DOOM

REINFORCEMENT LEARNING

ADAM KANIASTY, HUBERT KOWALSKI, NORBERT FRYDRYSIAK,
IGOR KOŁODZIEJ, KRZYSZTOF SAWICKI

AGENDA

01 CO TO DOOM

02 MODELE

03 ARCHITEKTURA SIECI

04 PREPROCESSING OBRAZU

05 FUNKCJA NAGRODY I METRYKI

06 PROCES UCZENIA

07 REZULTATY

DOOM

klasyczna gra komputerowa typu first-person shooter (FPS), która została stworzona przez id Software i po raz pierwszy wydana w 1993 roku. W naszym projekcie agent stara się pokonać poziom ”Death Corridor”

MODELE

WBUDOWANY

A2C

Gradient polityki jest aktualizowany na podstawie advantage:

$$\nabla_\theta \log \pi_\theta(a_t|s_t) A(s_t, a_t)$$

CUSTOMOWY

PPO

Aktualizujemy politykę na podstawie advantage, ale z ograniczeniem za pomocą klipowania.

$$\nabla_\theta \log \pi_\theta(a_t|s_t) \min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} A(s_t, a_t), \text{clip}(\epsilon, 1-\epsilon)\right)$$

ARCHITEKTURA SIECI

Zastosowaliśmy sieci typu CNN (Convolutional Neural Network) zawierającej warstwy konwulacyjne pozwalające na processing obrazu

TYP SIECI

Architektura sieci obejmowała warsty konwulacyjne, warstwy typu pool, oraz warstwy liniowe. Za funkcje aktywacji posłużyła Relu

ARCHITEKTURA

HIPERPARAMETRY

n_steps

8192

4096

clip_range

0.01

gamma

0.95

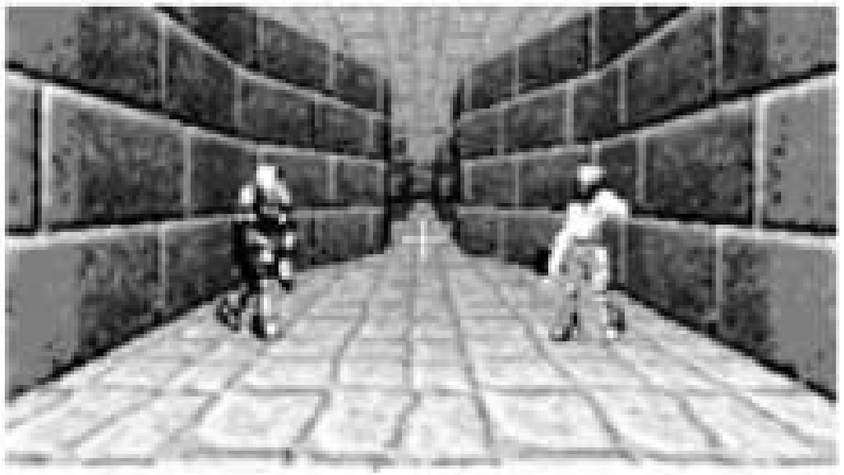
entropy_regularization

0.05

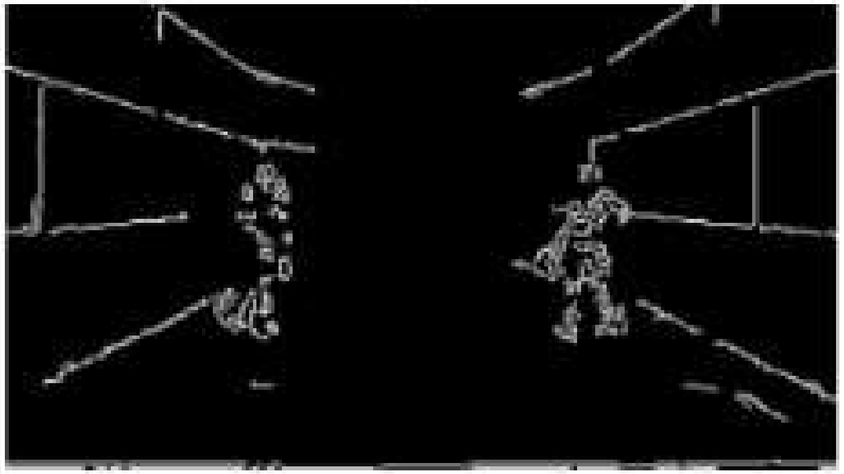
PREPROCESSING OBRAZU

Original Image

Processed Screen



Edges



METRYKI I FUNKCJA NAGRODY

METRYKI

Aby zweryfikować poprawność procesu uczenia zostały wprowadzone metryki mające na celu monitorować postępy. Wraz z obserwacją rozgrywki dawały one pełny wgląd na umiejętności modelu. Wprowadzone metryki:
Ammo, Distance, Killcount, Return,Steps,Timestep Reward

Zaimplementowana funkcja nagrody zwracała uwagę na kilka czynników. Adresowane zmiany stanu gry:

- Zdrowie
- Amunicja
- Kills
- X change

NAGRODA

FAZA PIERWSZA

- Scenariusz: Deadly Corridor

- Model PPO, trenowany przez 100K time stepów na poziomie trudności 2 i testowany również na 2. Jedynie nagroda customowa była brana pod uwagę.

FAZA PIERWSZA - REWARD

nagroda:

-11 za każdy stracony punkt życia

+210 za każde zabicie przeciwnika

-2 za każdy użyty nabój

+1 za każdą jednostkę w OSI OX w stronę kamizelki

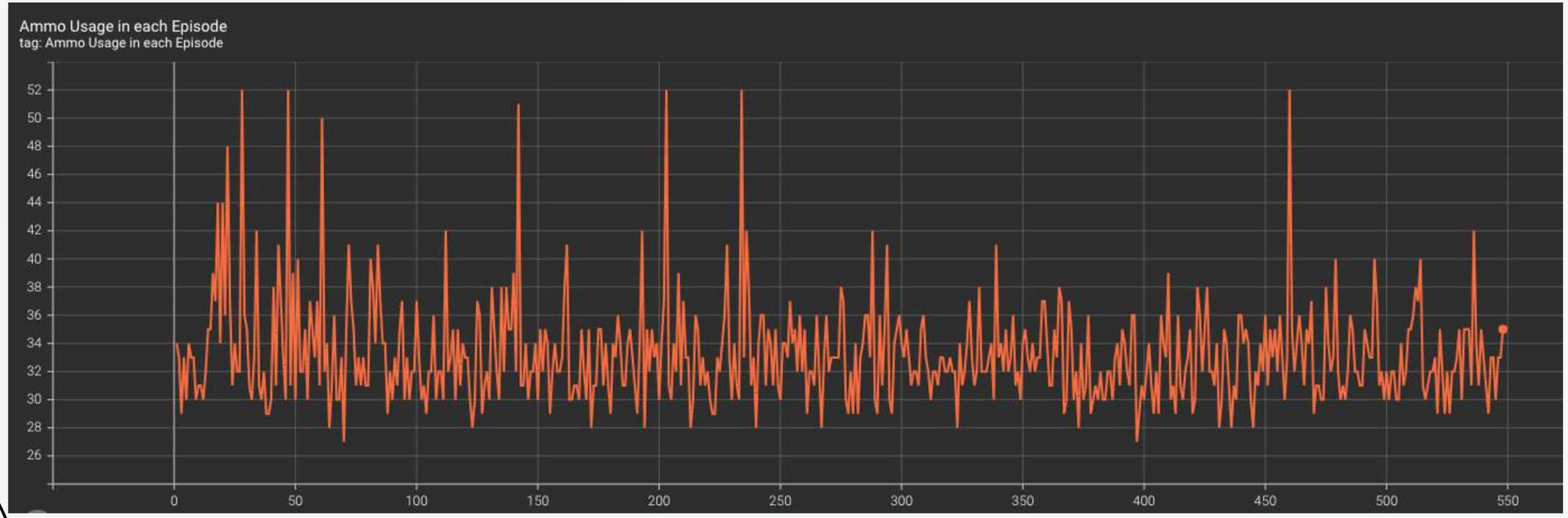
-100 za śmierć

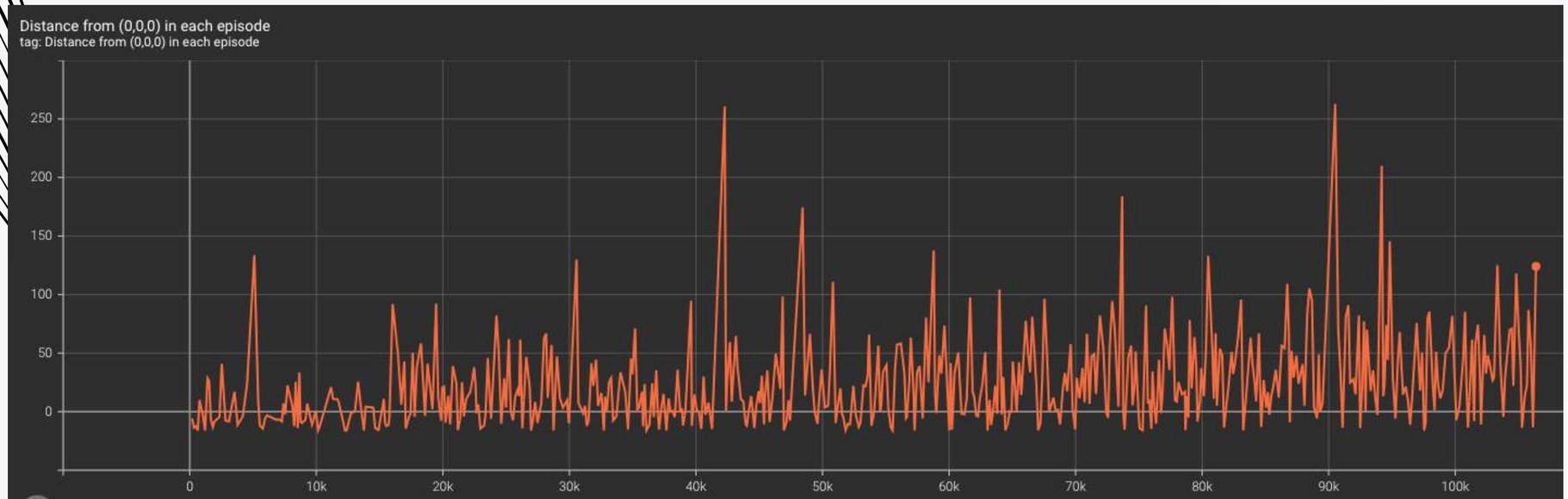
FAZA PIERWSZA

- Agent miał tendencje do jak najszybszego dostania się na koniec poziomu, nie podejmując walki z wrogami.

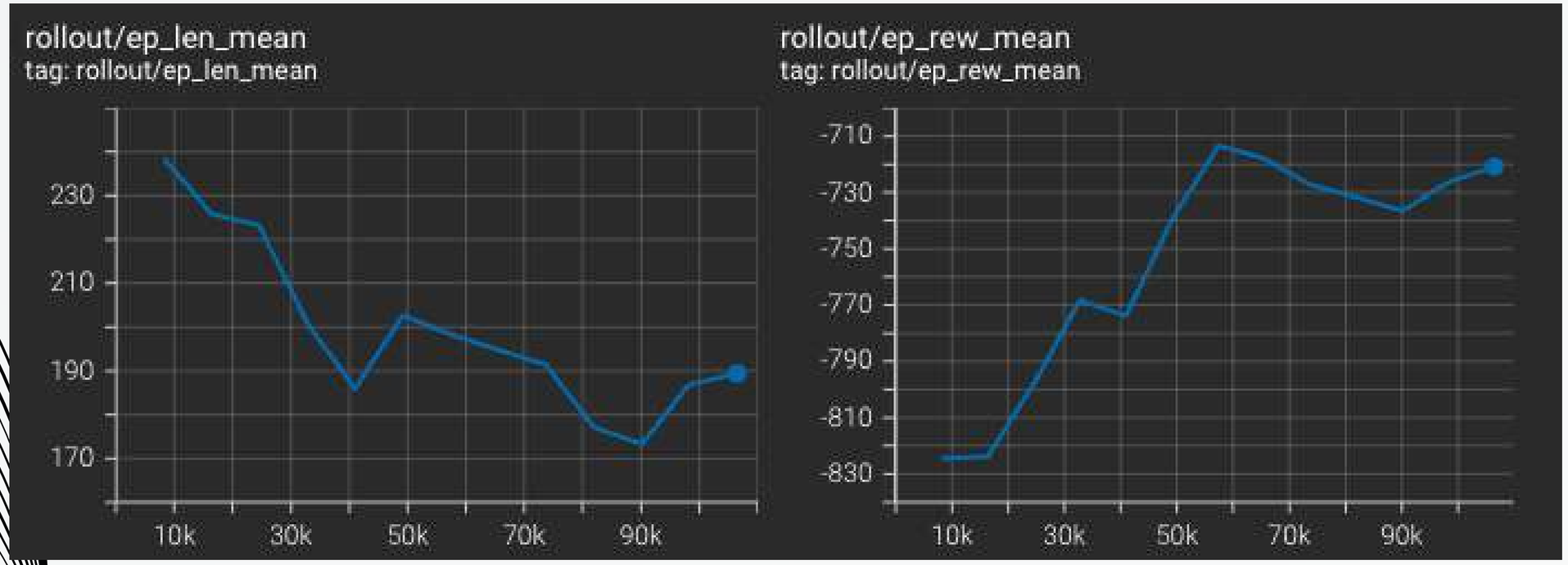
- **Liczba epizodów: 549**
- **Liczba śmierci: 545**
- **Liczba wygranych: 0**
- **Liczba zabójstw: 294**

Wyniki na metrykach - Faza 1

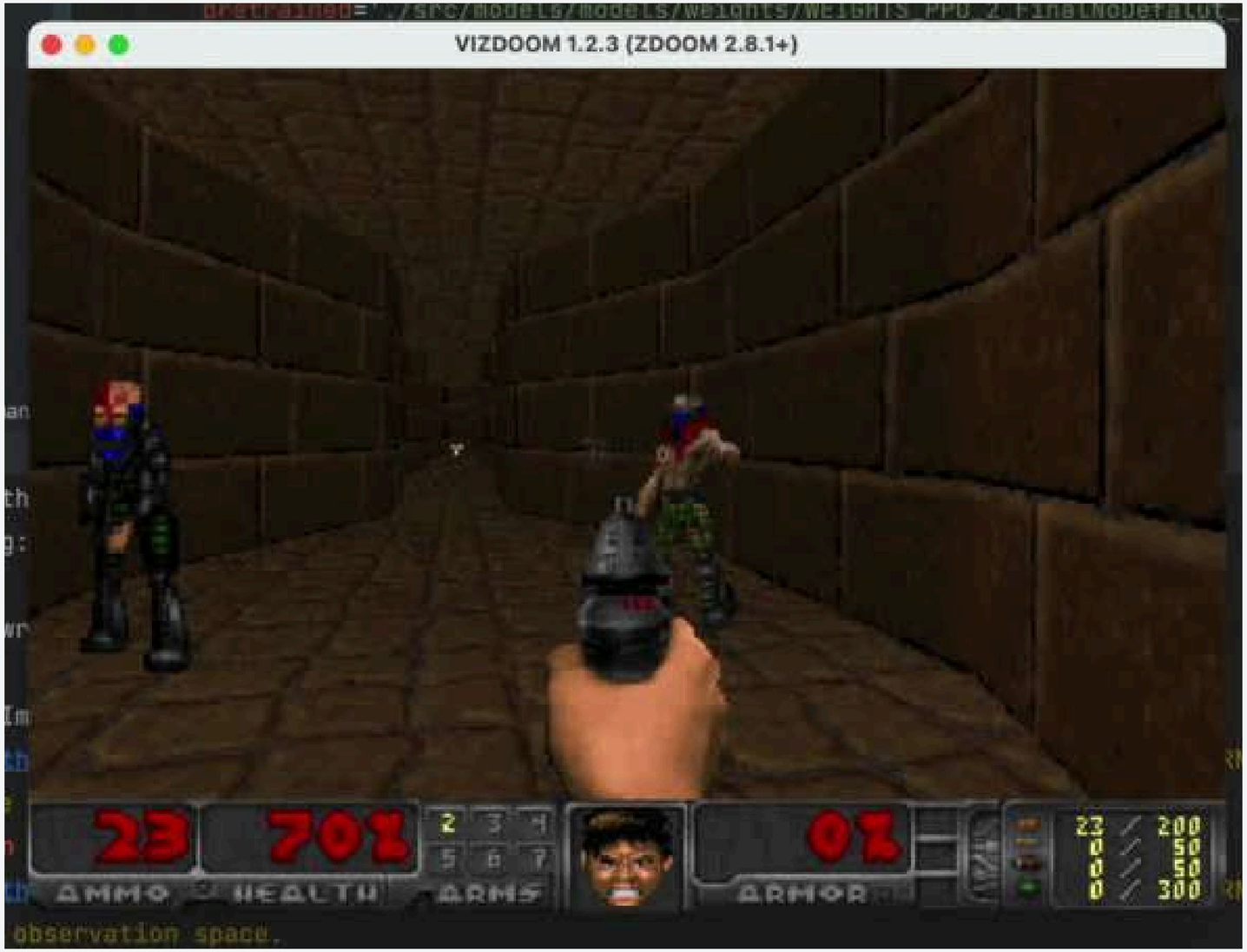




Wyniki na metrykach - Faza 1



Faza 1 - Polityka Stochastyczna



Faza 1- Polityka Deterministyczna

Test na nagrodzie podstawowej:
Średnia wartość nagrody z 10 epizodów wynosiła 553, wariancja 96.

FAZA DRUGA

- Model PPOtrenowany na 600k stepach na poziomie trudności 2. Prezentacja rozgrywki na poziomie trudności 1. Nagroda była sumą nagrody customowej i defaultowej.

- Kolejna iteracja funkcji nagrody wymusiła na agencie podejmowanie walki z wrogami, lecz nie poruszał się on do przodu.

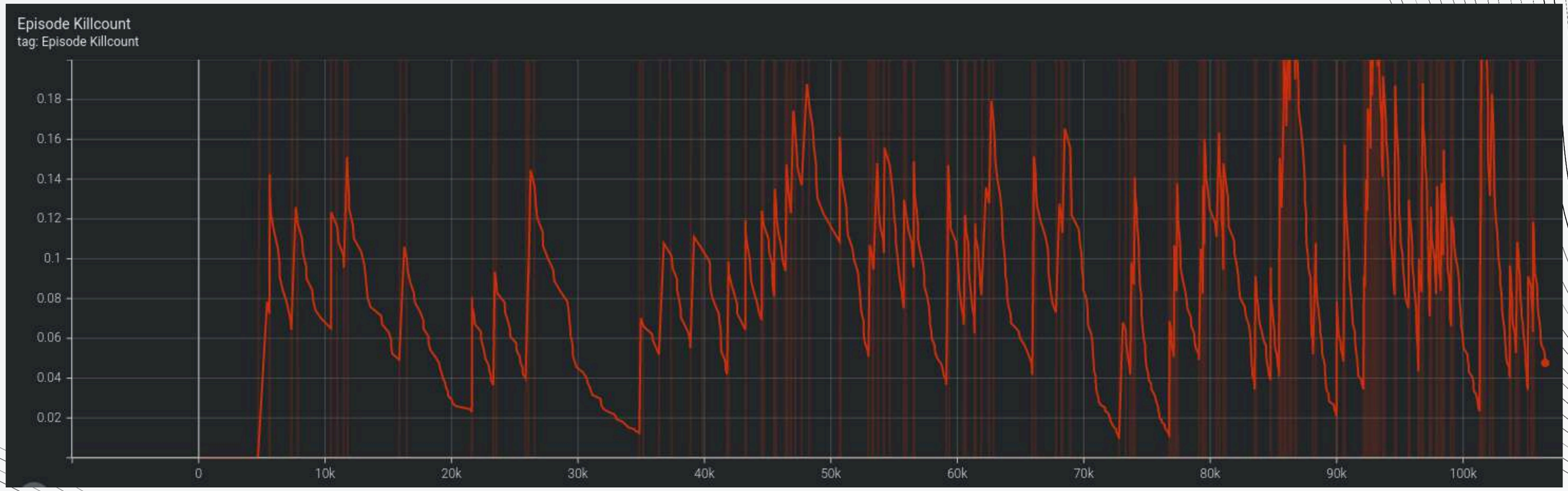
Prezentacja rozgrywki - polityka stochastyczna

FAZA TRZECIA

- Opisać uczenie które zadziało

- bla bla blabla bla blabla bla blabla bla blabla bla blabla bla blabla bla bla

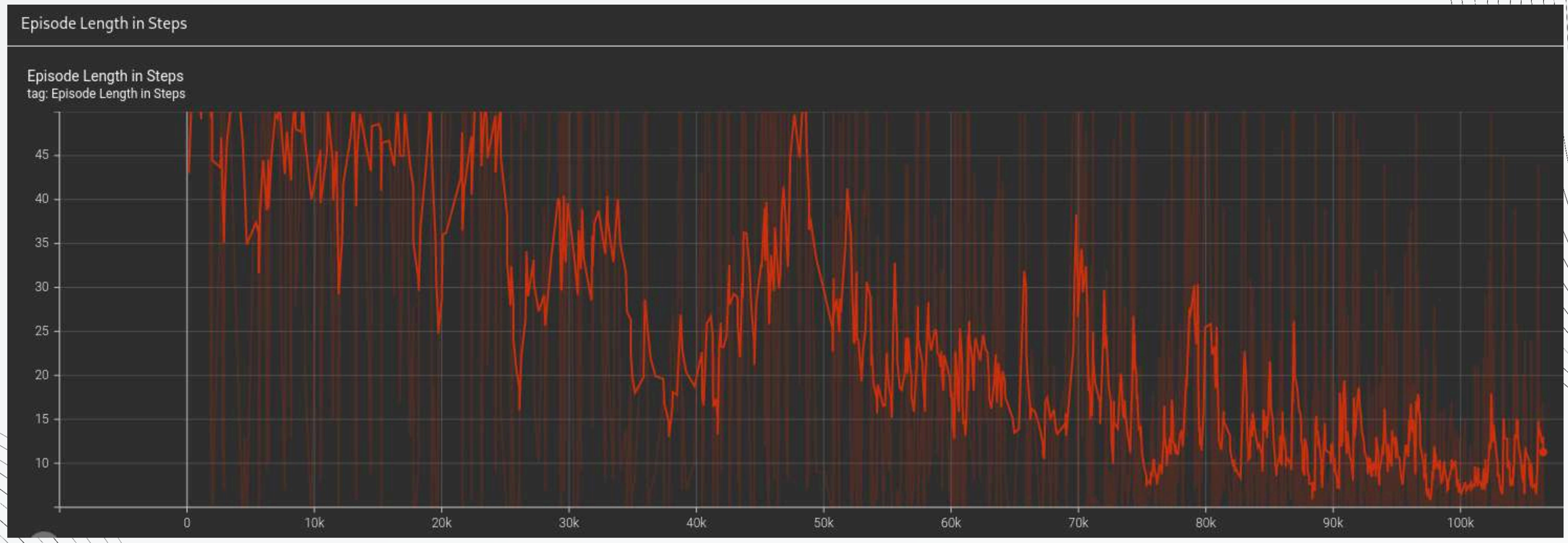
SCENARIUSZ 1 - BASIC



SCENARIUSZ 1

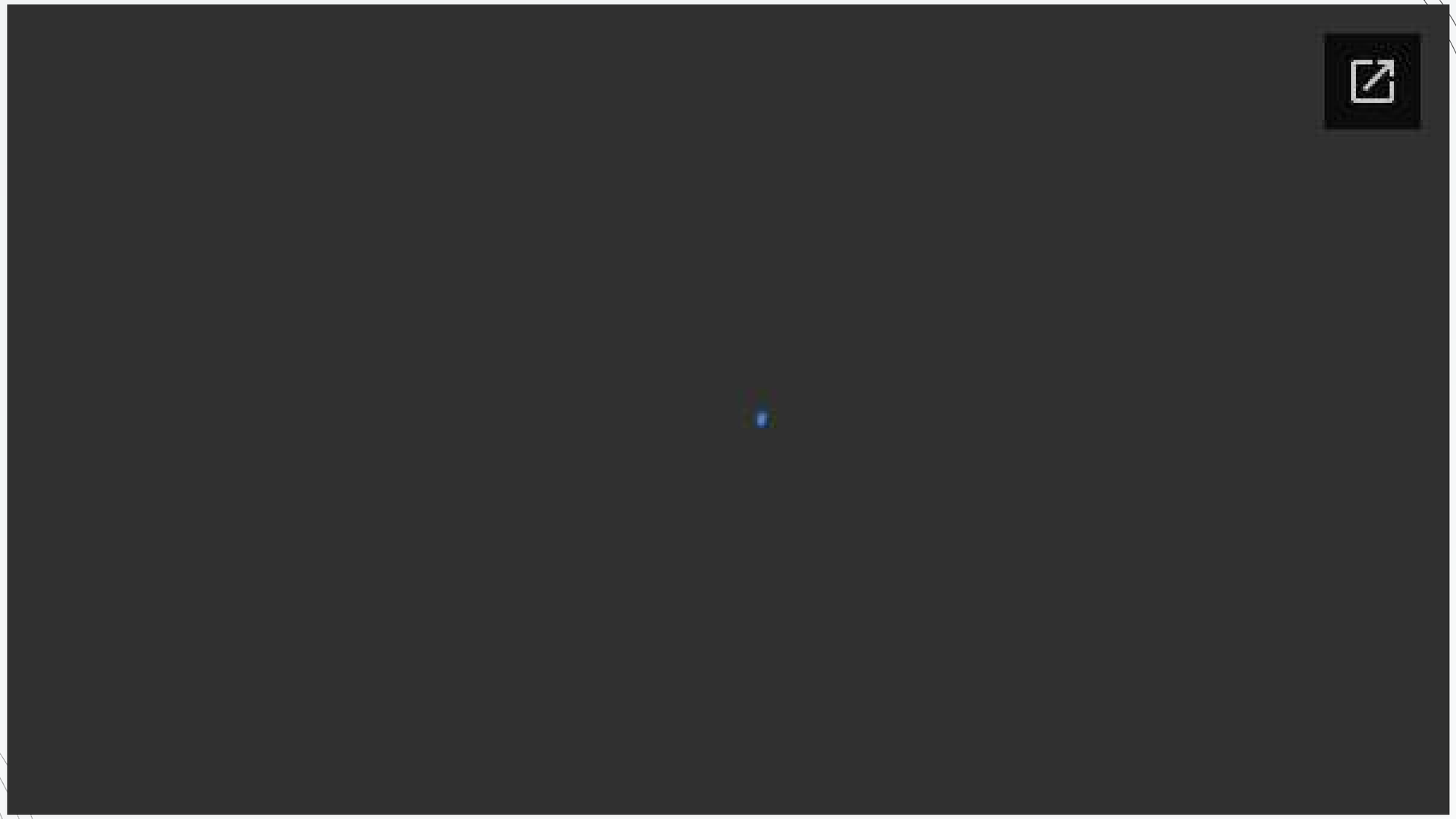
SCENARIUSZ 1

SCENARIUSZ 1



SCENARIUSZ 1

SCENARIUSZ 1

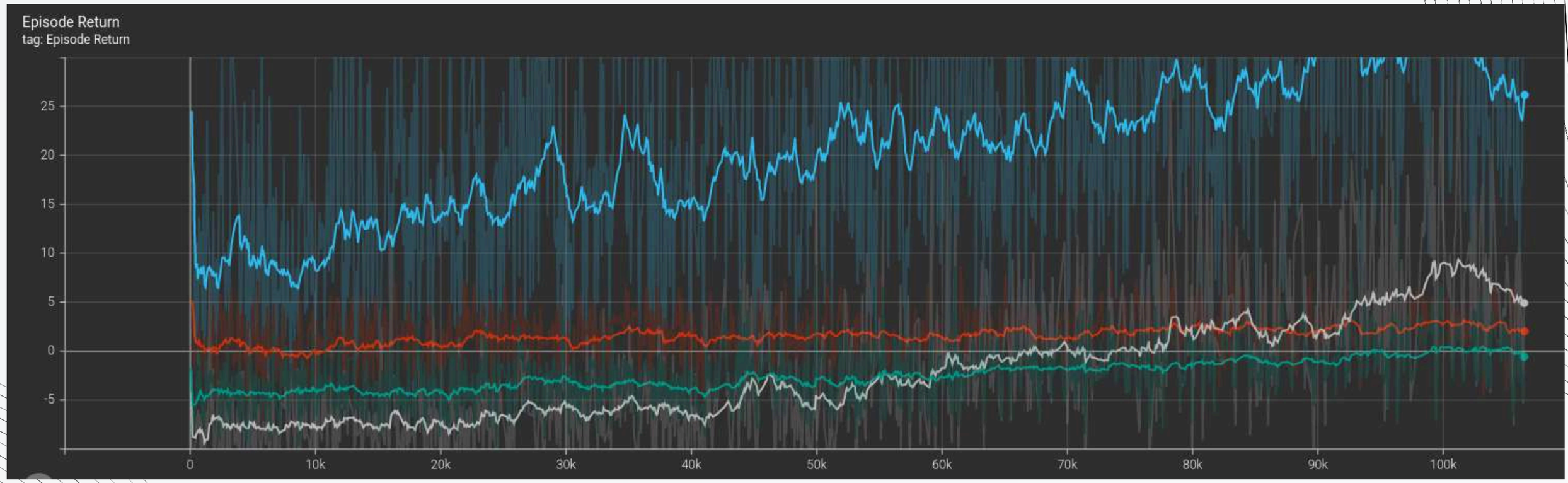


SCENARIUSZ 2 - DEFEND CENTER

SCENARIUSZ 2 - DEFEND CENTER

SCENARIUSZ 2 - DEFEND CENTER

SCENARIUSZ 2 - DEFEND CENTER



SCENARIUSZ 2 - DEFEND CENTER

SCENARIUSZ 2 - 100K

SCENARIUSZ 2 - 200K

DEADLY CORRIDOR

nagroda:

-1 za każdy stracony punkt życia

+200 za każde zabicie przeciwnika

-5 za każdy użyty nabój

+1 za każdą jednostkę w OSI OX w stronę kamizelki

-100 za śmierć

A2C 200K

deterministic=False

deterministic=True

poziom trudności 3

A2C 500K

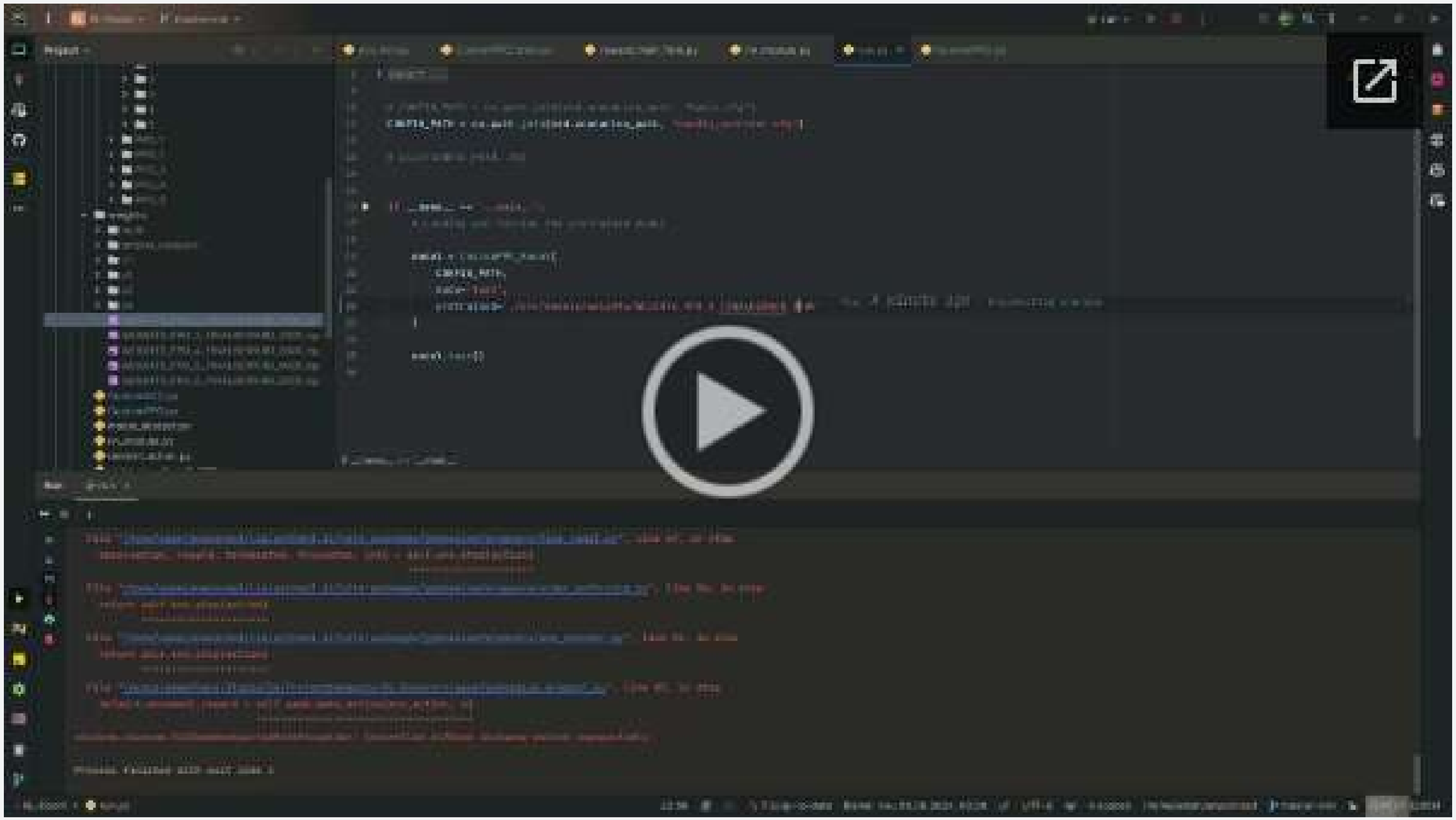
deterministic=False

deterministic=True

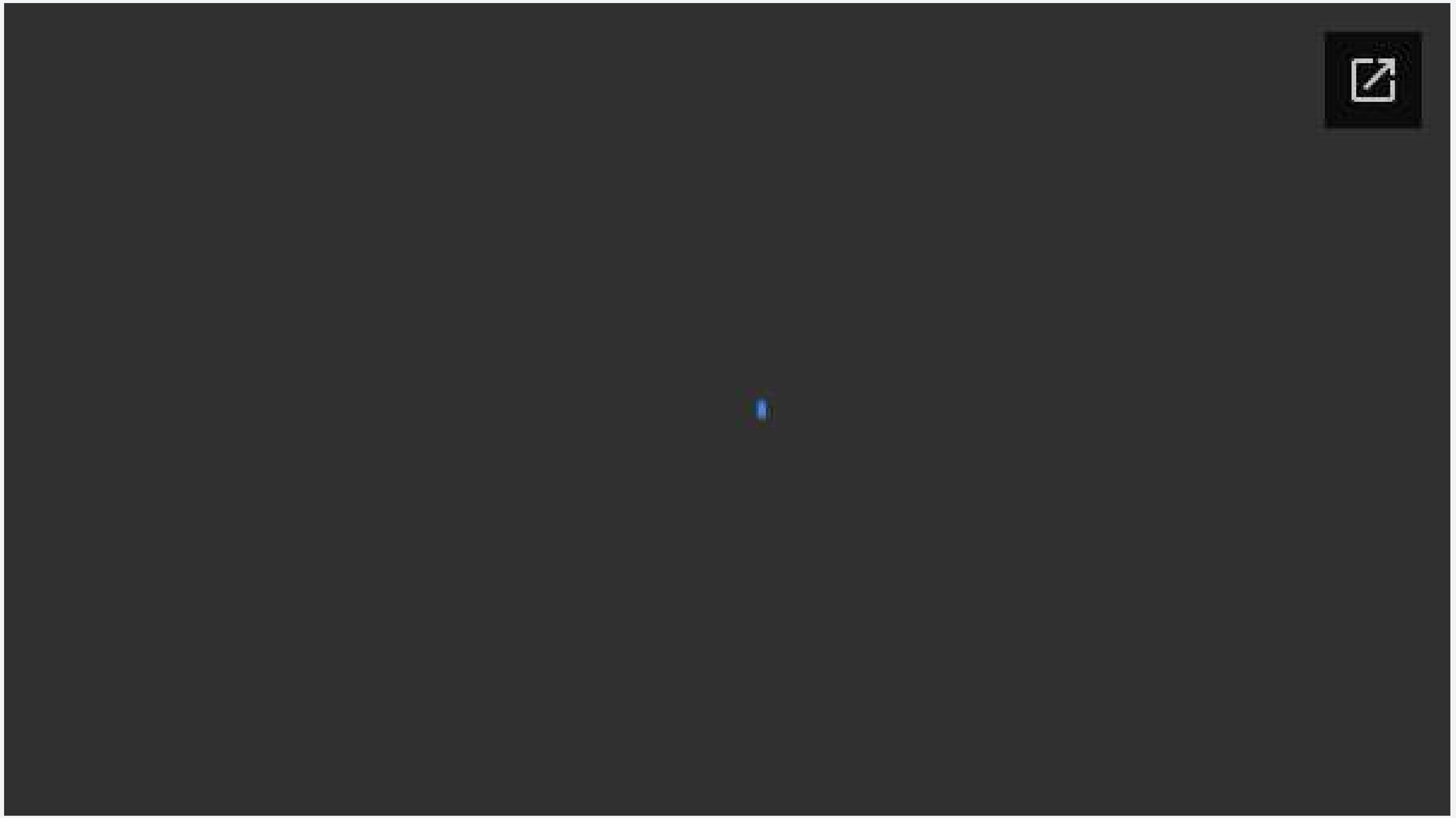
poziom trudności 3

PPO LEVEL 3 100K

PPO LEVEL 3 200K

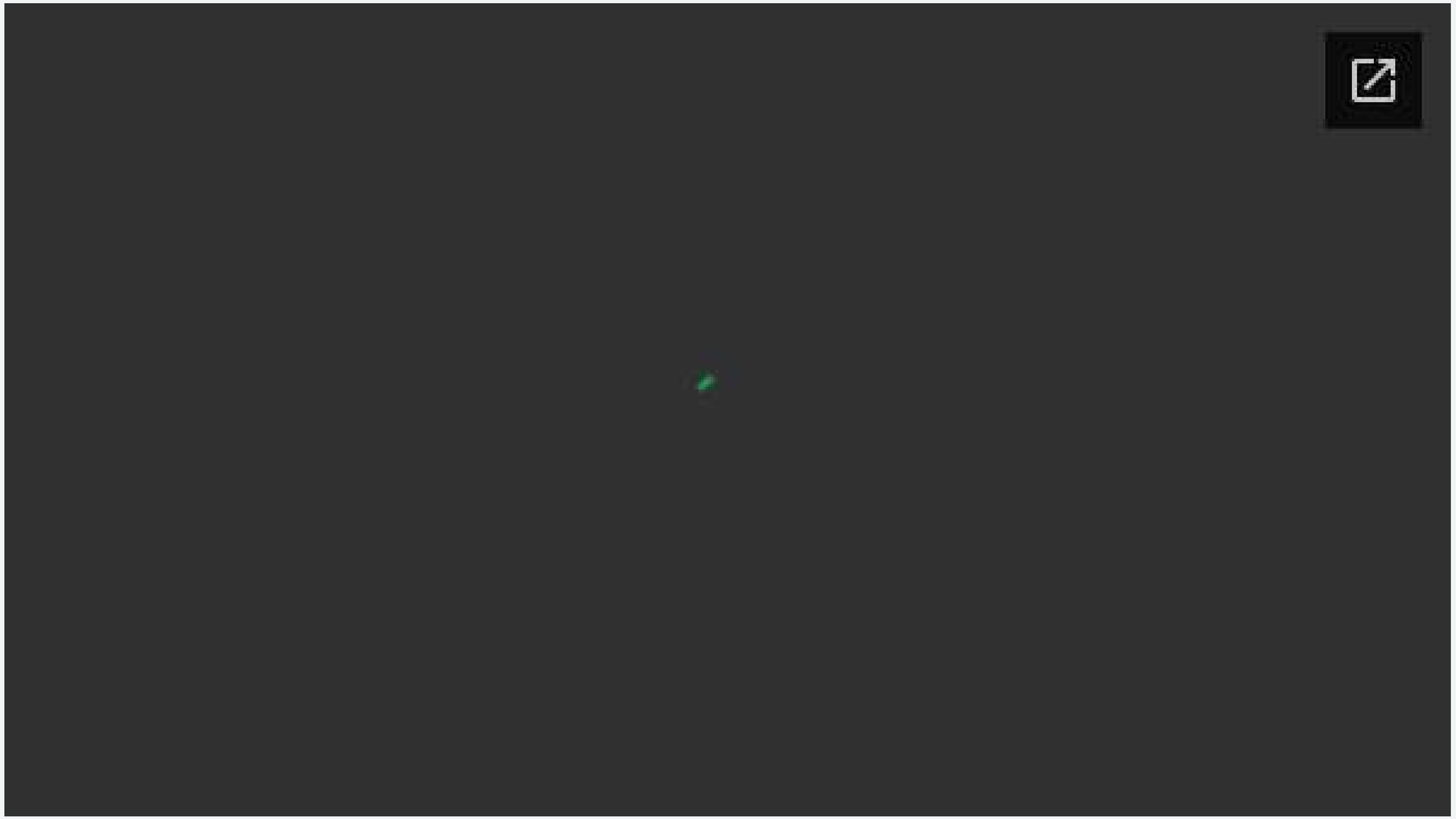


PPO LEVEL 4 300K

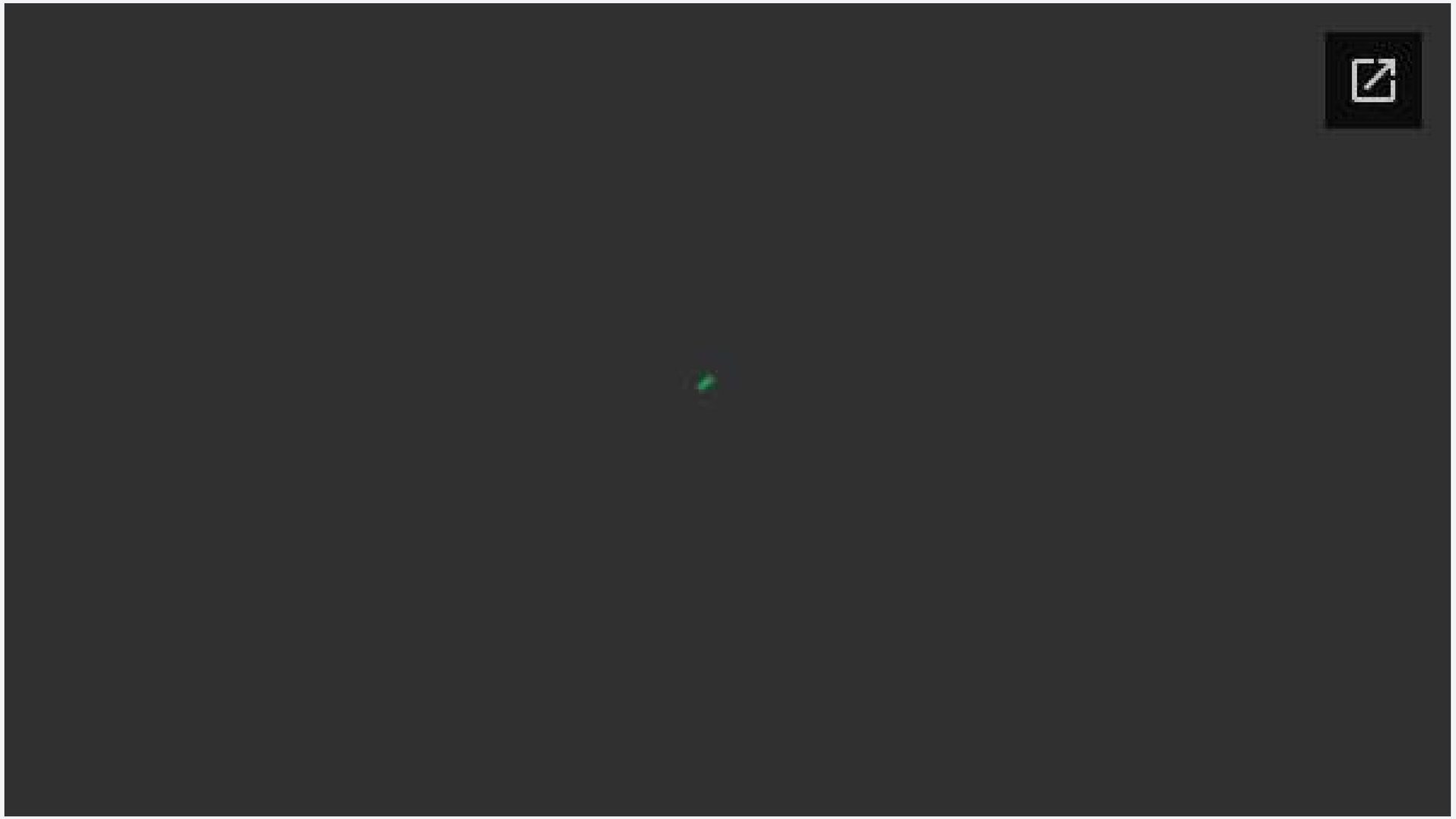


PPO LEVEL 5 400K

PPO LEVEL 5 500K

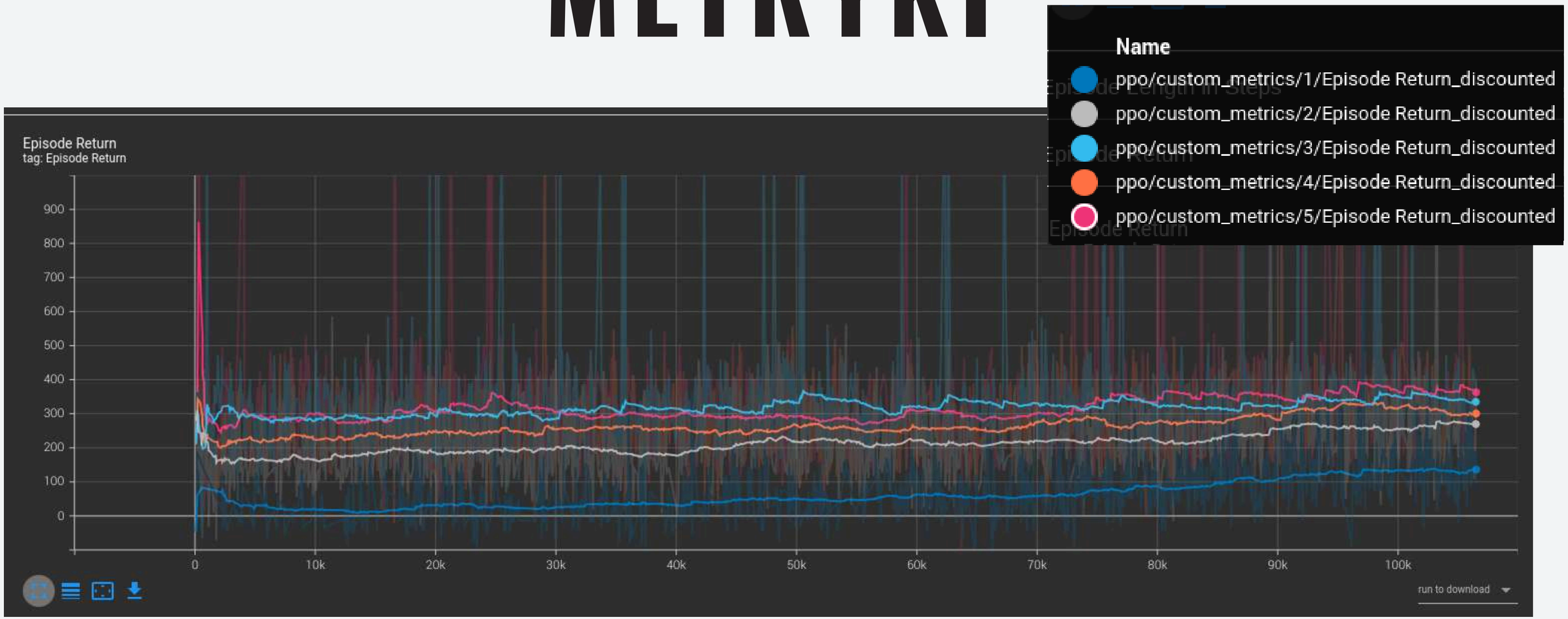


PPO LEVEL 5 500K

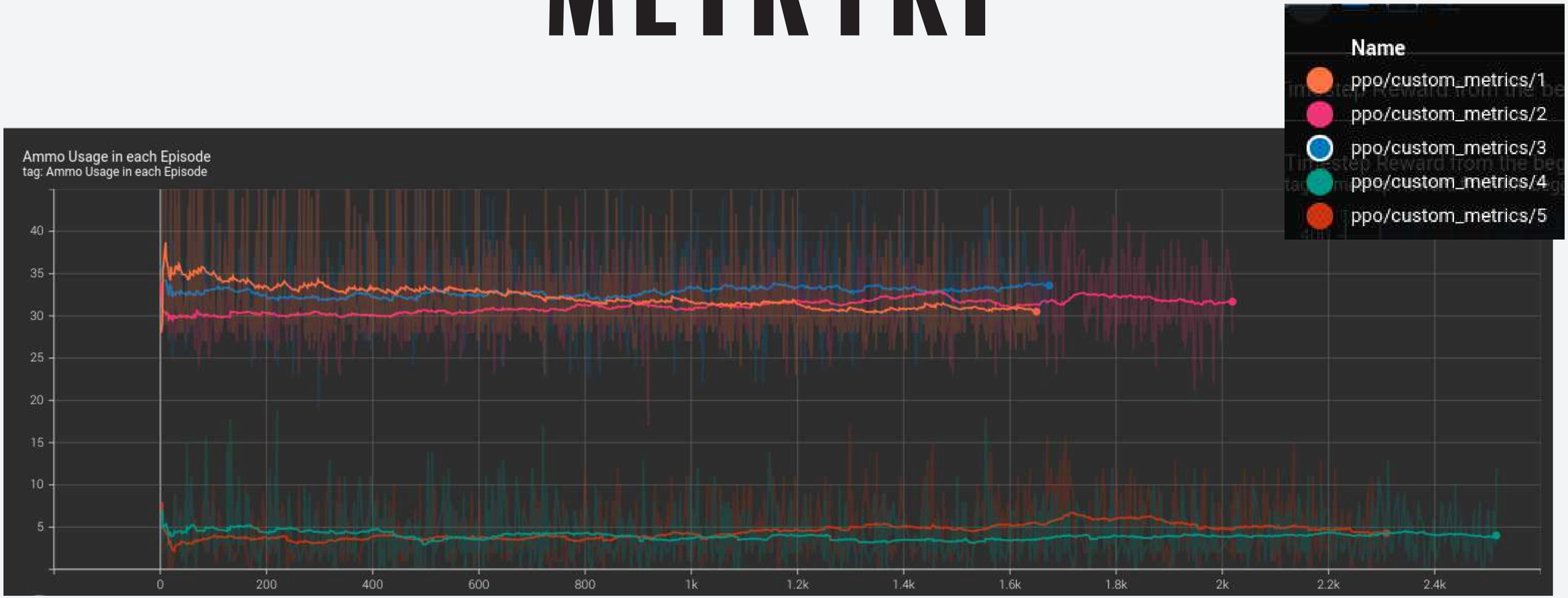


PPO LEVEL 5 800K

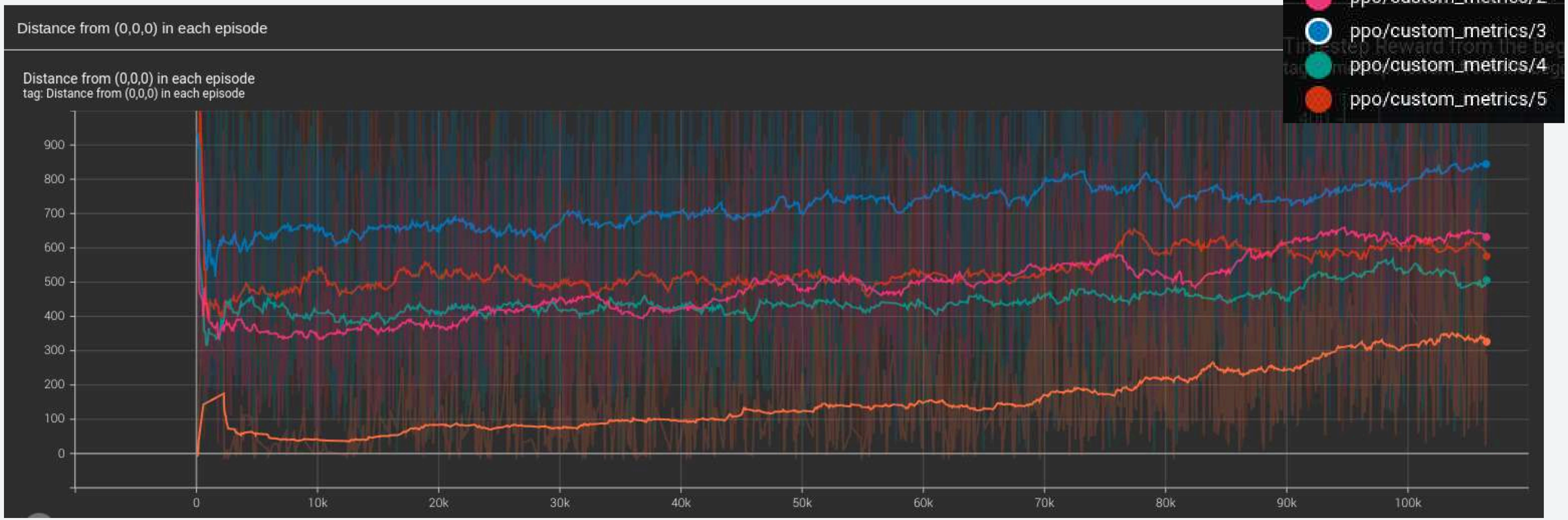
METRYKI


Name
ppo/custom_metrics/1/Episode Return_discounted
ppo/custom_metrics/2/Episode Return_discounted
ppo/custom_metrics/3/Episode Return_discounted
ppo/custom_metrics/4/Episode Return_discounted
ppo/custom_metrics/5/Episode Return_discounted
Episode Return
tag: Episode Return
900
800
700
600
500
400
300
200
100
0
0
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
run to download

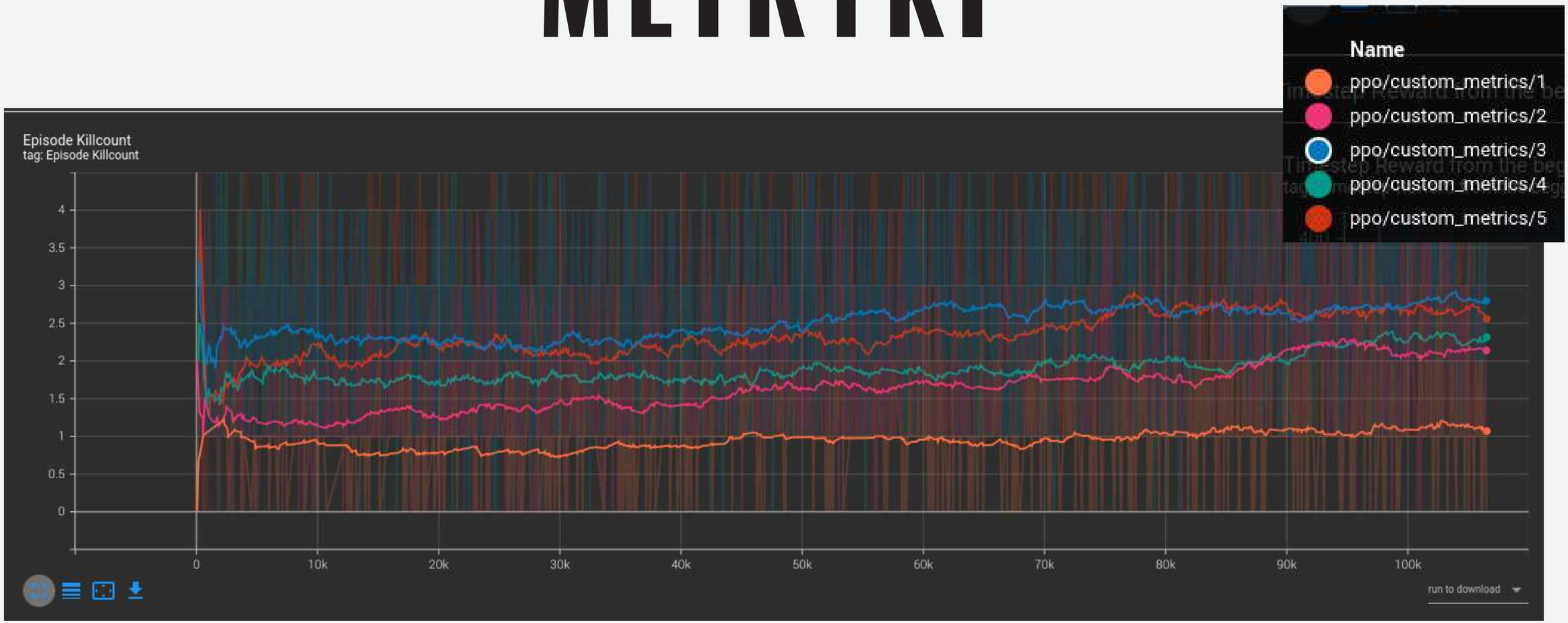
METRYKI


Name
ppo/custom_metrics/1
ppo/custom_metrics/2
ppo/custom_metrics/3
ppo/custom_metrics/4
ppo/custom_metrics/5
Ammo Usage in each Episode
tag: Ammo Usage in each Episode
40
35
30
25
20
15
10
5
0
200
400
600
800
1k
1.2k
1.4k
1.6k
1.8k
2k
2.2k
2.4k

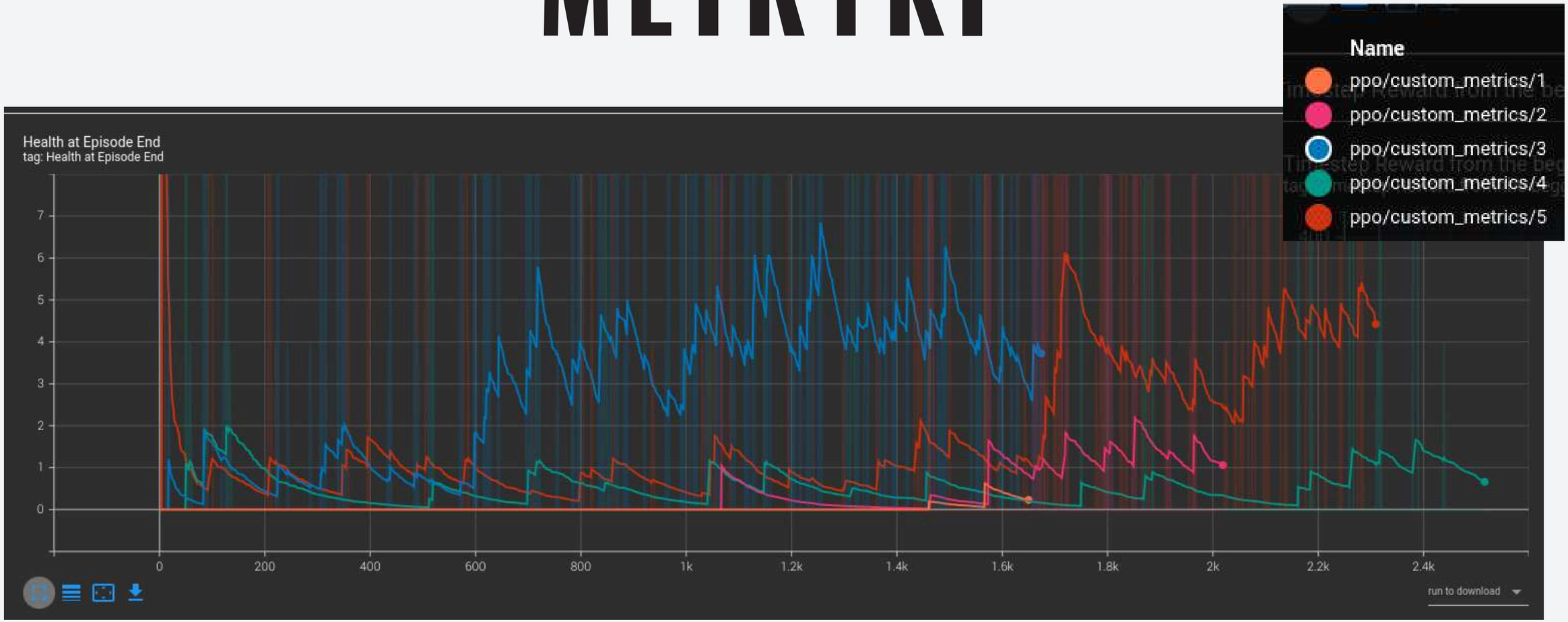
METRYKI


Distance from (0,0,0) in each episode
Distance from (0,0,0) in each episode
tag: Distance from (0,0,0) in each episode
900
800
700
600
500
400
300
200
100
0
0
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
Name
ppo/custom_metrics/1
ppo/custom_metrics/2
ppo/custom_metrics/3
ppo/custom_metrics/4
ppo/custom_metrics/5

METRYKI


Episode Killcount
tag: Episode Killcount
4
3.5
3
2.5
2
1.5
1
0.5
0
0
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
run to download
Name
ppo/custom_metrics/1
ppo/custom_metrics/2
ppo/custom_metrics/3
ppo/custom_metrics/4
ppo/custom_metrics/5

METRYKI


Health at Episode End
tag: Health at Episode End
7
6
5
4
3
2
1
0
0
200
400
600
800
1k
1.2k
1.4k
1.6k
1.8k
2k
2.2k
2.4k
run to download
Name
ppo/custom_metrics/1
ppo/custom_metrics/2
ppo/custom_metrics/3
ppo/custom_metrics/4
ppo/custom_metrics/5

DZIĘKUJEMY ZA UWAGĘ