



PONTIFICIA  
UNIVERSIDAD  
CATÓLICA DE  
VALPARAÍSO

[pucv.cl](http://pucv.cl)

# Robótica e inteligencia artificial

Módulo 4  
Inteligencia artificial S26

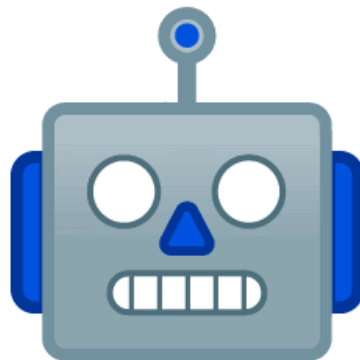
# INTELIGENCIA ARTIFICIAL

## SESIÓN 26

# Aprendizaje Reforzado (Reinforcement Learning)

El aprendizaje reforzado es una de las ramas más importantes del aprendizaje profundo. El objetivo es construir un modelo con un agente que mejora su rendimiento, basándose en la recompensa obtenida del entorno con cada interacción que se realiza. La recompensa es una medida de lo correcta que ha sido una acción para obtener un objetivo determinado. El agente utiliza esta recompensa para ajustar su comportamiento futuro, con el objetivo de obtener la recompensa máxima.

## Online Reinforcement Learning



Agent



Environment

# Aprendizaje Reforzado (Reinforcement Learning)

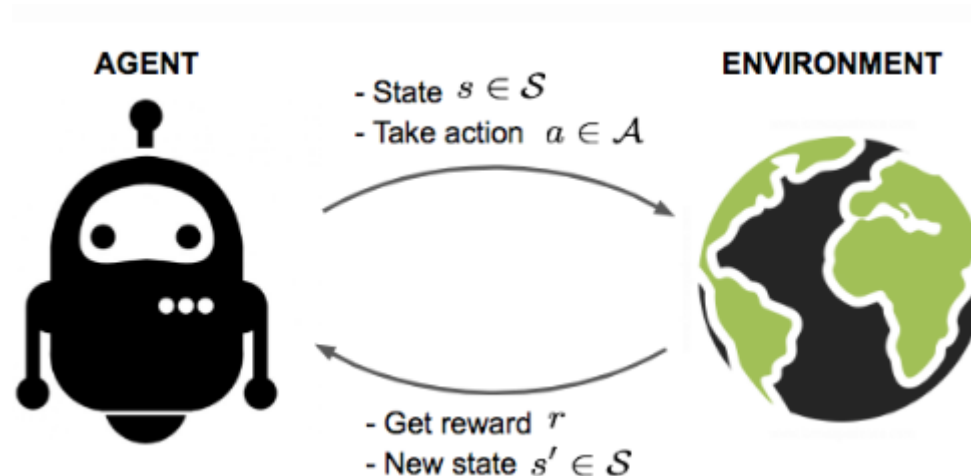
Todo problema de aprendizaje por refuerzo está compuesto por un agente y un entorno.

El agente debe percibir su entorno y entrenarse en él hasta alcanzar el objetivo y desempeño deseado.

El entorno estructurado como una serie de alternativas desde donde la gente recibe la información necesaria para su aprendizaje.

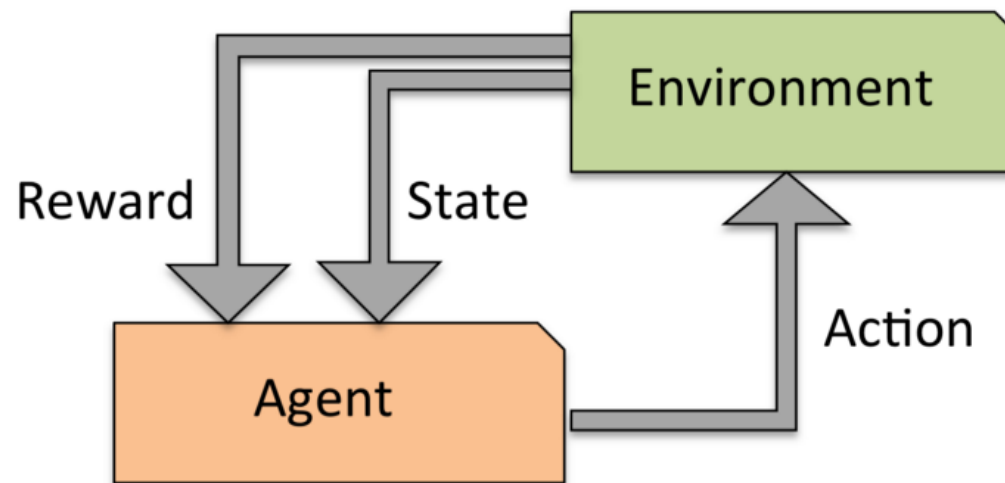
Un agente tiene distintos Estados en los que puede estar dependiendo del entorno y la acción realizada anteriormente.

El agente en cada momento tiene una serie de acciones disponibles dependiendo el estado en el que se encuentre.



# Aprendizaje Reforzado (Reinforcement Learning)

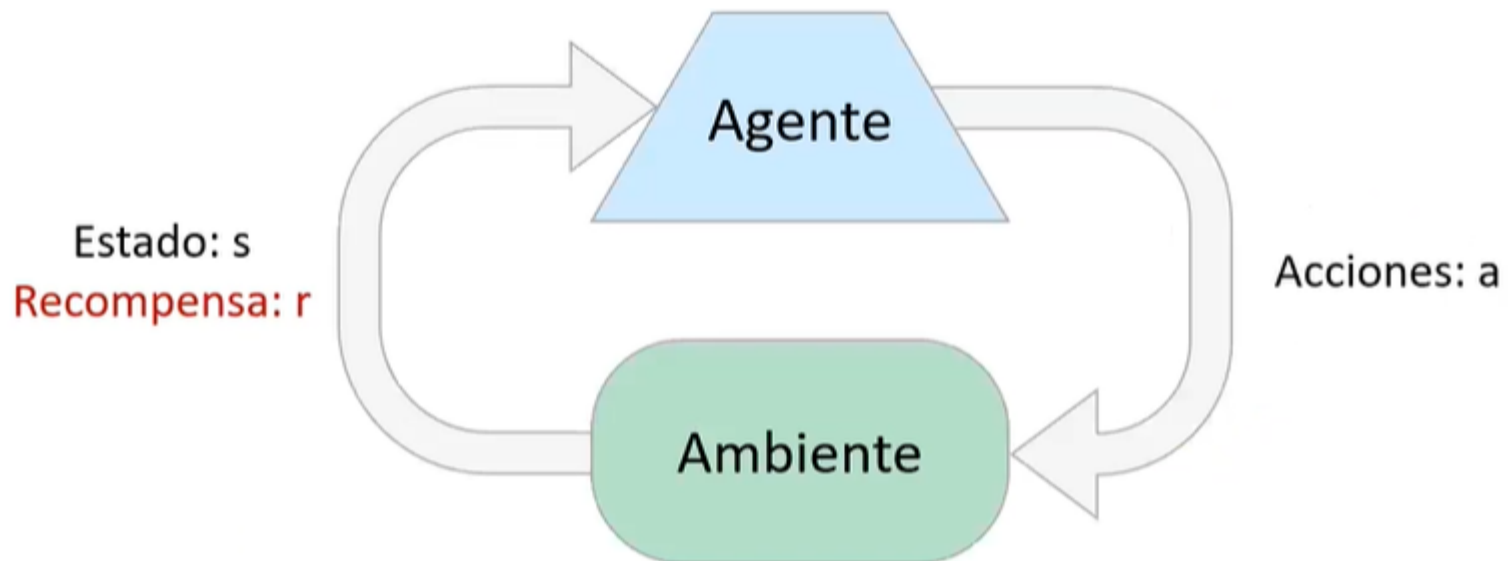
Un ejemplo común es una máquina de ajedrez, donde el agente decide entre una serie de posibles acciones, dependiendo de la disposición del tablero (que es el estado del entorno) y la recompensa se recibe según el resultado de la partida.



# Aprendizaje Reforzado (Reinforcement Learning)

Idea básica:

- Recibir feedback en la forma de recompensas.
- La función de utilidad del agente está definida por la función de recompensa.
- Debe actuar para maximizar las recompensas esperadas.
- Todo el aprendizaje está basado en muestras de observaciones de resultados.



# Aprendizaje Reforzado (Reinforcement Learning)

Ejemplo: aprendiendo a caminar



Inicial



Entrenando



Después de aprender  
[1K Intentos]

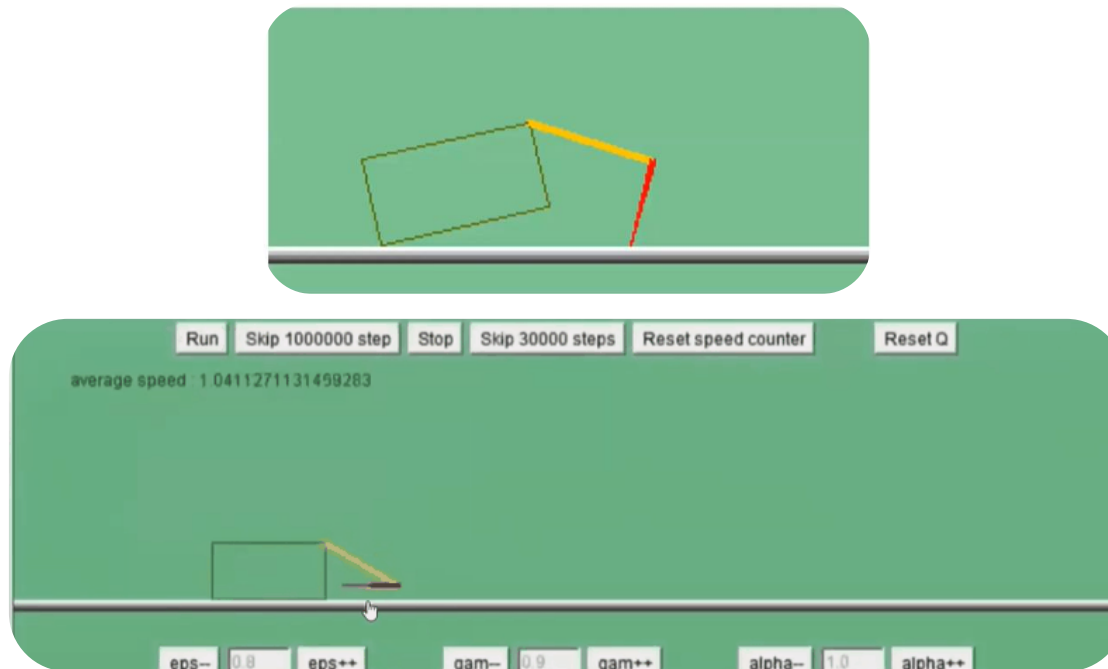


# Aprendizaje Reforzado (Reinforcement Learning)

Otro ejemplo denominado el arrastrado, ilustra como un cuerpo sencillo, el cual posee el objetivo de desplazarse hacia adelante logra aprender por medio de una gran cantidad de intentos.

Este agente posee solo dos variables a controlar, las cuales son los ángulos que se pueden formar con la extremidad de 2 GDL.

La recompensa positiva es cuando se registra un avance hacia la derecha, y la recompensa negativa cuando este se desplaza a la izquierda





# Aprendizaje Reforzado (reinforcement Learning)

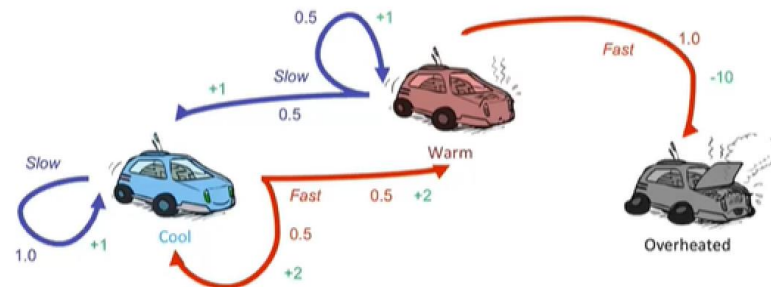
MDP (Markov decision process)

En matemáticas, un proceso de decisión de Markov es un proceso de control estocástico en tiempo discreto. Proporciona un marco matemático para modelar la toma de decisiones en situaciones donde los resultados son en parte aleatorios y en parte están bajo el control de quien toma las decisiones.

Se define:

- Un set de estados  $s \in S$
- Un set de acciones (por estado)  $A$
- Un modelo de transiciones  $T(s,a,s')$
- Una función de recompensa  $R(s,a,s')$

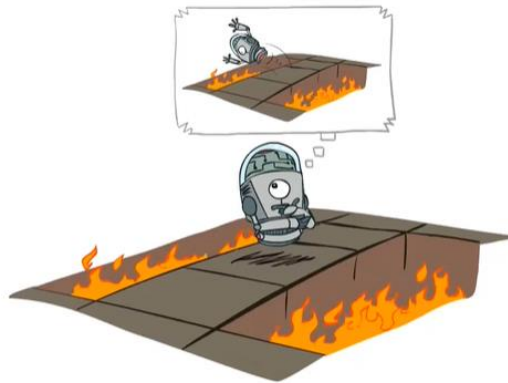
Se busca una política óptima que permita resolver el problema con la mayor eficiencia posible.



# Aprendizaje Reforzado (Reinforcement Learning)

En el aprendizaje reforzado no se conoce el modelo de transición  $T(s,a,s')$  ni la función de recompensa  $R(s,a,s')$ .

Se debe probar acciones y estados que permitan la obtención de información y posteriormente el aprendizaje.



Solución  
Offline

MDP

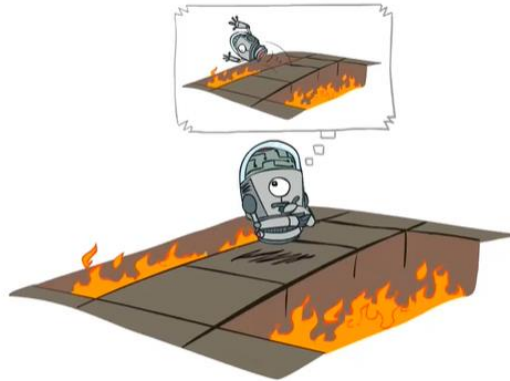


Aprendizaje  
Online

RL

# Aprendizaje Reforzado (Reinforcement Learning)

Básicamente en el aprendizaje reforzado se debe adquirir información desde la experiencia, ejecutando acciones que permitan identificar cómo alcanzar o no las recompensas.



Solución  
Offline

MDP



Aprendizaje  
Online

RL

# Aprendizaje basado en modelos

Idea basada en modelos:

- Aprender un modelo aproximado basado en experiencias.
- resuelve para valores como si el modelo aprendido fuese correcto.

Paso 1: Aprende un modelo MDP empírico

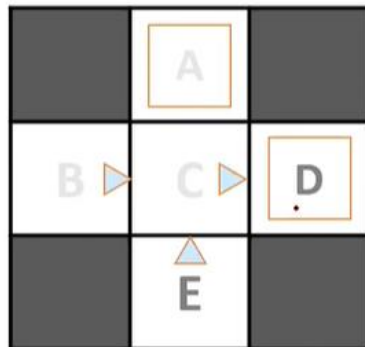
- Cuenta las salidas  $s'$  para cada  $s, a$
- Normaliza para obtener una estimación de función de transición  $T(s, a, s')$
- Descubre cada función de recompensas  $R(s, a, s')$  cuando experimenta  $(s, a, s')$

# Aprendizaje basado en modelos

Ejemplo: Aprendizaje basado en modelos

Cada episodio es una nueva experiencia ejecutada para obtener información útil para superar el desafío.

Política de  
Entrada  $\pi$



Asume:  $\gamma = 1$

Episodios Observados (Entrenamiento)

Episodio 1

B, este, C, -1  
C, este, D, -1  
D, salida, x, +10

Episodio 2

B, este, C, -1  
C, este, D, -1  
D, salida, x, +10

Episodio 3

E, norte, C, -1  
C, este, D, -1  
D, salida, x, +10

Episodio 4

E, norte, C, -1  
C, este, A, -1  
A, salida, x, -10

Modelo Aprendido

$\hat{T}(s, a, s')$

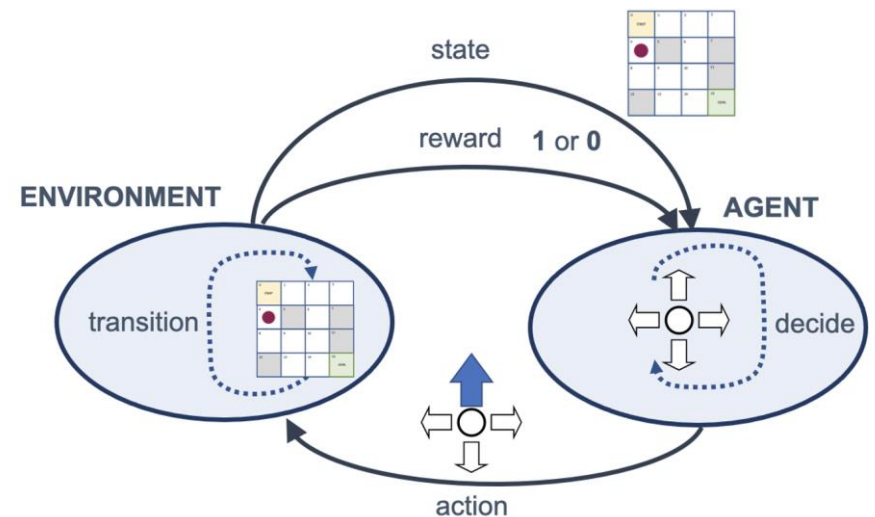
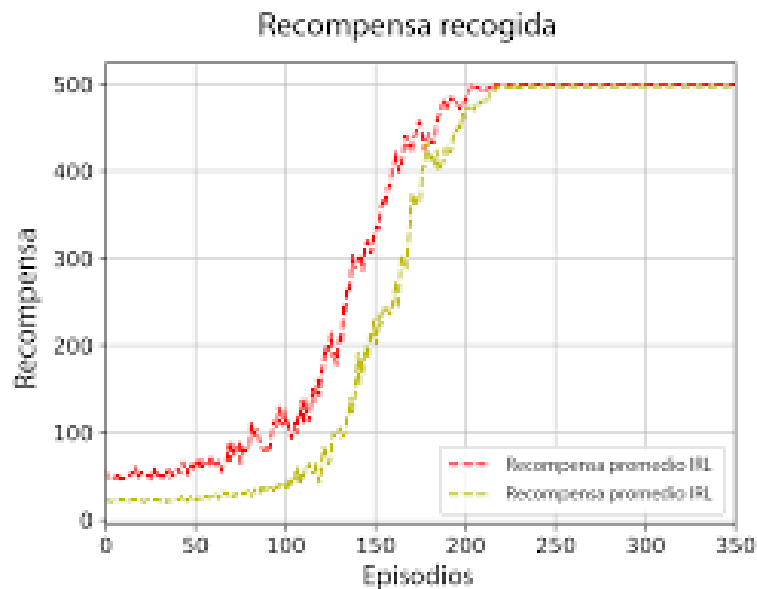
$T(B, \text{este}, C) = 1.00$   
 $T(C, \text{este}, D) = 0.75$   
 $T(C, \text{este}, A) = 0.25$   
...

$\hat{R}(s, a, s')$

$R(B, \text{este}, C) = -1$   
 $R(C, \text{este}, D) = -1$   
 $R(D, \text{salida}, x) = +10$   
...

# Aprendizaje reforzado

El agente no sabe a priori el estado que alcanzará, la recompensa que recibirá. Estos solo dependen del Estado actual y de la acción tomada. Entonces se trata de que la gente vaya a cumplir dando su conocimiento del entorno a medida que acumula recompensas, de manera que encuentre una secuencia de acciones que le proporcione la mayor recompensa acumulada.

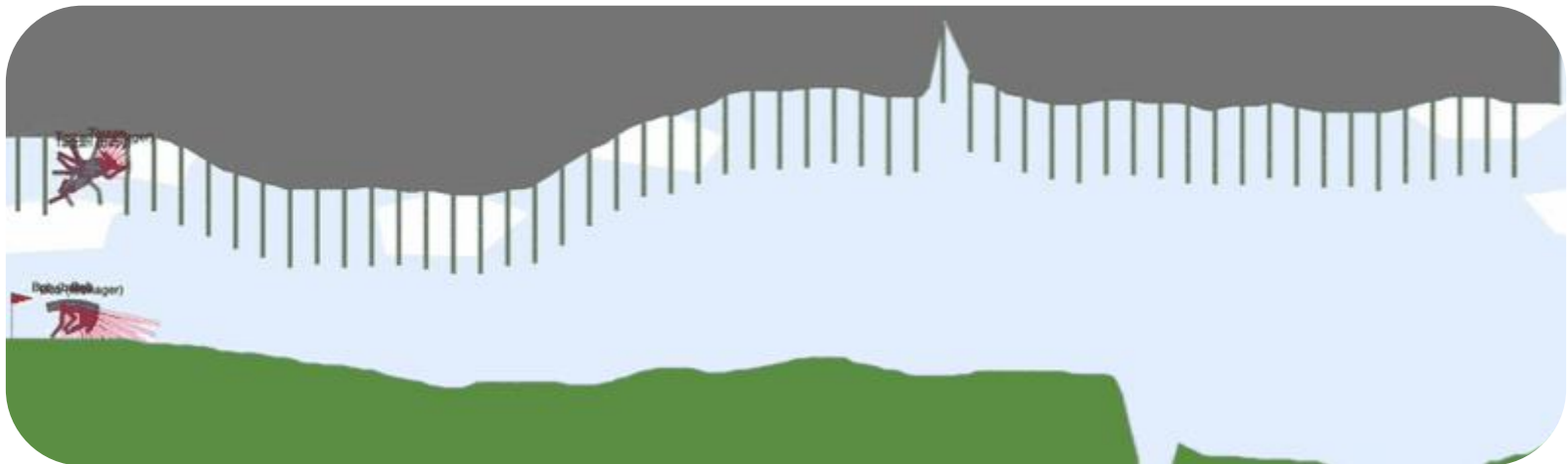


# Aprendizaje reforzado

Existen algunas plataformas web que permiten la interacción con sistemas de estudio del aprendizaje reforzado.

a continuación se presenta una plataforma web que permite Disponer distintos tipos de agentes con distintas morfologías, con el objetivo de desplazarse en un terreno 2D irregular.

los agentes han sido entrenados con diferentes algoritmos para aprender con éxito el movimiento necesario para acumular la máxima recompensa.



# Aprendizaje reforzado

La herramienta permite crear escenarios propios para identificar como el agente ya entrenado puede alcanzar el objetivo de cruzar hasta el extremo derecho.



Fuente: [https://developmentalsystems.org/Interactive\\_DeepRL\\_Demo/](https://developmentalsystems.org/Interactive_DeepRL_Demo/)



# Q learning

El Q-learning es una técnica de aprendizaje por refuerzo que busca encontrar una política óptima (una secuencia de acciones) maximizando el valor de la recompensa total sobre cada paso desde el estado actual.

Una tabla q es una tabla de búsqueda simple donde se calcula las máximas recompensas futuras esperadas para esa acción en cada estado.

Brinda información sobre la mejor acción a realizar en cada estado.

Inicialmente, sus valores tendrán el valor de cero e Irán actualizándose con el entrenamiento.

**Game Board:**



Current state (s):  
0 0 0  
0 1 0

**Q Table:**

$\gamma = 0.95$

	0 0 0 1 0 0	0 0 0 0 1 0	0 0 0 0 0 1	1 0 0 0 0 0	0 1 0 0 0 0	0 0 1 0 0 0
↑	0.2	0.3	1.0	-0.22	-0.3	0.0
↓	-0.5	-0.4	-0.2	-0.04	-0.02	0.0
→	0.21	0.4	-0.3	0.5	1.0	0.0
←	-0.6	-0.1	-0.1	-0.31	-0.01	0.0

# Q learning

La función Q usa la ecuación de Bellman y lo recibe dos entradas, un estado “s”, y una acción “a”. El objetivo es maximizar la función Q.

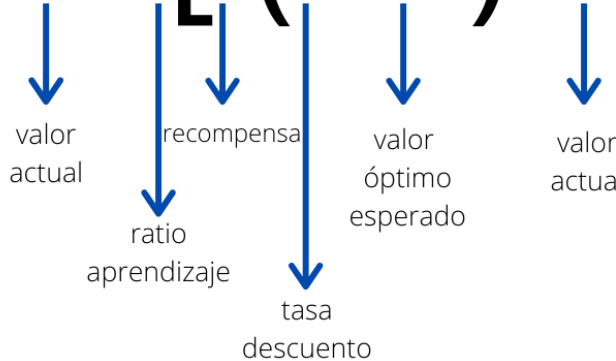
$$\hat{Q}(s,a) = Q(s,a) + \alpha \left[ R + \left( \lambda \max_{s'} Q(s',a) \right) - Q(s,a) \right]$$


Diagram illustrating the components of the Q-learning Bellman equation:

- $Q(s,a)$ : valor actual
- $\alpha$ : ratio aprendizaje
- $R$ : recompensa
- $\lambda$ : tasa descuento
- $\max_{s'} Q(s',a)$ : valor óptimo esperado
- $Q(s,a)$  (inside brackets): valor actual

Usando esta función se obtendrán los valores de Q para cada celda . inicialmente todos los valores son cero.

a medida que el aprendizaje avanza los valores de Q se Irán actualizando, dando cada vez una mejor aproximación.

# Q learning

Cuando comienza a ser probado cada episodio, se debe definir dos instancias, la primera correspondiente a la instancia de aprendizaje donde la gente realizará movimientos que permitan obtener información del ambiente y por ende identificar qué acciones permitirán alcanzar el objetivo, y luego la instancia de ejecución, para alcanzar el objetivo concretamente.

para cambiar entre una instancia y otra se debe disminuir el ratio de aprendizaje  $\alpha$ .

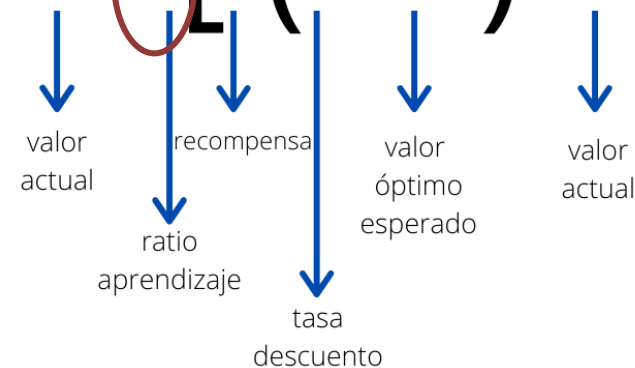
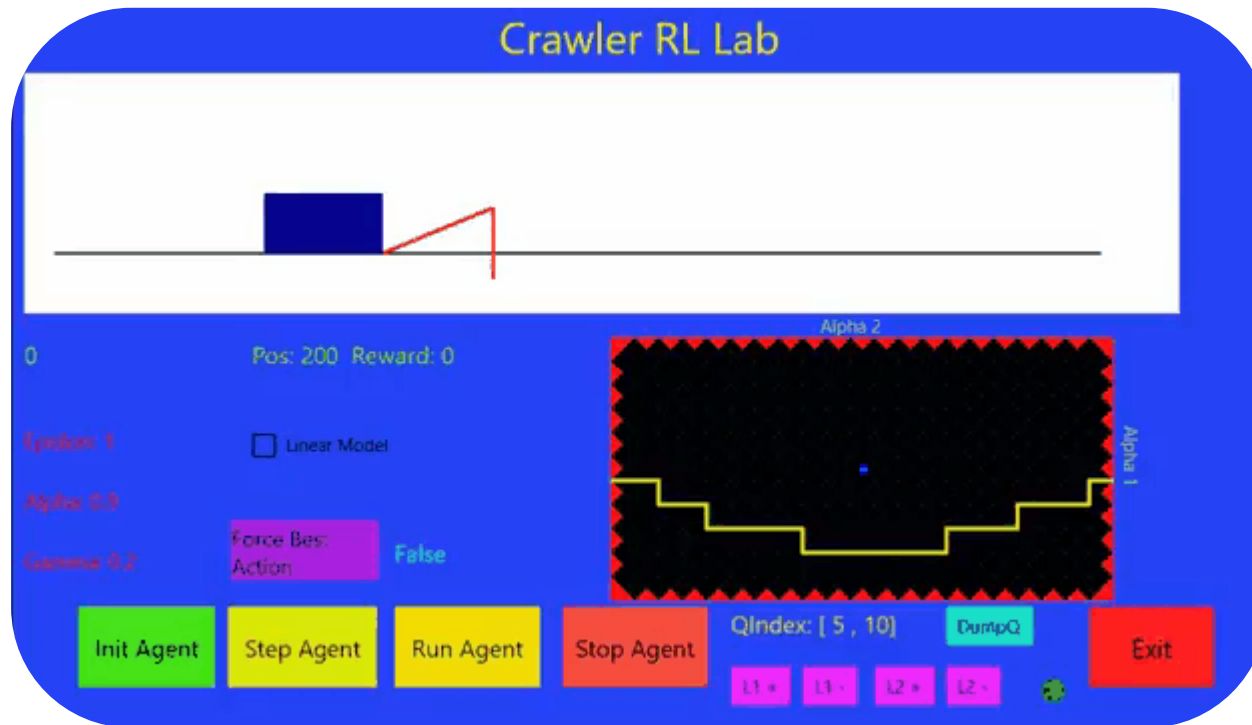
$$\hat{Q}(s,a) = Q(s,a) + \alpha [R + (\lambda \max_{s'} Q(s',a)) - Q(s,a)]$$


Diagram illustrating the Q-learning update equation with variable mappings:

- $\hat{Q}(s,a)$ : valor actual
- $Q(s,a)$ : valor actual
- $\alpha$ : ratio aprendizaje
- $R$ : recompensa
- $\lambda$ : tasa descuento
- $\max_{s'} Q(s',a)$ : valor óptimo esperado
- $Q(s,a)$  (inside brackets): valor actual

# Q learning

El siguiente video muestra cómo un objeto denominado “crawler” el cual tiene por objetivo arrastrarse hacia la derecha con el uso de una extremidad con dos grados de libertad (alpha 1 y alpha 2), Comienza a generar acciones que permiten completar la tabla Q.



# Q learning

La tabla de valores Q se irá actualizando a medida que y la gente obtenga información del entorno.

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0 R: 1 t	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0