

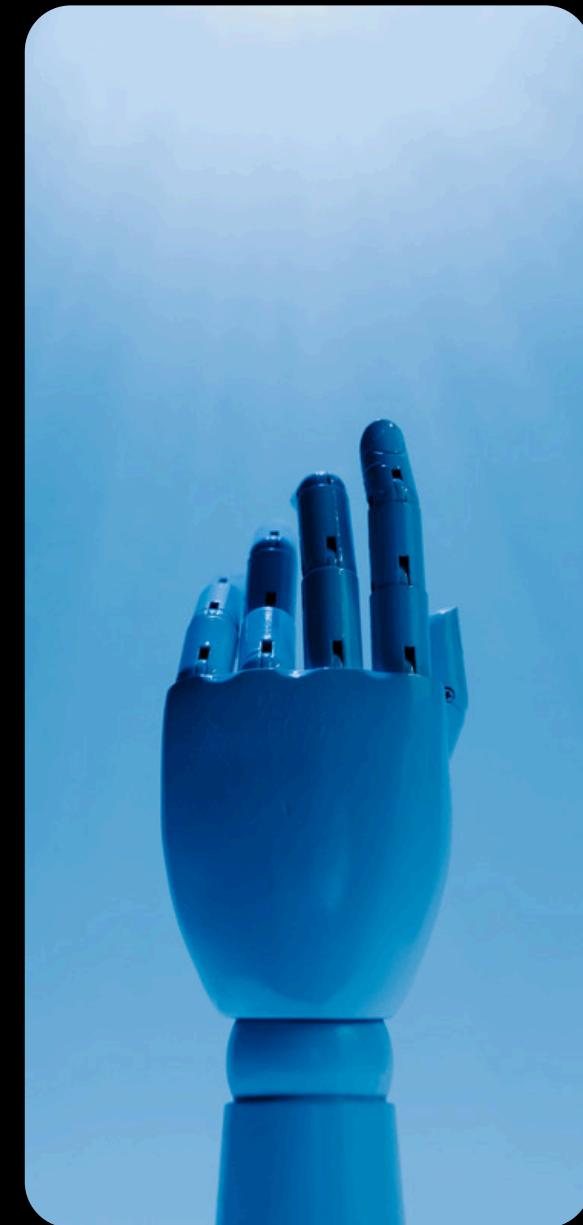
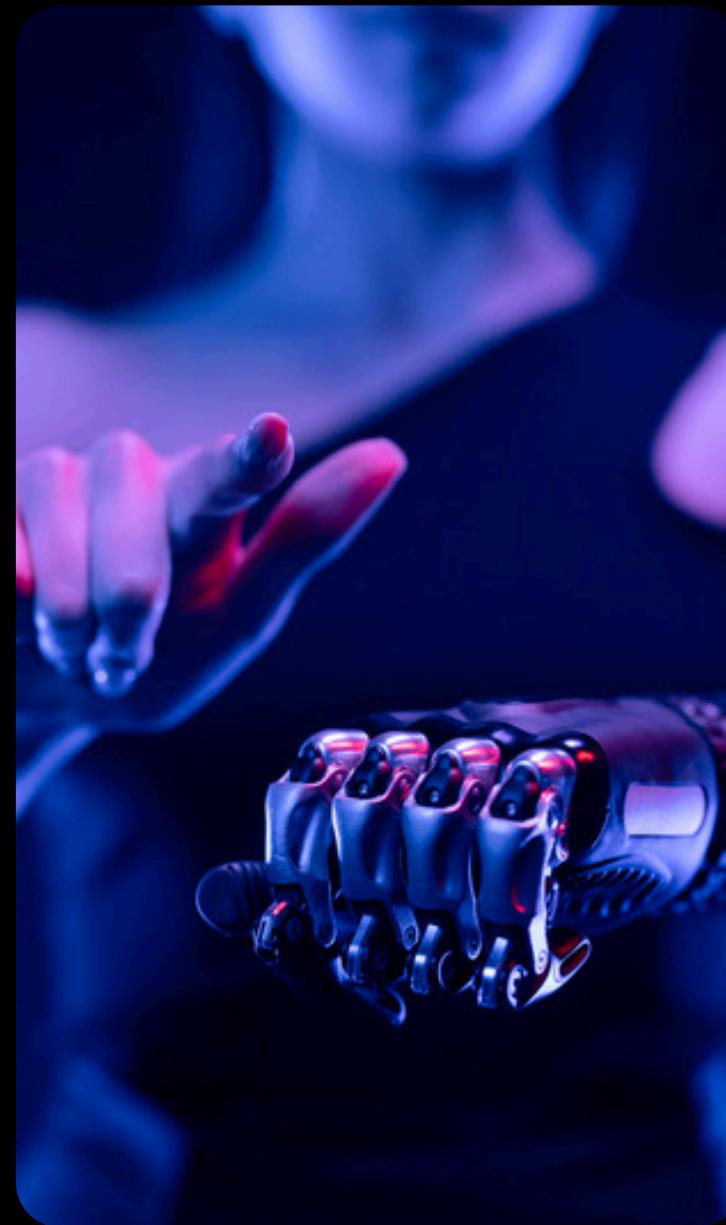
# SIMULATION

AMBIENTES MULTIAGENTES

Oswaldo Ilhuicatzi Mendizabal  
Mario Ignacio Frías Piña

A01781988  
A01782559





# Índice de contenidos

- Introducción
- Marco teórico
- Metodología
- Resultados



# Introducción

Reinforcement Learning (RL, por sus siglas en inglés)

Es una forma de enseñar a las computadoras a tomar decisiones aprendiendo de la experiencia.

Es como entrenar a una mascota: prueba cosas, aprende de los resultados y se vuelve mejor con el tiempo. RL se usa para resolver problemas complejos como jugar videojuegos, controlar robots o simular ecosistemas.



# Sistema multiagente

Se trata de un ambiente multiagente entrenado en el que las entidades deben cooperar para lograr su objetivo sin interponerse en el camino de los demás.





# Ambiente: Navigation

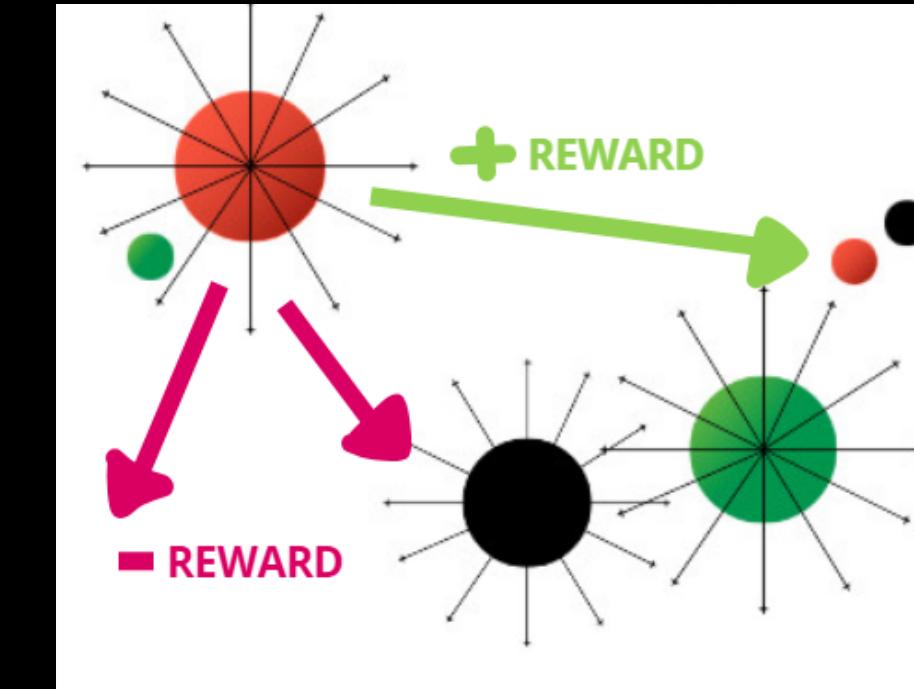
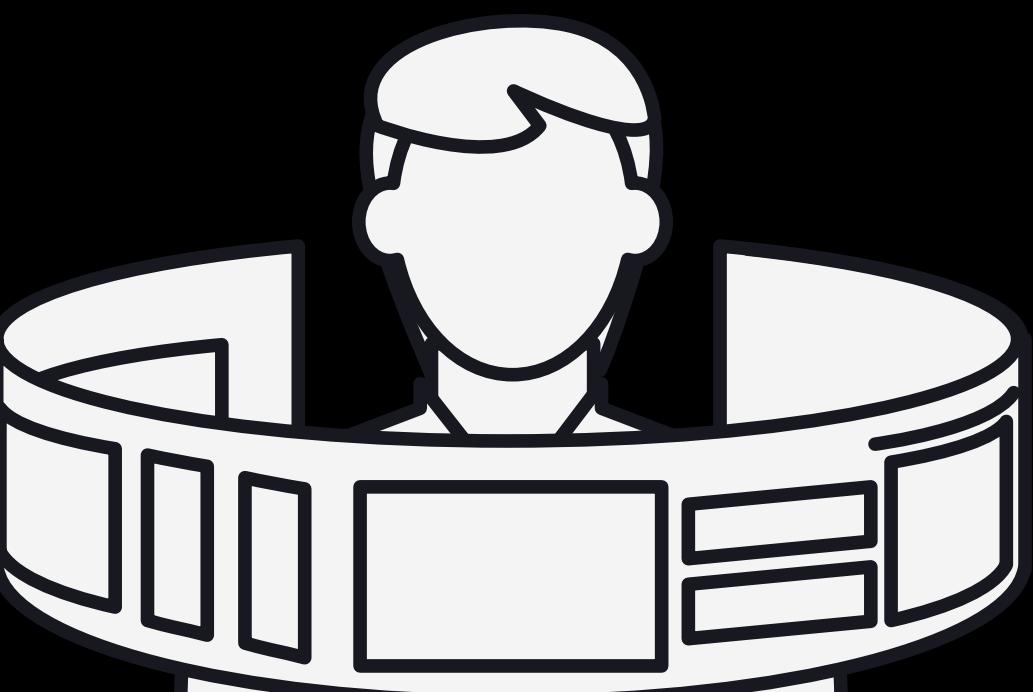
## ESPACIO DE OBSERVACIÓN

- Posición del agente [x, y] (bidimensional = 2)
- Velocidad del agente [vx, vy] (bidimensional = 2)
- Posiciones relativas de las metas: 2 \* n agentes si el argumento es verdadero, de lo contrario, solo 2, es decir, solo la posición relativa del agente con su meta.
- Lecturas del sonar solo si las colisiones están activas.

En total, las dimensiones del vector de observación son:

$$\text{dim(Observation)} = 4 + g + l$$

Donde  $g$  = posiciones relativas de las metas,  
 $l$  = lecturas del sonar.



## ESPACIO DE ACCIÓN

Espacio discreto con 9 acciones:

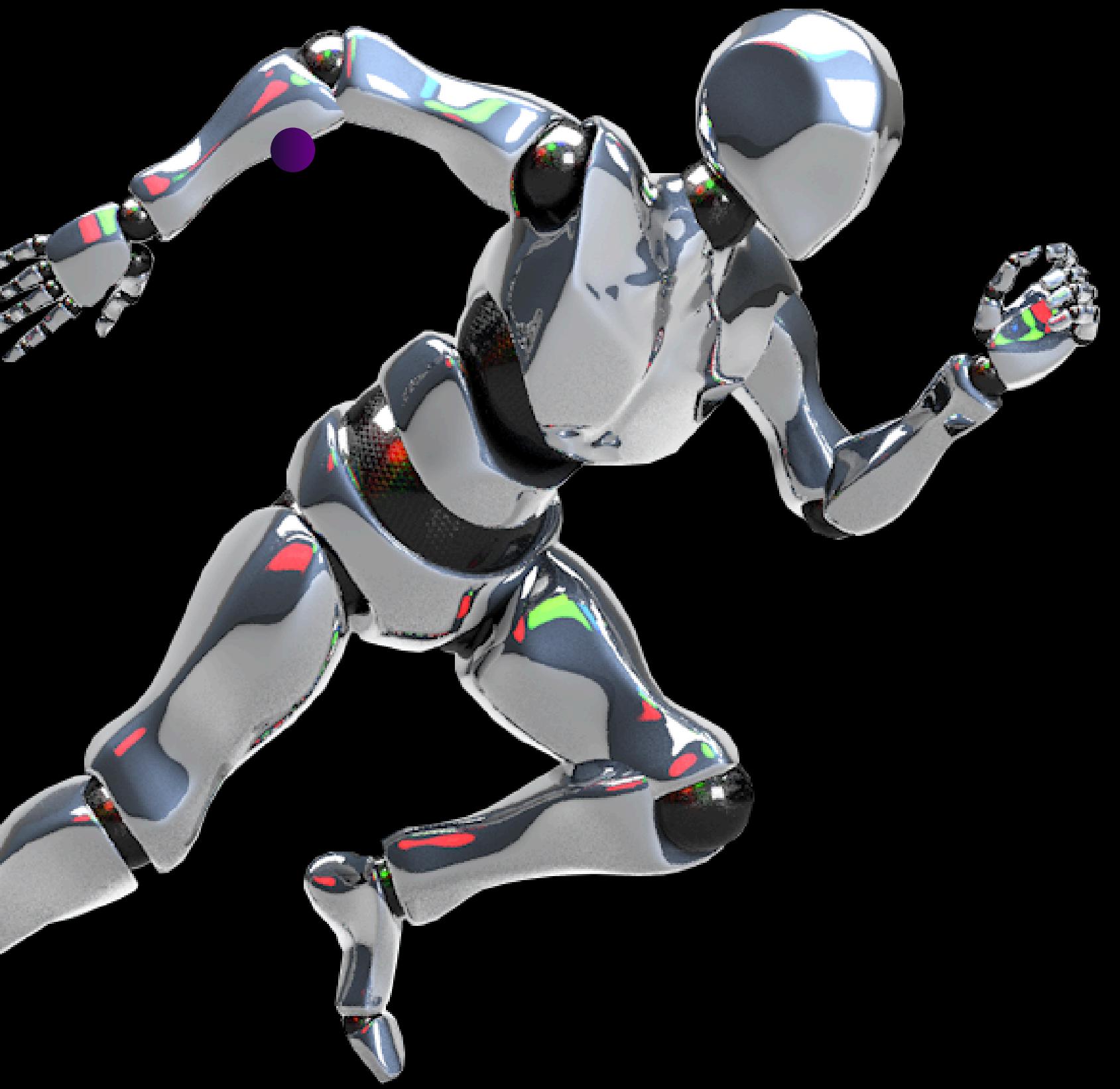
- Arriba
- Abajo
- Izquierda
- Derecha
- Diagonal arriba izquierda
- Diagonal arriba derecha
- Diagonal abajo izquierda
- Diagonal abajo derecha
- No tomar acción

# Marco teórico



## PPO: Proximal Policy Optimization

- Utiliza gradiente descendente para optimizar una política.
- La política aprende qué debe hacer en un estado específico. (Model-Free)
- Se restringe la actualización para evitar actualizar hacia una dirección incorrecta.
- Compuesta por un actor y un crítico.
- Busca encontrar un balance entre exploración y explotación agregando entropía a las acciones.



# PPO: Proximal Policy Optimization

- El algoritmo de PPO busca maximizar la recompensa acumulada en un episodio.
- Se busca optimizar el valor esperado de la combinación de 3 valores:
  - Pérdida de la política
  - Pérdida del valor
  - Pérdida de la entropía
- Las pérdidas son calculadas a partir de una serie de recorridos en el ambiente.

$$\mathbb{E}_t [L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)]$$

PPO Loss

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

$$L_t^{CLIP}(\theta) = \mathbb{E}_t \left[ \min \left( r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right]$$

Clipped Surrogate Objective

$$L_t^{VF} = \text{MSE} \left( r_t + \gamma r_{t+1} + \dots + \gamma^{T-t+1} r_{T-1} - V(s_t), V(s_t) \right)$$

Value Loss

$$S[\pi_\theta](s_t) = - \int \pi_\theta(a_t \mid s_t) \log(\pi_\theta a_t \mid s_t) da_t$$

Entropy Regularization

# PPO: Proximal Policy Optimization

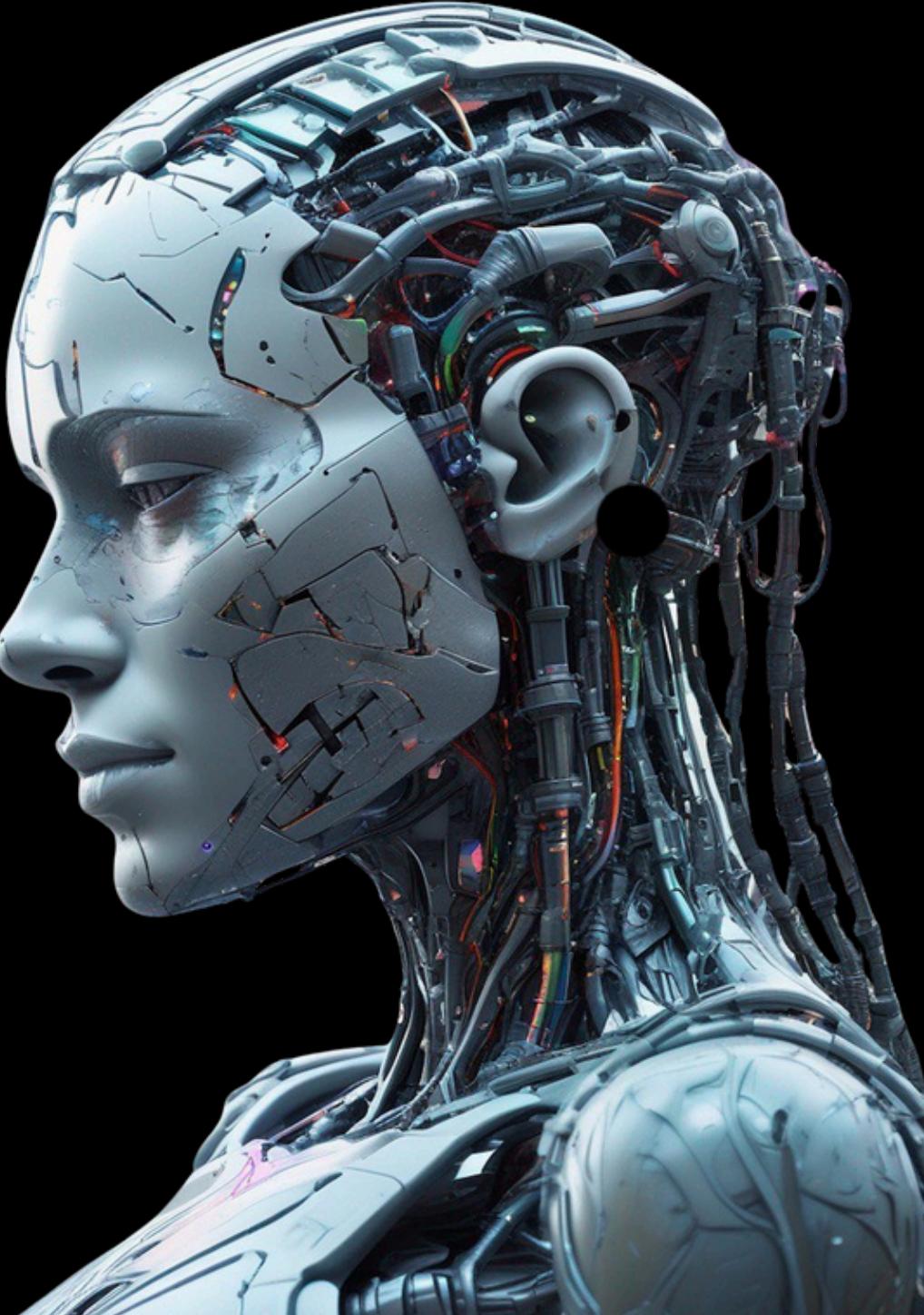


# GAE: Generalized Advantage Estimation

$$\hat{A}_t = \delta_t + \gamma \lambda \hat{A}_{t+1}$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

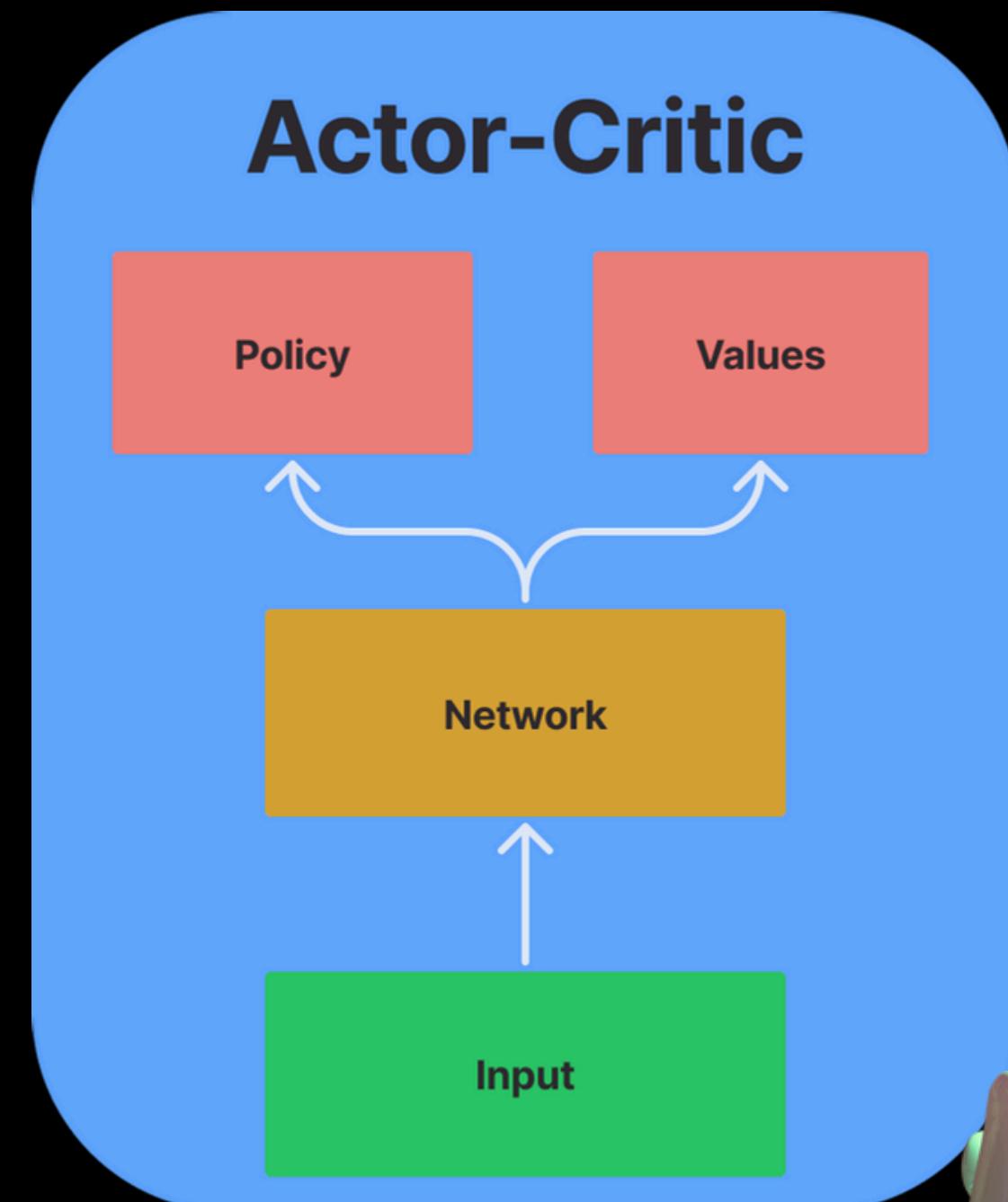
- Para obtener la pérdida de la política se necesita encontrar la ventaja de la acción tomada en un estado.
- La ventaja da dirección a la actualización, indicando qué acciones son mejores o peores de lo esperado.
- GAE busca encontrar un balance entre la varianza de *temporal difference* y el sesgo de  $\lambda$ -*return*.



# Actor-Critic

La red “Actor” produce la política  $\pi\theta(a|s)$ , la cual es una distribución de probabilidad sobre las acciones dadas las observaciones del estado  $s$ .

En un espacio discreto, la salida son las probabilidades para cada acción (usando Softmax).



La red “Critic” Estima el valor esperado  $V\phi(s)$ . La salida representa un escalar  $v(s)$ , siendo este la estimación del valor del estado actual.

# Centralized Learning

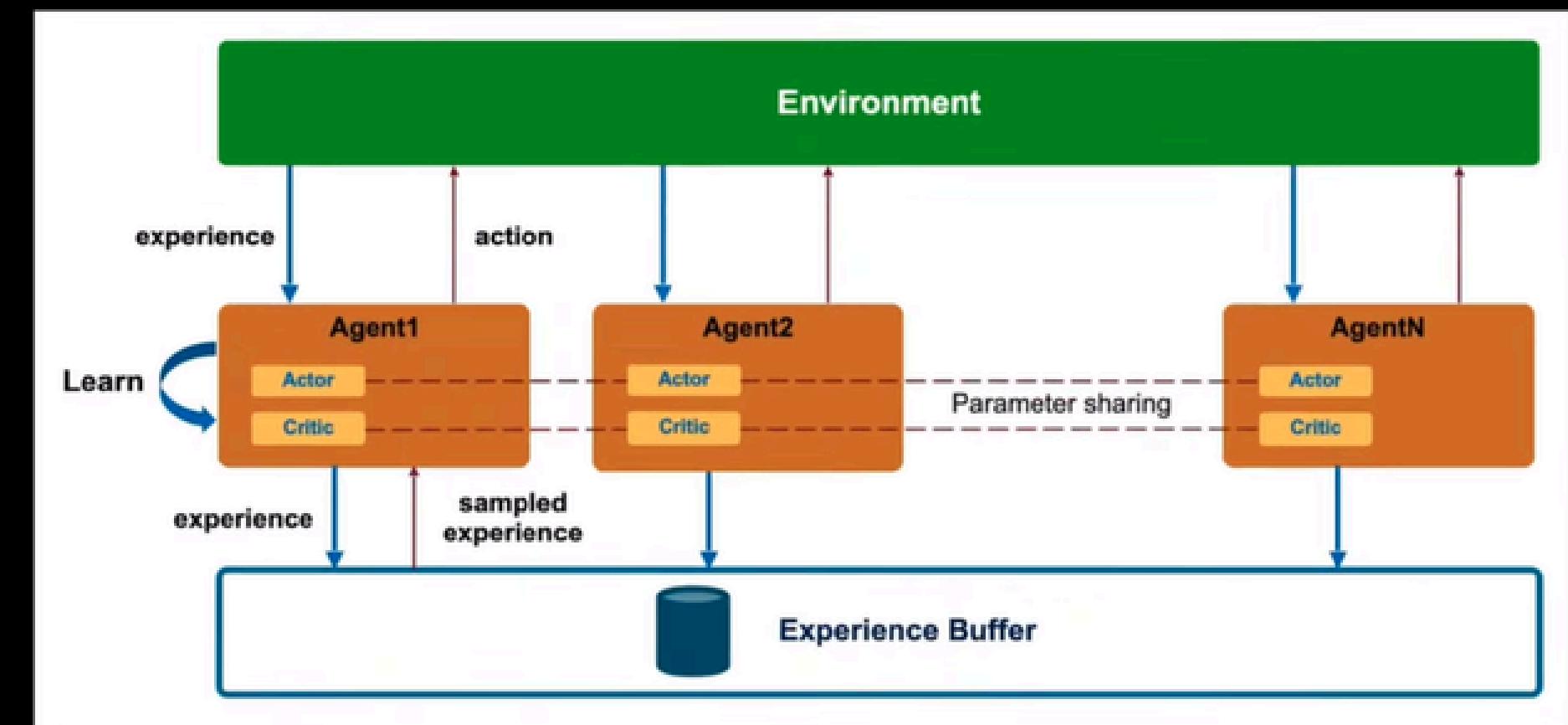
Ventajas:

- Aprendizaje más rápido.
- Solo se necesita tener una sola política.
- Los agentes toman en cuenta el comportamiento de los otros agentes.

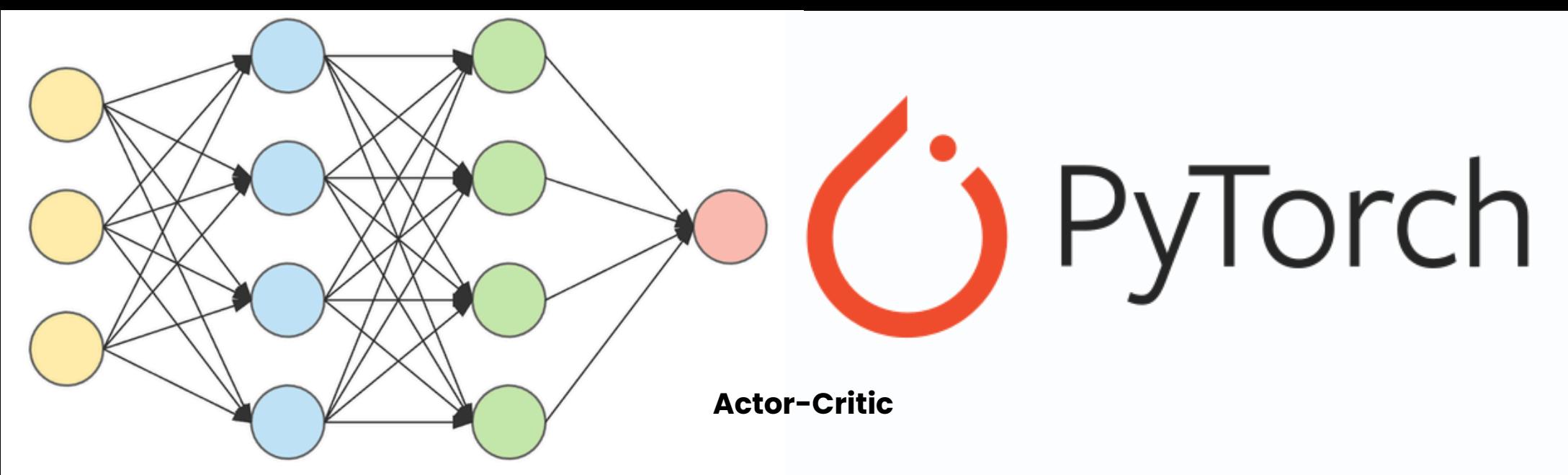
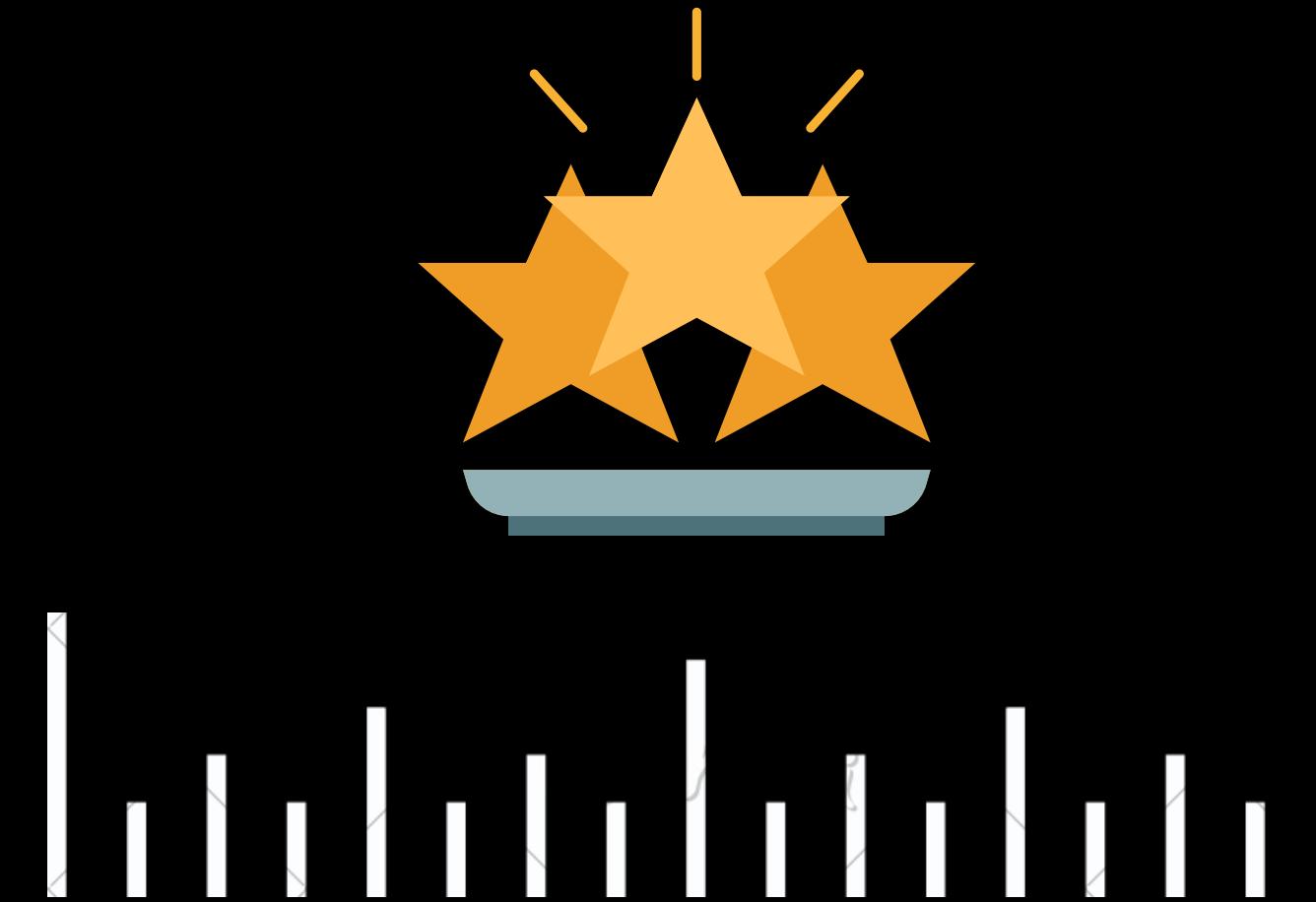
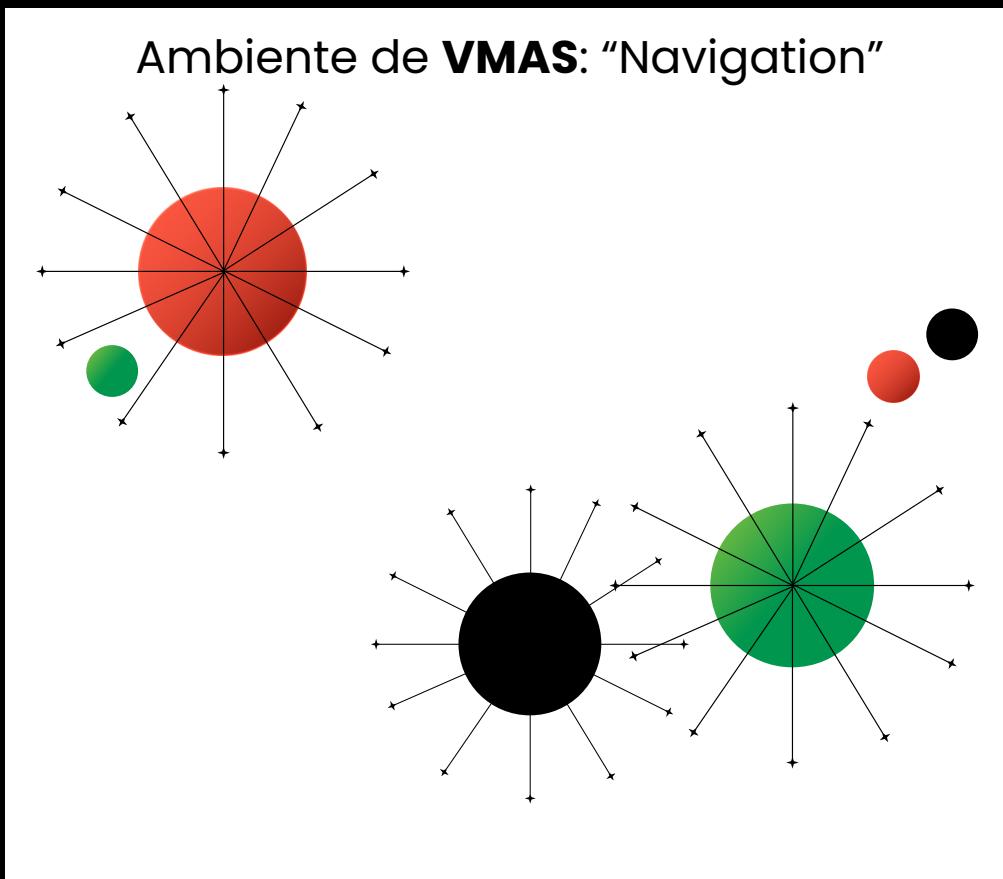
Desventajas:

- Los agentes necesitan tener observaciones y acciones iguales.
- Difícil de decir a qué agente pertenecen las recompensas compartidas.

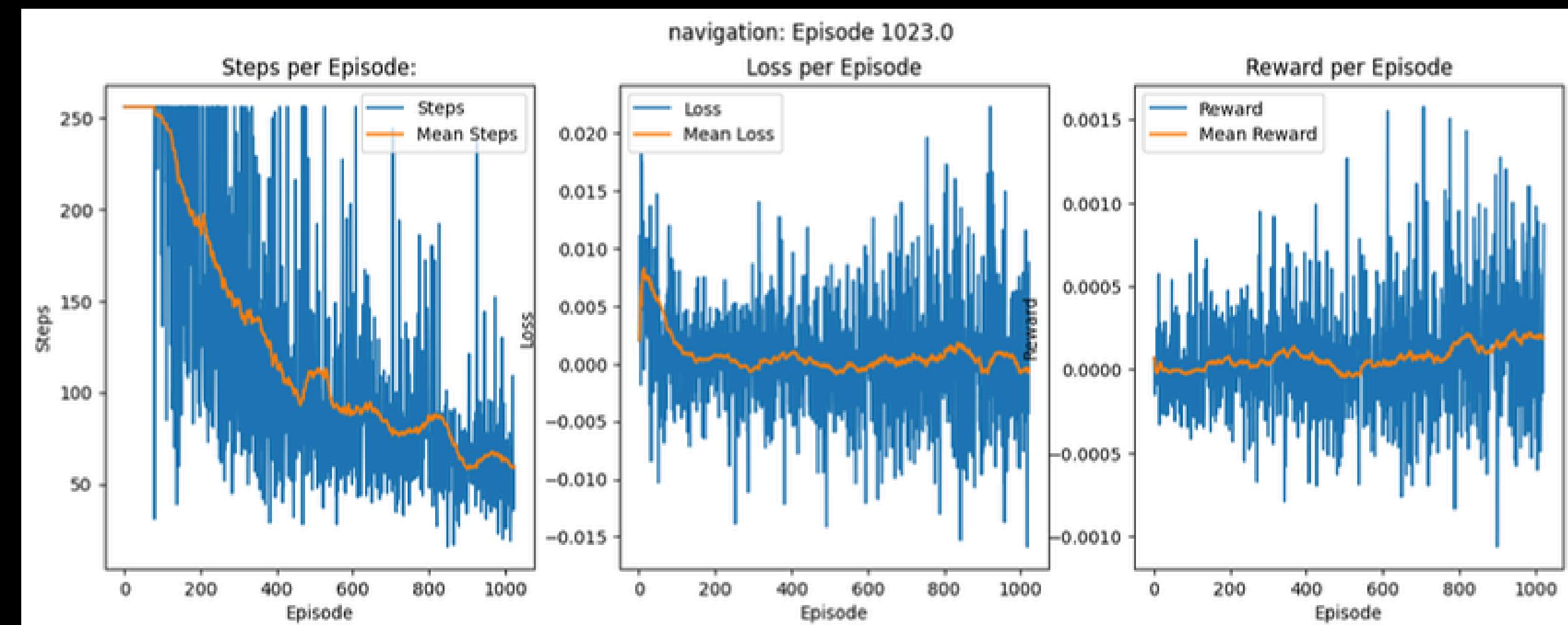
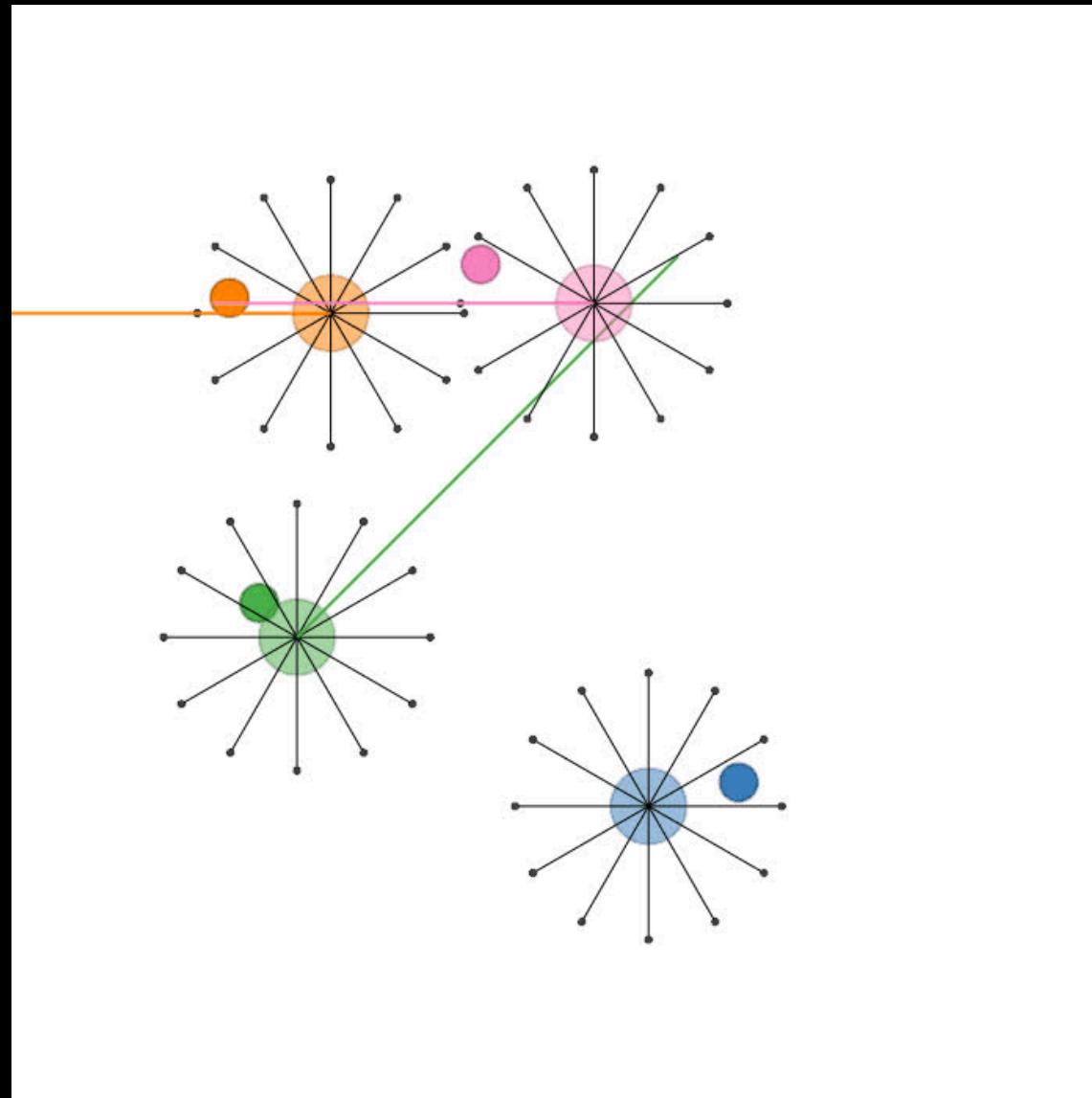
Centralized



# Metodología



# Resultados



# GRACIAS

P O R S U A T E N C I Ó N

Oswaldo Ilhuicatzi Mendizabal  
Mario Ignacio Frías Piña

A01781988  
A01782559

