

# Aprendizaje por refuerzo

## Introducción

---

Antonio Manjavacas Lucas

[manjavacas@ugr.es](mailto:manjavacas@ugr.es)

1. Contexto histórico
2. Aprendizaje por refuerzo en la IA
3. Interacción agente–entorno
4. ¿Por qué RL?
5. Bibliografía recomendada

# Contexto histórico

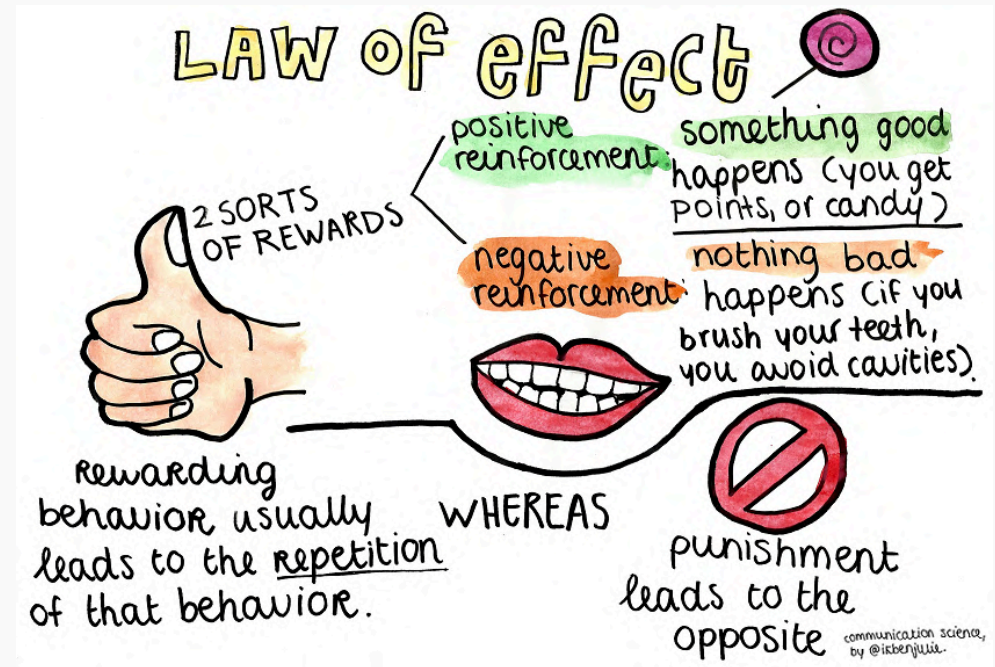
---

# Contexto histórico

**1850.** El filósofo y matemático **Alexander Bain** plantea el aprendizaje animal como un ejercicio basado en **prueba y error**.

**1911.** El psicólogo **Edward Thorndike** plantea la *ley del efecto*.

- Propone la existencia de **eventos de refuerzo** durante el aprendizaje animal que determinan el comportamiento.



# Contexto histórico

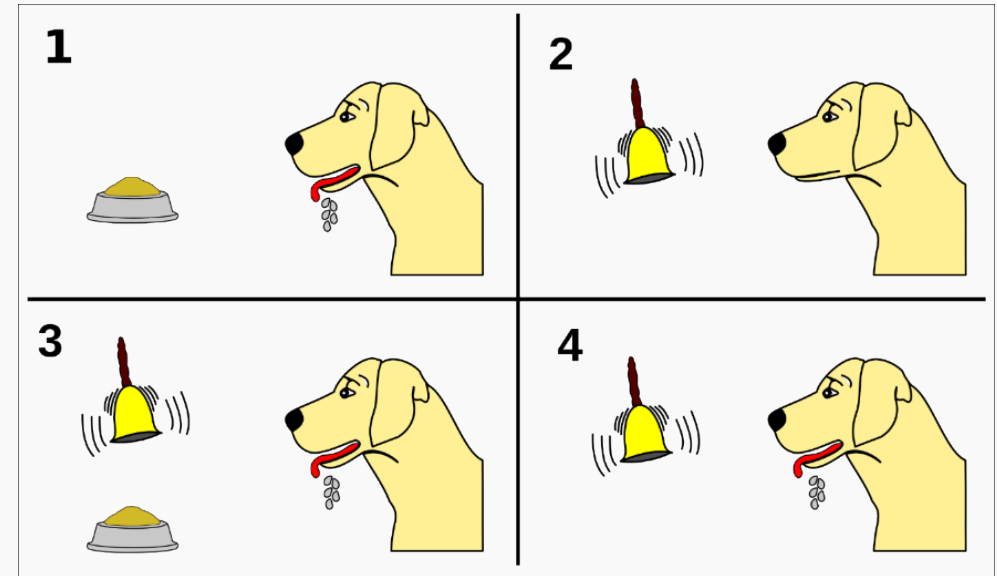
**1927.** Iván Pávlov formaliza el concepto de **refuerzo**.

## Refuerzo

Estímulo positivo o negativo que influencia un patrón de comportamiento.

- Determinados **estímulos** pueden incrementar la probabilidad de que un animal realice ciertas acciones.
- **Aprendizaje por refuerzo.** Modificación de la conducta a partir de estímulos/refuerzos.

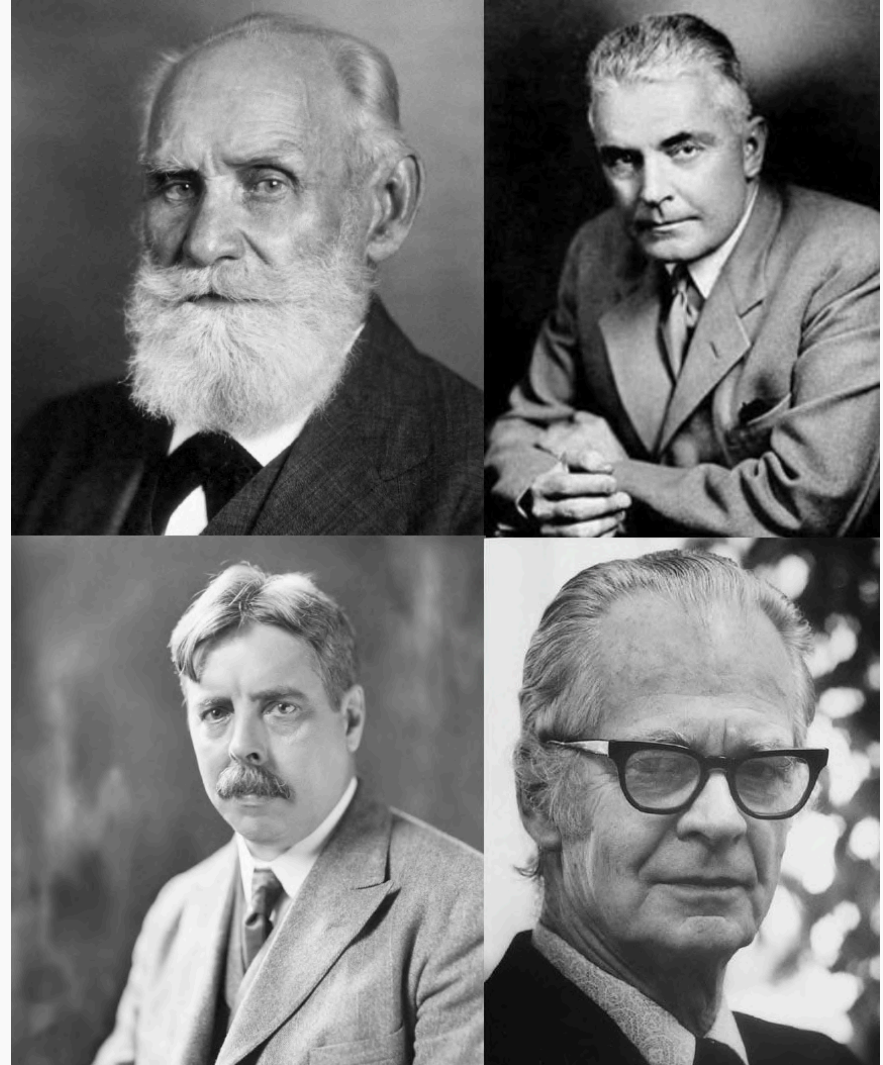
Experimento del **perro de Pávlov**:



# Contexto histórico

A partir de las ideas de **Pávlov**, **Watson**, **Thonrdike** y **Skinner** surge la psicología conductista, o **conductismo**.

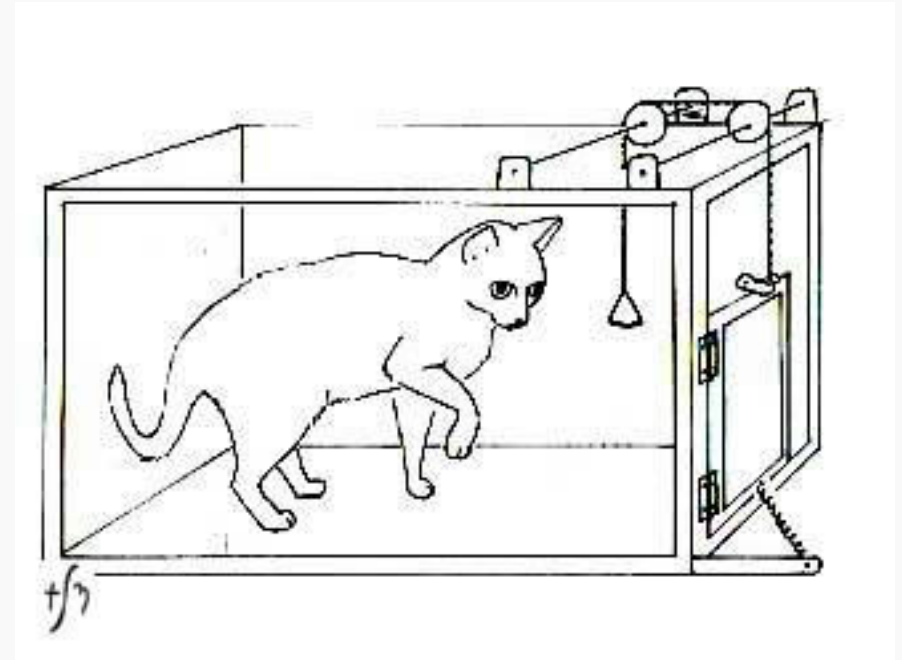
- Estudio de las leyes comunes que determinan el comportamiento humano y animal.



# Contexto histórico

El aprendizaje por refuerzo procede de los estudios sobre comportamiento animal, concretamente del **condicionamiento operante** / **aprendizaje instrumental**.

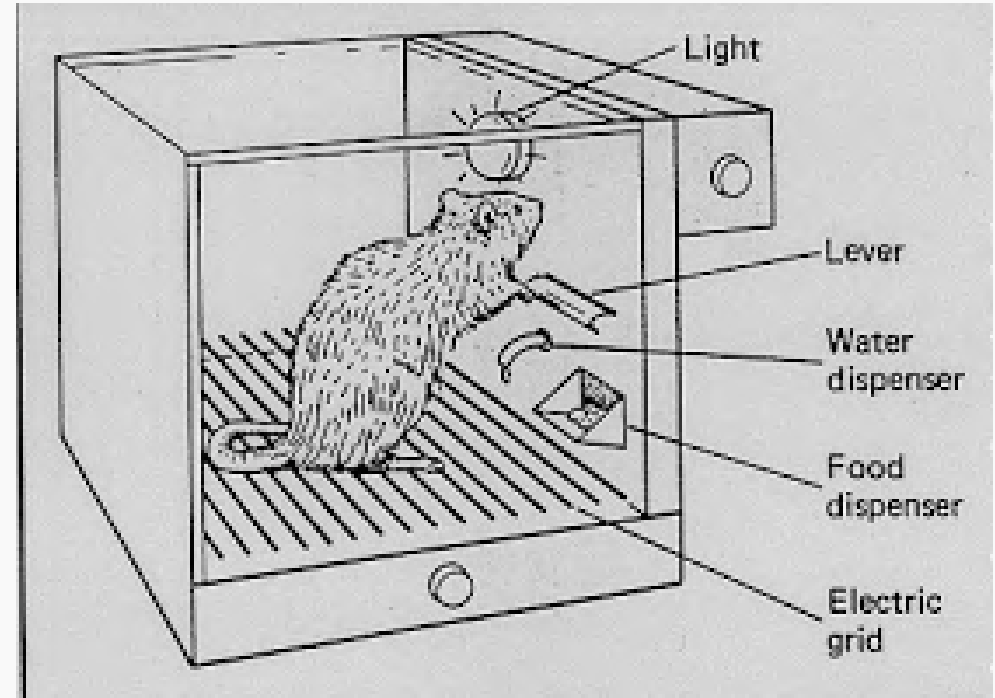
La *Ley del Efecto de Thorndike* sugería que conductas con **consecuencias satisfactorias** tienden a repetirse, mientras que aquellas que producen **consecuencias negativas** tienen menos probabilidades de volverse a realizar.



# Contexto histórico

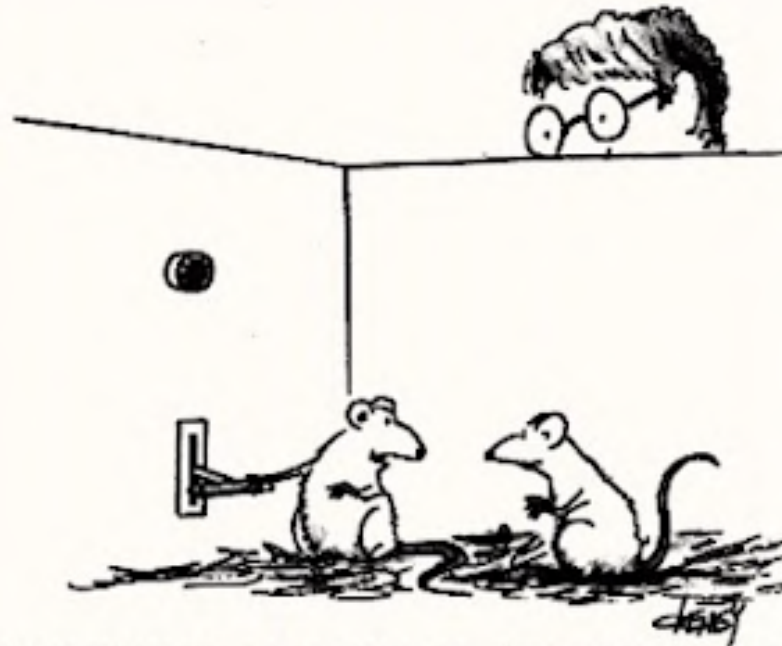
Skinner sostenía que los **refuerzos positivos** o **negativos** pueden ser empleados para modificar el comportamiento, tanto en animales como humanos.

- Experimento de la **caja de Skinner**.
- Proyectos *Paloma* y *ORCON*.





# Contexto histórico



It's a rather interesting phenomenon. Every time I press this lever, that post-graduate student breathes a sigh of relief.

¿Pueden los ordenadores aprender así?

# Aprendizaje por refuerzo en la IA

---

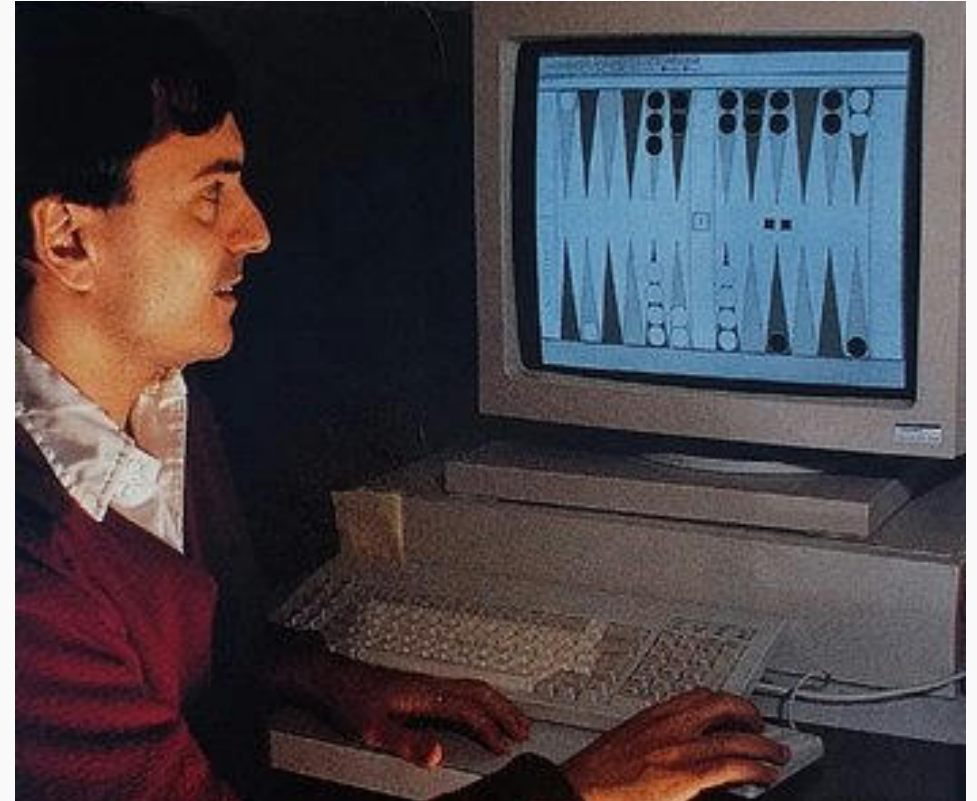
# Aprendizaje por refuerzo en la IA

- **1950s.** Control óptimo, procesos de decisión de Markov, programación dinámica.
  - **Richard Bellman, Ronald Howard.**
- **1970s.** Planteamiento del aprendizaje mediante “prueba y error” en el marco de la inteligencia artificial.
  - **Marvin Minsky, Harry Klopff, Robert Rescorla, Allan Wagner.**
- **1980s.** *Temporal difference learning*, algoritmo *Q-learning*.
  - **Richard Sutton, Andrew Barto, Christopher Watkins, Peter Dayan.**



# Aprendizaje por refuerzo en la IA

**1992.** IBM desarrolla *TD-Gammon*, alcanzando un nivel de habilidad humano en el juego del backgammon.



**2013.** Investigadores de DeepMind desarrollan el algoritmo *DQN*, capaz de superar al ser humano en 22 juegos de la consola Atari.



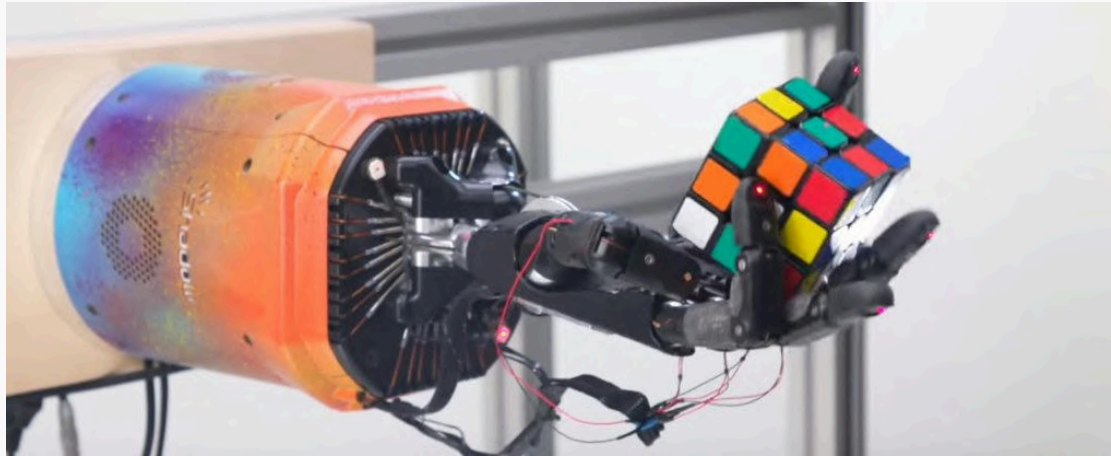
# Aprendizaje por refuerzo en la IA

- **2015–2017.** *AlphaGo*, desarrollado por DeepMind, vence a varios de los mejores jugadores de Go del mundo.
- **2017.** *AlphaZero*, vence 100–0 a su predecesor y alcanza un nivel superhumano en ajedrez.
- **2017–2019.** *OpenAI Five* vence a jugadores profesionales de Dota 2.
- **2019.** *AlphaStar* vence a los mejores jugadores de StarCraft II.



# Aprendizaje por refuerzo en la IA

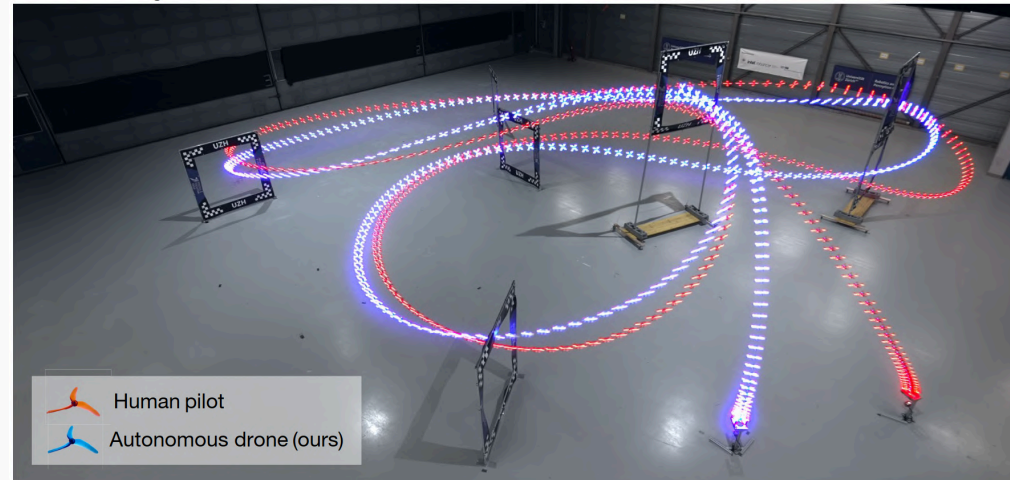
- **2019.** OpenAI desarrolla un brazo robótico capaz de resolver cubos de rubik.
- **2020.** *MuZero* aprende a dominar Go, ajedrez, shogi y Atari simultáneamente.
- **2022.** Control magnético de plasma en reactores nucleares de fusión mediante *reinforcement learning*.





# Aprendizaje por refuerzo en la IA

- **2023**. *Swift* logra vencer a varios campeones mundiales en pilotaje de drones.
- **2023–**. Aprendizaje por refuerzo aplicado a mejorar las respuestas de agentes conversacionales como *ChatGPT*.



En los últimos años, encontramos algoritmos de *reinforcement learning* aplicados a...

- Robótica
- Ciencias naturales
- Sistemas de recomendación
- Conducción autónoma
- Energía
- Economía e inversión
- Aceleradores de partículas
- Distribución eléctrica
- Telecomunicaciones
- ...

En los últimos años, encontramos algoritmos de *reinforcement learning* aplicados a...

- Robótica
- Ciencias naturales
- Sistemas de recomendación
- Conducción autónoma
- Energía
- Economía e inversión
- Aceleradores de partículas
- Distribución eléctrica
- Telecomunicaciones
- ...

¿PERO CÓMO FUNCIONAN?

# Interacción agente-entorno

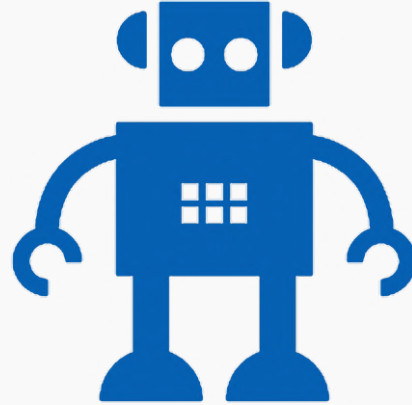
---

# Componentes del aprendizaje por refuerzo

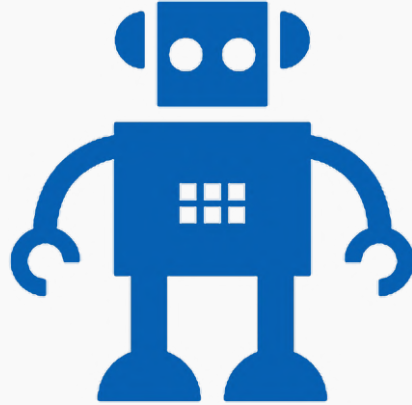
Un problema de *reinforcement learning* (RL) está compuesto por múltiples elementos.

Agente, entorno, estado, acción, recompensa, política...

Veamos en detalle en qué consiste cada uno de ellos...

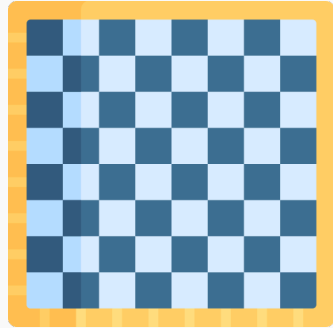


Persigue un determinado **objetivo**. Observa y actúa sobre un entorno.



Persigue un determinado **objetivo**. Observa y actúa sobre un entorno.

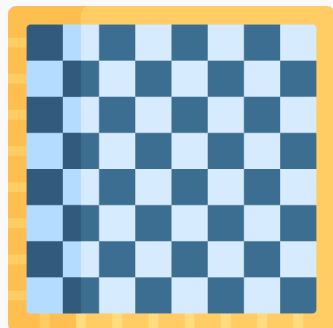
- *Colocar una pieza en un tablero.*
- *Moverse a una determinada posición.*
- *Aumentar/reducir la velocidad de un vehículo.*



Sistema dinámico con el que el agente interactúa, recibiendo información o alterándolo.

El agente percibe el **estado** del entorno, y utiliza esta información para elegir qué **acciones** realizar.





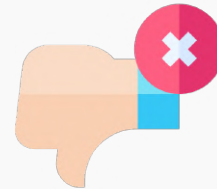
Sistema dinámico con el que el agente interactúa, recibiendo información o alterándolo.

El agente percibe el **estado** del entorno, y utiliza esta información para elegir qué **acciones** realizar.

- *Posición de las piezas en un tablero de ajedrez.*
- *Proximidad de un vehículo a los límites de la carretera.*

# Recompensa

Valor que indica **cómo de buena o mala** es una *acción* o *estado* para el agente.



# Recompensa

Valor que indica **cómo de buena o mala** es una *acción* o *estado* para el agente.



- $R = +1$  por cada instante de tiempo en que un robot se mantiene en pie (**refuerzo positivo**).

# Recompensa

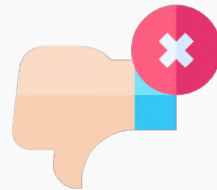
Valor que indica **cómo de buena o mala** es una *acción* o *estado* para el agente.



- $R = +1$  por cada instante de tiempo en que un robot se mantiene en pie (**refuerzo positivo**).
- $R = -1$  por cada instante de tiempo que el agente tarda en salir de un laberinto (**refuerzo negativo**).

# Recompensa

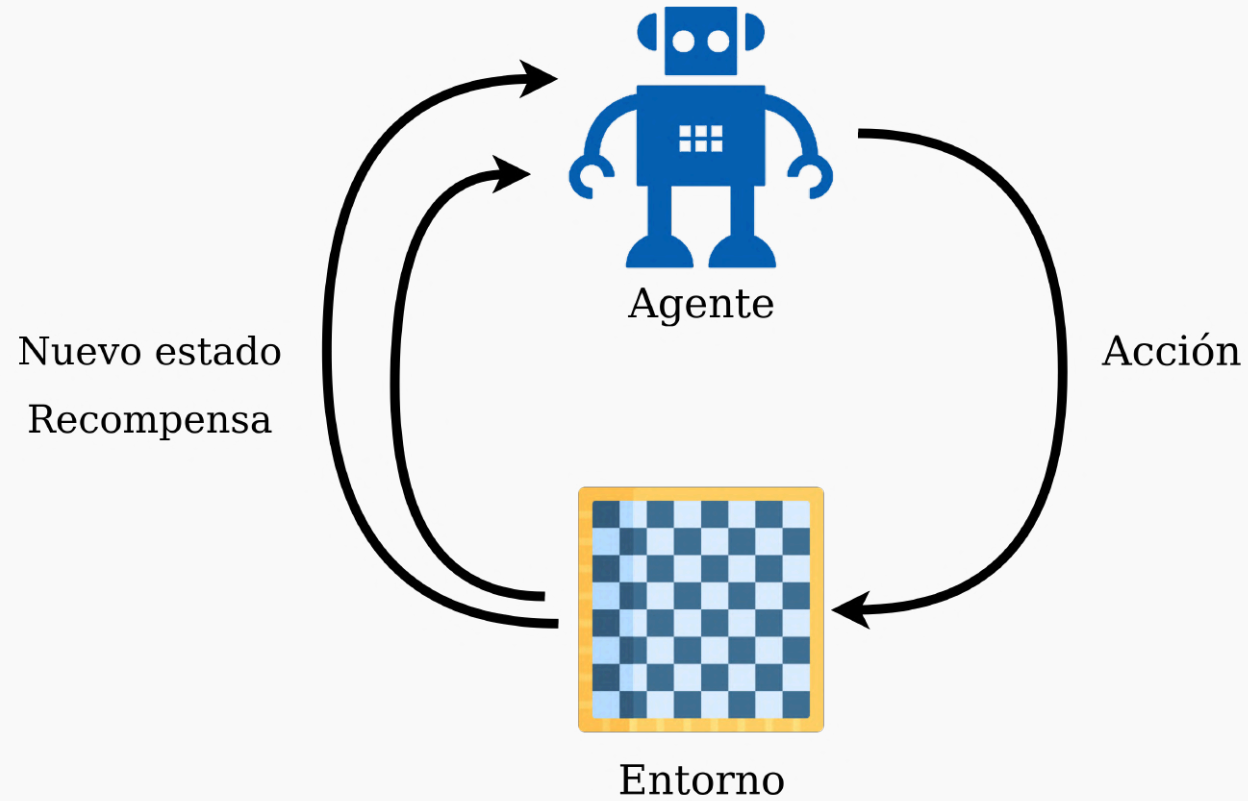
Valor que indica **cómo de buena o mala** es una *acción* o *estado* para el agente.



- $R = +1$  por cada instante de tiempo en que un robot se mantiene en pie (**refuerzo positivo**).
- $R = -1$  por cada instante de tiempo que el agente tarda en salir de un laberinto (**refuerzo negativo**).
- $R = +1$  si el agente recomienda un producto y el usuario lo compra;  $R = +0$  si no lo compra;  $R = -1$  si lo compra y lo devuelve.

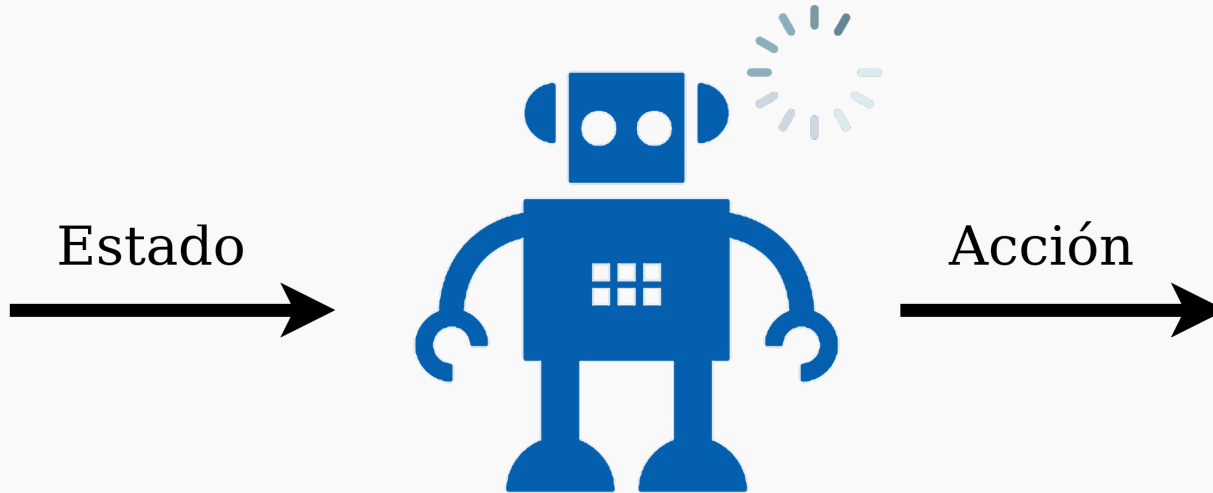
Si combinamos todos estos elementos...

# Procesos de decisión de Markov



# Acciones y estados

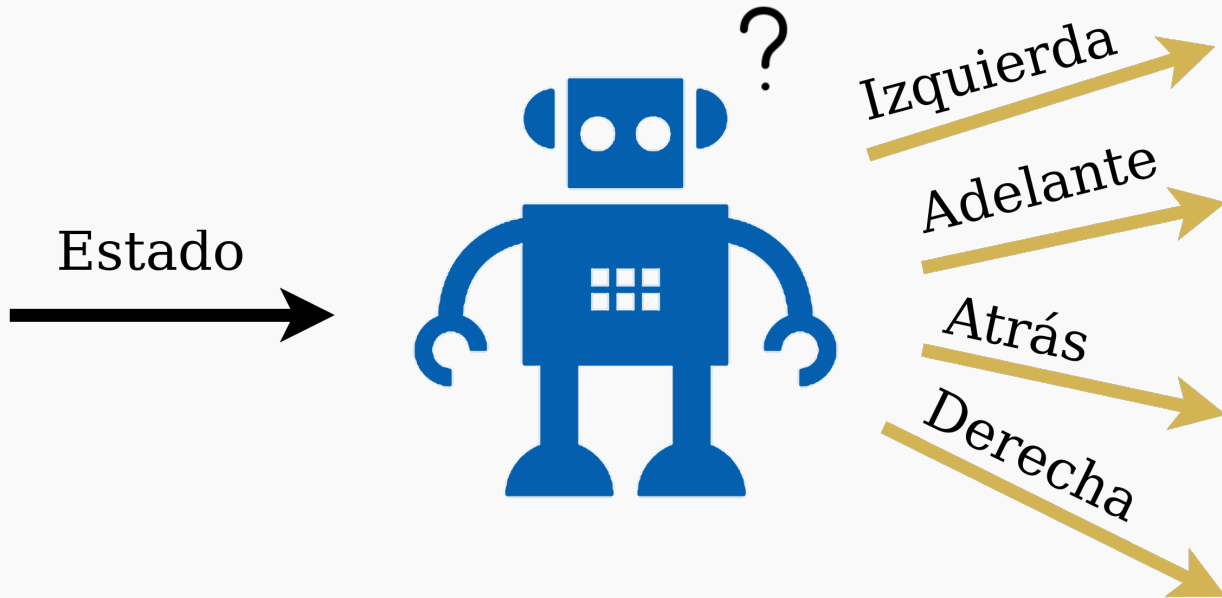
Internamente, el agente establece una *correspondencia* entre estados y acciones:





# Acciones y estados

¿Pero cómo determina qué acción es más apropiada ante un determinado estado?





Define el comportamiento del agente, y establece una **correspondencia** entre *estados y acciones*.

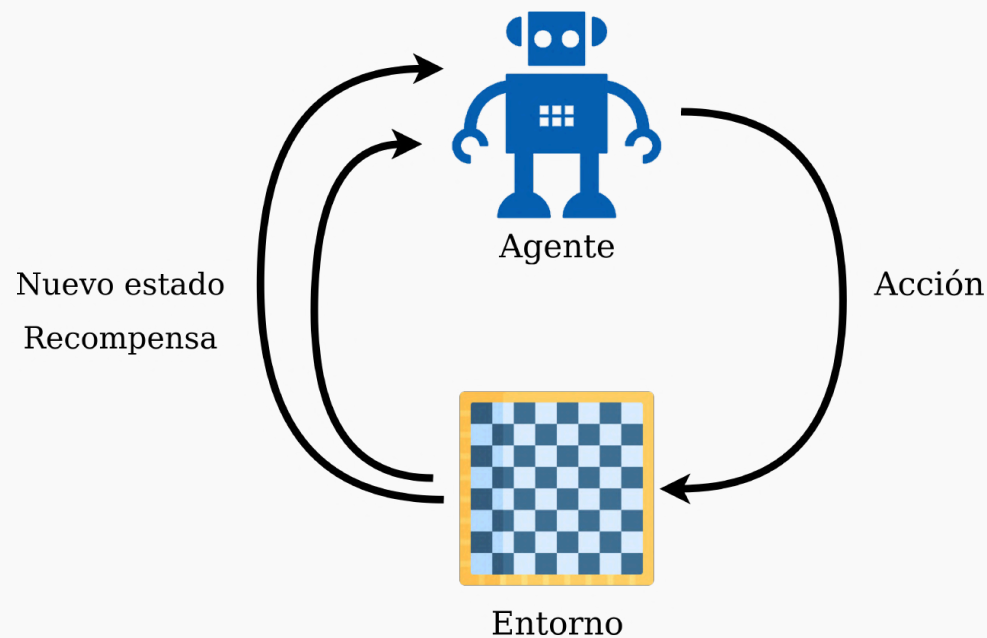
Puede componerse de reglas simples, conocimiento experto, o requerir una gran cantidad de cálculos (ej. redes neuronales).

El **objetivo** del agente es aprender una **política de comportamiento óptima**.

- Aquella que le conduzca a maximizar las recompensas obtenidas.
- Aprendizaje basado en “prueba y error”.

# En resumen...

- Método de aprendizaje computacional basado en la interacción de un **agente** con su **entorno**.
- El agente percibe el **estado** del entorno y ejecuta **acciones** que lo alteran.
- Proceso iterativo, basado en prueba y error, donde el objetivo del agente es la maximización de una señal de **recompensa**.
- Aprendizaje de una **política** de comportamiento óptima.

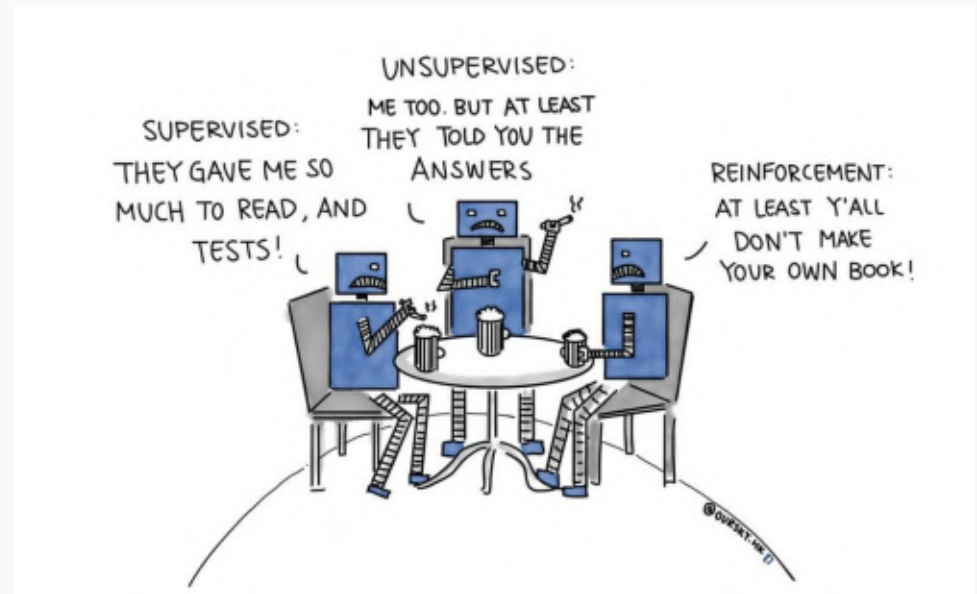


¿Por qué RL?

---

# ¿Por qué RL?

- Interés por emular el **aprendizaje animal** basado en causa-efecto.
- Diferencias relevantes con respecto al **aprendizaje supervisado** y **no supervisado**.
- RL no es solamente un conjunto de algoritmos destinados a resolver un problema, también incluye el **planteamiento** y **formalización** de dicho problema.



## Formalización del problema

- ¿Problema episódico o continuado?
- Modelo del entorno
  - Variables observadas
  - Función de recompensa
  - *Model free vs. model-based*
- Definición del espacio de acciones (discreto o continuo)

## Planteamiento de la solución

- Sistemas de ecuaciones
- Programación dinámica
- Monte Carlo
- *TD-learning* (ej. *Q-learning*)
- Métodos basados en gradiente
- Aproximación de funciones (ej. *tile coding*, *Deep Reinforcement Learning*)

# RL vs. Aprendizaje Supervisado

- **Aprendizaje supervisado** → aprendizaje a partir de un conjunto de datos previamente **etiquetados** por un **supervisor externo**.
- Cada ejemplo está compuesto por un conjunto de **características** ( $X_1, X_2, X_3, \dots$ ) y una **etiqueta** o **valor** a predecir ( $Y$ ).
- El objetivo de estos algoritmos es aprender a **generalizar** más allá de los datos empleados en su entrenamiento.

El **aprendizaje por refuerzo** permite aprender en base a la propia experiencia del agente, sin necesidad de un supervisor externo (vs. aprendizaje supervisado).

No obstante, veremos cómo pueden combinarse.



# RL vs. Aprendizaje Supervisado

- **Aprendizaje supervisado** → aprendizaje a partir de un conjunto de datos previamente **etiquetados** por un **supervisor externo**.
- Cada ejemplo está compuesto por un conjunto de **características** ( $X_1, X_2, X_3, \dots$ ) y una **etiqueta** o **valor** a predecir ( $Y$ ).
- El objetivo de estos algoritmos es aprender a **generalizar** más allá de los datos empleados en su entrenamiento.

El **aprendizaje por refuerzo** permite aprender en base a la propia experiencia del agente, sin necesidad de un supervisor externo (vs. aprendizaje supervisado).

No obstante, veremos cómo pueden combinarse.

"Reinforcement learning is supervised learning on optimized data"

*What makes RL challenging is that, unless you're doing imitation learning, actually acquiring that "good data" is quite challenging.*

 <https://bair.berkeley.edu/blog/2020/10/13/supervised-rl/>

# RL vs. Aprendizaje Supervisado

APRENDIZAJE SUPERVISADO	APRENDIZAJE POR REFUERZO
Se dispone de un conjunto de datos de entrenamiento preparados por un <b>supervisor externo</b> .	El conocimiento/comportamiento óptimo se obtiene a través de <b>prueba y error</b> . No hay supervisión alguna.
El aprendizaje está dirigido por la diferencia entre predicción y objetivo ( <b>error de predicción</b> ).	Los refuerzos ( <b>recompensas</b> ) modifican la probabilidad de seleccionar determinadas acciones.
El <b>feedback</b> es <b>instantáneo</b> : el objetivo es inmediatamente conocido.	El <b>feedback</b> puede no ser inmediato ( <b>delayed feedback</b> ). No podemos saber inicialmente qué acciones llevaron a una determinada recompensa.
Los ejemplos de entrenamiento se seleccionan aleatoriamente del <i>training set</i> .	A medida que el comportamiento mejora, los datos observados varían.

# RL vs. Aprendizaje No Supervisado

- **Aprendizaje no supervisado** → aprendizaje de la estructura intrínseca de un conjunto de datos no etiquetados.
- Tanto RL como ANS carecen de supervisión externa, pero sus objetivos son diferentes.


Aunque tanto el aprendizaje no supervisado como el aprendizaje por refuerzo carecen de ejemplos de comportamiento “correcto”, el **aprendizaje por refuerzo** trata de **maximizar una señal de recompensa**, en lugar de intentar encontrar **patrones comunes en conjuntos de datos**.

# RL vs. Aprendizaje No Supervisado

- **Aprendizaje no supervisado** → aprendizaje de la estructura intrínseca de un conjunto de datos no etiquetados.
- Tanto RL como ANS carecen de supervisión externa, pero sus objetivos son diferentes.

Aunque tanto el aprendizaje no supervisado como el aprendizaje por refuerzo carecen de ejemplos de comportamiento “correcto”, el **aprendizaje por refuerzo** trata de **maximizar una señal de recompensa**, en lugar de intentar encontrar **patrones comunes en conjuntos de datos**.

*“Uncovering structure in an agent’s experience can certainly be useful in reinforcement learning, but by itself does not address the reinforcement learning problem of maximizing a reward signal. **We therefore consider reinforcement learning to be a third machine learning paradigm**, alongside supervised learning and unsupervised learning and perhaps other paradigms.”*

 Sutton & Barto. Reinforcement Learning. An introduction (2nd ed.)

# Evaluar vs. instruir

Los algoritmos de RL **evalúan**, no **instruyen**.

# Evaluar vs. instruir

Los algoritmos de RL **evalúan**, no **instruyen**.

**Instruir** (*instructive feedback*) consiste en indicar directamente cuál es la mejor decisión.

- Siempre se toma la acción que se asume óptima.
- Se basa en información completa del problema.
- Se atribuye al aprendizaje supervisado.

# Evaluar vs. instruir

Los algoritmos de RL **evalúan**, no **instruyen**.

**Instruir** (*instructive feedback*) consiste en indicar directamente cuál es la mejor decisión.

- Siempre se toma la acción que se asume óptima.
- Se basa en información completa del problema.
- Se atribuye al aprendizaje supervisado.

**Evaluar** (*evaluative feedback*) consiste en indicar cómo buena o mala ha sido una decisión, pero no si ha sido la mejor o peor posible.

- Motiva la exploración.
- Búsqueda de un comportamiento óptimo en base a información parcial del problema.
- Es propio del aprendizaje por refuerzo.

\* Ambos tipos de aprendizaje pueden combinarse.

## Bibliografía recomendada

---



## Libros

- Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT press.
- Morales, M. (2020). Grokking deep reinforcement learning. Manning Publications.
- Zai, A., & Brown, B. (2020). Deep reinforcement learning in action. Manning Publications.
- Szepesvári, C. (2010). Algorithms for reinforcement learning. Synthesis lectures on artificial intelligence and machine learning, 4(1), 1-103.

## Recursos web

- <https://github.com/huggingface/deep-rl-class>
- <https://spinningup.openai.com/en/latest/index.html>
- <https://julien-vitay.net/deeprl/>

## Cursos

- <https://youtu.be/2pWv7GOvuf0>
- <https://youtu.be/TCCjZe0y4Qc>
- <http://rll.berkeley.edu/deeprlcourse/>
- <https://youtu.be/nyjbcRQ-uQ8>
- <https://www.coursera.org/specializations/reinforcement-learning>

# Aprendizaje por refuerzo

## Introducción

---

Antonio Manjavacas Lucas

[manjavacas@ugr.es](mailto:manjavacas@ugr.es)