Aprendizaje por refuerzo multiagente en entornos competitivos Atari: Transferencia de conocimiento de Pong a Quadrapong

Juan Manuel Camacho Lugo

Trabajo Final de Máster (Área 4)

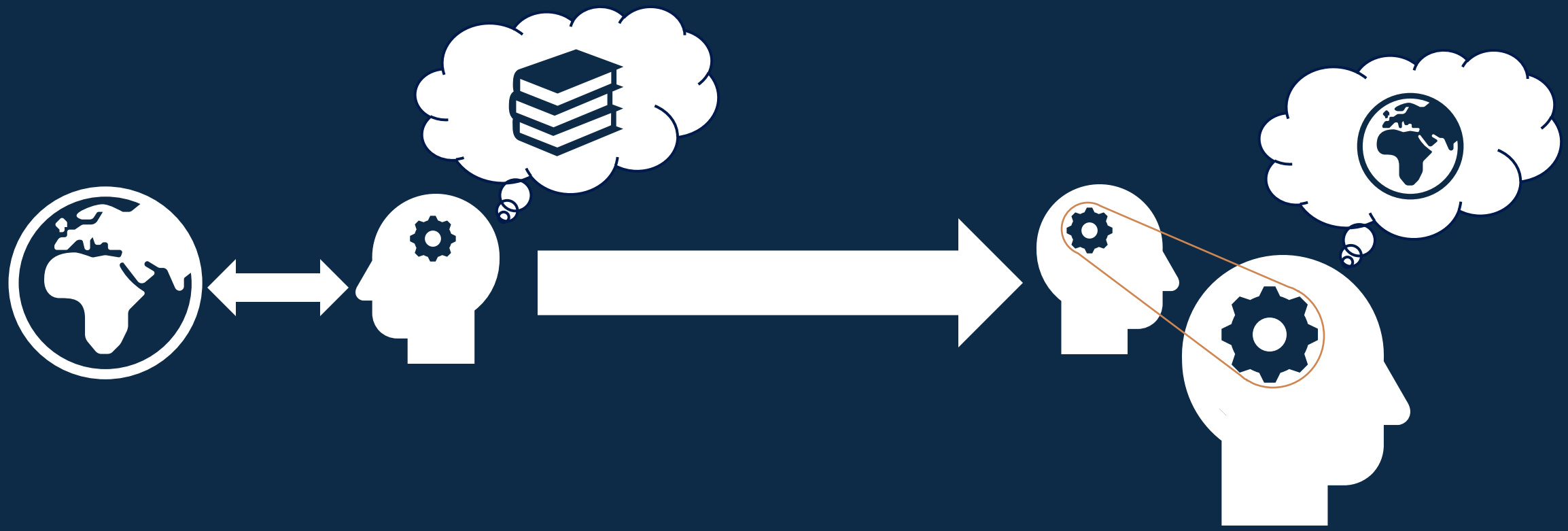
PDC: Luis Esteve Elfau

PRA: Ismael Benito Altamirano

Máster Universitario de Ciencia de Datos (UOC)

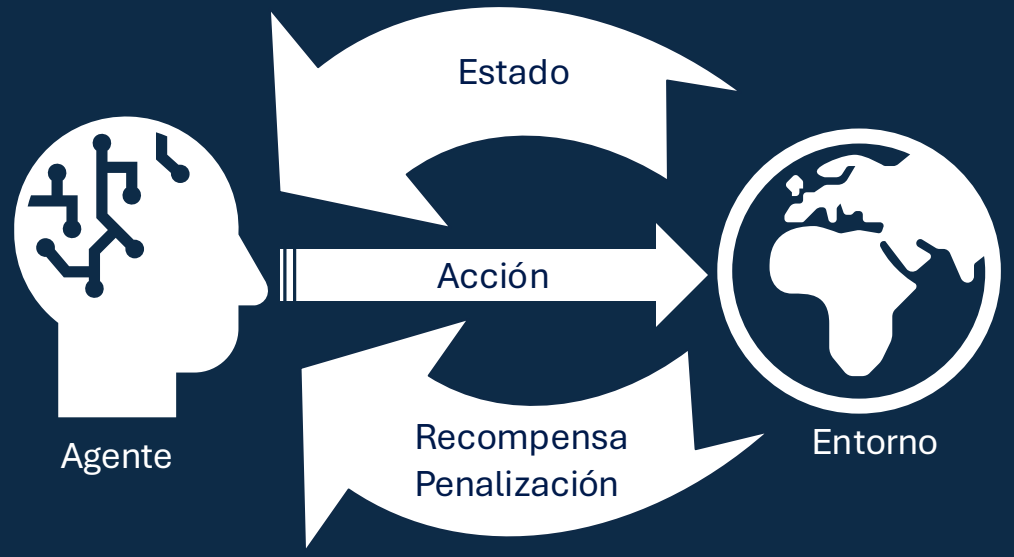
¿Qué problema queremos resolver?

Queremos enseñar a varios agentes a trabajar en un entorno complejo sin tener que aprender como hacerlo desde cero , partiendo de lo aprendido en un entorno simple.



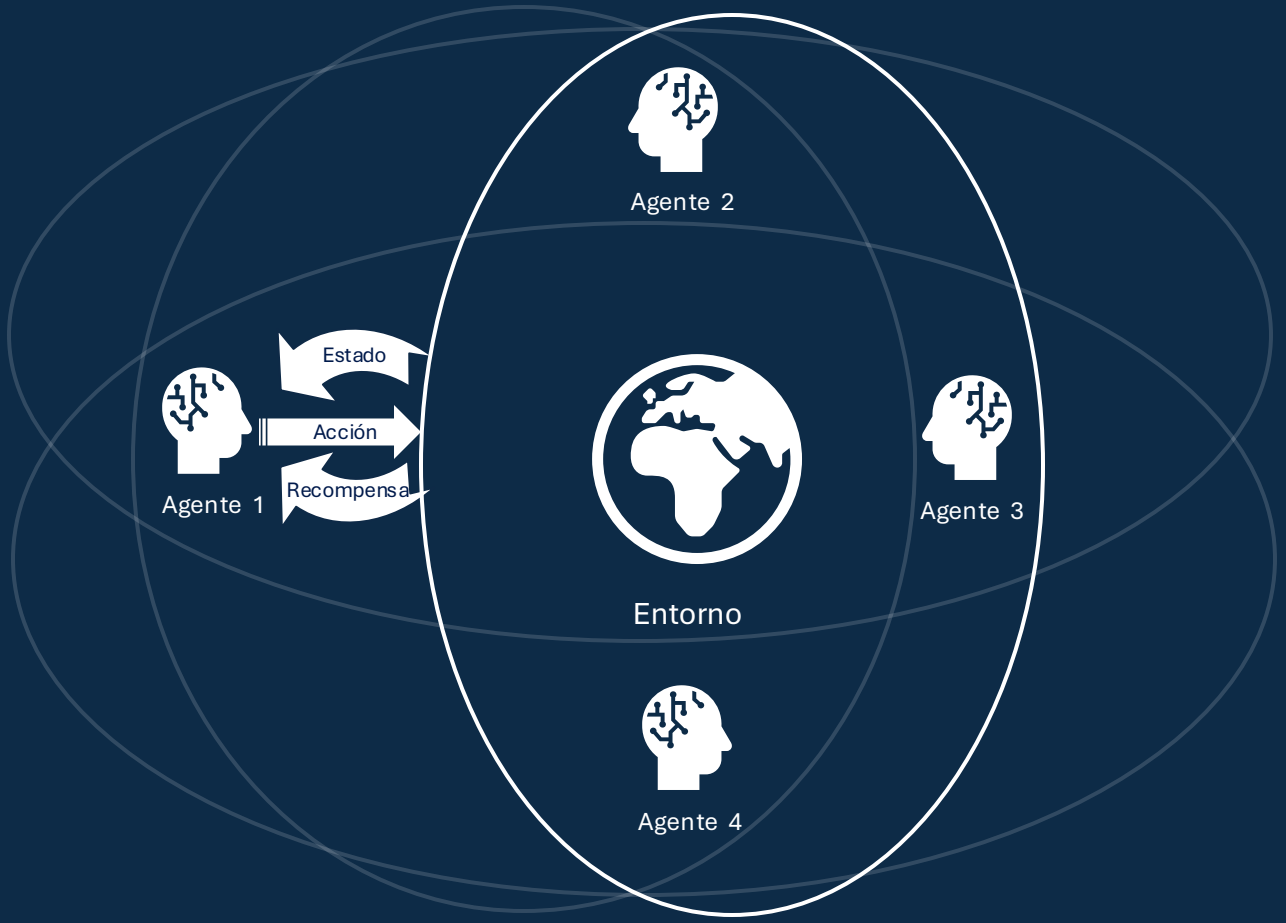
El aprendizaje por refuerzo permite a los agentes aprender mediante recompensas.

- Interactuar
- Maximizar la recompensa acumulada
- Aprender la mejor estrategia o política



El aprendizaje por refuerzo multiagente, introduce coordinación, cooperación y nuevos desafíos.

- Varios agentes, mismo entorno
- Metas por equipos o globales
- Encontrar equilibrio de los agentes

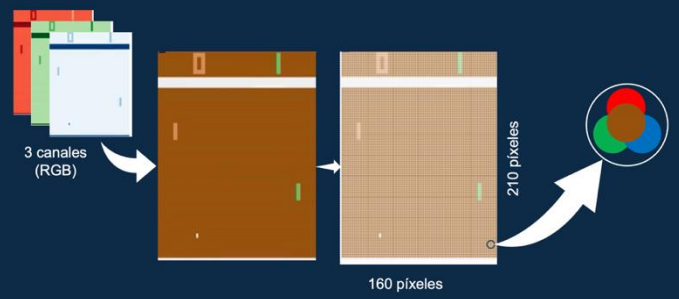


La transferencia de conocimiento acelera el aprendizaje en nuevos entornos.

- Reducir tiempo de entrenamiento
- Facilitar la adaptación al nuevo entorno
- Mejorar la exploración del nuevo entorno

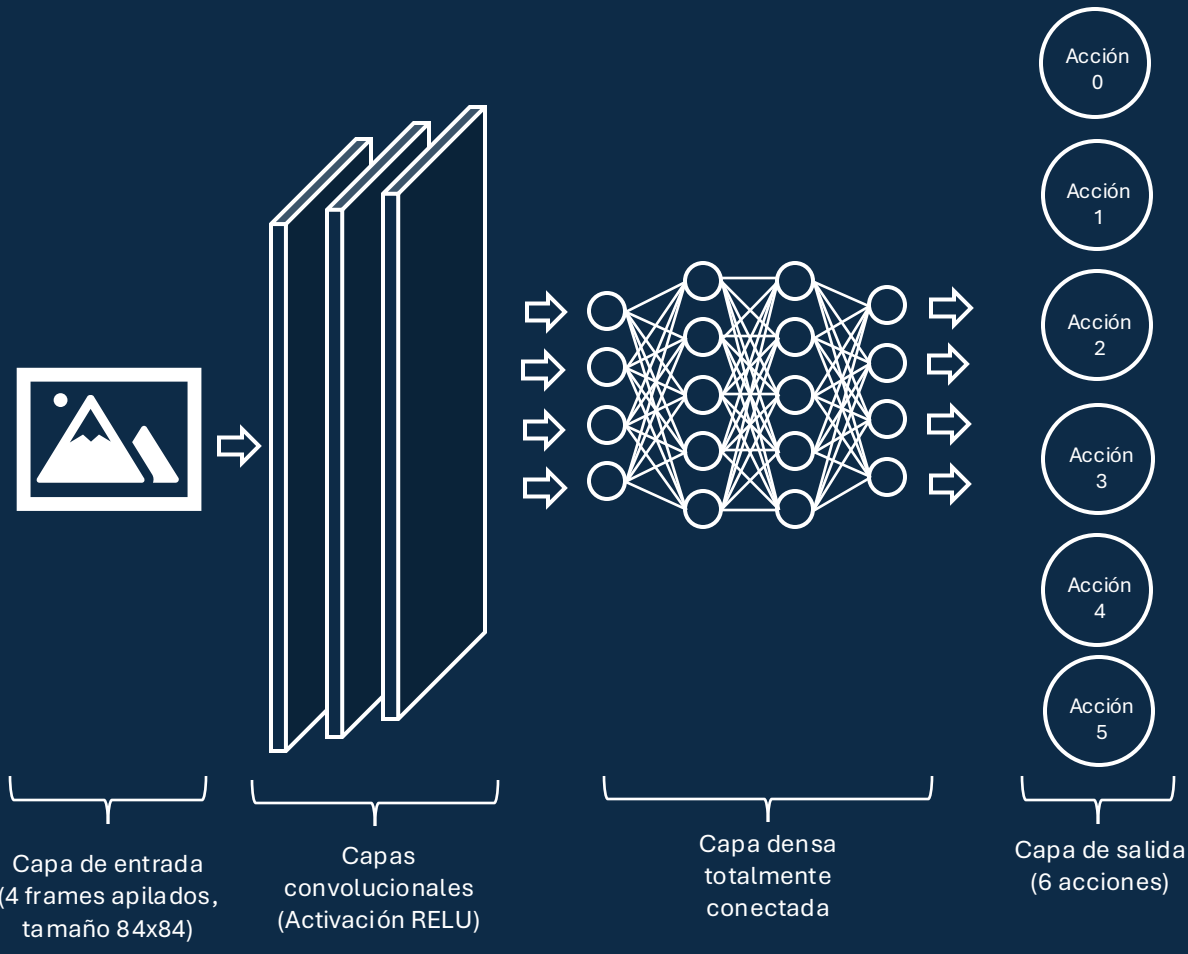
Entorno Pong / Quadrapong

ESPACIO DE OBSERVACIONES	ESPACIO DE ACCIONES DE PONG	ESPACIO DE ACCIONES DE QUADRAPONG
• Imagen RGB (210,160,3) • 210 píxeles alto • 160 píxeles ancho • 3 canales de color (rojo, verde, azul)	• 0: no hacer nada • 1: disparar / “sacar o servir la pelota” • 2: mover hacia la derecha • 3: mover hacia la izquierda • 4: mover hacia la derecha mientras saca • 5: mover hacia la izquierda mientras saca	• 0: no hacer nada • 1: disparar / “sacar o servir la pelota” • 2: mover hacia arriba • 3: mover hacia la derecha • 4: mover hacia la izquierda • 5: mover hacia abajo



Modelo empleado Pong / Quadrapong

- Implementación optimizada de PPO, para redes neuronales convolucionales (modelo propio por agente)
- Wrappers (reducir color, redimensionar, apilar frames)

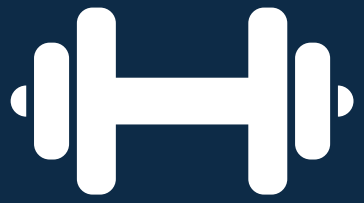


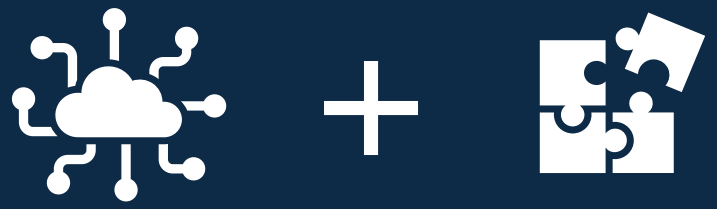
Entrenamiento

DISEÑO	CICLO APRENDIZAJE	RECOMPENSAS
• Simétrico • Competitivo • Alternando entre agentes • Self-Play • 5000 rondas (Pong) • 2000 rondas (Quadrapong)	• Carga de modelos preentrenados • Observación del estado • Selección de acción • Interacción con el entorno • Almacenamiento de experiencia • Actualización de política • Repetición de ciclo	• Punto a favor, recibe +1 • Punto en contra, recibe -1 • Stalling (no sacar tras dos segundos), recibe -1

Entrenamiento Quadrapong

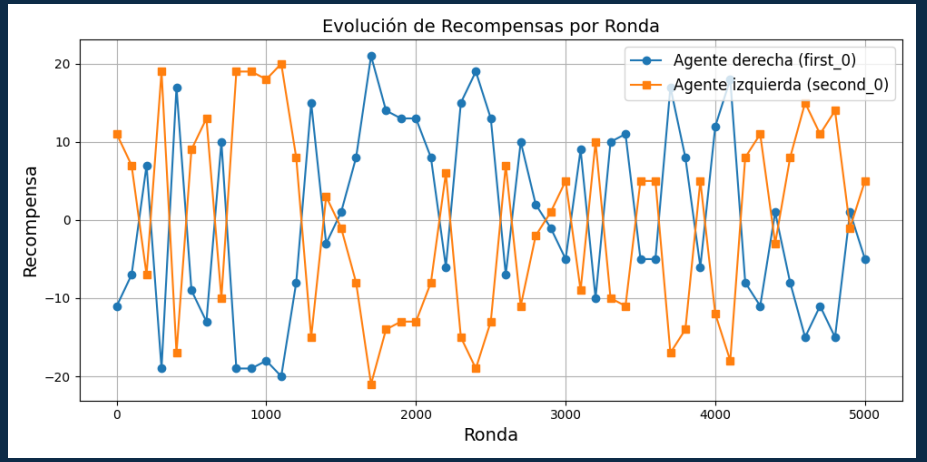
MODELO SIN TRANSFERENCIA DE CONOCIMIENTO	MODELO CON TRANSFERENCIA DE CONOCIMIENTO
• Entrenamiento desde cero	• TRANSFER LEARNING: (reconocer objetos, seguir la pelota, devolverla, ...) • FINE-TUNING: Entrenamiento adicional (adaptación a entorno / tareas)

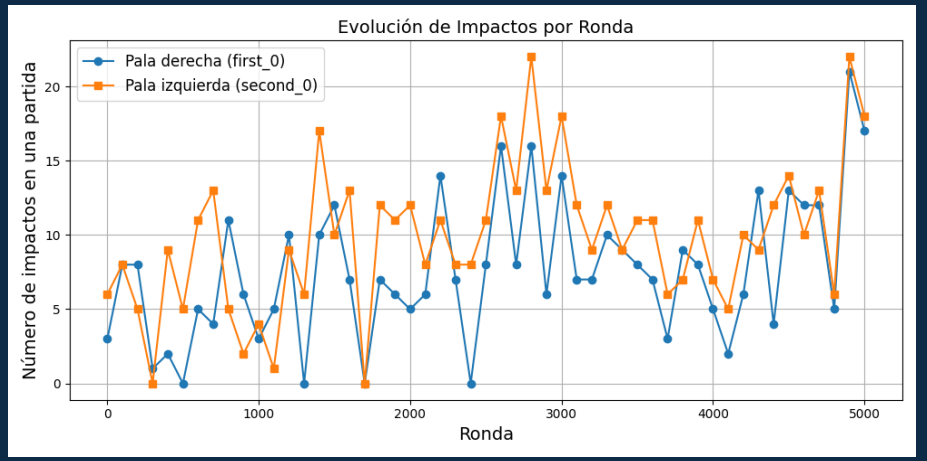




RESULTADOS PONG (EVALUACIÓN CADA 100 RONDAS)

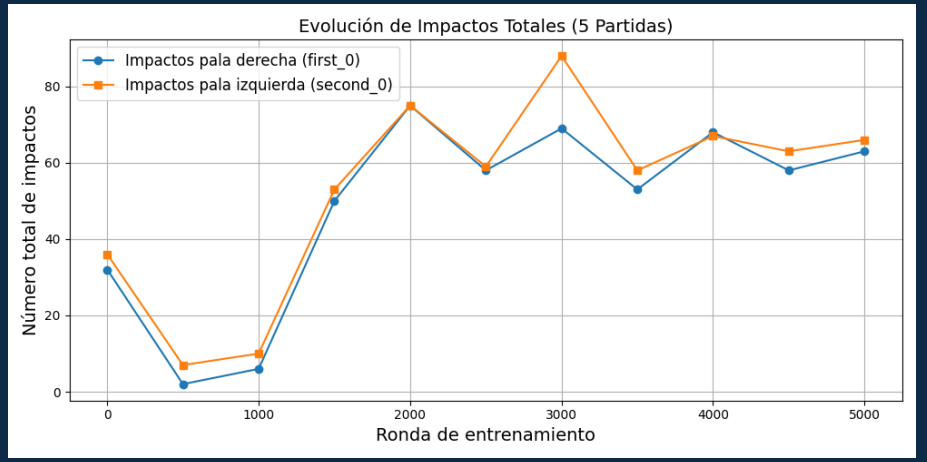
- Recompensas altamente relacionadas de forma inversa
- Impactos con tendencia ascendente (”progreso en aprendizaje”)

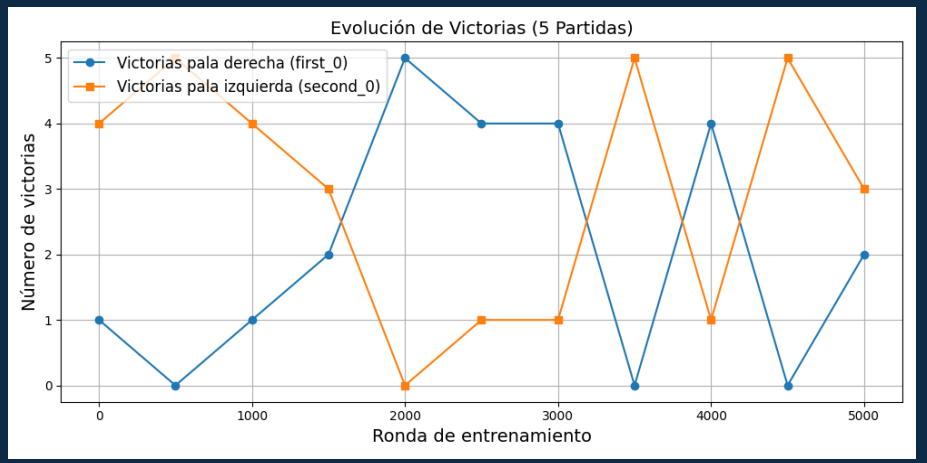




RESULTADOS PONG (EVALUACIÓN EN 5 PARTIDAS)

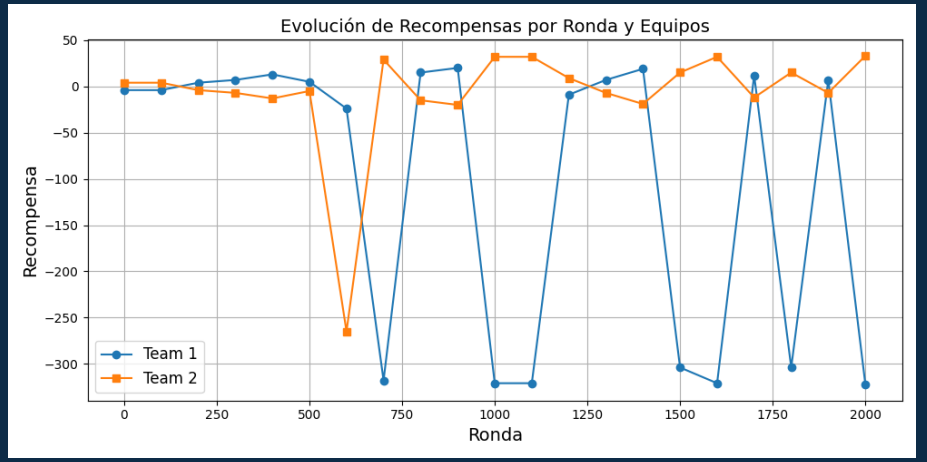
- Impactos con tendencia al aumento (“mejora en la capacidad de predicción”)
- Victorias con patrón oscilante (“alcanzando habilidades similares”)

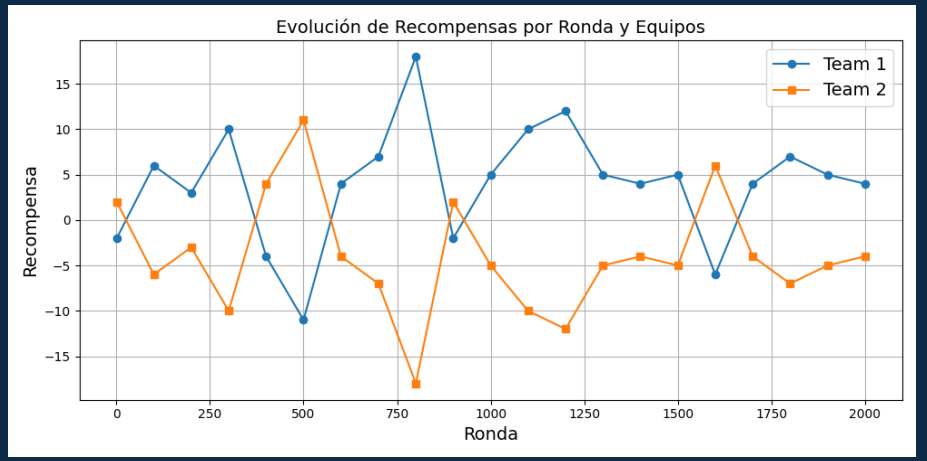




RESULTADOS QUADRAPONG

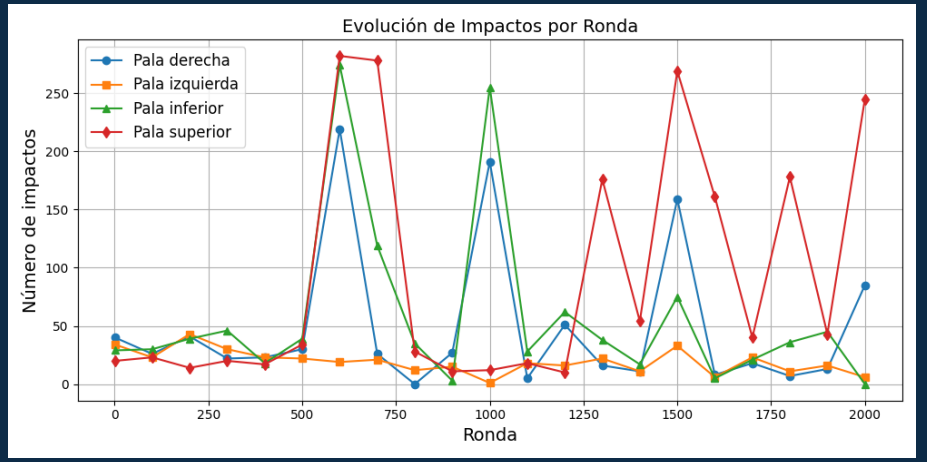
- Inestabilidad en el rendimiento (equipo 1, comportamientos extraños)
- Evolución más estable (“cierto beneficio de la transferencia de conocimiento”)

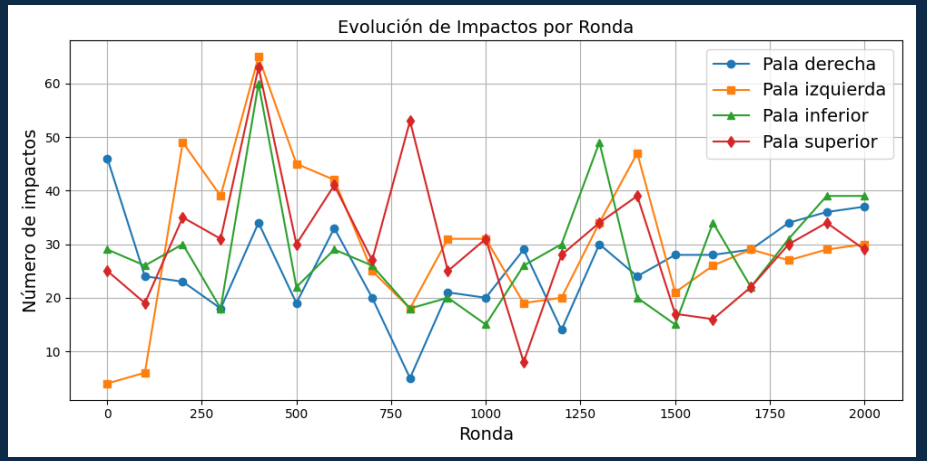




RESULTADOS QUADRAPONG

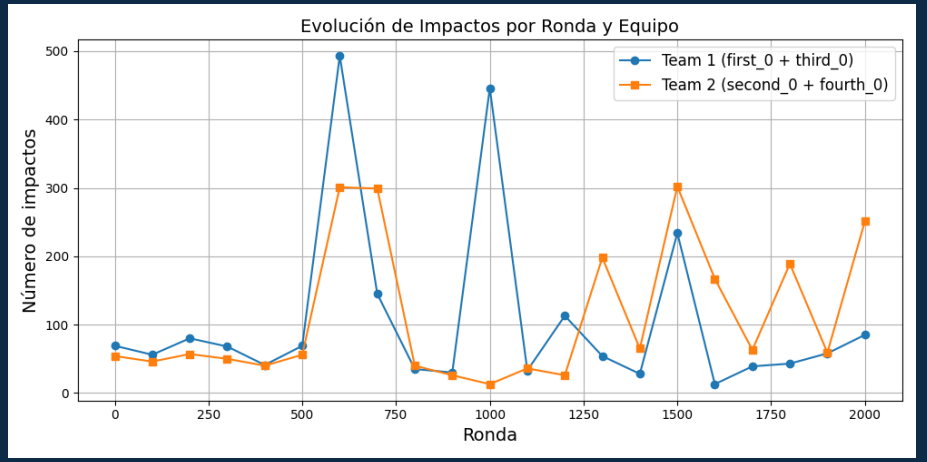
- Impactos con gran variabilidad e inestabilidad (“3 con comportamientos irregulares ”)
- Patrones de impactos con cierto equilibro, ausencia de picos extremos (“estabilizándose al final, posible presencia de algo de equilibrio”)

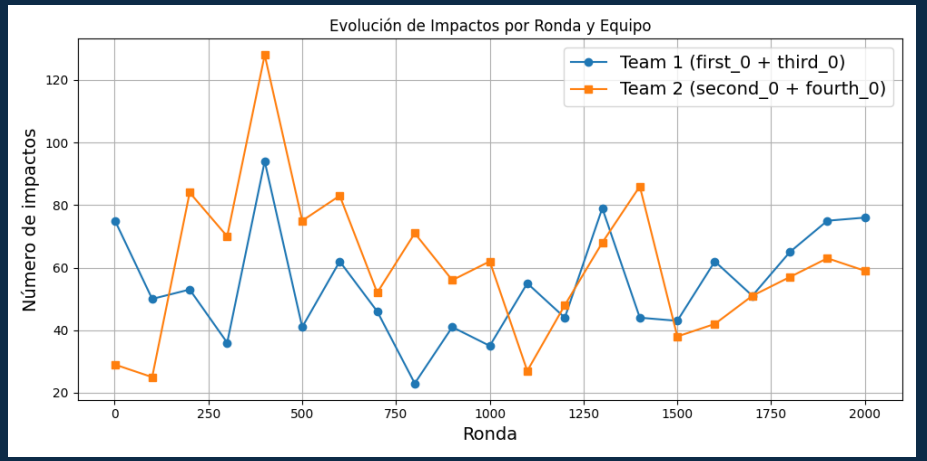




RESULTADOS QUADRAPONG

- A nivel de equipos se mantiene inestabilidad en los impactos (“posible aprendizaje al final del entrenamiento”)
- Distribución más estable y uniforme (“parece evolucionar controladamente, posible inicio de convergencia”)





CONCLUSIONES

LOGRO DE OBJETIVOS	PLANIFICACIÓN Y METODOLOGÍA
• No podemos confirmar la eficacia de la transferencia de conocimiento (resultados no concluyentes) • Objetivo abordado, pero no cumplido completamente	• Modelo inicial DQN sin resultados funcionales tras múltiples intentos (nuevo enfoque PPO) • Entrenamientos extremadamente lentos (tiempo insuficiente para entrenamientos más largos) • Alta demanda de recursos • Dificultad para realizar repeticiones

TRABAJOS FUTUROS

AMPLIAR RONDAS DE ENTRENAMIENTO

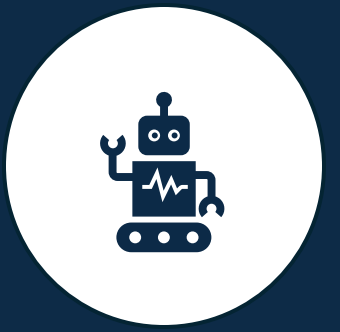
Consolidar estrategias robustas

EXPLORAR FUNCIONES DE RECOMPENSA ALTERNATIVAS

Fomentar cooperación o especialización

ESTUDIAR AGENTES HETEROGÉNEOS

Modelos o algoritmos distintos



APLICAR A ENTORNO REALES

Robótica, logística o gestión del tráfico

Muchas gracias por su atención

Juan Manuel Camacho Lugo
jcamacholu@uoc.edu