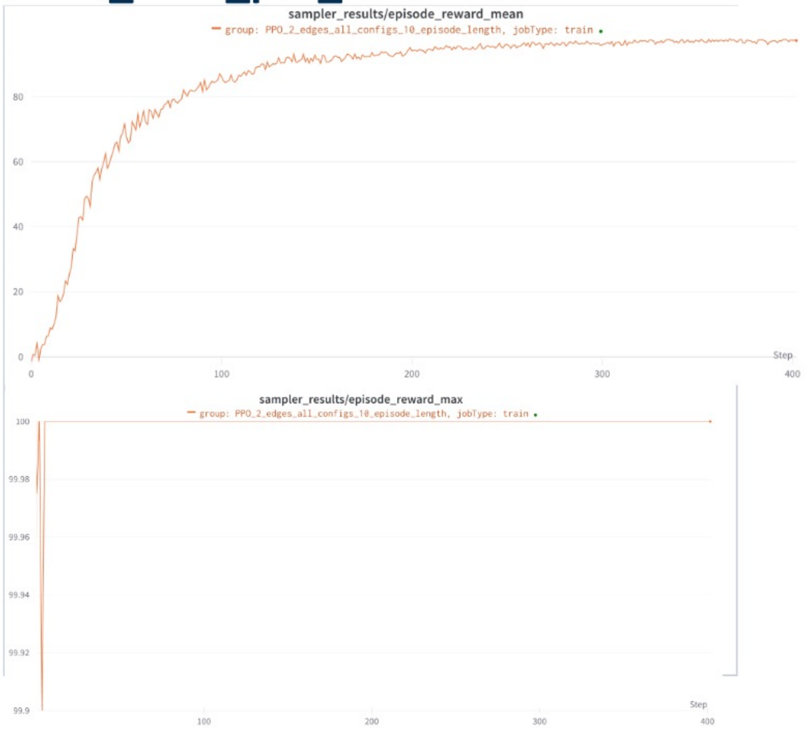
Valleys of Death

Influence of Parameter number of environments per worker

‘num_envs_per_worker’ = 1

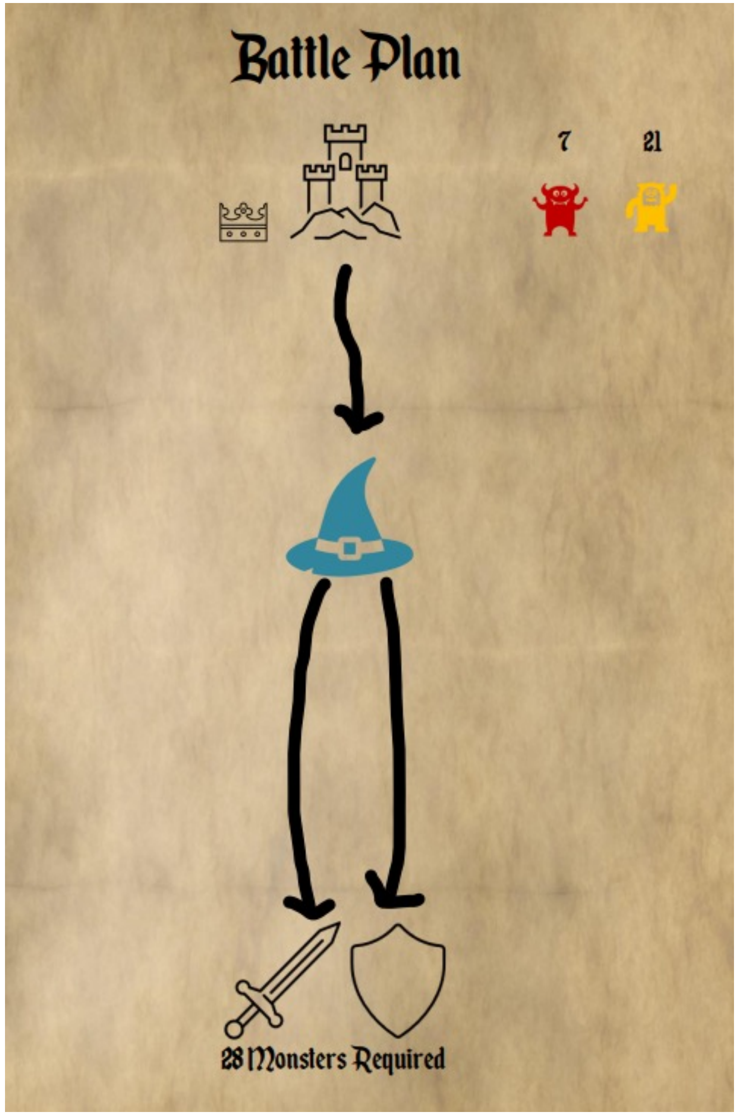


‘num_envs_per_worker’ = 10



16

Gamification Concept


Battle Plan
7
21
28 Monsters Required