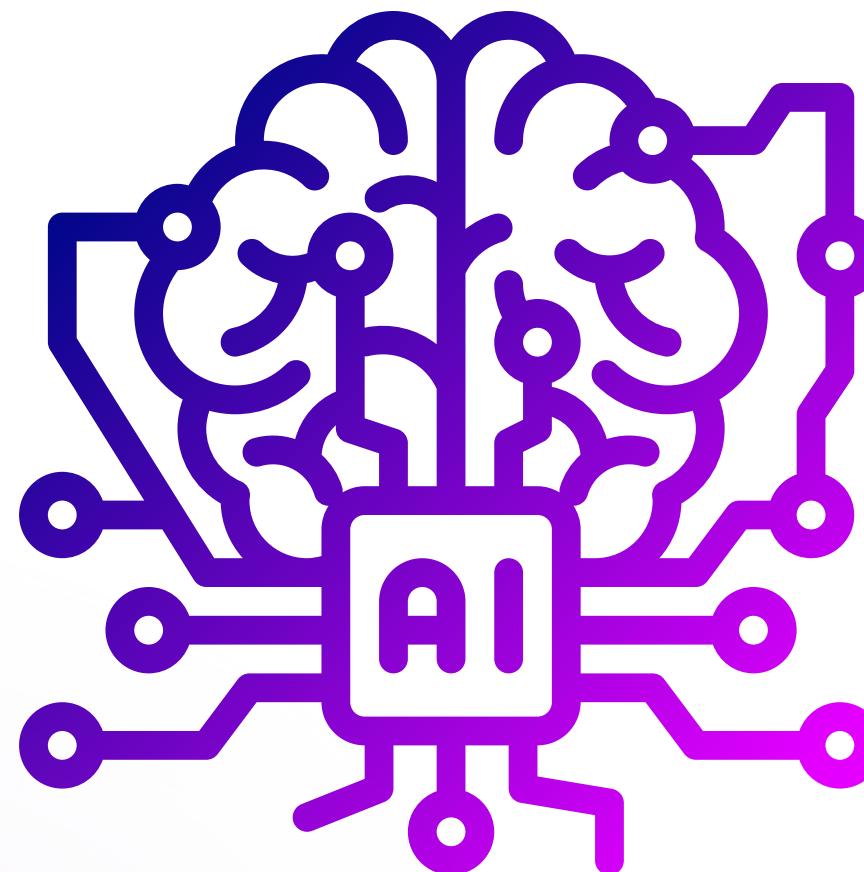




UNIVERSIDAD AUTONOMA DE BAJA CALIFORNIA  
FACULTAD DE CIENCIA QUIMICAS E INGENIERIA



# APRENDIZAJE POR REFUERZO

ALUMNA: SANDI GUILMA ROBLERO ESCALANTE  
MATERIA: INTELIGENCIA ARTIFICIAL  
MATRICULA: 1274524



# INDICE

01

## Aprendizaje por Refuerzo

**1.1 Definición**

**1.2 Estructura**

**1.3 Proceso de decisión de Márkov**

**1.4 Método de Aprendizaje por refuerzo**

02

## Aprendizaje de Refuerzo Profundo

**2.1 Definición**

**2.2 Limites de la tabla Q**

**2.3 Deep Q-learning**

03

## Referencias bibliográficas



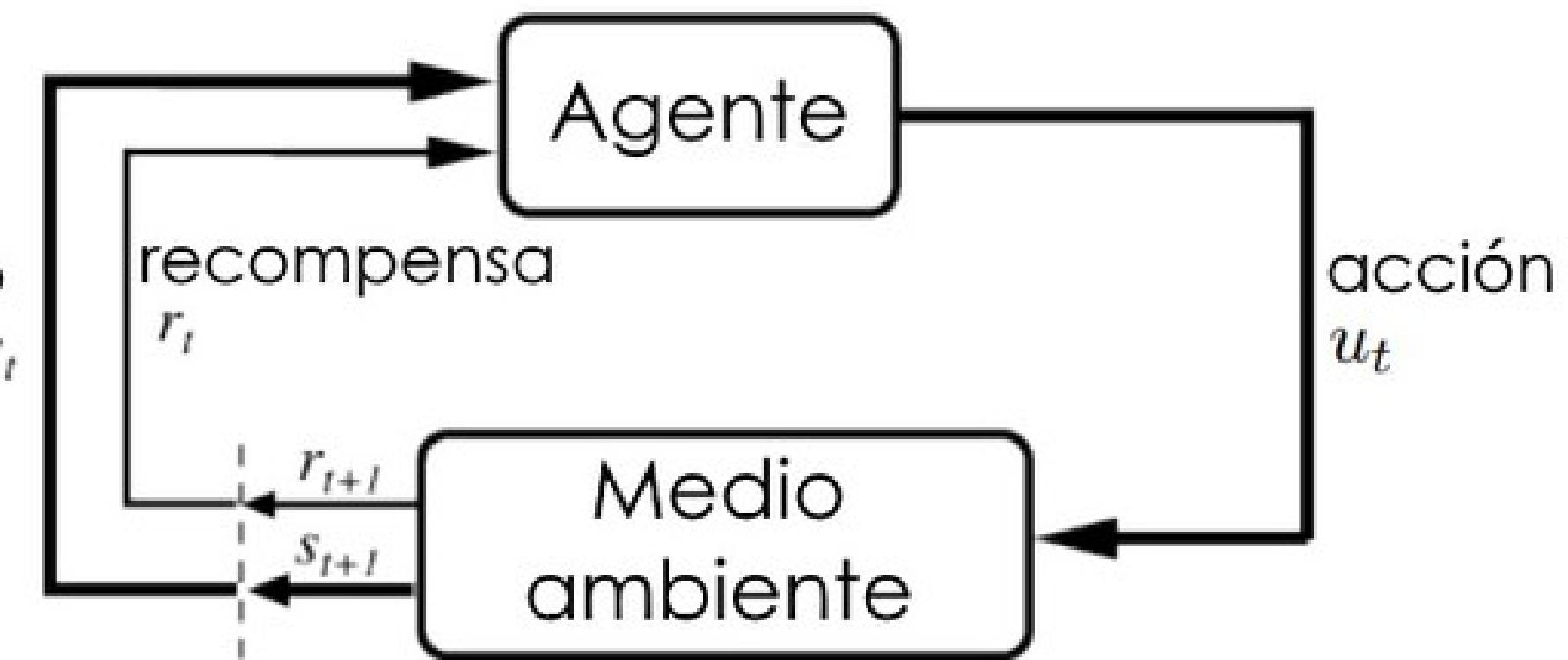


# Aprendizaje por Refuerzo

## Definición

El **aprendizaje por refuerzo** es una rama del machine learning en la cual la máquina guía su propio aprendizaje a través de recompensas y castigos. Es decir, consiste en un sistema de instrucción autónomo cuyo camino es indicado según sus aciertos y errores.

estado  
 $s_t$



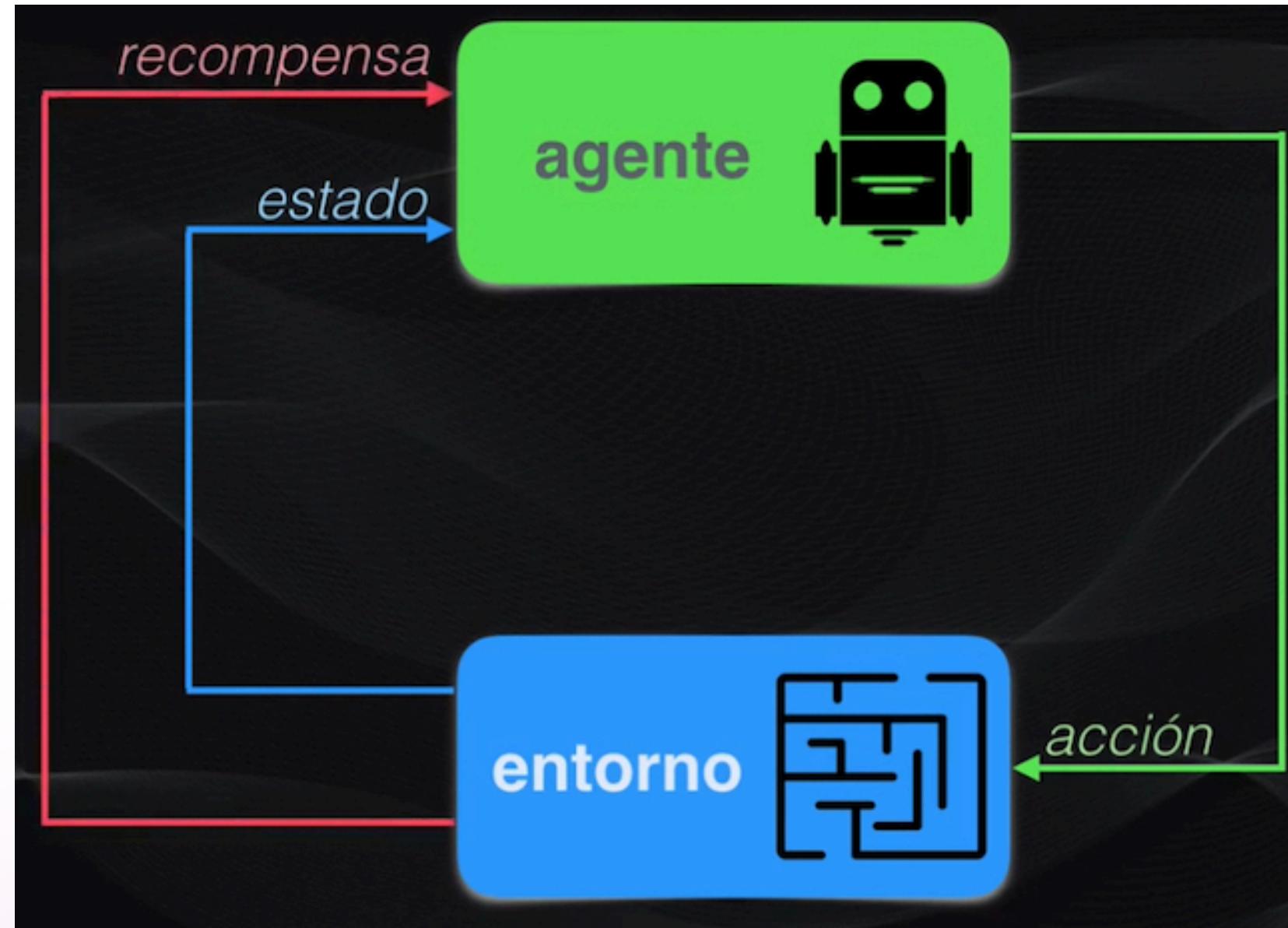
[Figure source: Sutton & Barto, 1998]

Un agente puede hacerse experto en un entorno desconocido, únicamente a partir de sus propias percepciones y recompensas ocasionales.

**Modelo de interacción del aprendizaje por refuerzo**



# Estructura



01

## Entorno

Es el mundo en el que el agente de aprendizaje por refuerzo opera. Puede ser un juego, un robot en una fábrica, un sistema financiero, etc.

02

## Agente

Es el "aprendiz" que interactúa con el entorno. El objetivo del agente es aprender a realizar acciones que maximicen una recompensa.

03

## Estados

Los estados representan la situación actual del entorno. Por ejemplo, en un juego de ajedrez, un estado sería la disposición actual de las piezas en el tablero.

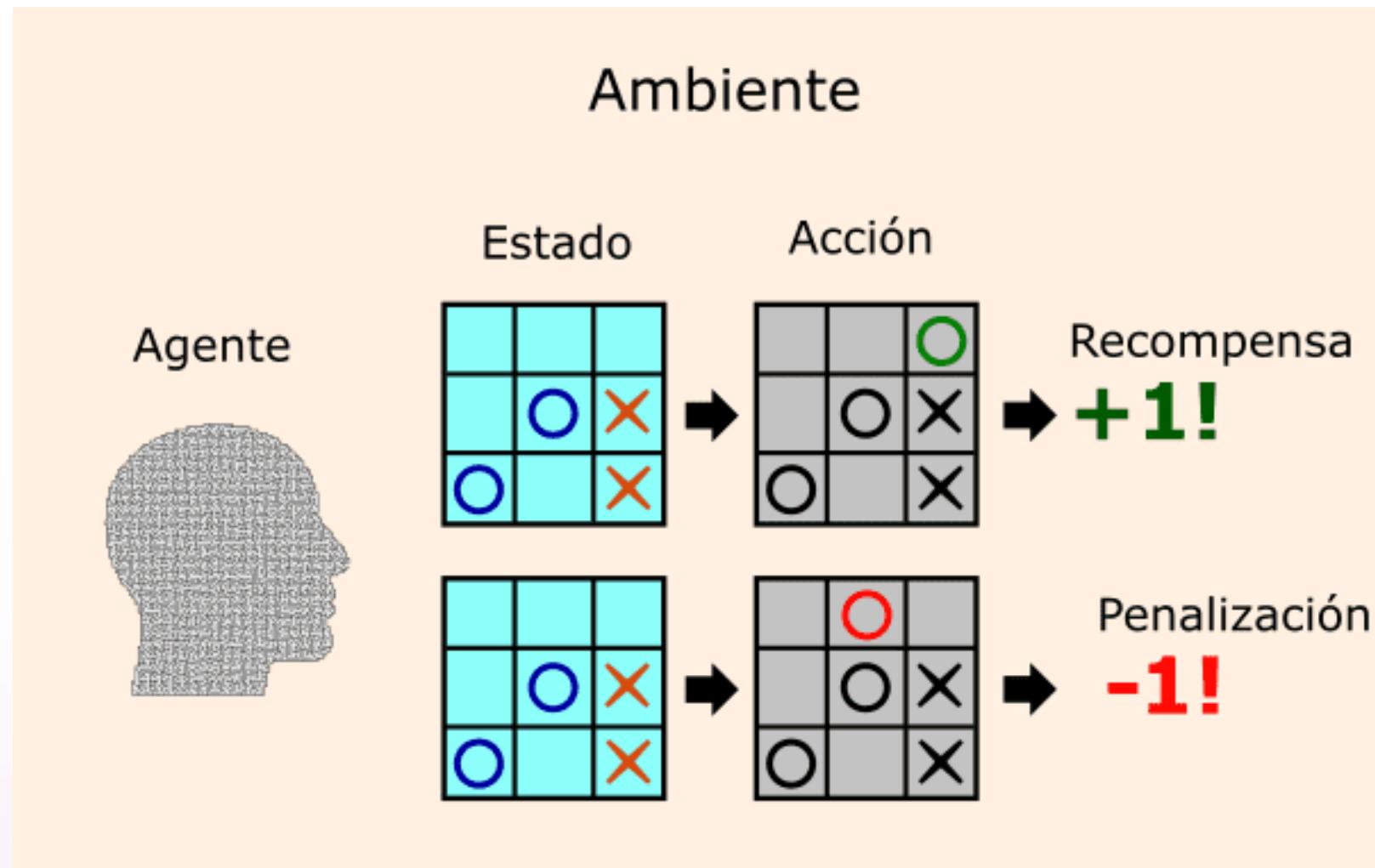
04

## Acciones

Son las posibles operaciones que el agente puede realizar en un determinado estado. En el ajedrez, esto sería mover una pieza a otra casilla.



# Estructura



05

## Recompensa

Es el retorno inmediato que recibe el agente tras realizar una acción

06

## Política

Es la estrategia que el agente sigue para decidir qué acción tomar en cada estado.

07

## Función de valor

Especifica el valor de un estado, que es la cantidad total de recompensa esperada a partir de ese estado

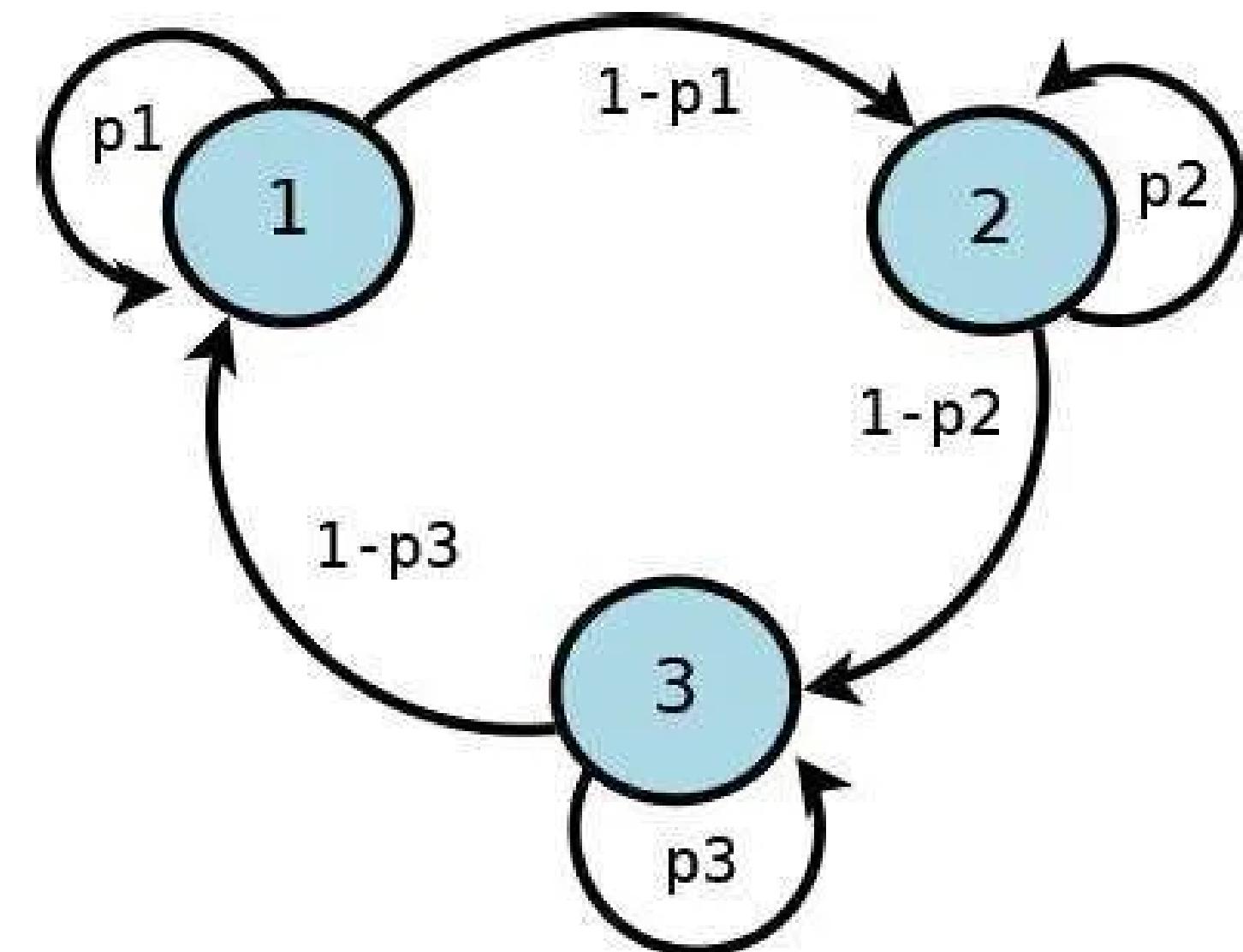


# Proceso de decisión de Márkov (MDP)

Llamado así por el matemático ruso Andrei Markov, es un modelo matemático mediante el cual podemos modelar la resolución de cierta clase de problemas relacionados con la toma de decisiones.

$$M = \langle S, A, P, R \rangle$$

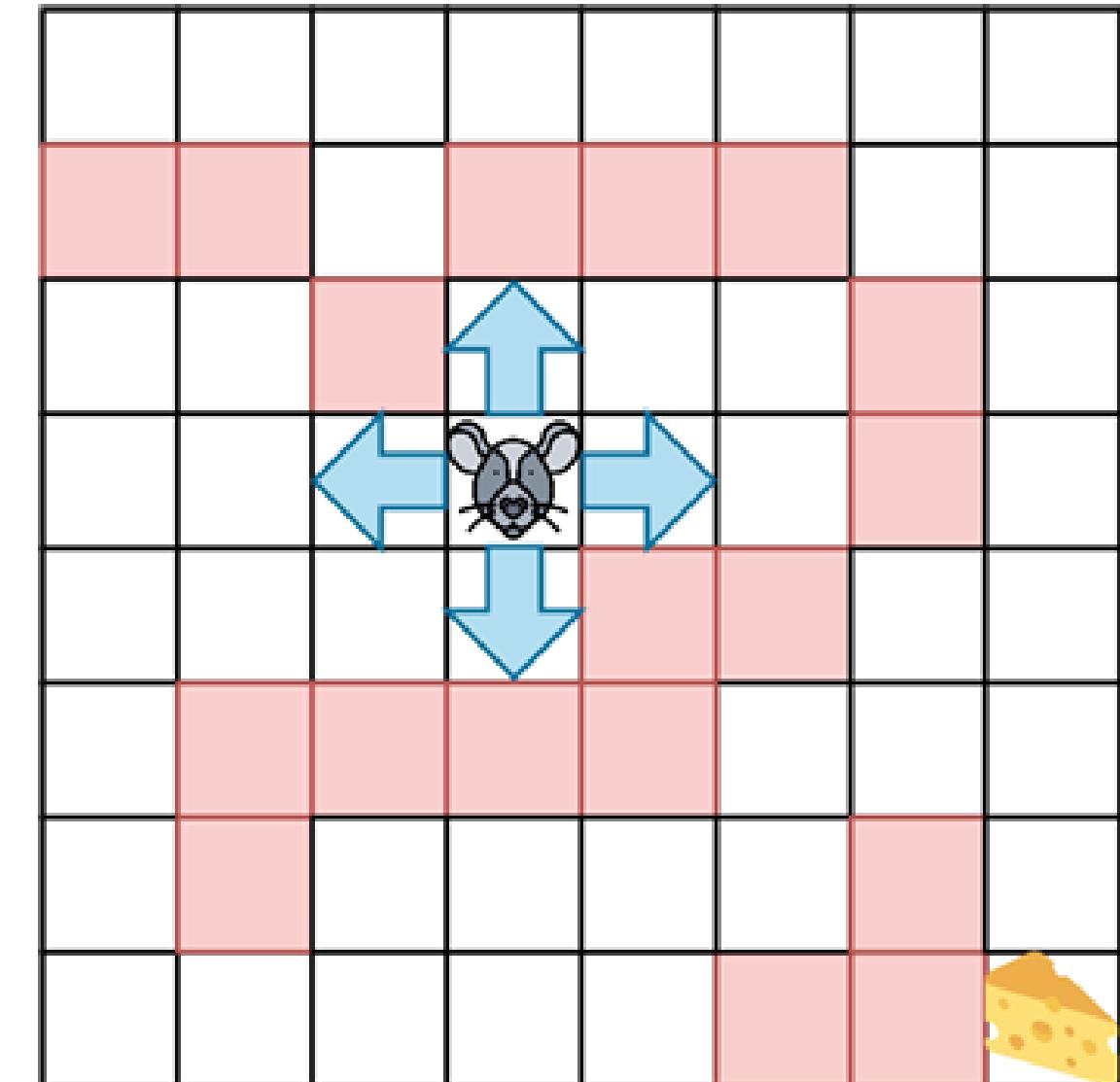
El objetivo de nuestro agente, el de encontrar una política óptima que nos ayude a tomar buenas decisiones acerca de qué acciones tomar en cada momento basándonos en el estado actual y conseguir maximizar nuestra "recompensa futura"





# Elementos

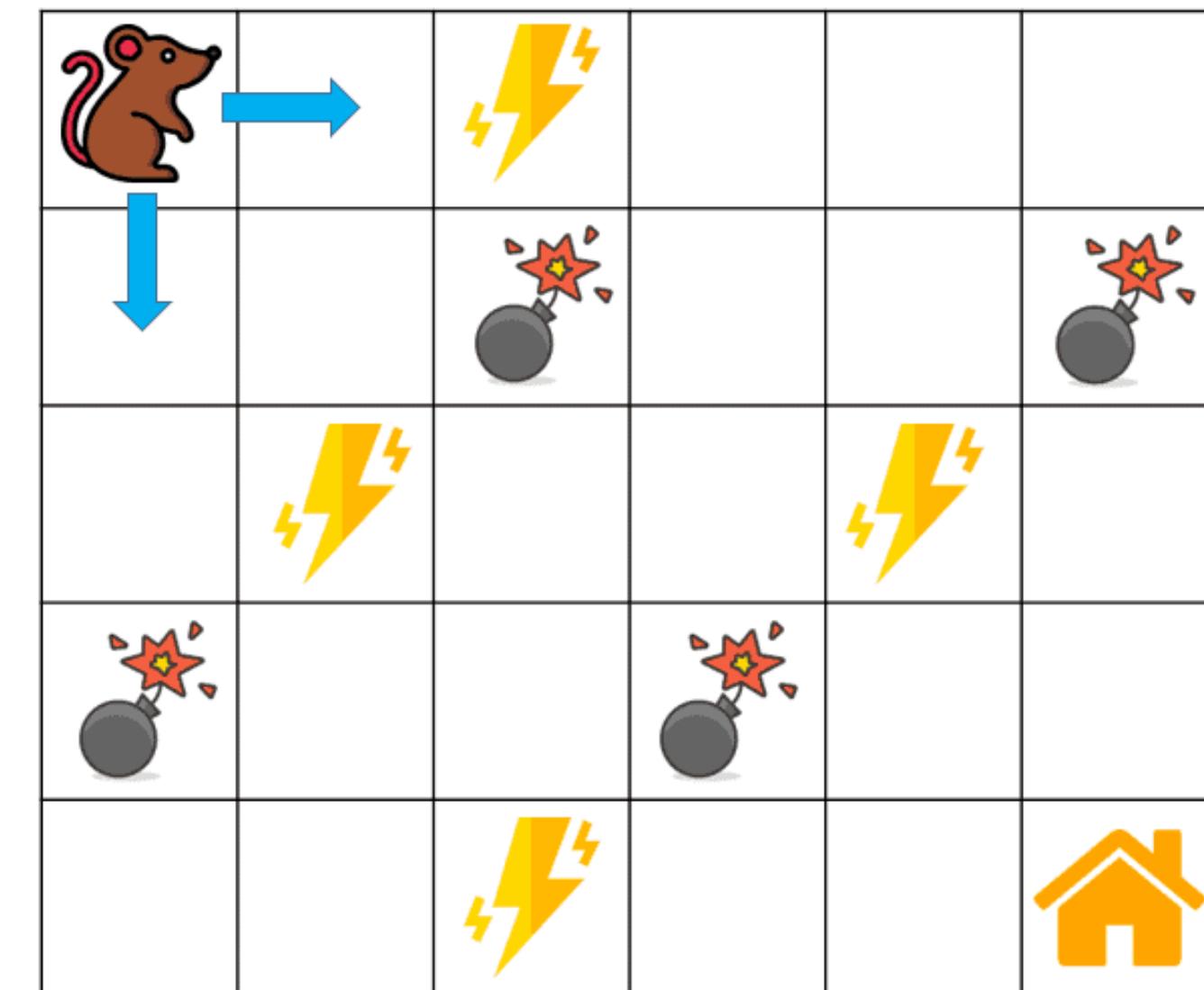
-  **S:** conjunto finito de estados en los que podemos encontrar en un determinado momento.
-  **A:** conjunto de acciones que pueden ser tomadas.
-  **Modelo de transición:** función que define el estado futuro ( $s'$ ) basándose en el estado actual ( $s$ ) y la acción tomada en dicho estado ( $a$ ). Es representado mediante la función  $T(s, a, s')$ .
-  **Recompensa:** “puntuación” positiva o negativa que se obtiene al tomar una acción en un determinado estado. Se representa mediante la función  $R(r | s, a)$ .





# Conceptos clave

- Paso: cuando un agente toma una acción y haga cambiar el entorno  
paso = (estado, acción, recompensa)
- Episodio: los pasos que suceden hasta llegar a un estado finalizado.





# ¿Cómo decidimos que acción es mejor tomar en cada estado?



Gracias a esto conseguimos que nuestra política aprenda a pensar, no solo en este momento, sino que accion tomar después de la misma.



# Problemas

Un inconveniente que podemos encontrar es que nuestra política piense demasiado a largo plazo y le lleve **un elevado número** de pasos encontrar una recompensa que merezca la pena, de modo que nuestra recompensa futura diverge.

## FACTOR DE DESCUENTO

Es un pequeño valor que se descuenta de la recompensa con el fin de evitar que, como acabamos de ver, nuestra política tome demasiados episodios antes de encontrar una recompensa positiva, Si las recompensas recibidas después de un tiempo  $t$  se denotan como:

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \dots + r_T$$

Si se tiene un punto terminal se llaman **tareas episódicas**, si no se tiene se llaman **tareas continuas**. En este último caso, la fórmula de arriba presenta problemas, ya que no podemos hacer el cálculo cuando  $T$  no tiene límite.

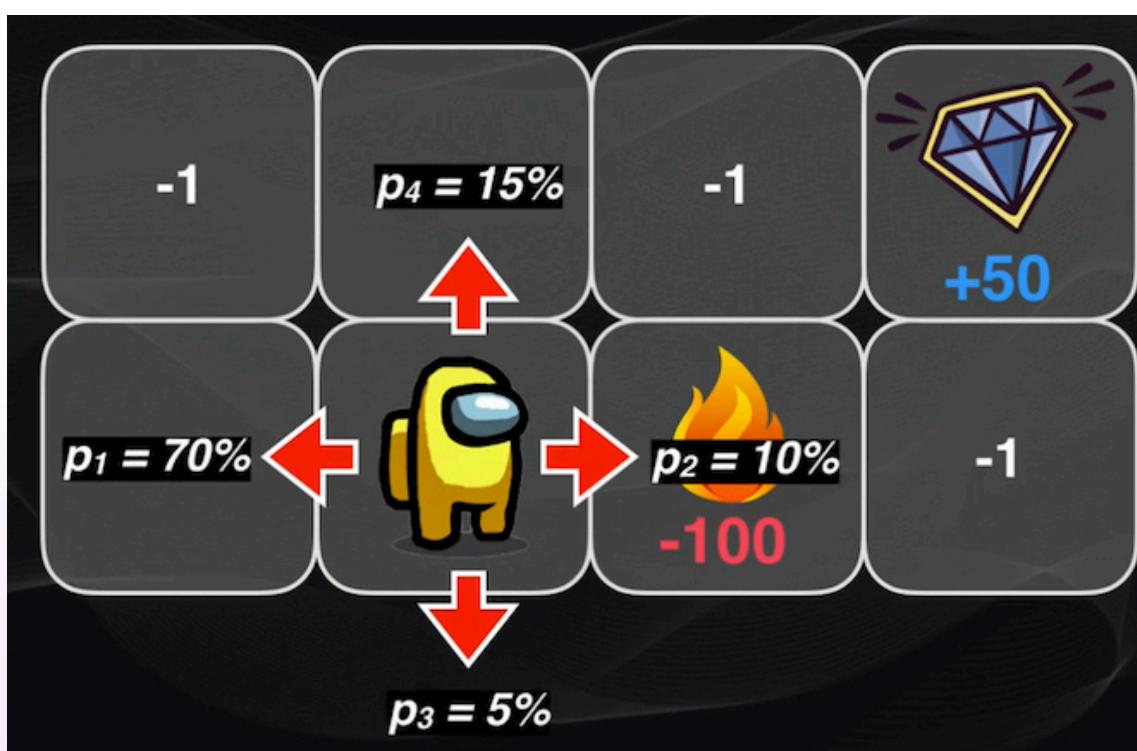
$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

donde ( $\gamma$ ) se conoce como la razón de descuento y está entre:  $0 \leq \gamma < 1$



# Problemas

**Exploración vs. Explotación**, nuestro agente contará con un hiperparámetro explotación, el cual determina con qué frecuencia el agente decide tomar una acción aleatoria (exploración) en lugar de aquella que le proporcione el valor más prometedor (explotación),



$\epsilon$ -greedy)

Es una estrategia que implica tomar una decisión en cada paso para tomar la acción registrada por el agente con una mayor recompensa o tomar una acción al azar. La probabilidad de que el agente tome una acción aleatoria se rige por el parámetro épsilon ( $\epsilon$ ).

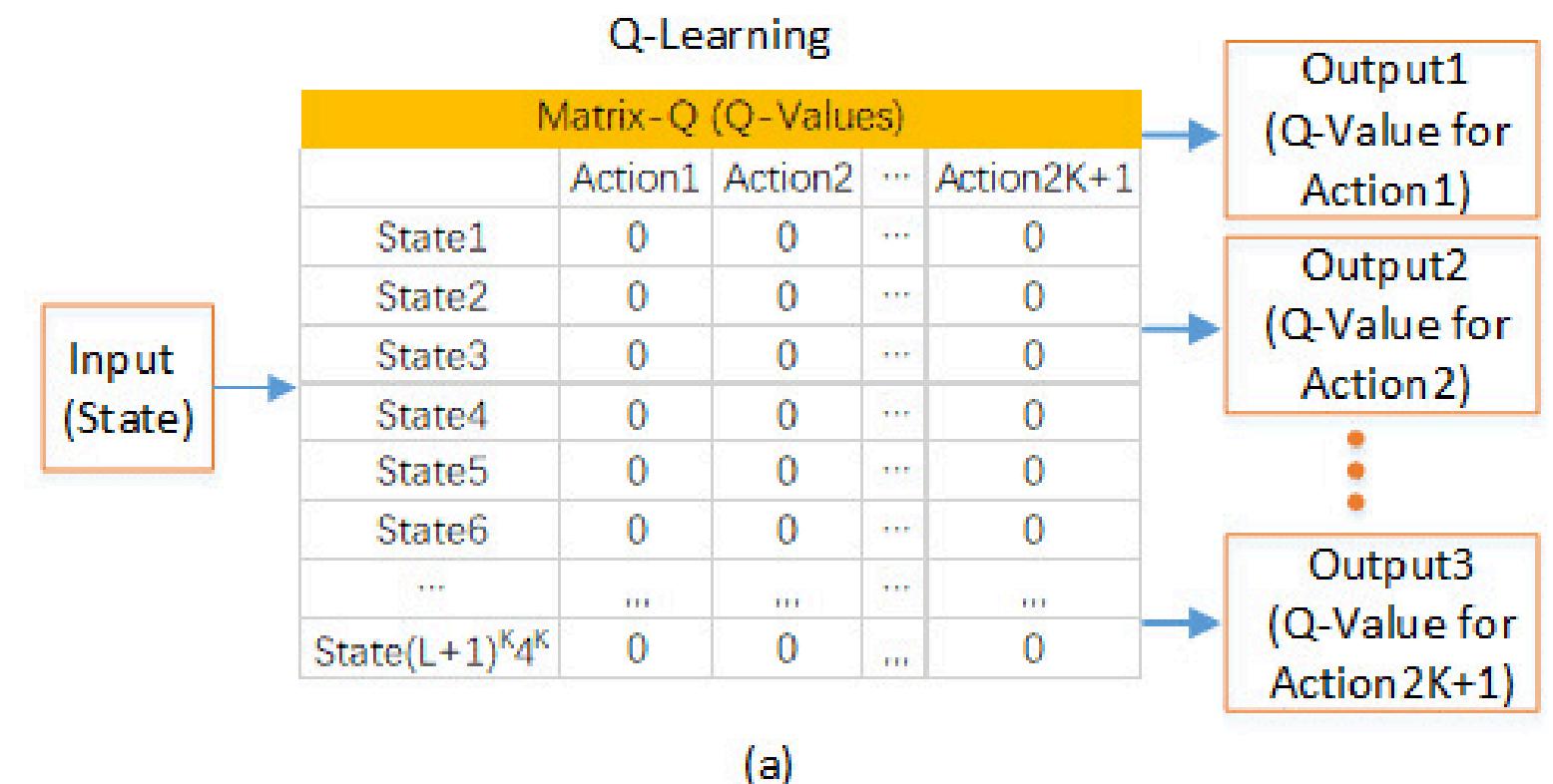
Se busca obtener un balance entre la exploración y explotación, con el objetivo de evitar que se explore demasiado, lo que conllevaría que nuestro agente no optimizase lo suficiente su política, ni que explote demasiado, lo que significaría caer en un mínimo local.

# Método de aprendizaje por refuerzo

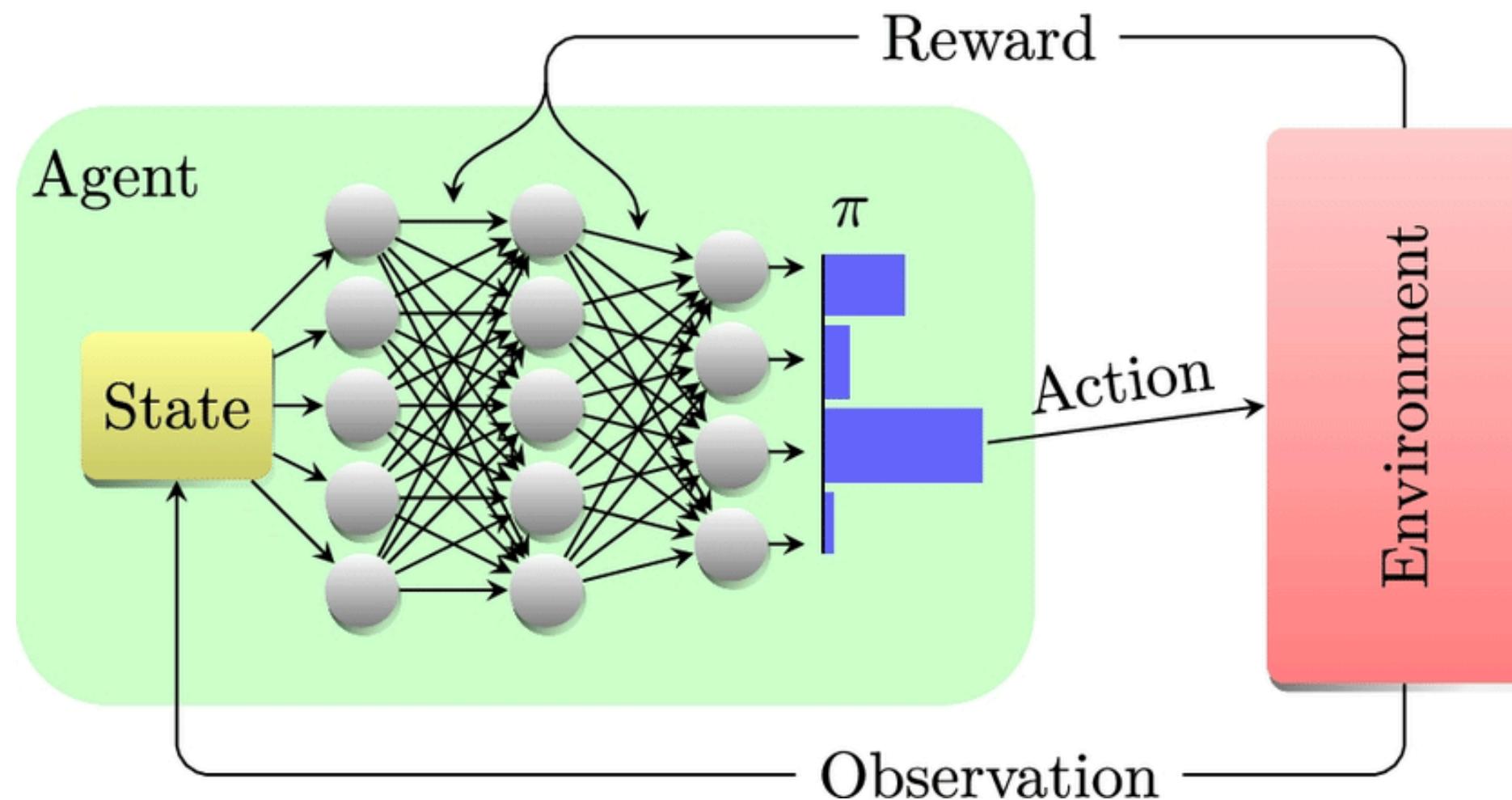
## Q-Learning

Se trata de un algoritmo *off-policy* con diferenciación temporal que busca encontrar una función acción-utilidad que nos dirá cómo de bueno es ejecutar una acción en un determinado estado. Los algoritmos *off-policy* son capaces de encontrar una política óptima independientemente de la política utilizada por el agente para elegir acciones, siempre que pase por todos los estados suficientes veces.

El algoritmo de Q-Learning cuenta con una tabla-Q con los estados posibles contemplados a partir del MDP, en la que se van almacenando las sumas de las posibles recompensas futuras. También conocidas como valores-Q, se predicen o actualizan usando el valor-Q del estado futuro  $s'$  y la acción a que más utilidad produzca.



$$Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_a' Q(s', a') - Q(s, a)]$$



# Aprendizaje de Refuerzo Profundo

## Definición

El aprendizaje de refuerzo profundo o Deep reinforcement learning (DRL) es un subcampo del aprendizaje automático que combina el aprendizaje reforzado con el aprendizaje profundo.

Los agentes de aprendizaje por refuerzo utilizan redes neuronales profundas para tomar decisiones en entornos complejos. Esto les permite aprender en situaciones en las que las reglas no son evidentes y las recompensas no son inmediatas, lo que otorga una gran capacidad de adaptación en tareas que requieren múltiples pasos o decisiones.



# Limites de la tabla Q

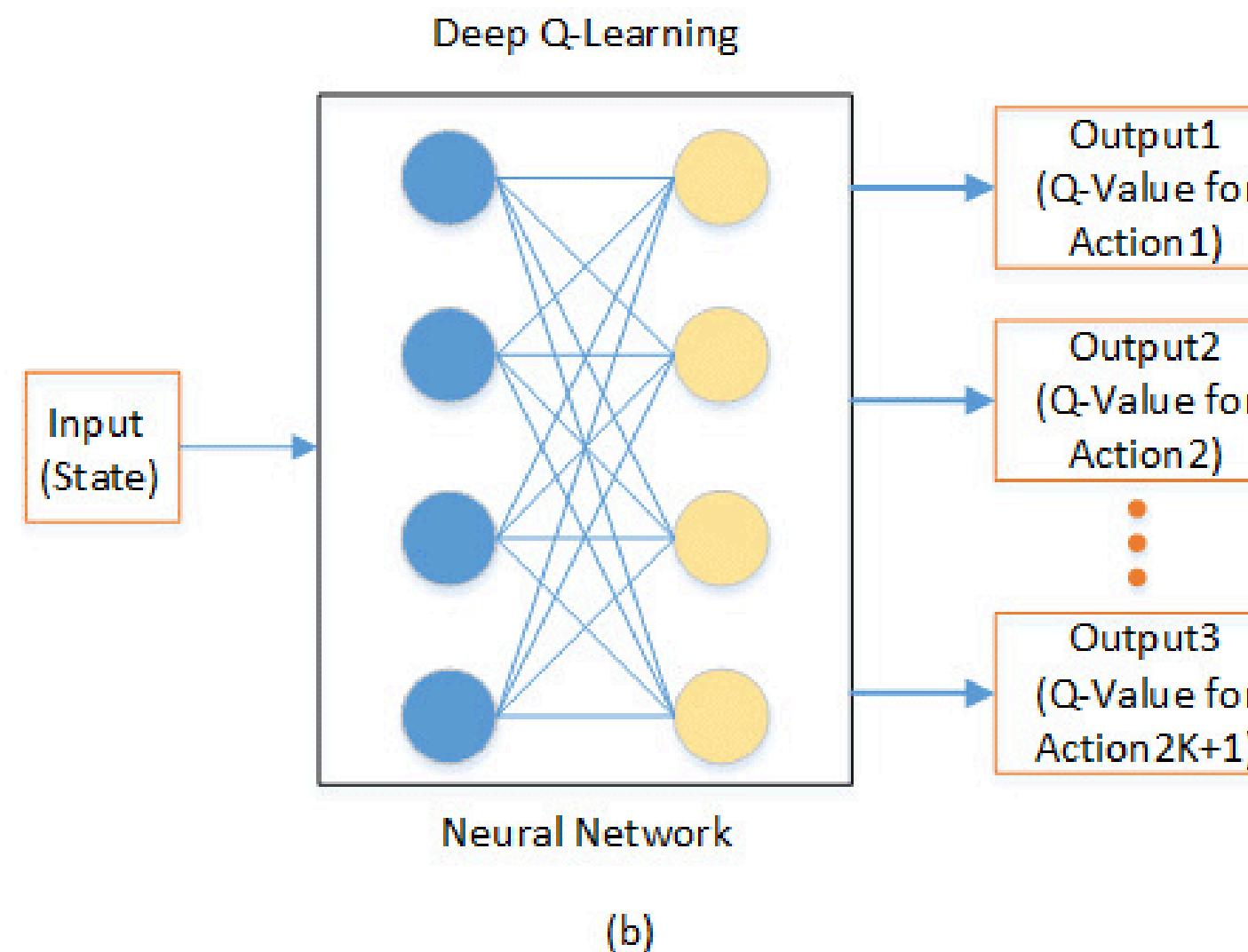
Este problema es conocido como la **maldición de la dimensión**, expresión que fue acuñada por Richard E. Bellman y que en el aprendizaje automático hace referencia al aumento exponencial en el tamaño de las tablas de estados en la memoria de los agentes, aunque también está presente en otros dominios como combinatoria, muestreo, optimización...

Por ejemplo, en el **juego de conecta 4** se dispone de un tablero de  $6 \times 7$  casillas, las cuales puedes estar vacías, ocupadas por una ficha roja u ocupadas por una ficha amarilla, lo que resulta en una combinación de  $3^{(7*6)} = 4531985219092$  estados diferentes.





# Deep Q-Learning



Al combinar estos dos enfoques, Deep Q-Learning utiliza redes neuronales para aproximar la función Q, lo que permite manejar problemas con espacios de estado y acción muy grandes y complejos.



# REFERENCIAS BIBLIOGRAFICAS

Blog de CEUPE. (2022, 4 de abril). Aprendizaje por refuerzo: Concepto, características y ejemplo. Ceupe. <https://www.ceupe.com/blog/aprendizaje-por-refuerzo.html>

Codificando Bits. (2021, 14 de febrero). El APRENDIZAJE REFORZADO: la guía DEFINITIVA [Video]. YouTube. <https://www.youtube.com/watch?v=qBtB-xcJp4c>

¿Cómo funciona el aprendizaje por refuerzo? (s.f.). OBS Business School. <https://www.obsbusiness.school/blog/como-funciona-el-aprendizaje-por-refuerzo>

El Aprendizaje por Refuerzo: guía introductoria. (s.f.). Codificando Bits. <https://codificandobits.com/blog/el-aprendizaje-reforzado-la-guia-introductoria/>

Merino, M. (2019, 27 de enero). Conceptos de inteligencia artificial: qué es el aprendizaje por refuerzo. Xataka - Tecnología y gadgets, móviles, informática, electrónica. <https://www.xataka.com/inteligencia-artificial/conceptos-inteligencia-artificial-que-aprendizaje-refuerzo>

(s.f.). Docta Complutense :: Home. <https://docta.ucm.es/rest/api/core/bitstreams/caa804ba-a240-4e45-aafc-182c05f80d71/content>