

Reinforcement Learning

Introduction

Antonio Ricciardi, PhD



GLADIA



SAPIENZA
UNIVERSITÀ DI ROMA

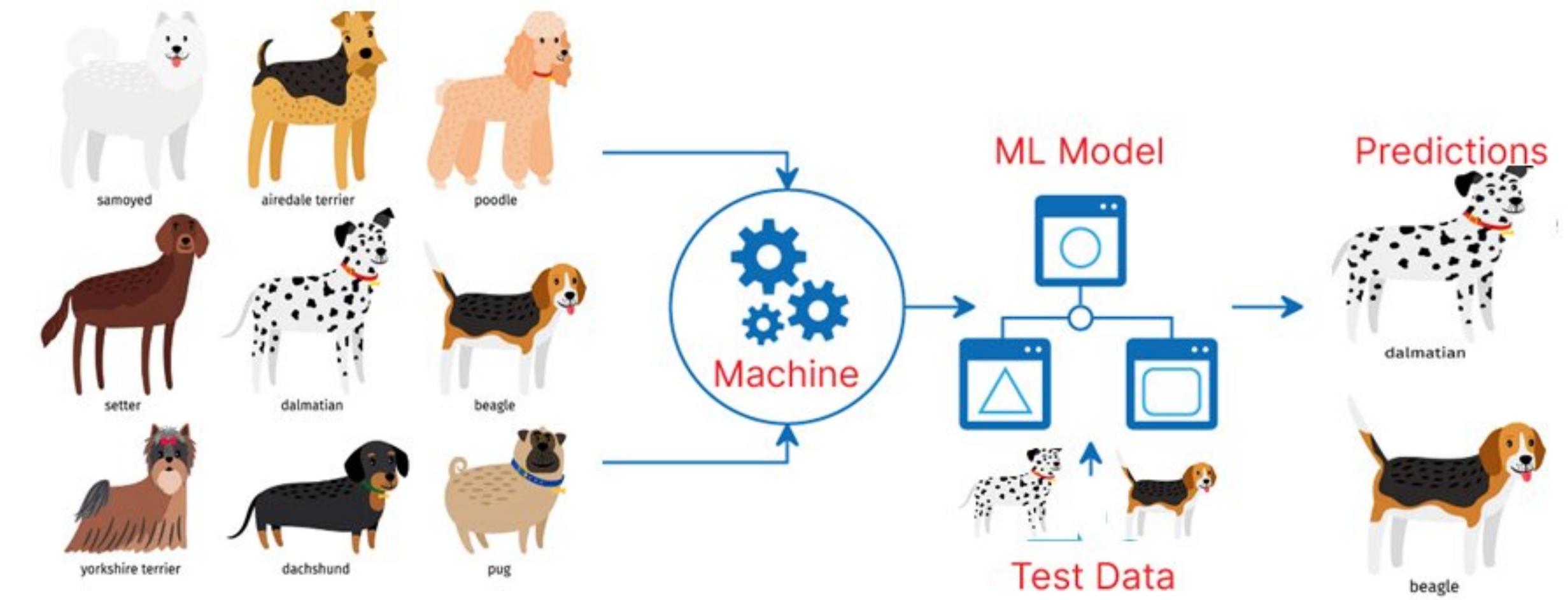
Machine Learning @ Mathematical Sciences for AI

Supervised Learning vs Reinforcement Learning

Supervised Learning

Learn using labeled data

Need pairs $\{x_i, y_i\}$



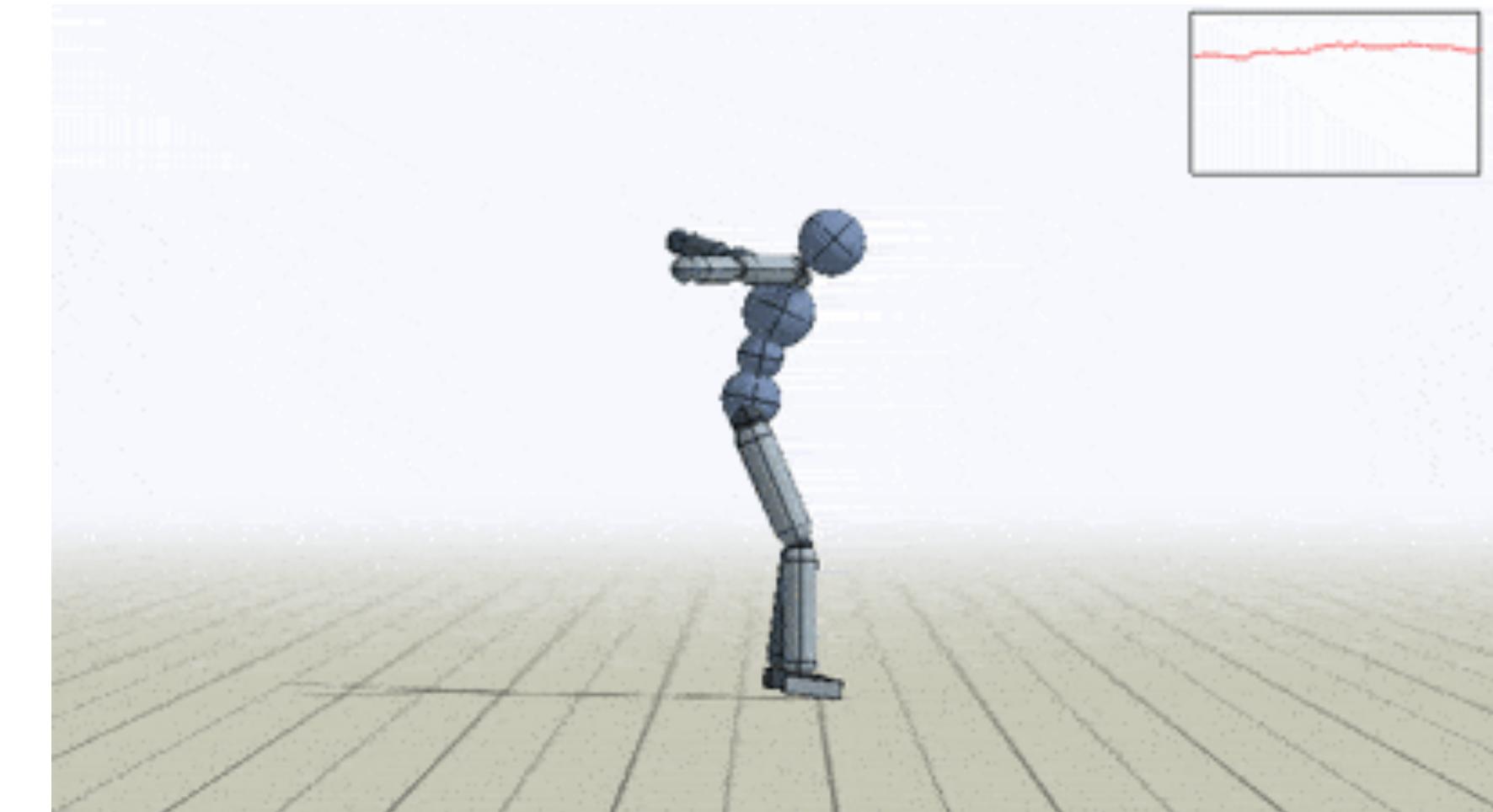
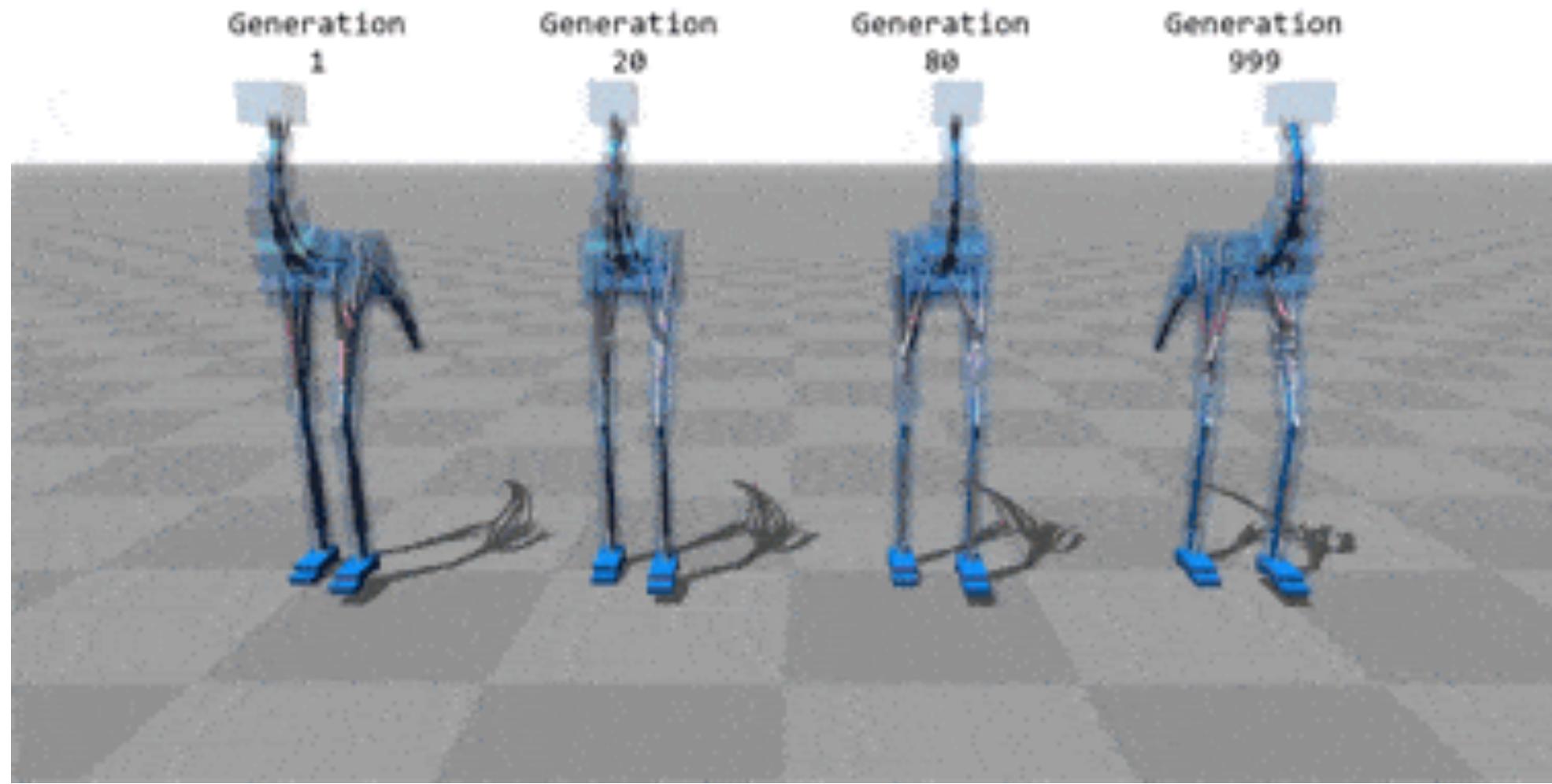
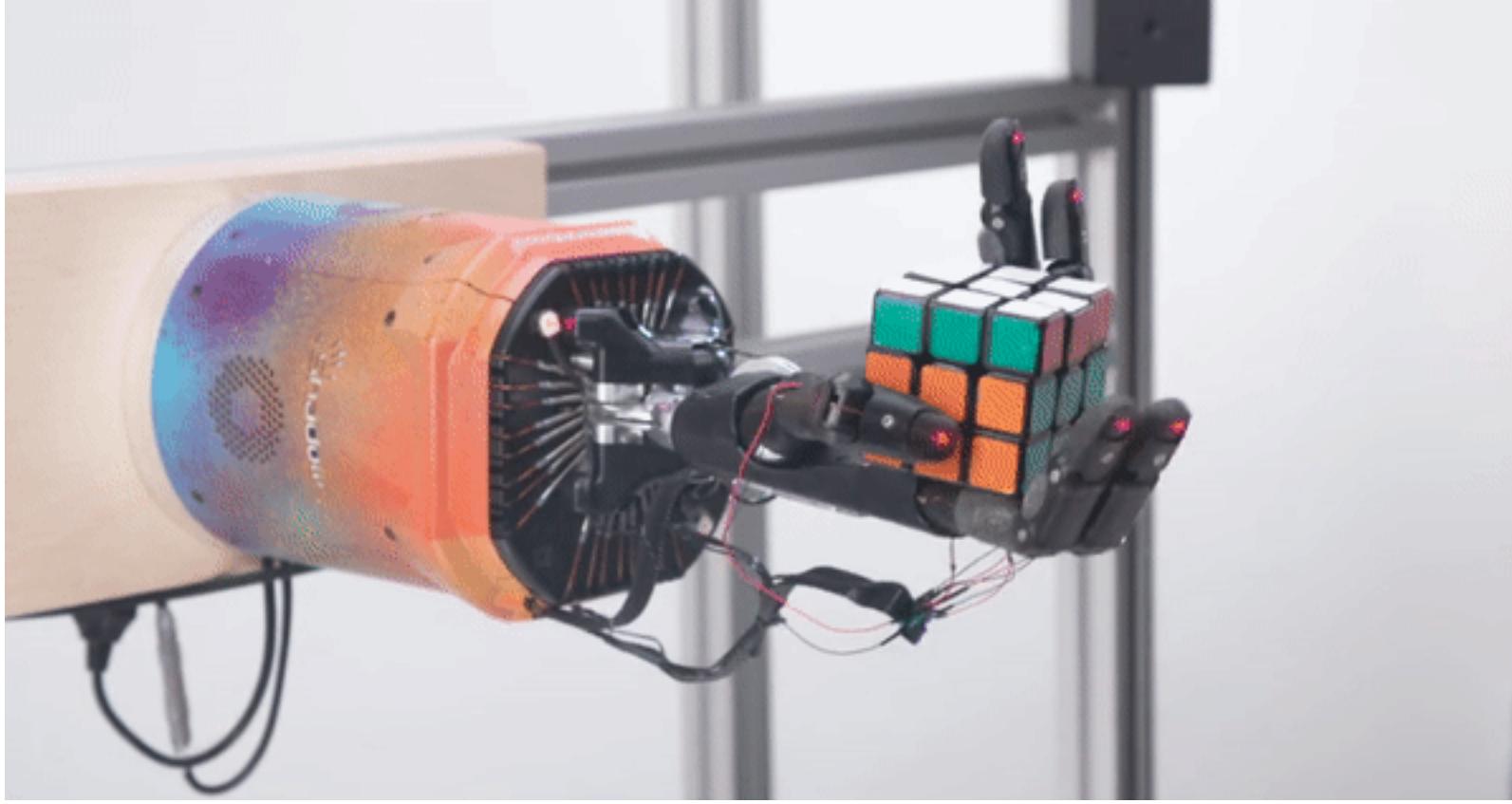
Reinforcement Learning

Framework for solving control tasks

Learn from the environment through trial and error

Receive rewards as feedback

Reinforcement Learning



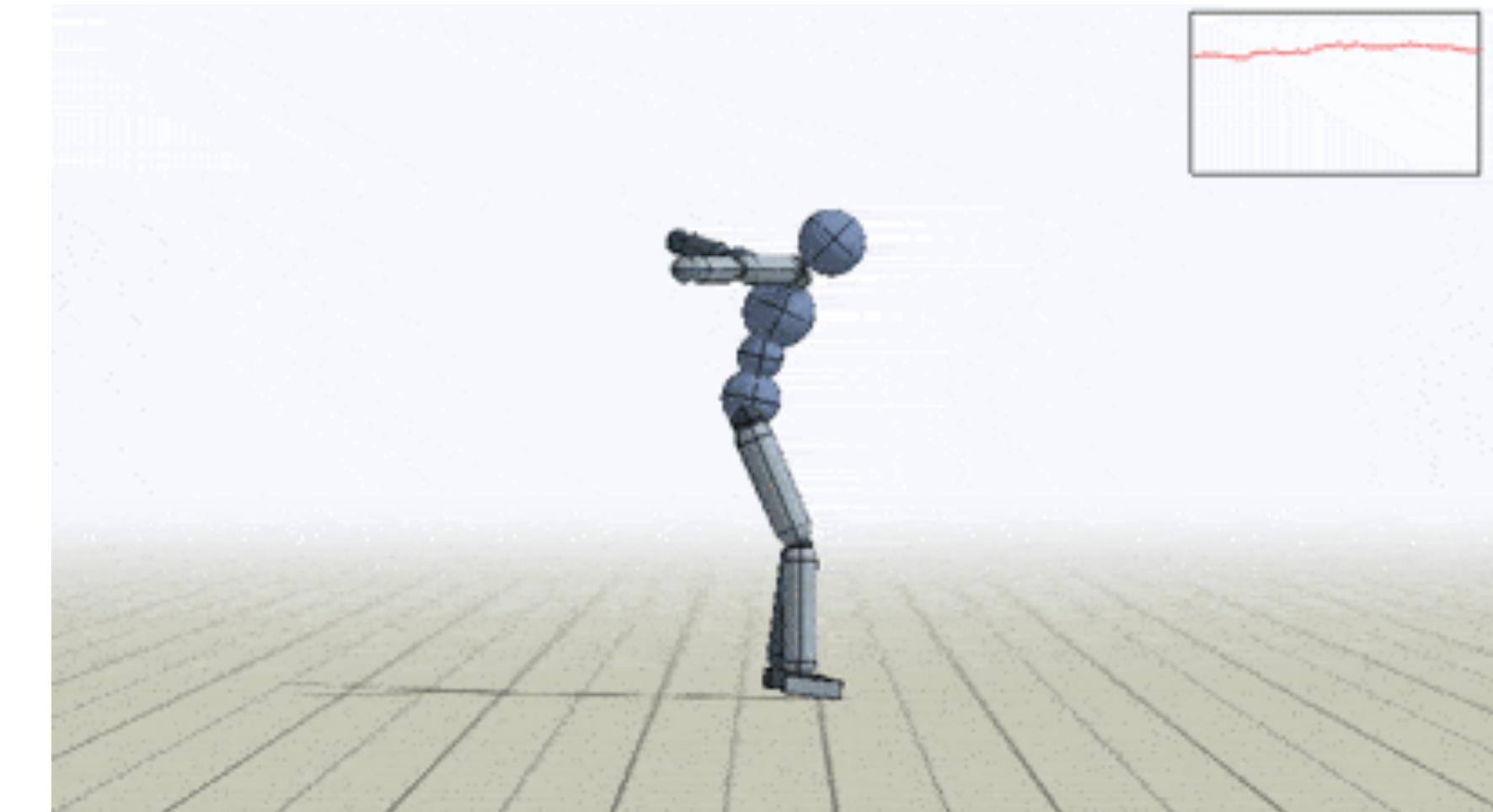
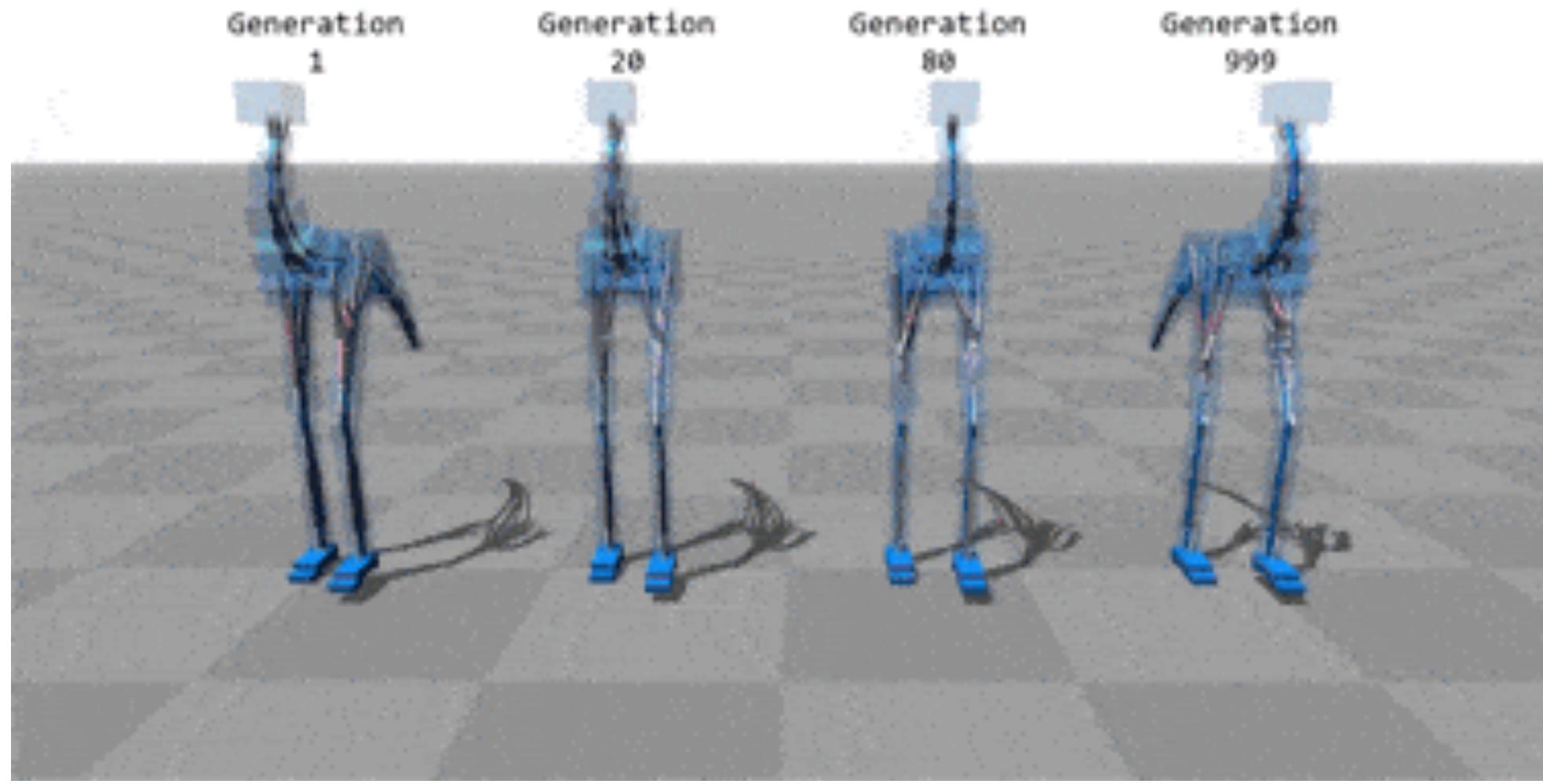
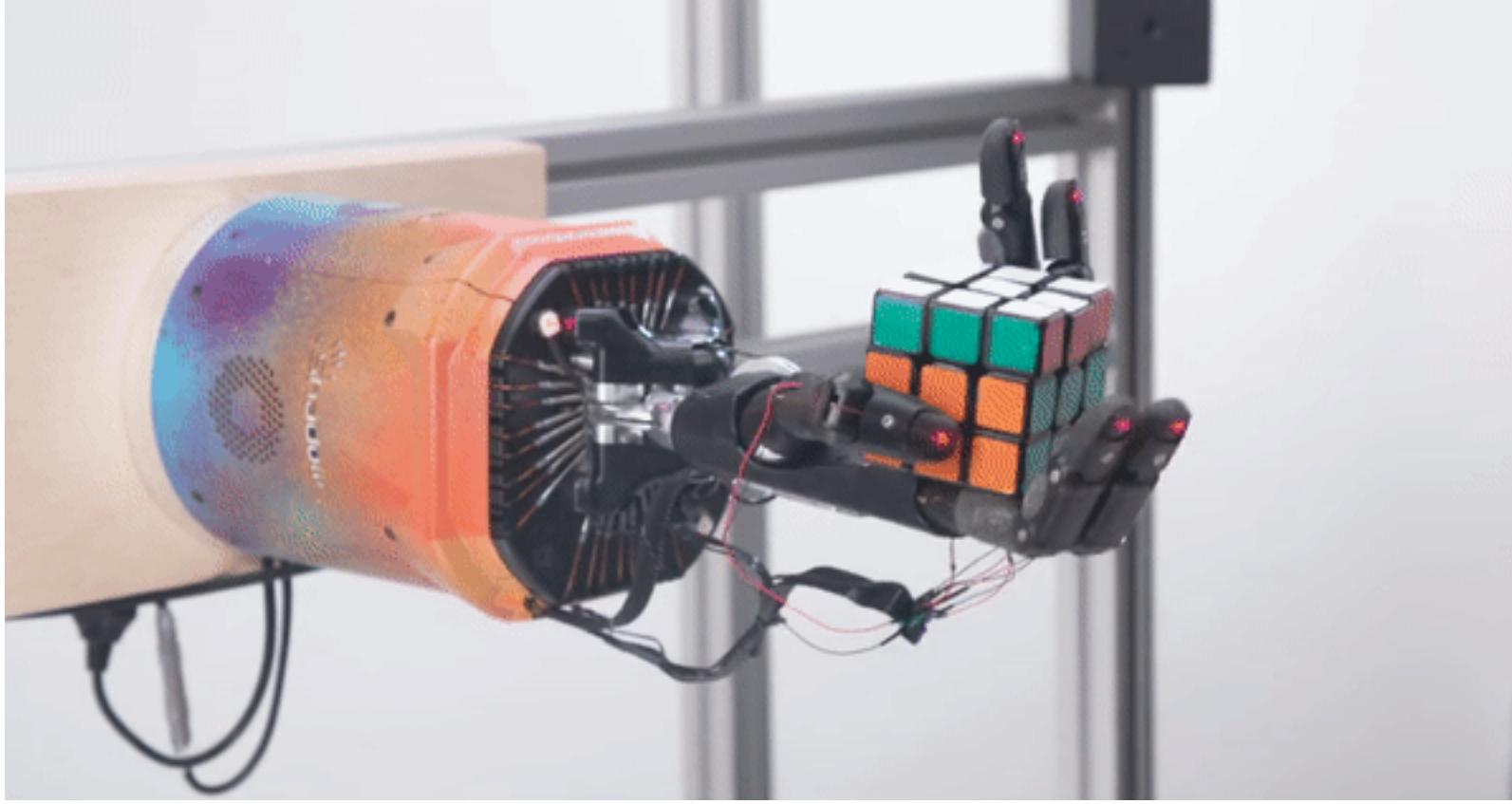
Akkaya, Ilge, et al. "Solving rubik's cube with a robot hand." arXiv preprint arXiv:1910.07113 (2019). (<https://openai.com/research/solving-rubiks-cube>)

Mnih, V., Kavukcuoglu, K., Silver, D. et al. Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015).

Geijtenbeek, Thomas, Michiel Van De Panne, and A. Frank Van Der Stappen. "Flexible muscle-based locomotion for bipedal creatures." ACM Transactions on Graphics (TOG) 32.6 (2013): 1-11.

Peng, Xue Bin, et al. "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills." ACM Transactions On Graphics (TOG) 37.4 (2018): 1-14.

Reinforcement Learning



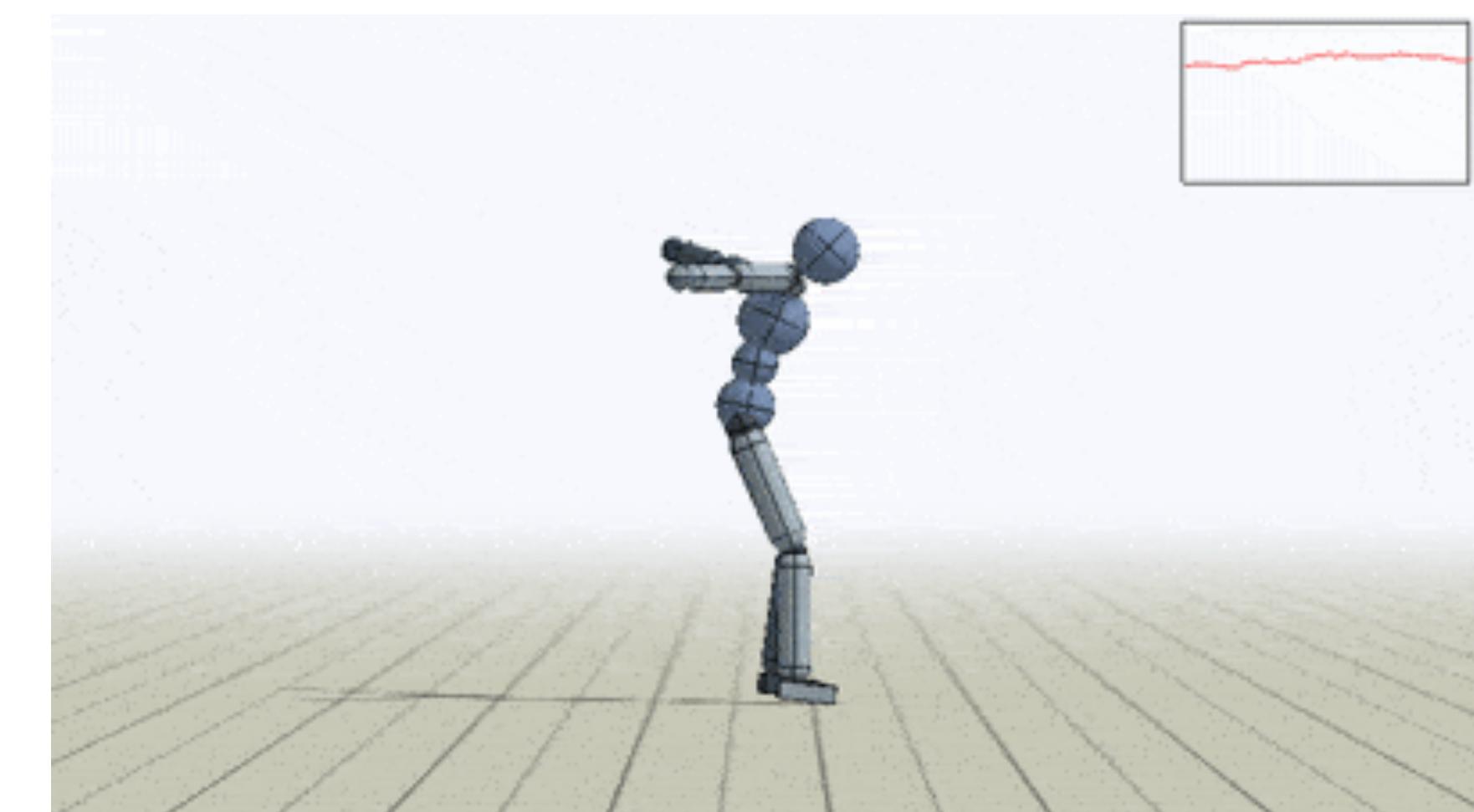
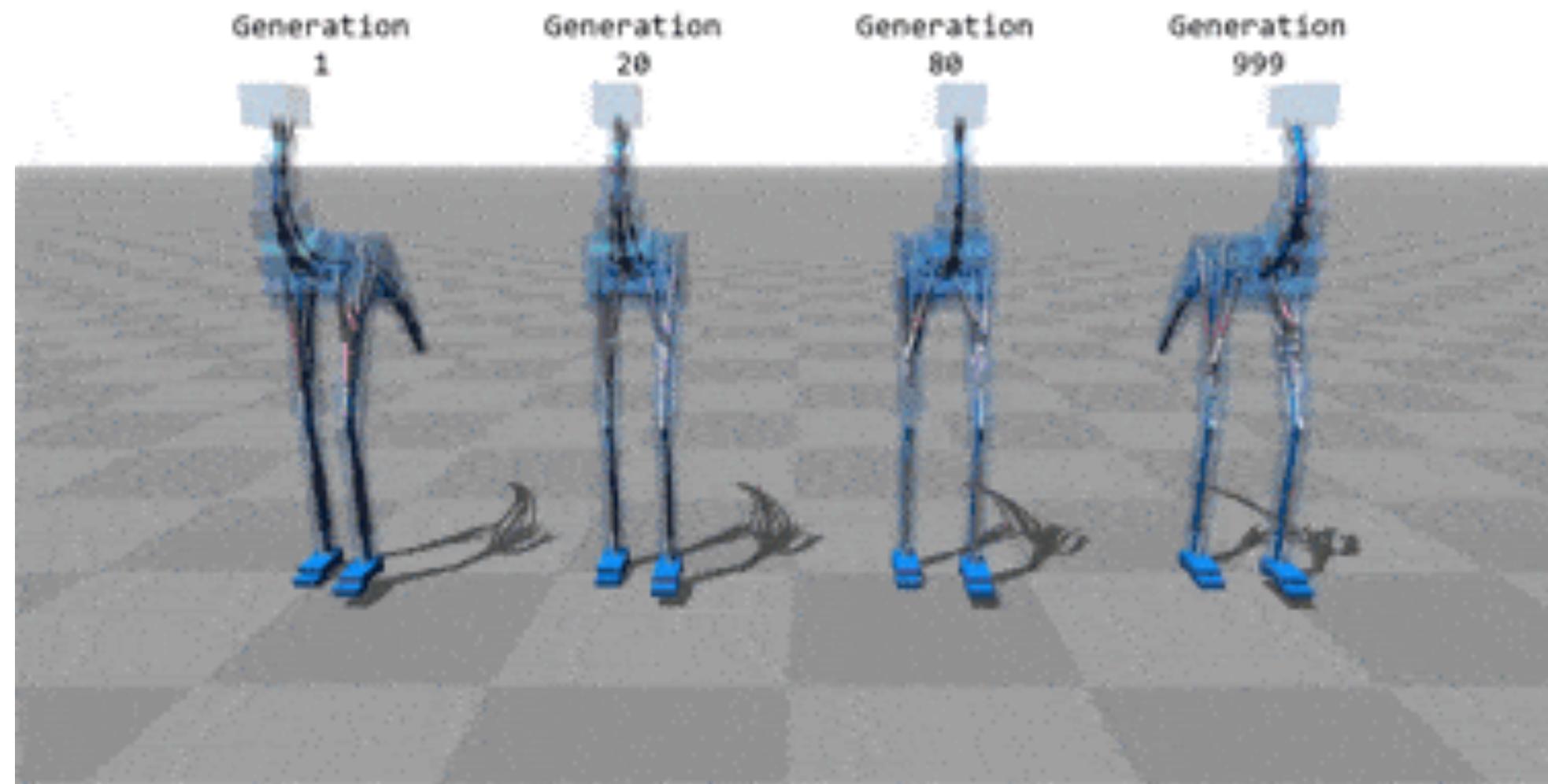
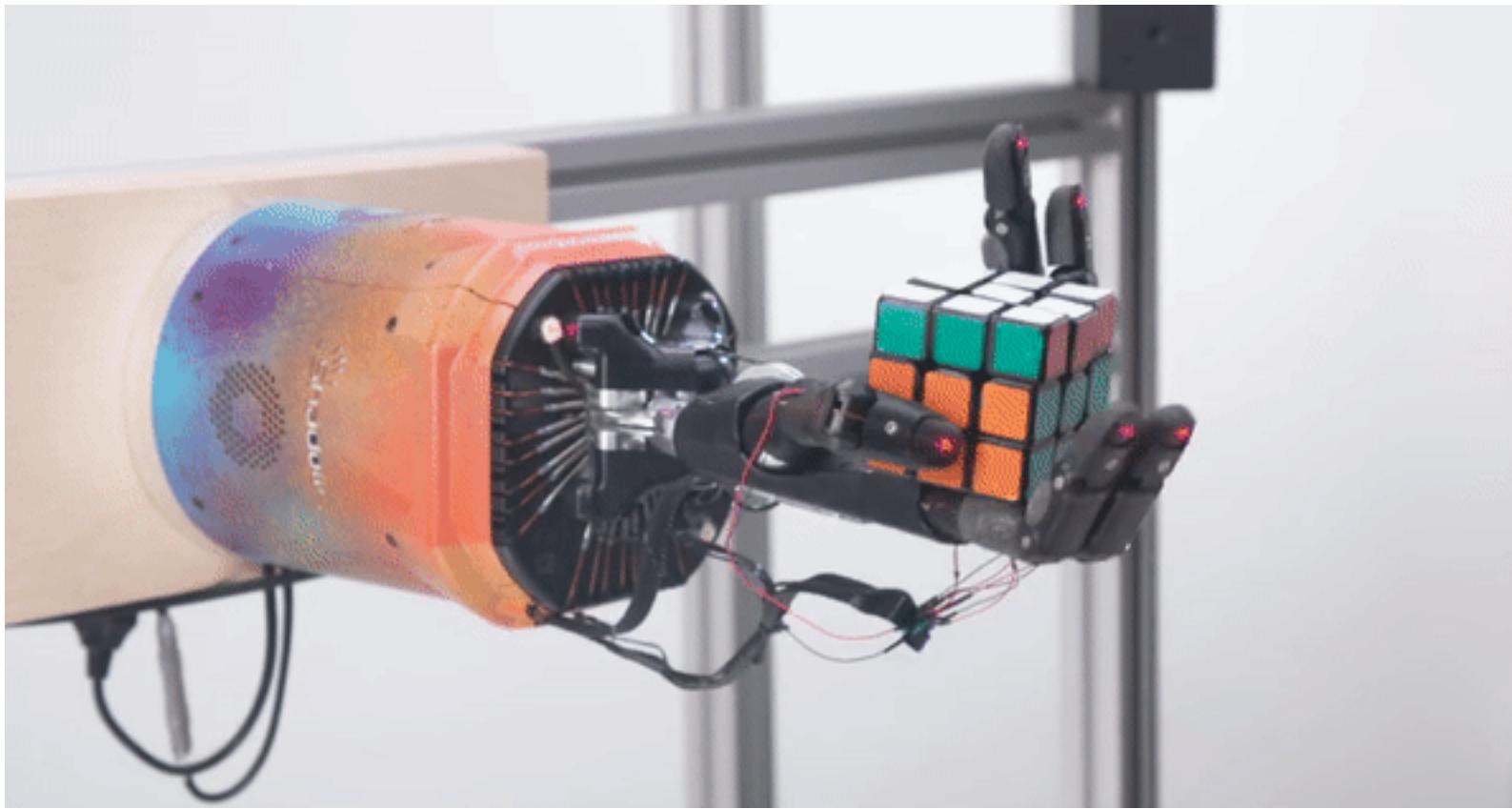
Akkaya, Ilge, et al. "Solving rubik's cube with a robot hand." arXiv preprint arXiv:1910.07113 (2019). (<https://openai.com/research/solving-rubiks-cube>)

Mnih, V., Kavukcuoglu, K., Silver, D. et al. Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015).

Geijtenbeek, Thomas, Michiel Van De Panne, and A. Frank Van Der Stappen. "Flexible muscle-based locomotion for bipedal creatures." ACM Transactions on Graphics (TOG) 32.6 (2013): 1-11.

Peng, Xue Bin, et al. "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills." ACM Transactions On Graphics (TOG) 37.4 (2018): 1-14.

Reinforcement Learning



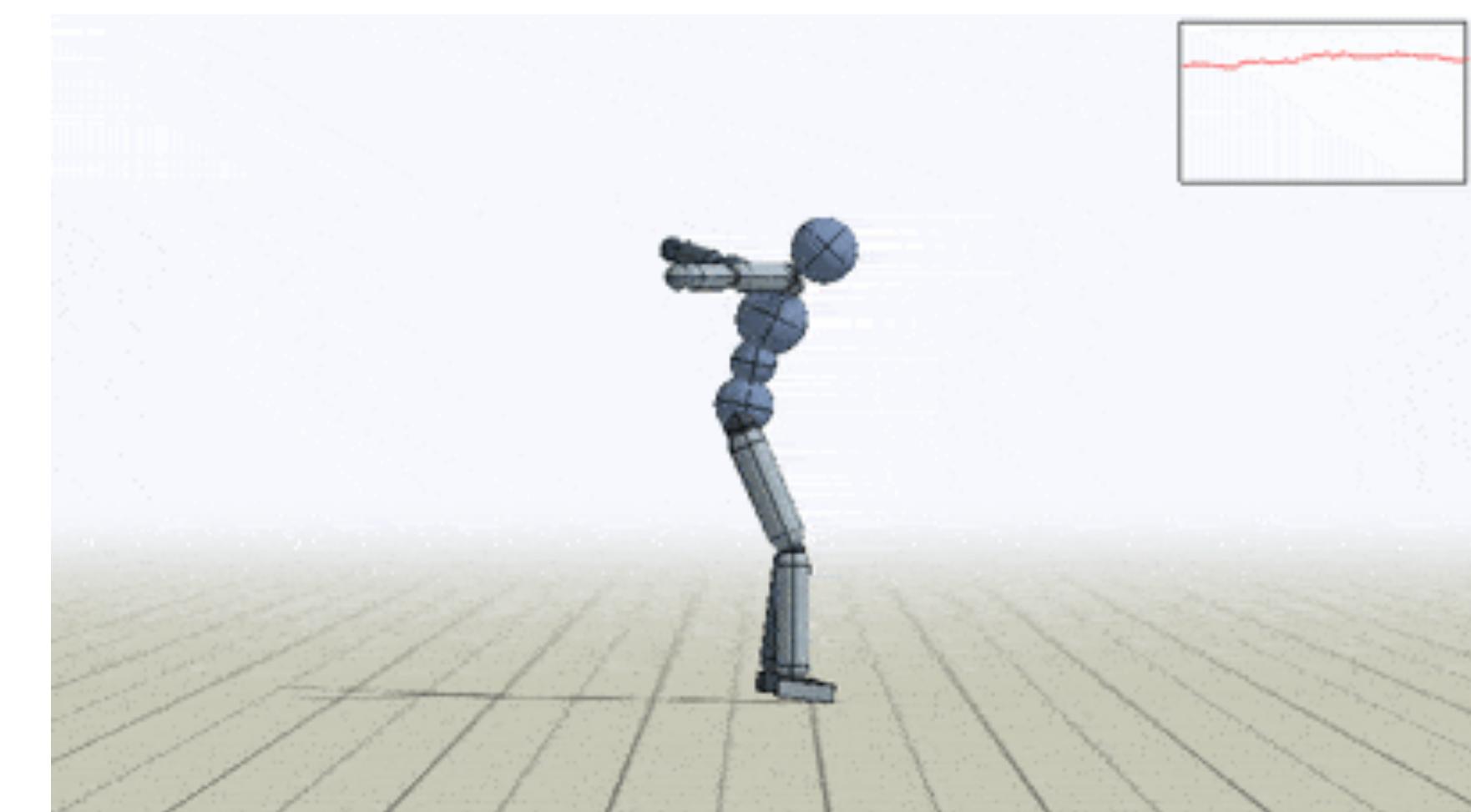
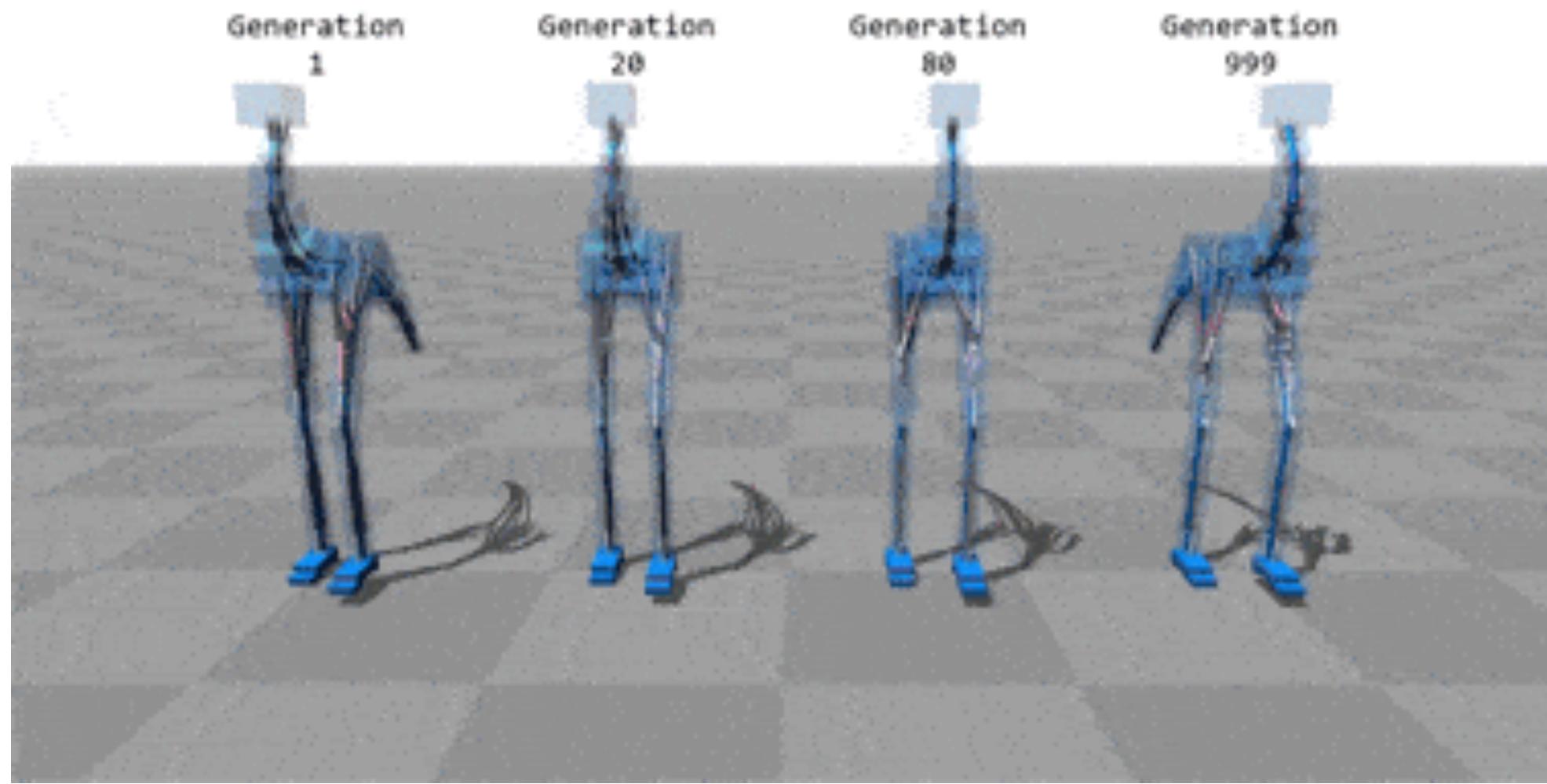
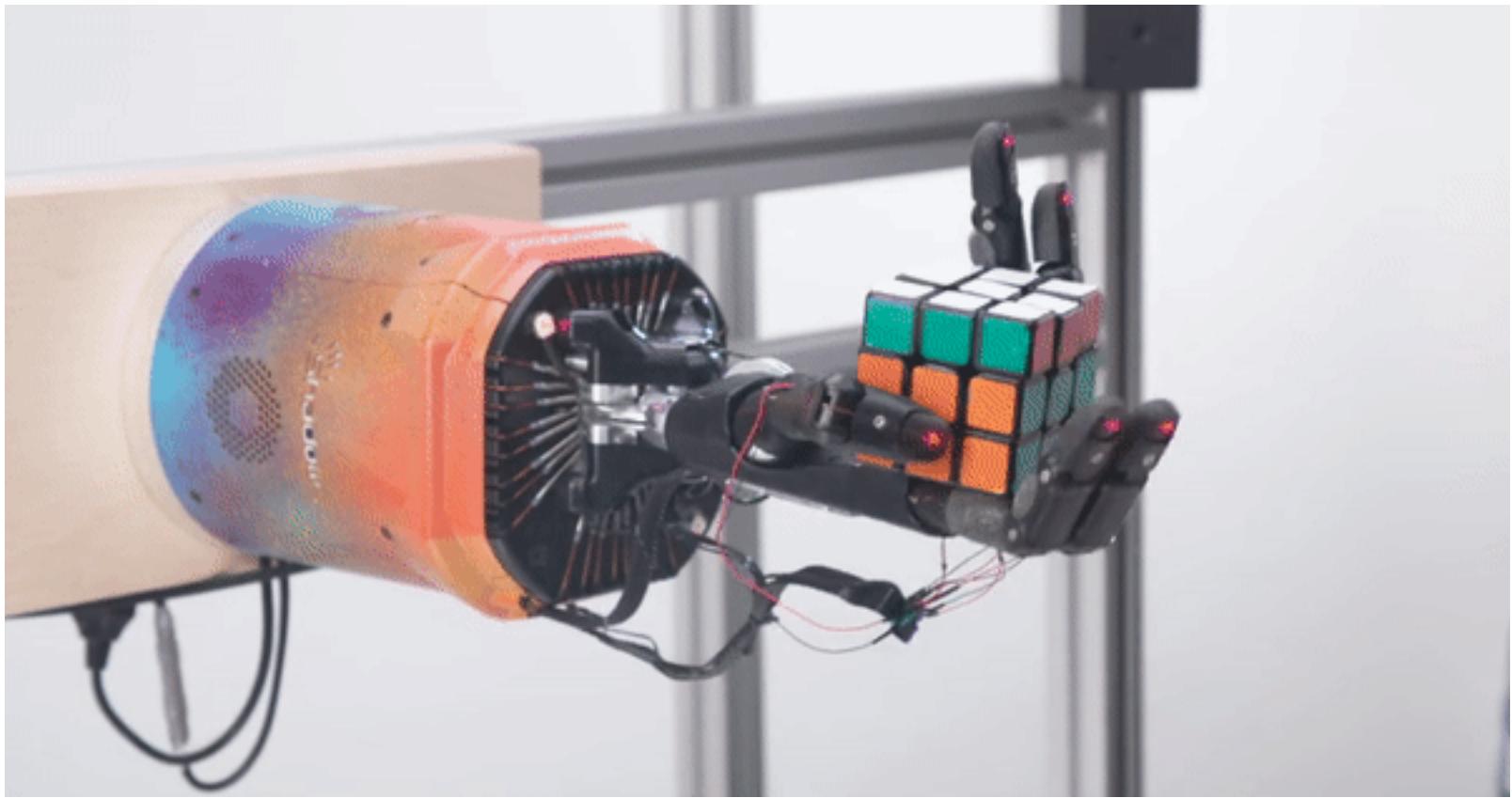
Akkaya, Ilge, et al. "Solving rubik's cube with a robot hand." arXiv preprint arXiv:1910.07113 (2019). (<https://openai.com/research/solving-rubiks-cube>)

Mnih, V., Kavukcuoglu, K., Silver, D. et al. Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015).

Geijtenbeek, Thomas, Michiel Van De Panne, and A. Frank Van Der Stappen. "Flexible muscle-based locomotion for bipedal creatures." ACM Transactions on Graphics (TOG) 32.6 (2013): 1-11.

Peng, Xue Bin, et al. "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills." ACM Transactions On Graphics (TOG) 37.4 (2018): 1-14.

Reinforcement Learning



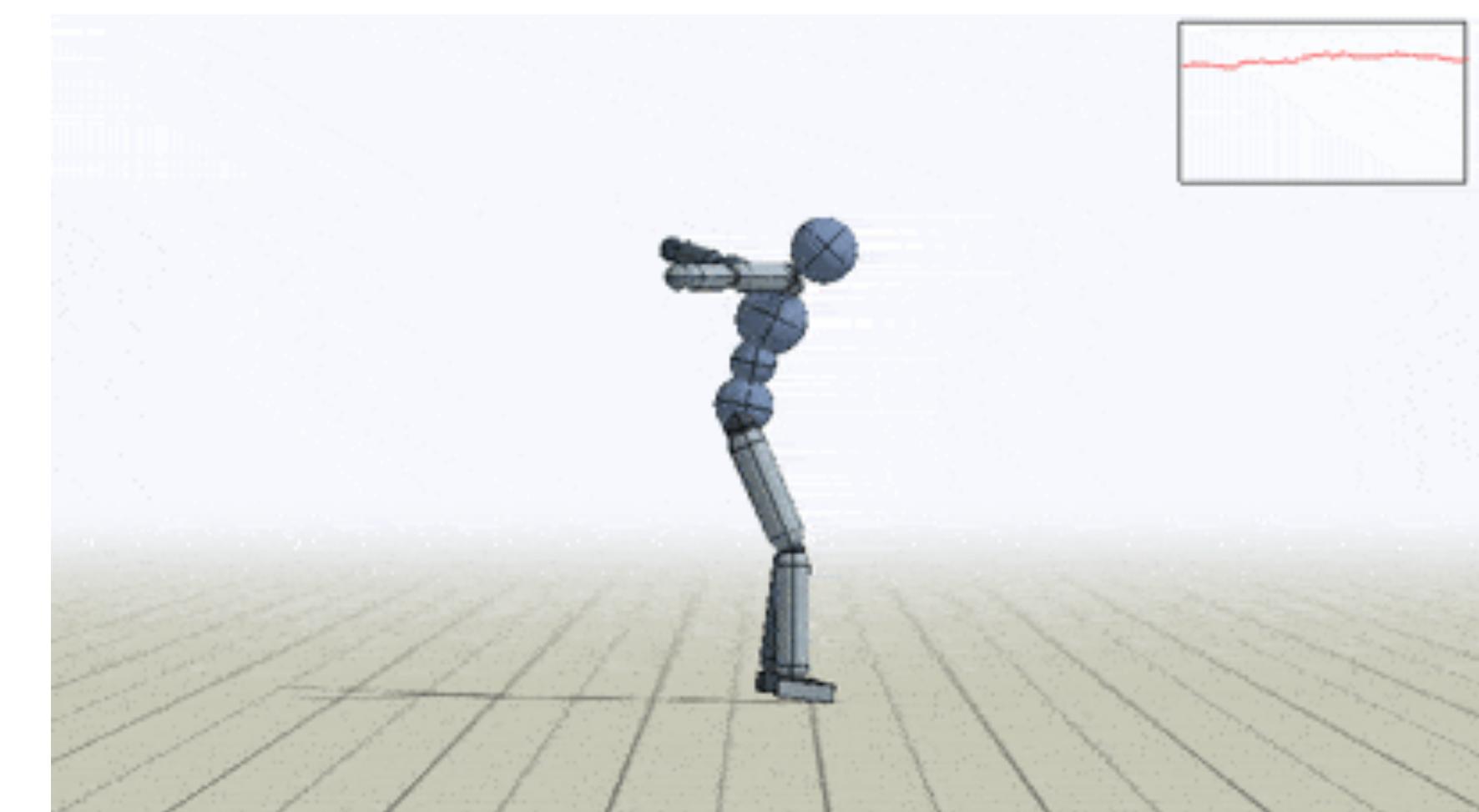
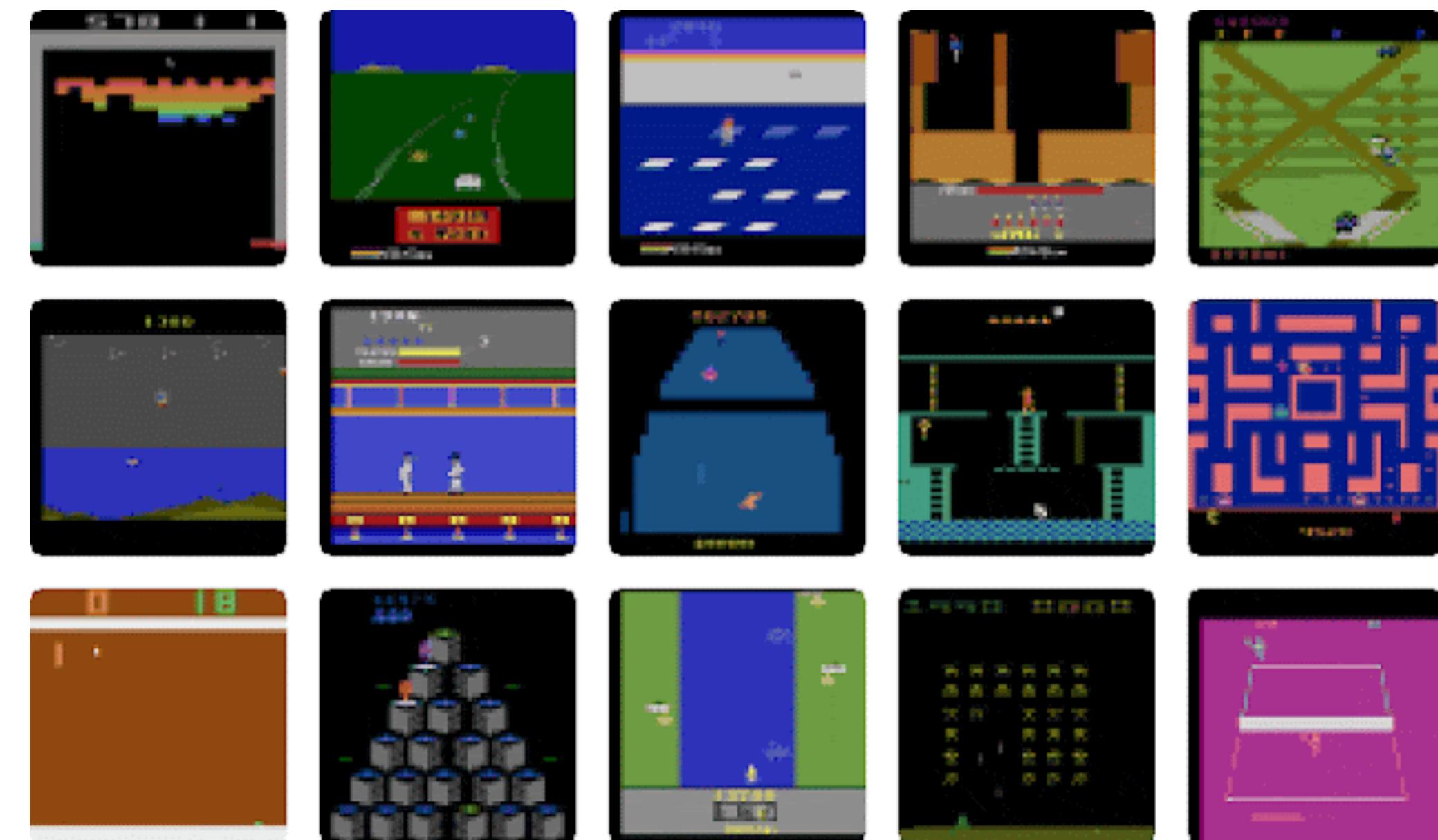
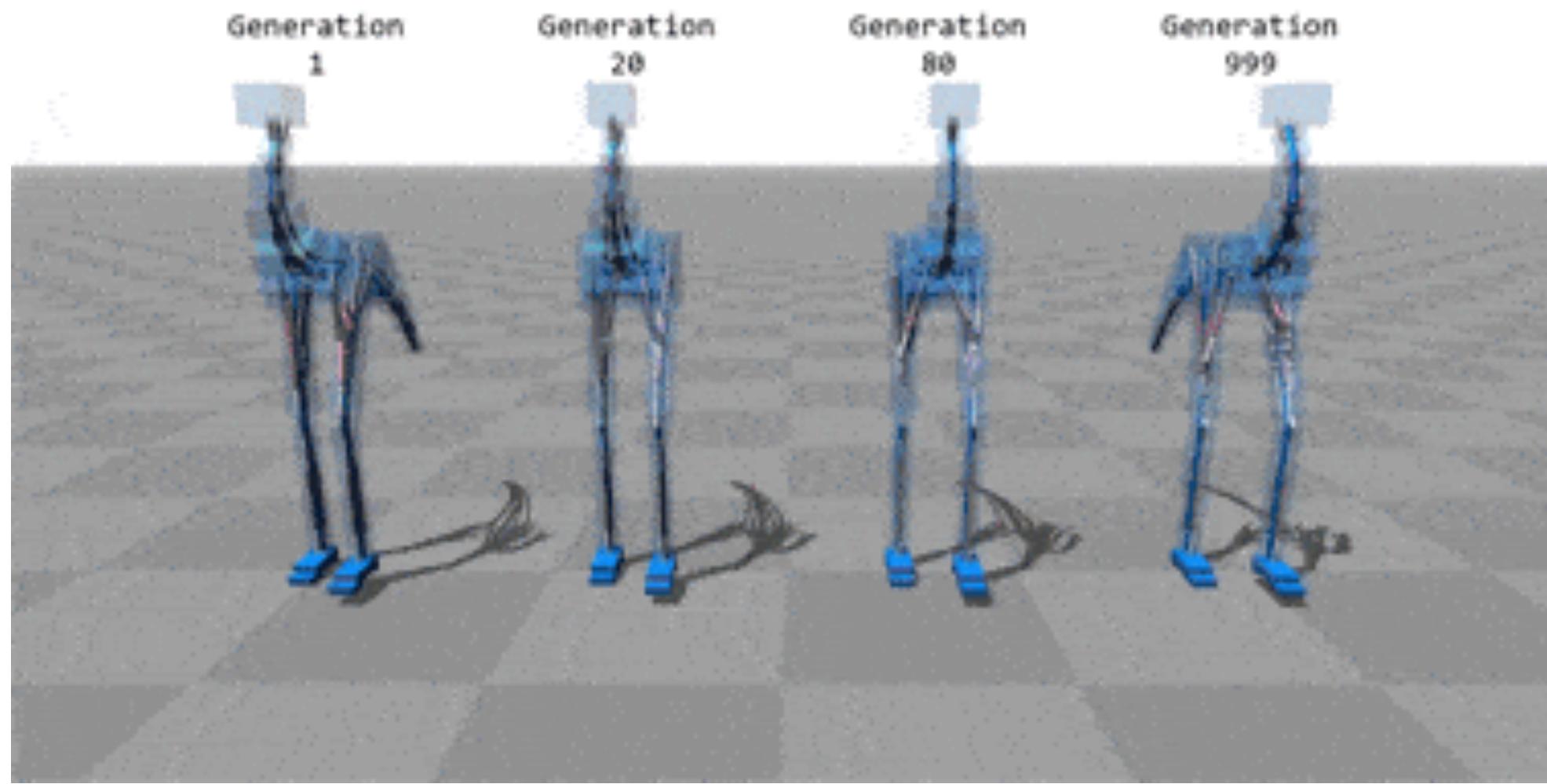
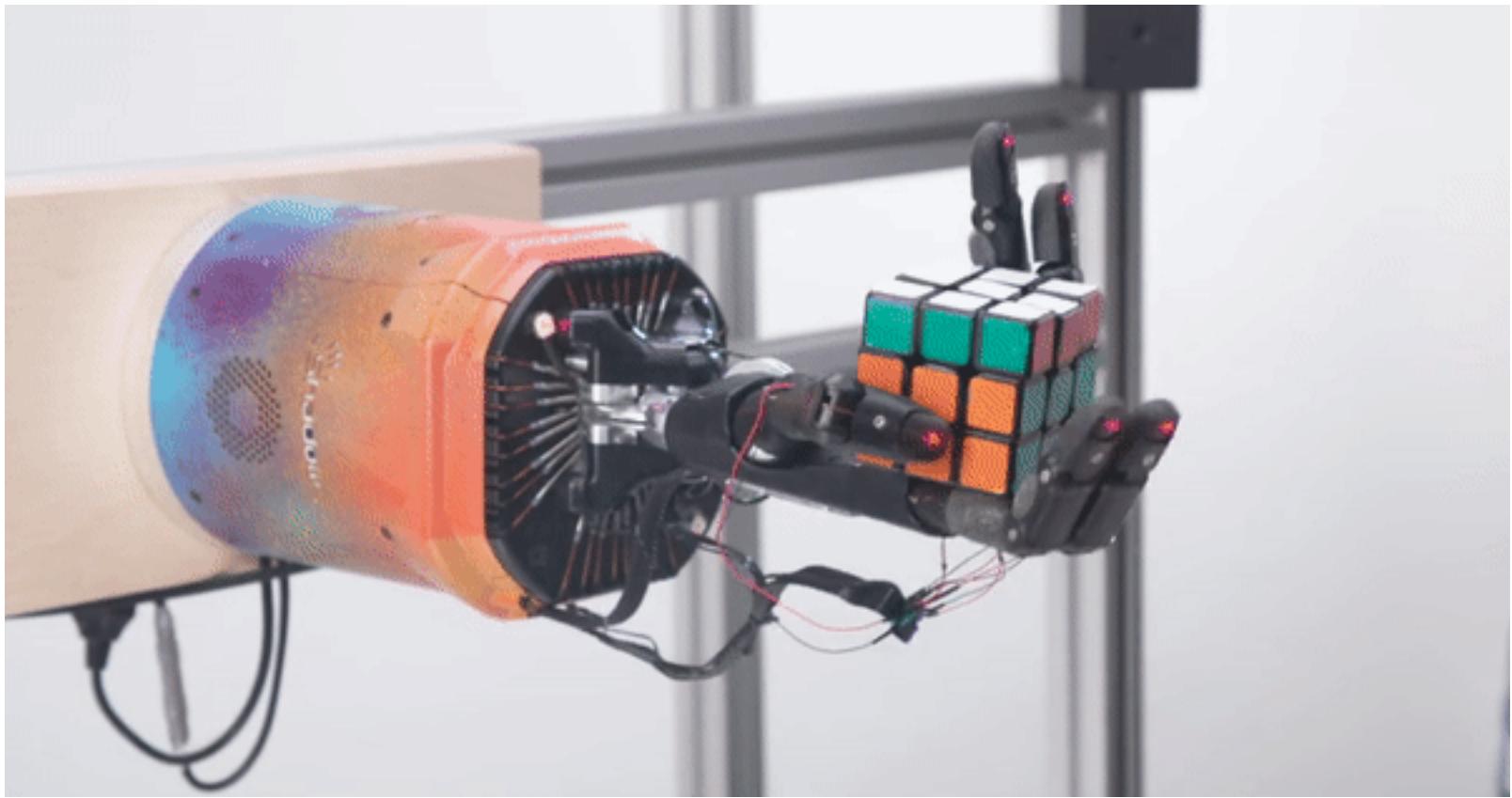
Akkaya, Ilge, et al. "Solving rubik's cube with a robot hand." arXiv preprint arXiv:1910.07113 (2019). (<https://openai.com/research/solving-rubiks-cube>)

Mnih, V., Kavukcuoglu, K., Silver, D. et al. Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015).

Geijtenbeek, Thomas, Michiel Van De Panne, and A. Frank Van Der Stappen. "Flexible muscle-based locomotion for bipedal creatures." ACM Transactions on Graphics (TOG) 32.6 (2013): 1-11.

Peng, Xue Bin, et al. "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills." ACM Transactions On Graphics (TOG) 37.4 (2018): 1-14.

Reinforcement Learning



Akkaya, Ilge, et al. "Solving rubik's cube with a robot hand." arXiv preprint arXiv:1910.07113 (2019). (<https://openai.com/research/solving-rubiks-cube>)

Mnih, V., Kavukcuoglu, K., Silver, D. et al. Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015).

Geijtenbeek, Thomas, Michiel Van De Panne, and A. Frank Van Der Stappen. "Flexible muscle-based locomotion for bipedal creatures." ACM Transactions on Graphics (TOG) 32.6 (2013): 1-11.

Peng, Xue Bin, et al. "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills." ACM Transactions On Graphics (TOG) 37.4 (2018): 1-14.

Reinforcement Learning



Reinforcement Learning



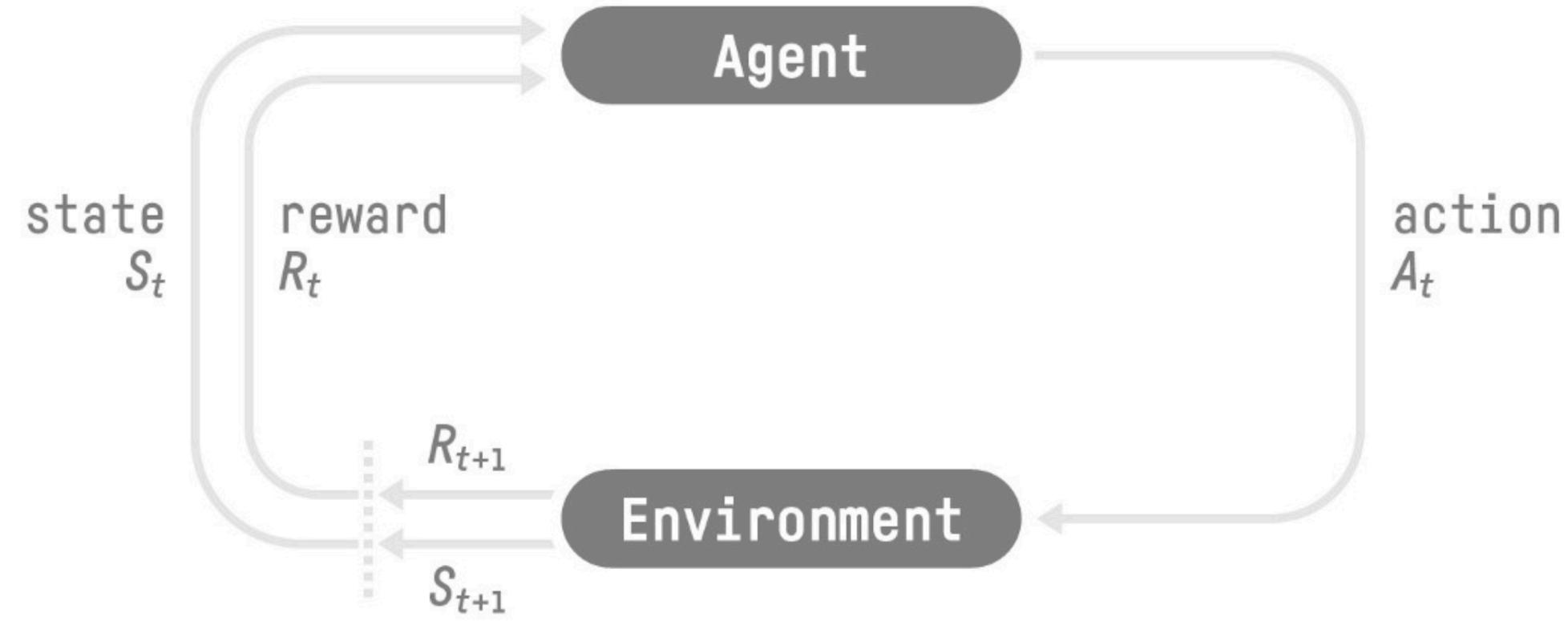
What you will learn

Basic concepts of RL

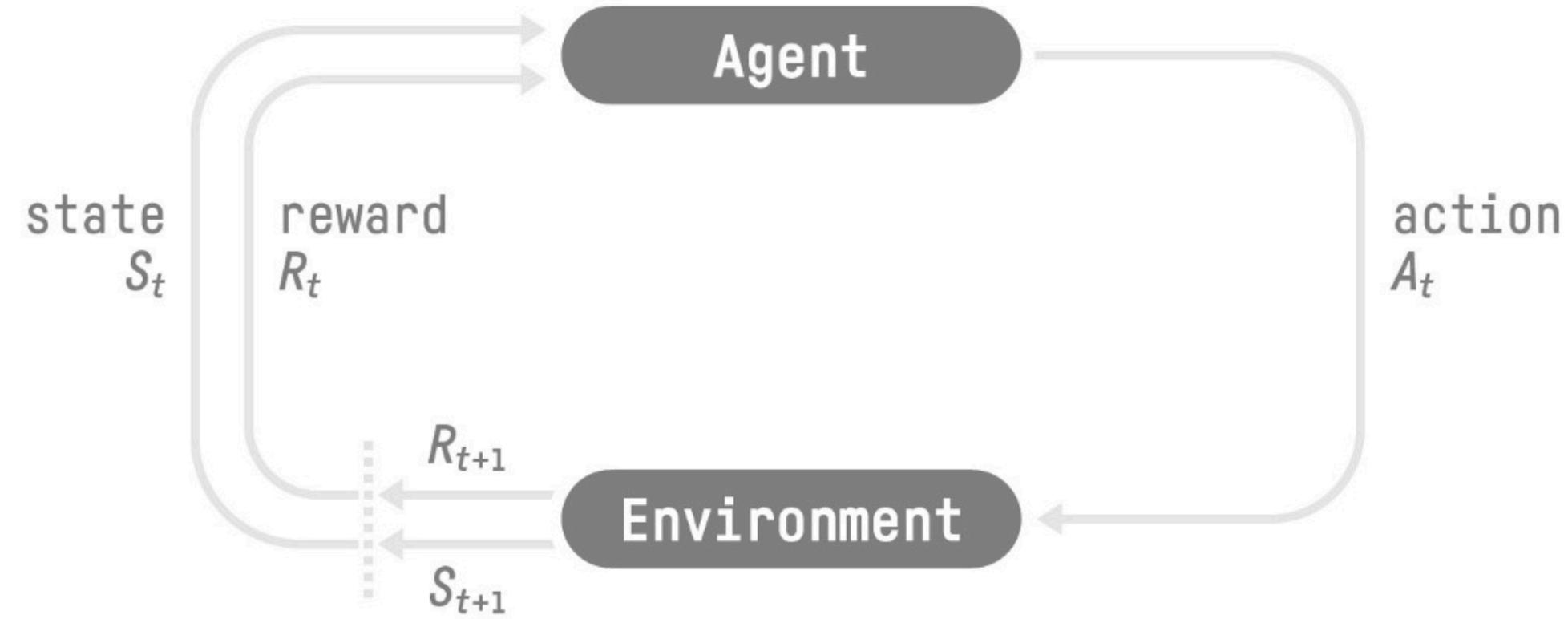
Q-Learning

Deep Q-Learning

Markov Decision Process (MDP) (finite)

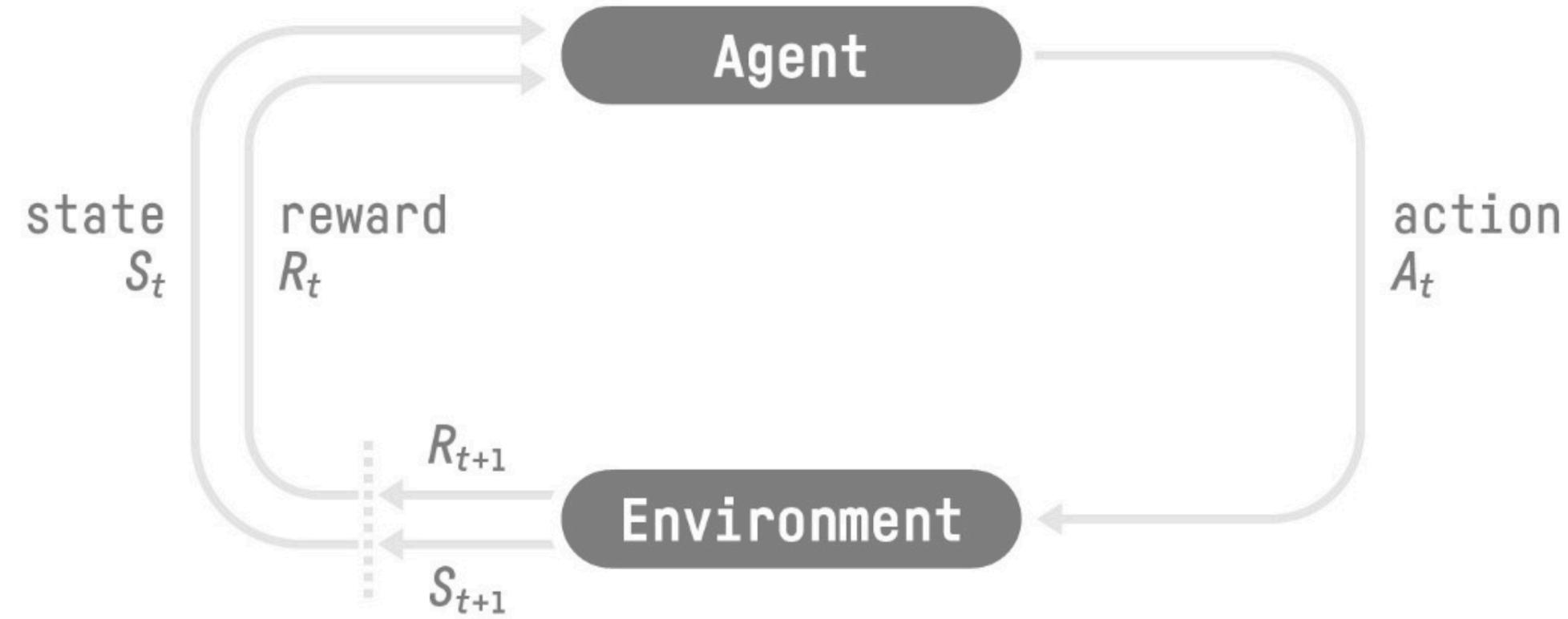


Markov Decision Process (MDP) (finite)



For $t = 0, 1, 2, 3 \dots$

Markov Decision Process (MDP) (finite)



For $t = 0, 1, 2, 3 \dots$

Trajectory: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \dots$

Agent Inputs

Agent Inputs

How do agents get information from the environment?

Agent Inputs

How do agents get information from the environment?

States

Agent Inputs

How do agents get information from the environment?

States

Observations

Agent Inputs

How do agents get information from the environment?

States

Complete description of the state of the world
(no hidden information)

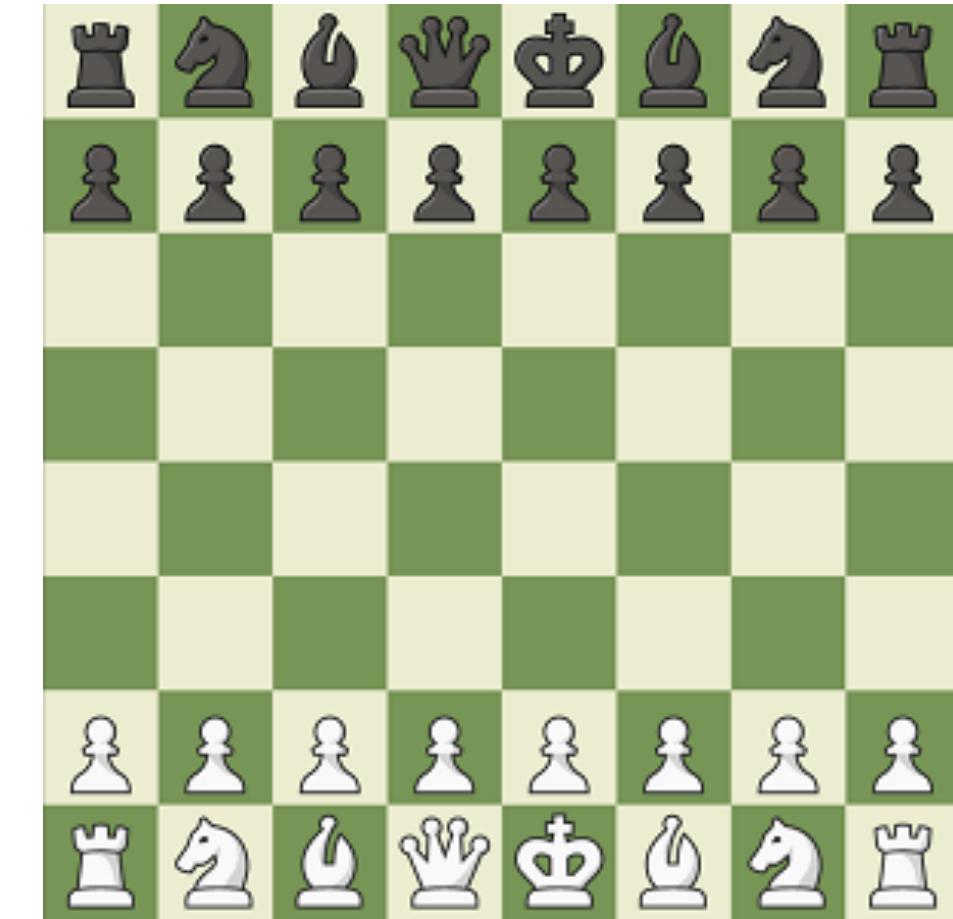
Observations

Agent Inputs

How do agents get information from the environment?

States

Complete description of the state of the world
(no hidden information)



Observations

Agent Inputs

How do agents get information from the environment?

States

Complete description of the state of the world
(no hidden information)



Observations

Partial description of the state of the world

Agent Inputs

How do agents get information from the environment?

States

Complete description of the state of the world
(no hidden information)



Observations

Partial description of the state of the world



Agent Outputs

Agent Outputs

How do agents interact with the environment?

Agent Outputs

How do agents interact with the environment?

Action space

Agent Outputs

How do agents interact with the environment?

Action space Set of all possible actions in an environment.

Agent Outputs

How do agents interact with the environment?

Action space Set of all possible actions in an environment.

Discrete

Agent Outputs

How do agents interact with the environment?

Action space Set of all possible actions in an environment.

Discrete

Continuous

Agent Outputs

How do agents interact with the environment?

Action space Set of all possible actions in an environment.

Discrete

The number of possible actions is *finite*

Continuous

Agent Outputs

How do agents interact with the environment?

Action space Set of all possible actions in an environment.

Discrete

The number of possible actions is *finite*



Continuous

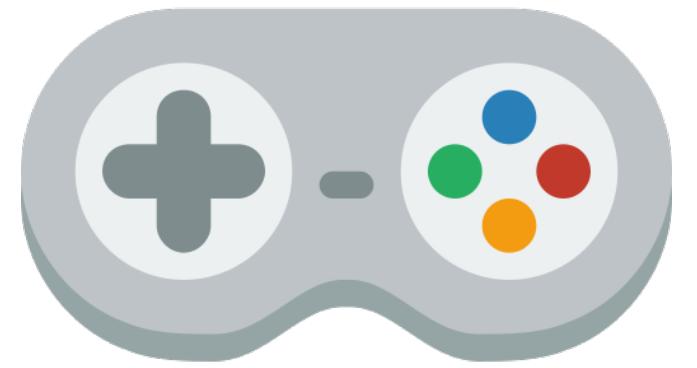
Agent Outputs

How do agents interact with the environment?

Action space Set of all possible actions in an environment.

Discrete

The number of possible actions is *finite*



Continuous

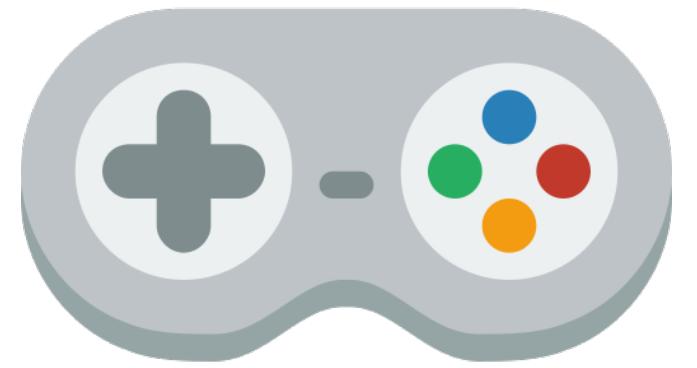
Agent Outputs

How do agents interact with the environment?

Action space Set of all possible actions in an environment.

Discrete

The number of possible actions is *finite*



Continuous

The number of possible actions is *continuous*

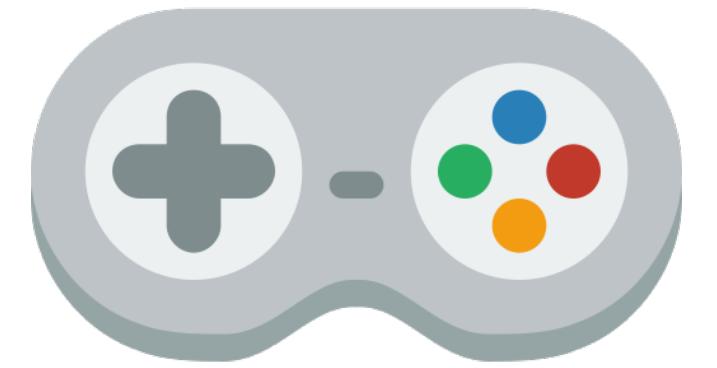
Agent Outputs

How do agents interact with the environment?

Action space Set of all possible actions in an environment.

Discrete

The number of possible actions is *finite*



Continuous

The number of possible actions is *continuous*



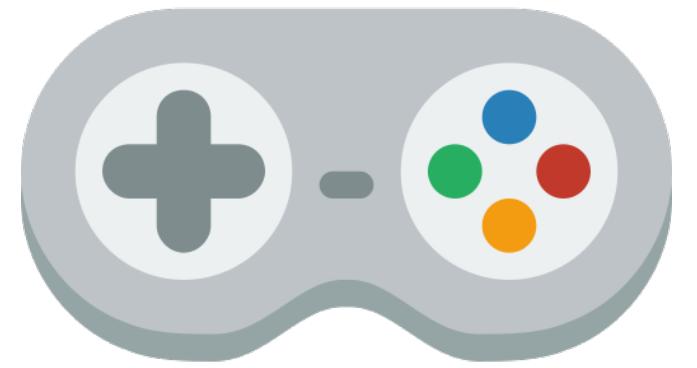
Agent Outputs

How do agents interact with the environment?

Action space Set of all possible actions in an environment.

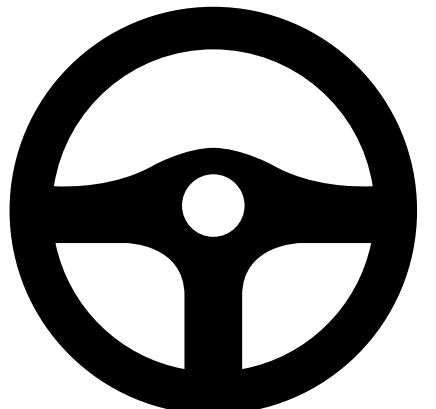
Discrete

The number of possible actions is *finite*

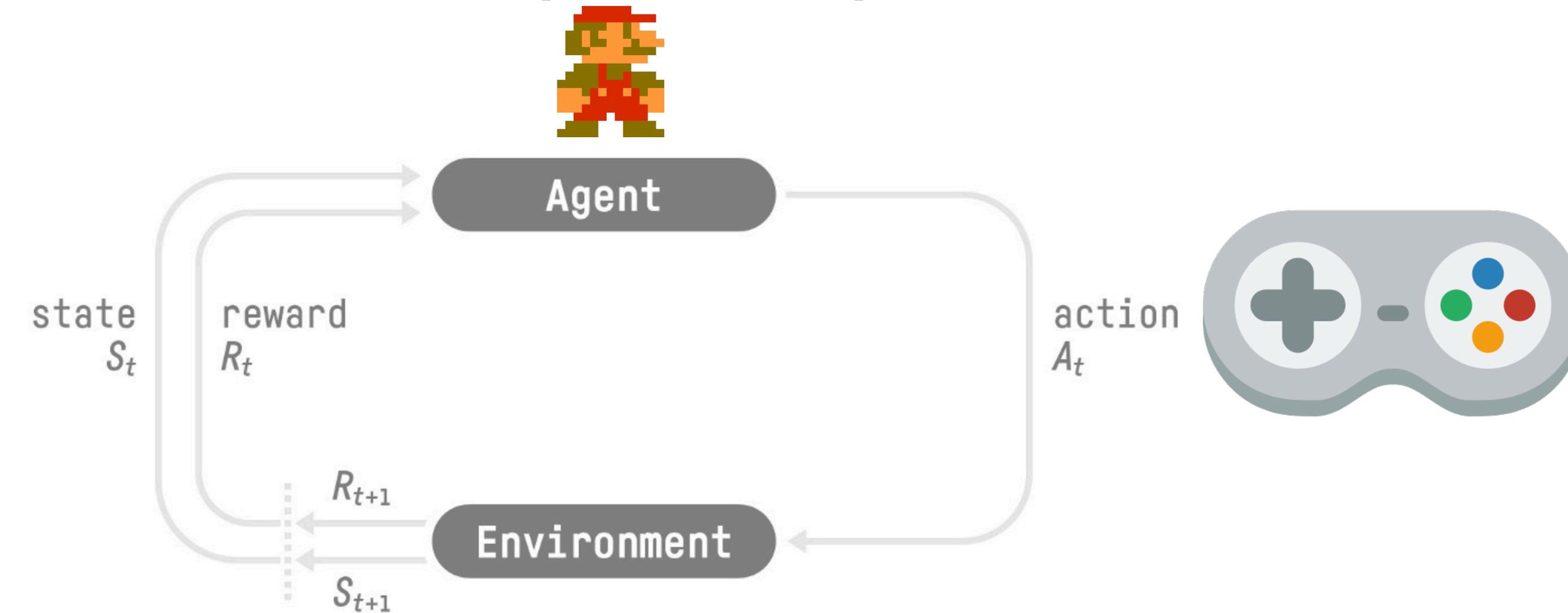
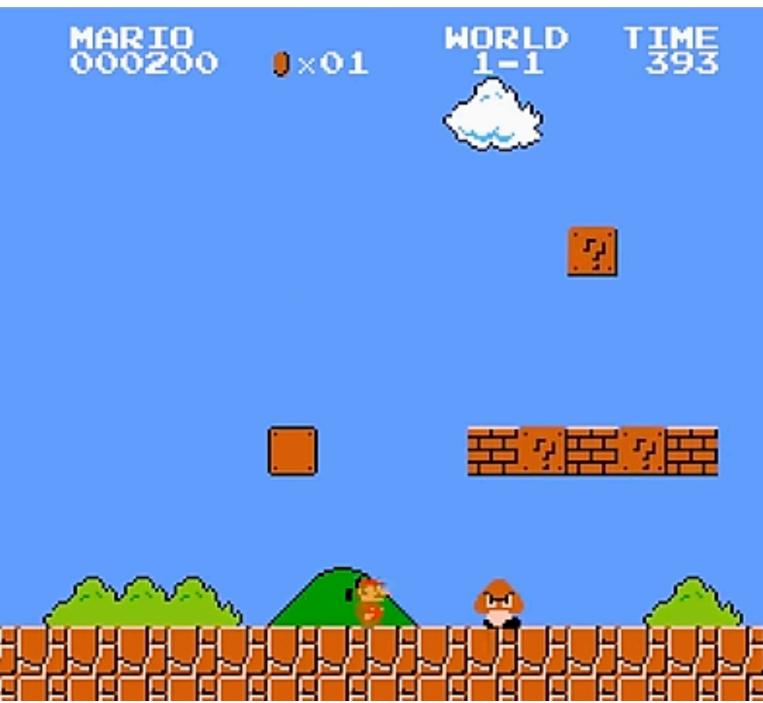


Continuous

The number of possible actions is *continuous*



Markov Decision Process (MDP)



For $t = 0, 1, 2, 3 \dots$

Paradigm



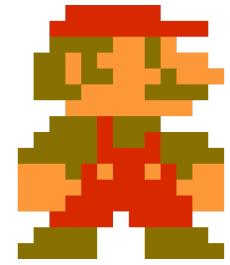
Paradigm

S_0



Paradigm

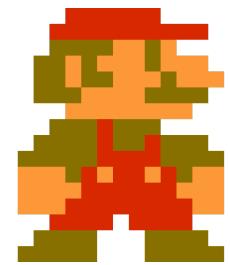
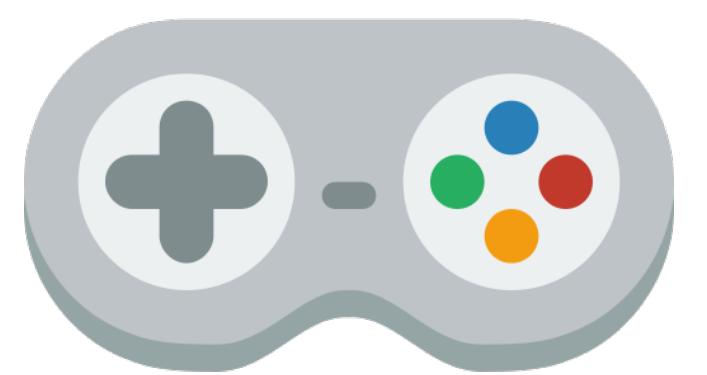
S_0



Policy π

Paradigm

S_0



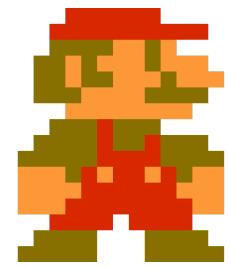
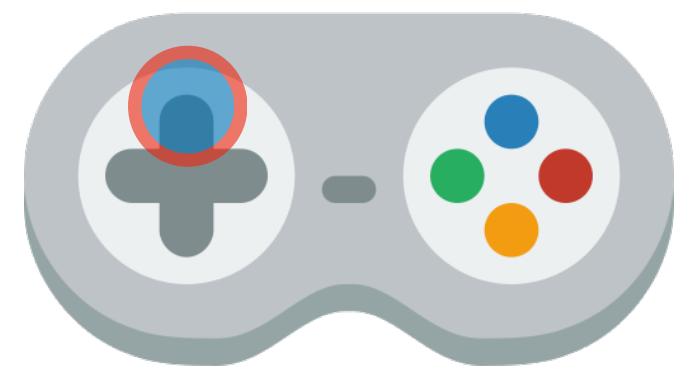
Policy π

Paradigm

S_0

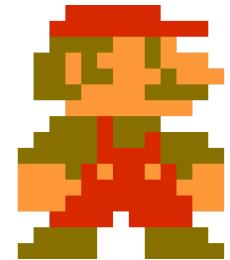
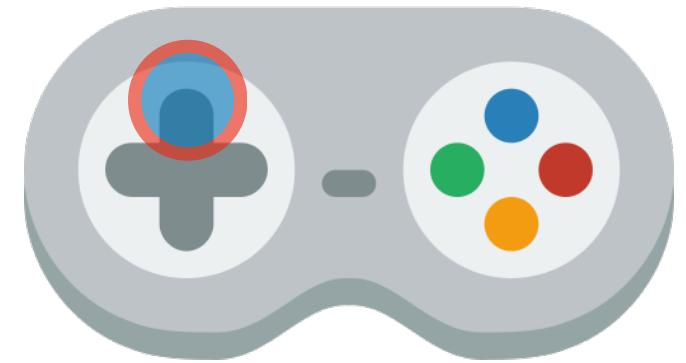


A_0



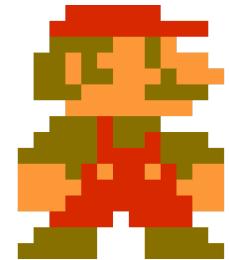
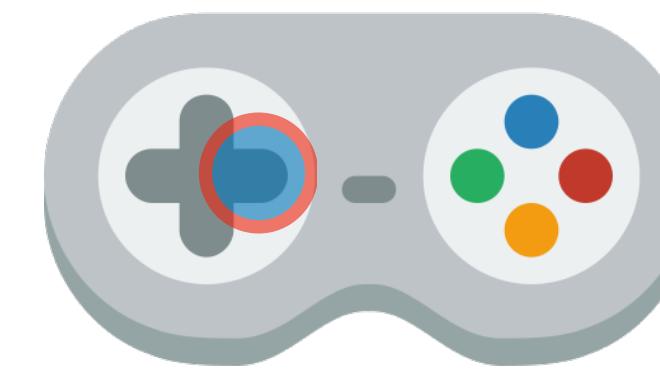
Policy π

Paradigm

 S_0  S_1 $r_1 = 0$  A_0 

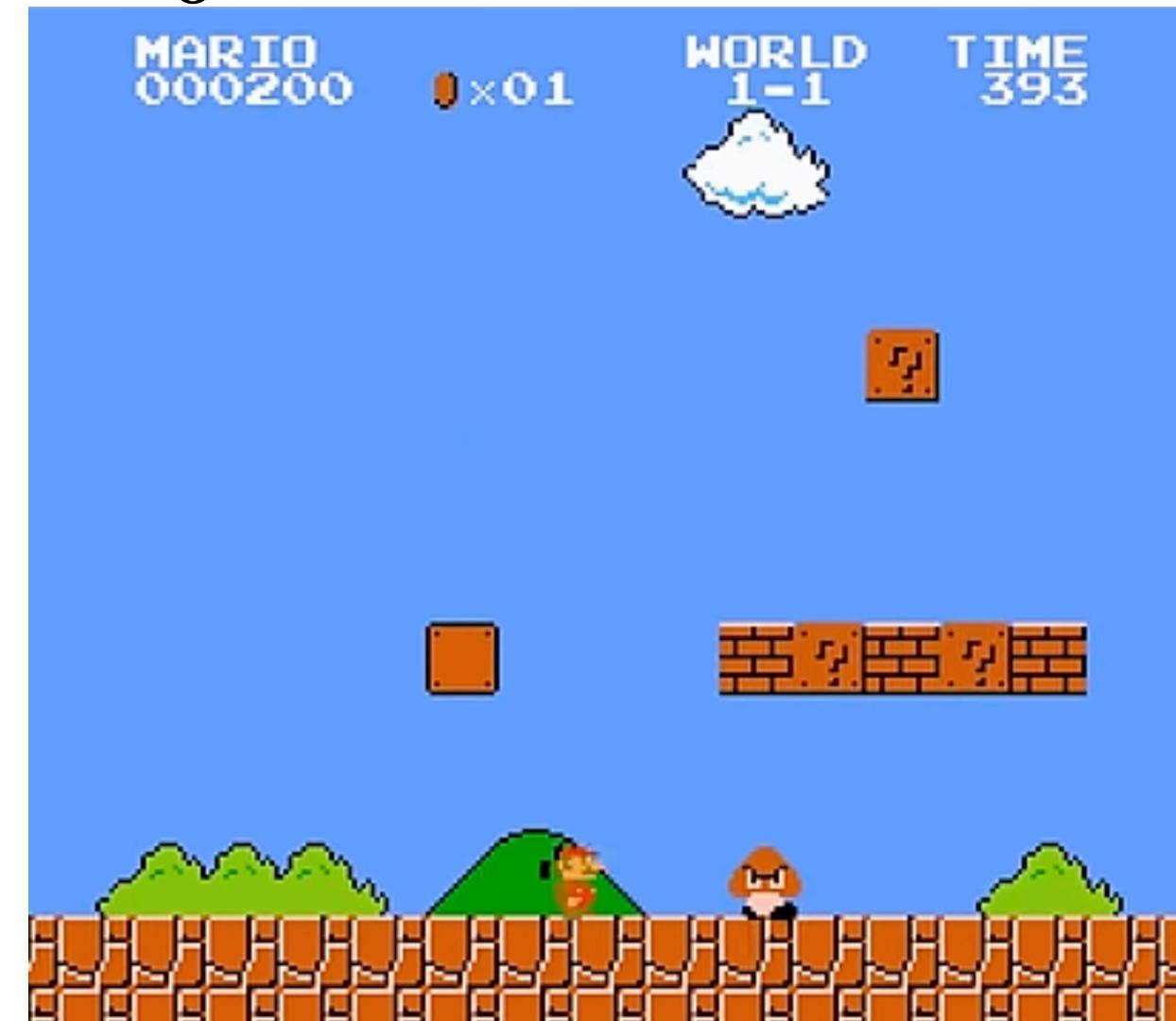
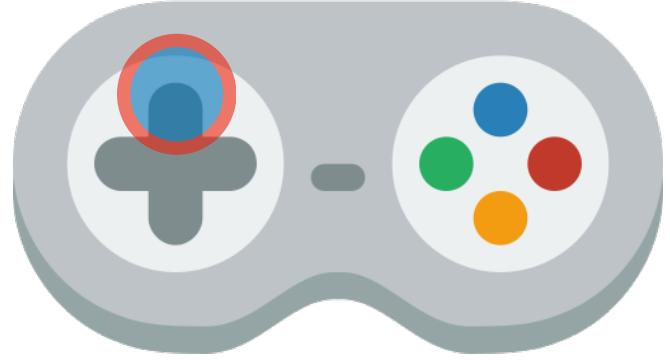
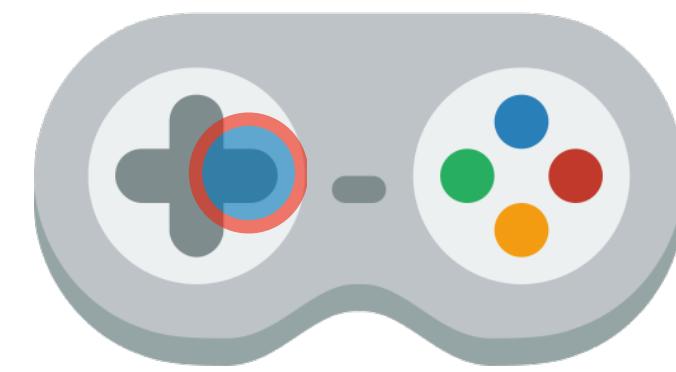
Policy π

Paradigm

 S_0  S_1  $r_1 = 0$ A_0  A_1 

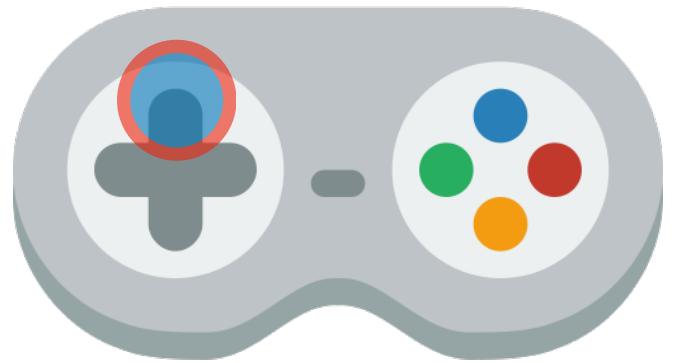
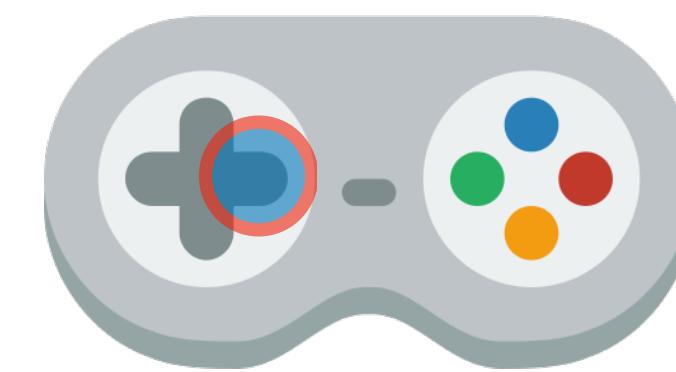
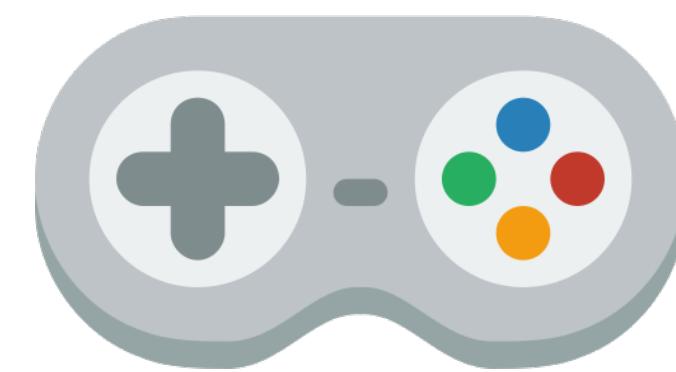
Policy π

Paradigm

 S_0  S_1  $r_1 = 0$ S_2  $r_2 = 100$ A_0  A_1 

Policy π

Paradigm

 S_0  S_1  $r_1 = 0$ S_2  $r_2 = 100$ A_0  A_1  A_2 

Policy π

Goal

Goal

Complete tasks successfully

Goal

Complete tasks successfully

Use rewards

Goal

Complete tasks successfully

Use rewards

Rewards communicate an agent *what* we want to achieve (NOT *how* to do it)

Goal

Complete tasks successfully

Use rewards

Rewards communicate an agent *what* we want to achieve (NOT *how* to do it)

Only feedback the agent receives

Goal

Complete tasks successfully

Use rewards

Rewards communicate an agent *what* we want to achieve (NOT *how* to do it)

Only feedback the agent receives

Maximize total amount of reward received

Goal

Complete tasks successfully

Use rewards

Rewards communicate an agent *what* we want to achieve (NOT *how* to do it)

Only feedback the agent receives

Maximize total amount of reward received

$$R_{t+1} + R_{t+2} + R_{t+3}, \dots + R_T$$

Goal

Complete tasks successfully

Use rewards

Rewards communicate an agent *what* we want to achieve (NOT *how* to do it)

Only feedback the agent receives

Maximize total amount of reward received (*return*)

$$R_{t+1} + R_{t+2} + R_{t+3}, \dots + R_T$$

Goal

Complete tasks successfully

Use rewards

Rewards communicate an agent *what* we want to achieve (NOT *how* to do it)

Only feedback the agent receives

Maximize total amount of reward received (*return*)

$$G_t = R_{t+1} + R_{t+2} + R_{t+3}, \dots + R_T$$

Episodic tasks

Continuing tasks

Episodic tasks

T is the final step. RL breaks into *episodes*.

Continuing tasks

Episodic tasks

T is the final step. RL breaks into *episodes*.

At the end of an episode, final step is reached

Continuing tasks

Episodic tasks

T is the final step. RL breaks into *episodes*.

At the end of an episode, final step is reached

The simulation resets to starting state

Continuing tasks

Episodic tasks

T is the final step. RL breaks into *episodes*.

At the end of an episode, final step is reached

The simulation resets to starting state

Continuing tasks

$T = \infty$, return could be infinite

Episodic tasks

T is the final step. RL breaks into *episodes*.

At the end of an episode, final step is reached

The simulation resets to starting state

Continuing tasks

$T = \infty$, return could be infinite

Use *discount rate* γ for convergence

Episodic tasks

T is the final step. RL breaks into *episodes*.

At the end of an episode, final step is reached

The simulation resets to starting state

Continuing tasks

$T = \infty$, return could be infinite

Use *discount rate* γ for convergence

$0 < \gamma < 1$

Episodic tasks

T is the final step. RL breaks into *episodes*.

At the end of an episode, final step is reached

The simulation resets to starting state

Continuing tasks

$T = \infty$, return could be infinite

Use *discount rate* γ for convergence

$0 < \gamma < 1$

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

Episodic tasks

T is the final step. RL breaks into *episodes*.

At the end of an episode, final step is reached

The simulation resets to starting state

Continuing tasks

$T = \infty$, return could be infinite

Use *discount rate* γ for convergence

$0 < \gamma < 1$

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

Policies and Value functions

Policies and Value functions

RL involves evaluating value functions

Policies and Value functions

RL involves evaluating value functions

Functions of states or state-action pairs

Policies and Value functions

RL involves evaluating value functions

Functions of states or state-action pairs

- How good is it to be in a certain state (or to take an action in a certain state)

Policies and Value functions

RL involves evaluating value functions

Functions of states or state-action pairs

- How good is it to be in a certain state (or to take an action in a certain state)

Policy π decide actions, maps states to actions

Policies and Value functions

RL involves evaluating value functions

Functions of states or state-action pairs

- How good is it to be in a certain state (or to take an action in a certain state)

Policy π decide actions, maps states to actions

At time t, $\pi(a | s)$ is the probability that $A_t = a$ given $S_t = s$

Policies and Value functions

RL involves evaluating value functions

Functions of states or state-action pairs

- How good is it to be in a certain state (or to take an action in a certain state)

Policy π decide actions, maps states to actions

At time t, $\pi(a | s)$ is the probability that $A_t = a$ given $S_t = s$

RL is about changing the policy as a result of its experience

Q-Learning

Q-Learning

One of the first major *online* RL algorithms

Q-Learning

One of the first major *online* RL algorithms

Bellman Equation for Q-Learning

Q-Learning

One of the first major *online* RL algorithms

Bellman Equation for Q-Learning

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Q-Learning

One of the first major *online* RL algorithms

Bellman Equation for Q-Learning

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Use samples from the environment

Q-Learning

One of the first major *online* RL algorithms

Bellman Equation for Q-Learning

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Use samples from the environment

Q-Learning converges to the optimal value function, as long as the agent continues to explore and samples all areas at the state-action space

Q-Learning

One of the first major *online* RL algorithms

Bellman Equation for Q-Learning

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Use samples from the environment

Q-Learning converges to the optimal value function, as long as the agent continues to explore and samples all areas at the state-action space

Q-Learning target policy is Greedy wrt its current values (max operator)

Q-Learning

One of the first major *online* RL algorithms

Bellman Equation for Q-Learning

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Use samples from the environment

Q-Learning converges to the optimal value function, as long as the agent continues to explore and samples all areas at the state-action space

Q-Learning target policy is Greedy wrt its current values (max operator)

Its behaviour policy can be anything that continues to visit all state actions pairs during learning (e.g. ε – greedy)

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

 Initialize S

 Loop for each step of episode:

 Choose A from S using policy derived from Q (e.g., ε -greedy)

 Take action A , observe R, S'

$$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$$

$S \leftarrow S'$

 until S is terminal

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

 Initialize S

 Loop for each step of episode:

 Choose A from S using policy derived from Q (e.g., ε -greedy)

 Take action A , observe R, S'

$$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$$

$$S \leftarrow S'$$

 until S is terminal

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

 Initialize S

 Loop for each step of episode:

 Choose A from S using policy derived from Q (e.g., ε -greedy)

 Take action A , observe R, S'

$$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$$

$S \leftarrow S'$

 until S is terminal

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Bellman equation for state-action values

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

 Initialize S

 Loop for each step of episode:

 Choose A from S using policy derived from Q (e.g., ε -greedy)

 Take action A , observe R, S'

$$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$$

$S \leftarrow S'$

 until S is terminal

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Bellman equation for state-action values

What is $Q(S, A)$?

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

 Initialize S

 Loop for each step of episode:

 Choose A from S using policy derived from Q (e.g., ε -greedy)

 Take action A , observe R, S'

$$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$$

$S \leftarrow S'$

 until S is terminal

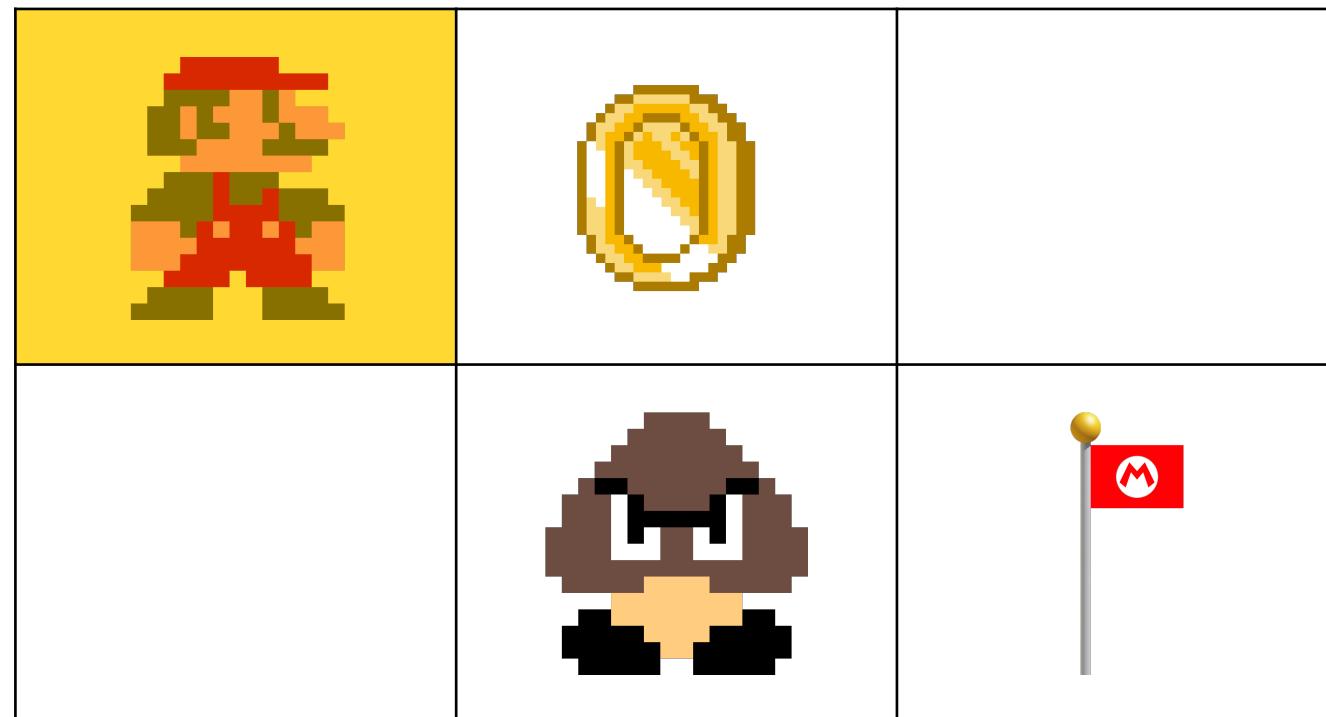
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Bellman equation for state-action values

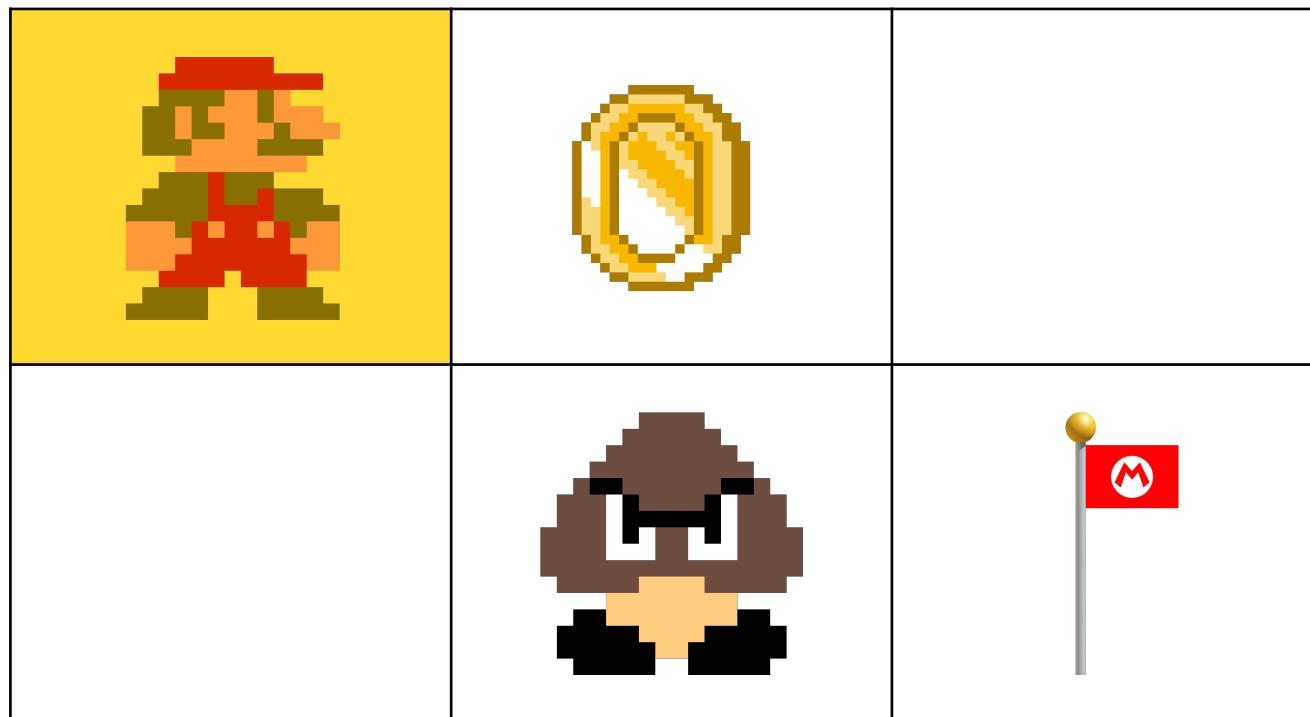
What is $Q(S, A)$?

Encode Q-values in a table, each cell corresponds to a state-action pair value.

Q-Learning example

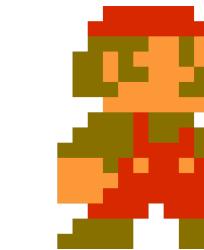
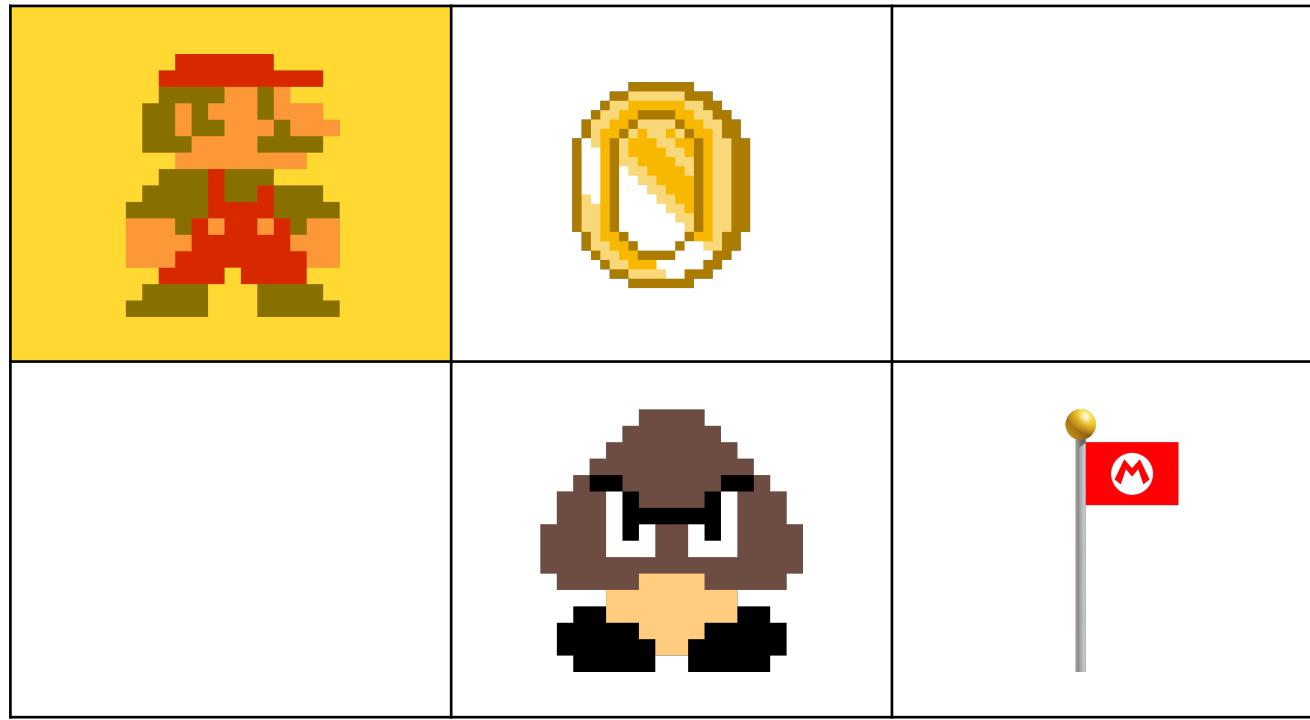


Q-Learning example



Policy π

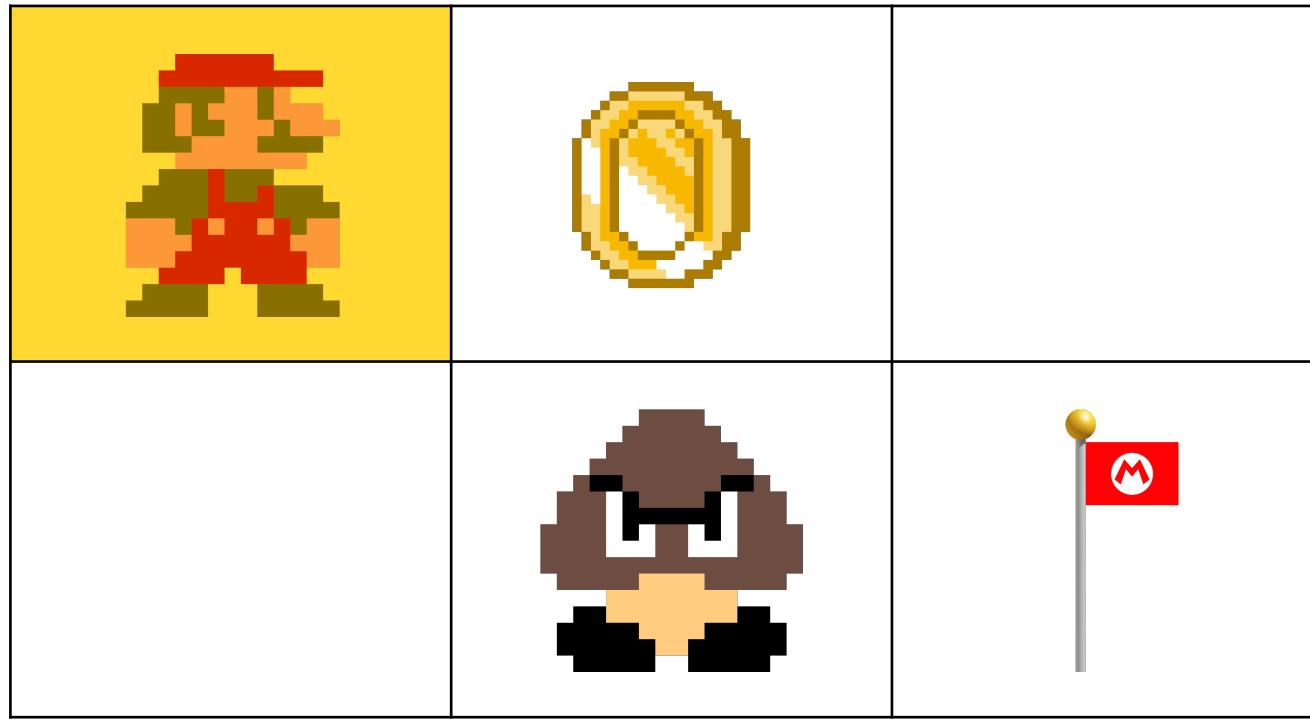
Q-Learning example



Policy π

Actions $\uparrow \downarrow \leftarrow \rightarrow$

Q-Learning example



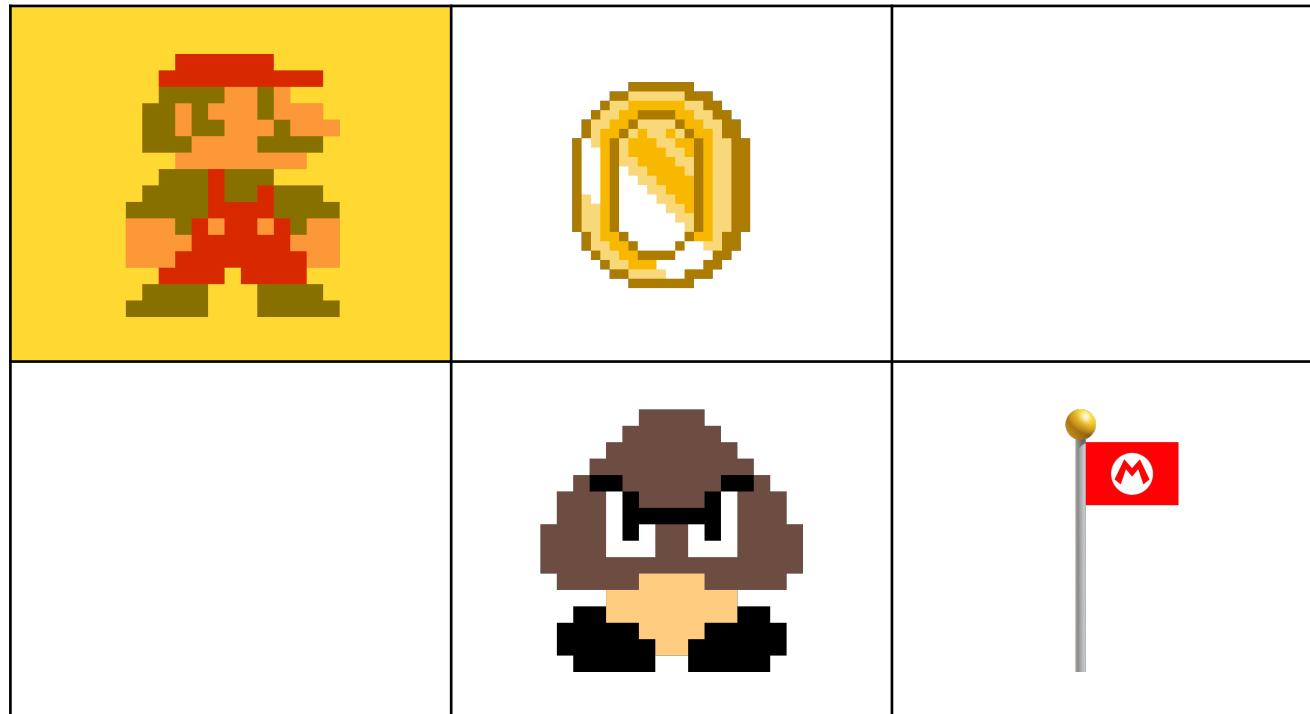
Policy π

Actions $\uparrow \downarrow \leftarrow \rightarrow$



+1

Q-Learning example



Policy π

Actions $\uparrow \downarrow \leftarrow \rightarrow$

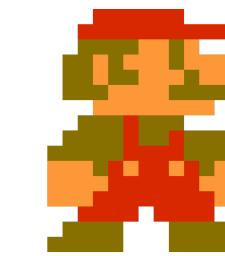
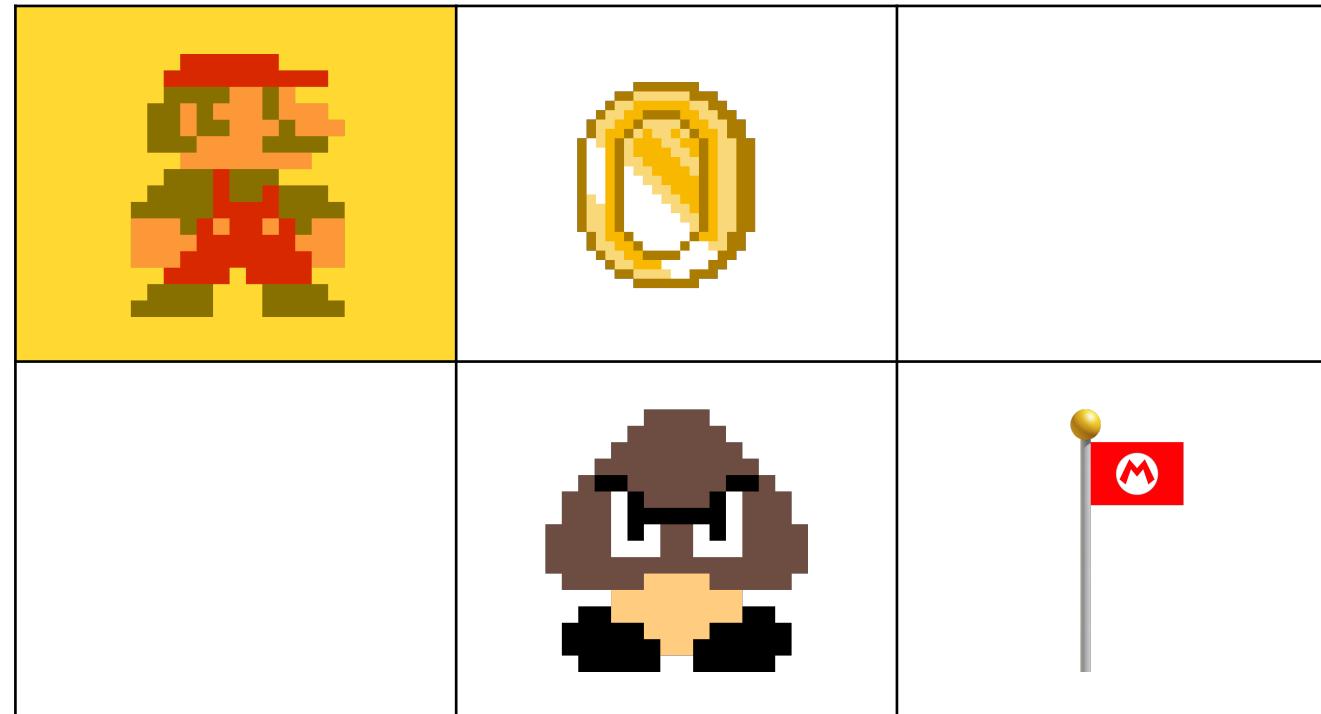


+1



-10

Q-Learning example

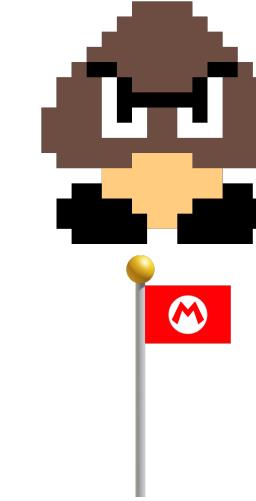


Policy π

Actions $\uparrow \downarrow \leftarrow \rightarrow$

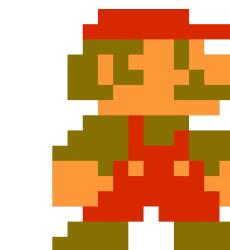
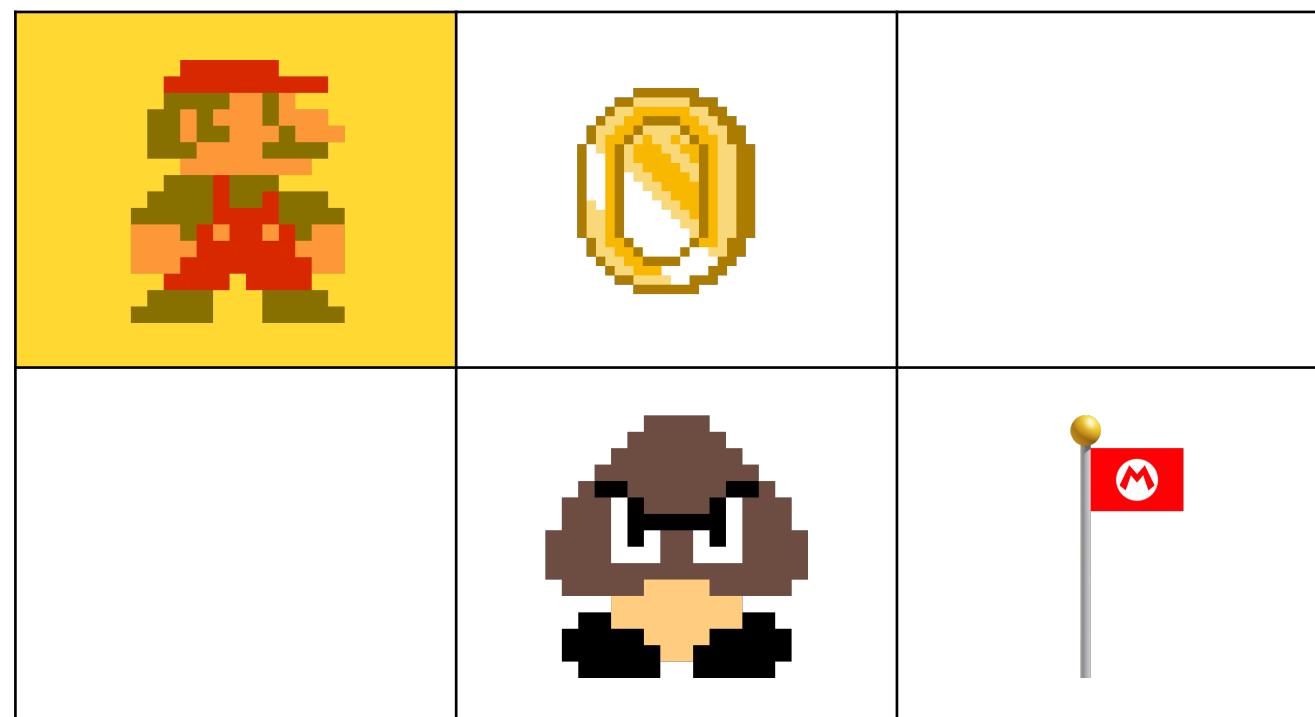


+1



-10
+10
(terminal)

Q-Learning example



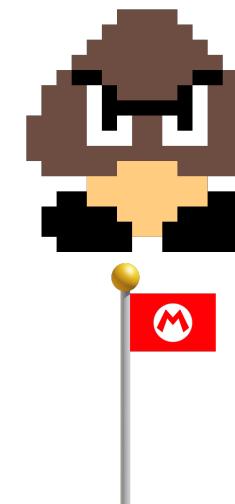
Policy π

Actions



+1

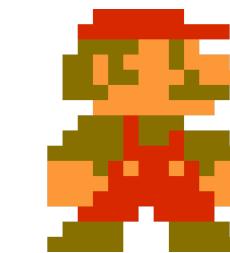
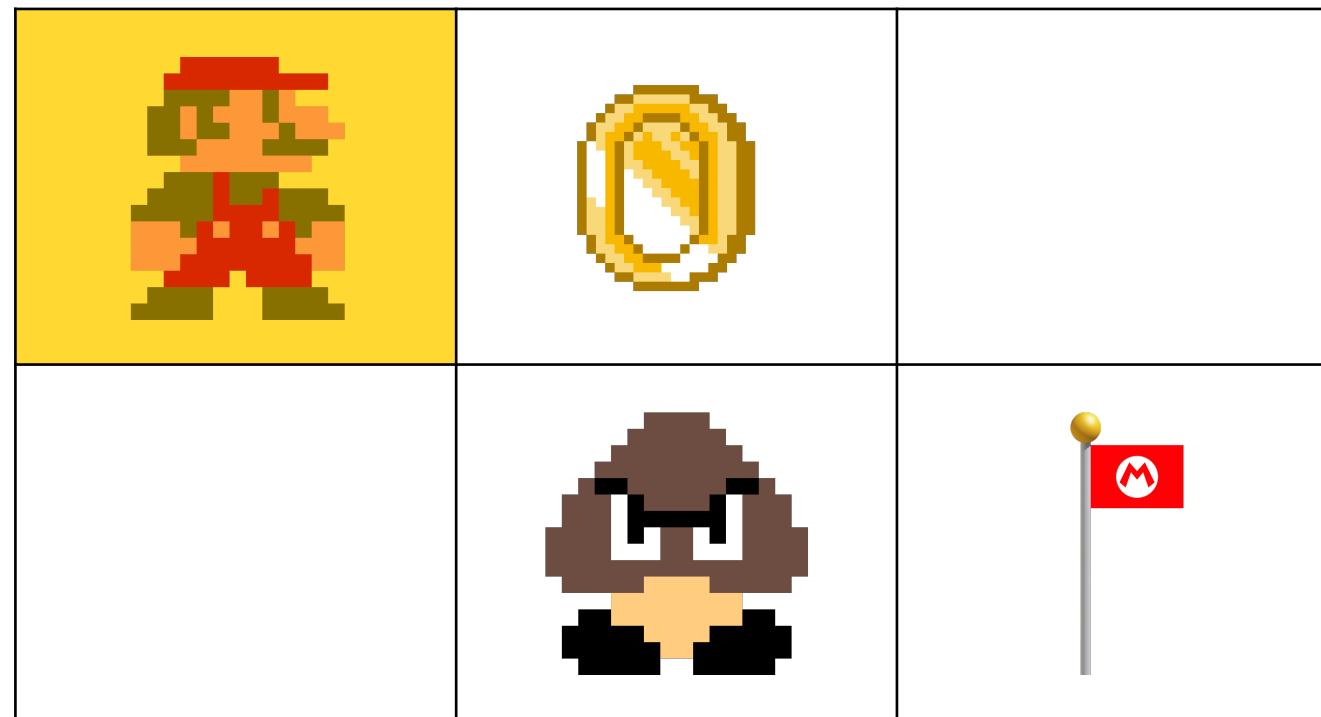
$\uparrow \downarrow \leftarrow \rightarrow$



-10
+10
(terminal)

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Q-Learning example

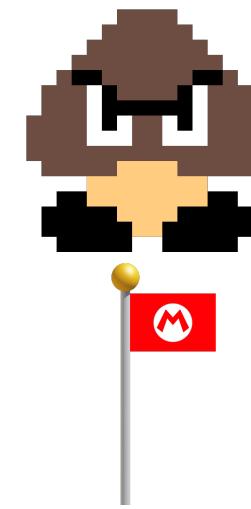


Policy π

Actions $\uparrow \downarrow \leftarrow \rightarrow$



$+1$

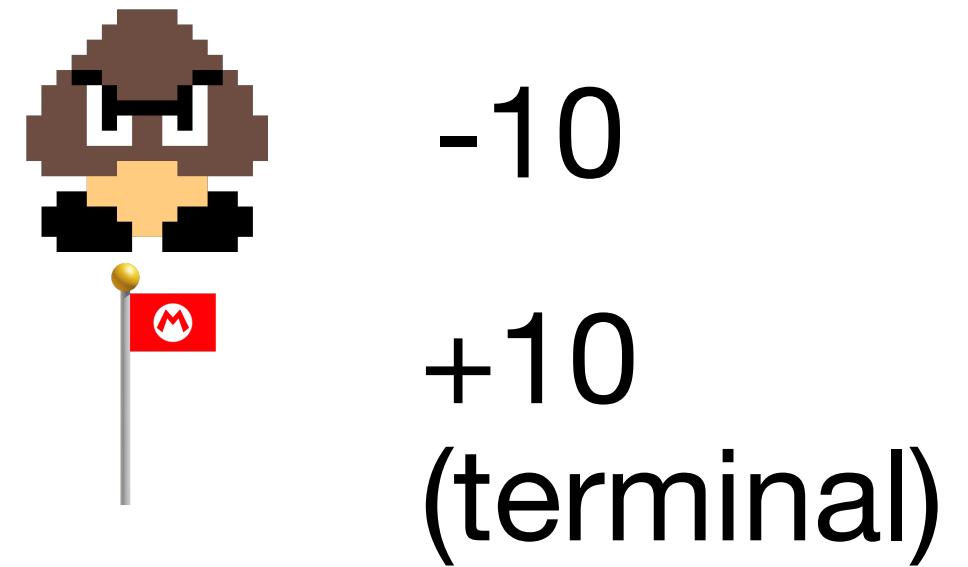
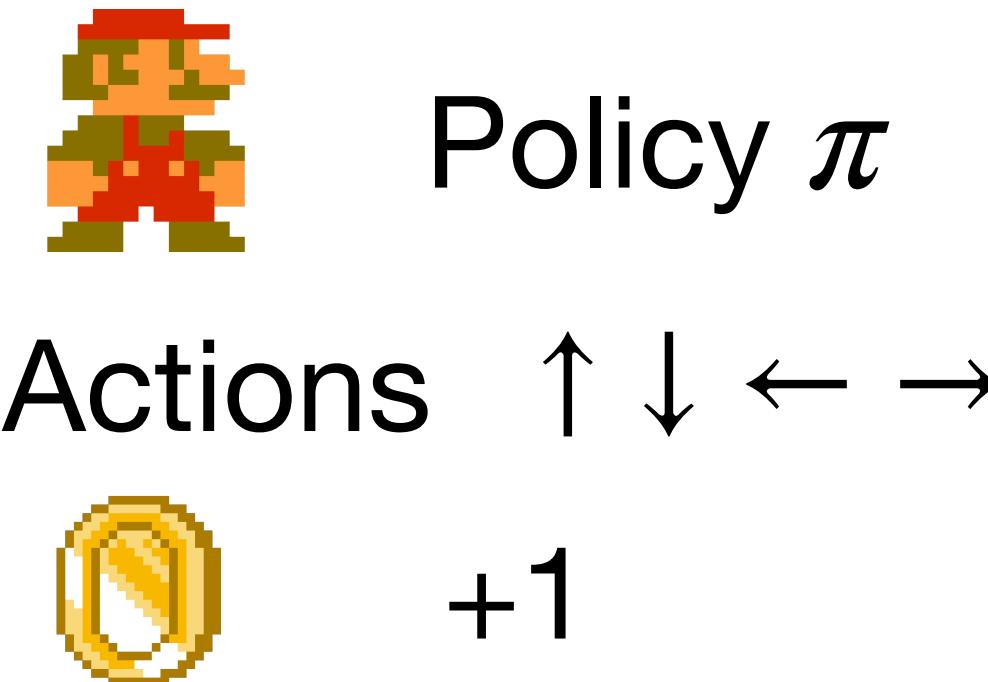
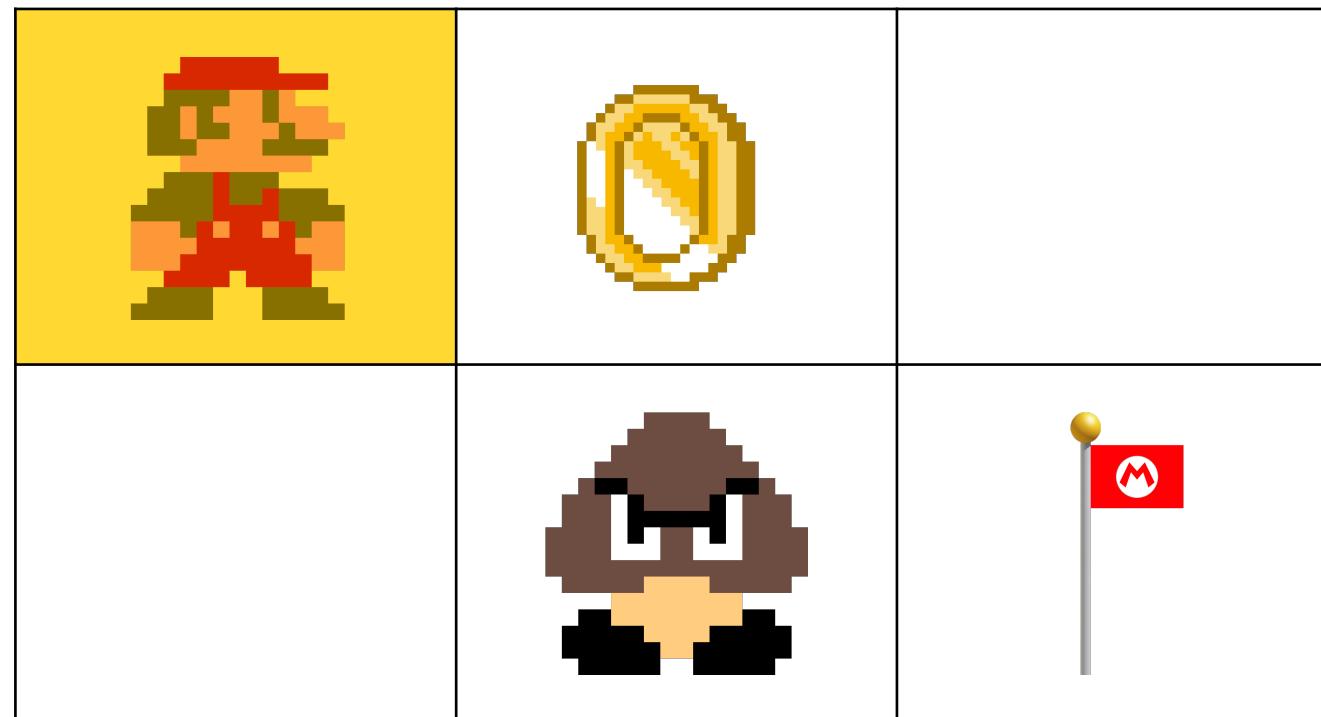


-10
 $+10$
(terminal)

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$\begin{aligned}\alpha &= 0.1 \\ \gamma &= 0.99\end{aligned}$$

Q-Learning example

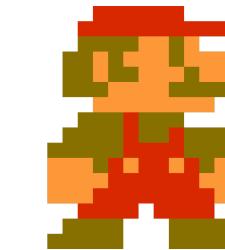
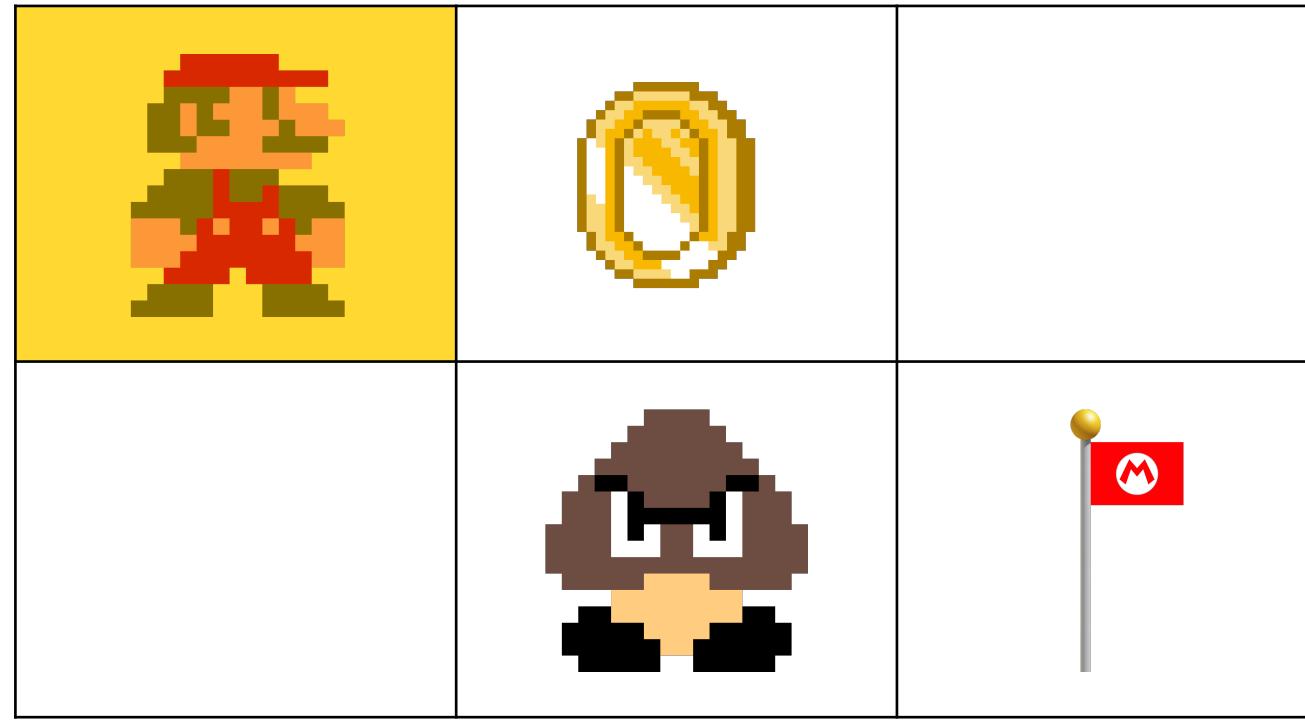


$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$\begin{aligned}\alpha &= 0.1 \\ \gamma &= 0.99\end{aligned}$$

How to represent the Q-Table?

Q-Learning example

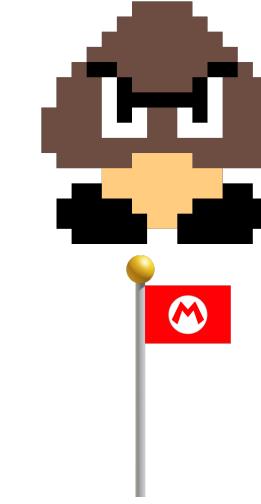


Policy π

Actions $\uparrow \downarrow \leftarrow \rightarrow$



+1



-10
+10
(terminal)

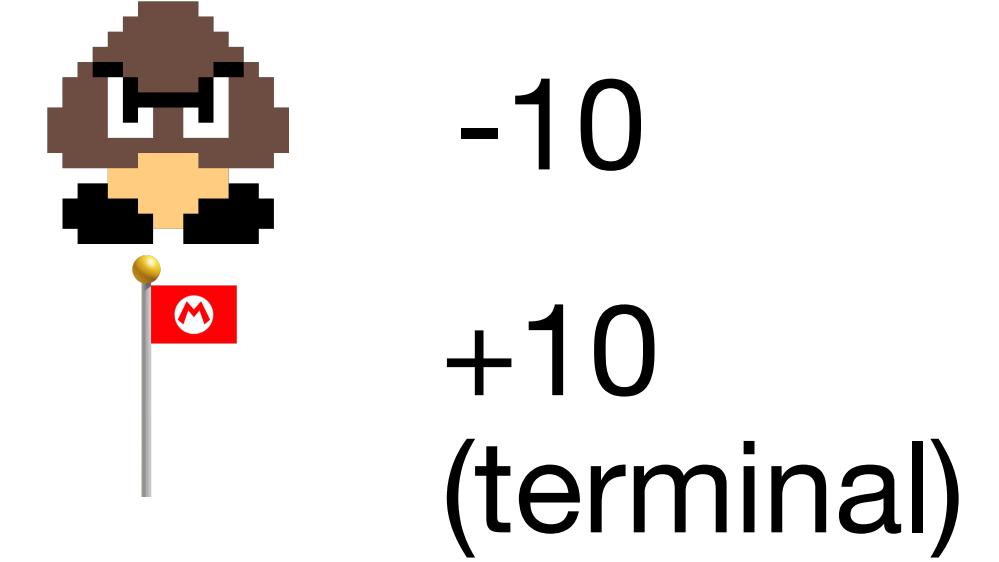
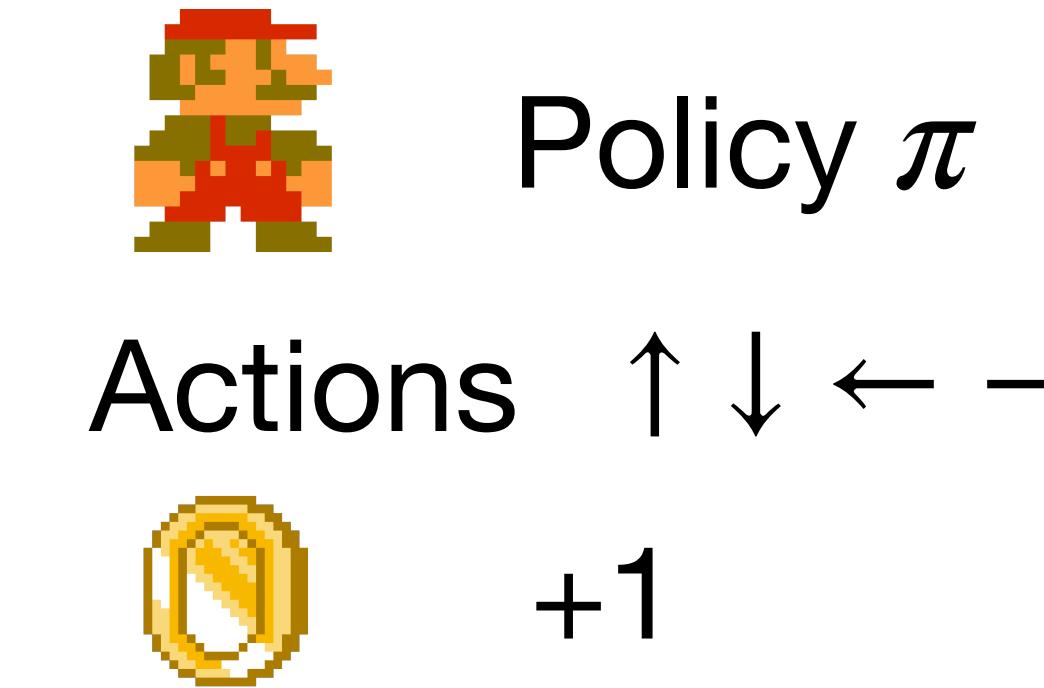
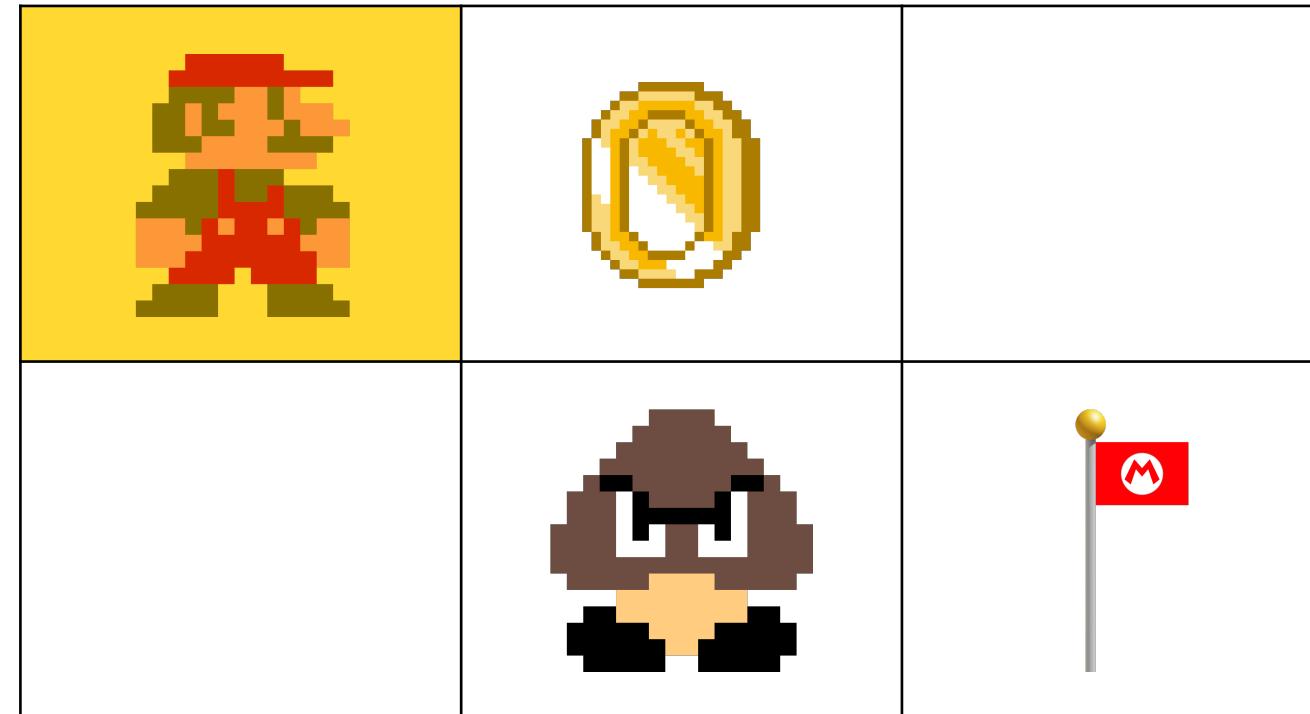
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$\begin{aligned}\alpha &= 0.1 \\ \gamma &= 0.99\end{aligned}$$

How to represent the Q-Table?

	\leftarrow	\rightarrow	\uparrow	\downarrow
Yellow				
Gold Coin				
Grey				
Goomba				
Flag				

Q-Learning example



$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

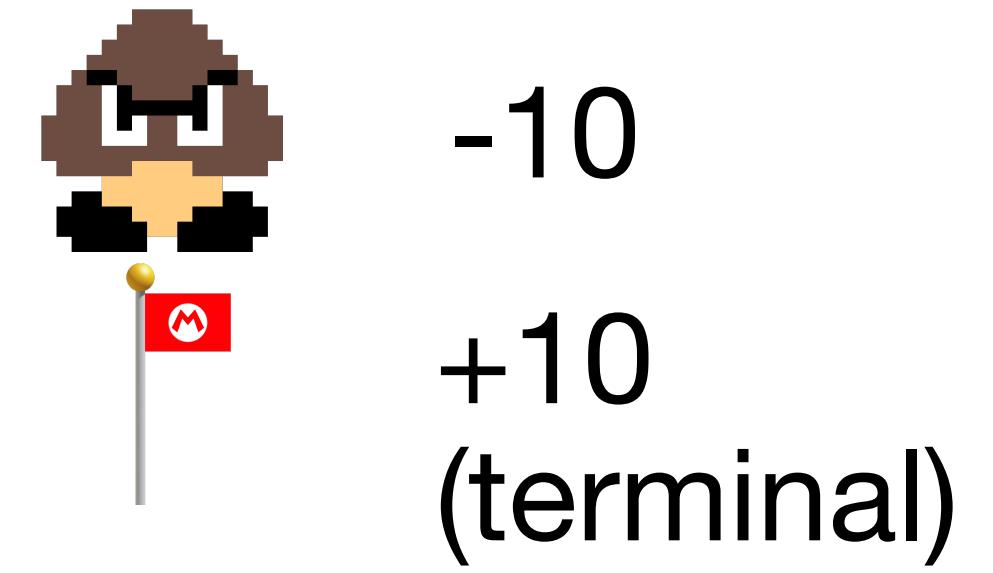
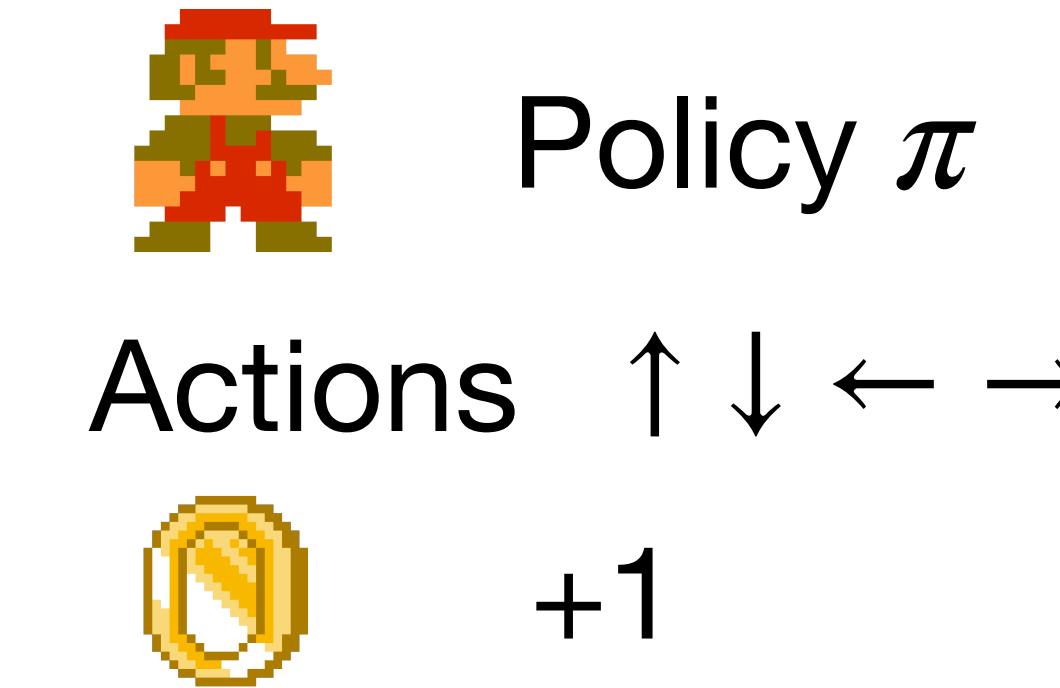
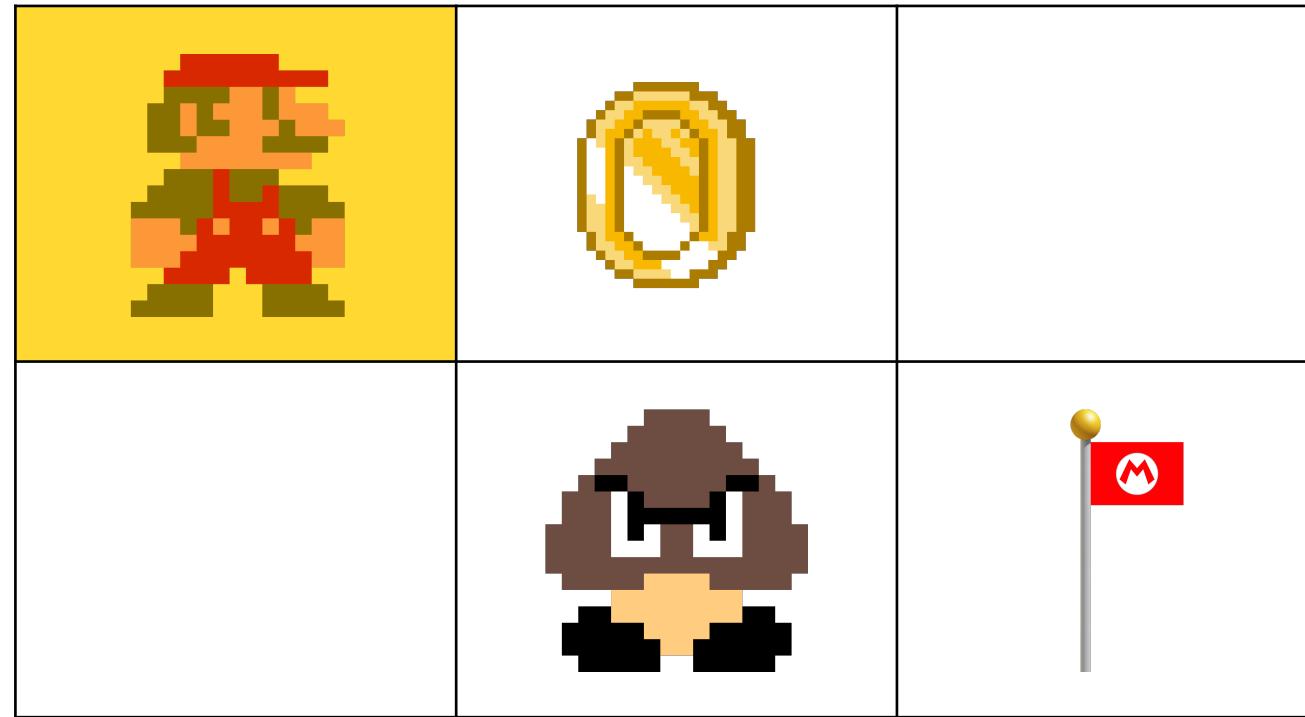
$$\begin{aligned}\alpha &= 0.1 \\ \gamma &= 0.99\end{aligned}$$

How to represent the Q-Table?

	\leftarrow	\rightarrow	\uparrow	\downarrow

How to move?

Q-Learning example



$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$\alpha = 0.1$$
$$\gamma = 0.99$$

How to represent the Q-Table?

	\leftarrow	\rightarrow	\uparrow	\downarrow
Mario				
Goomba				
Blank				
Flag				

How to move?

$$\pi = \operatorname{argmax}_A Q(S, A)$$

Q-Learning example

Q-Learning example

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+$, $a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Q-Learning example

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

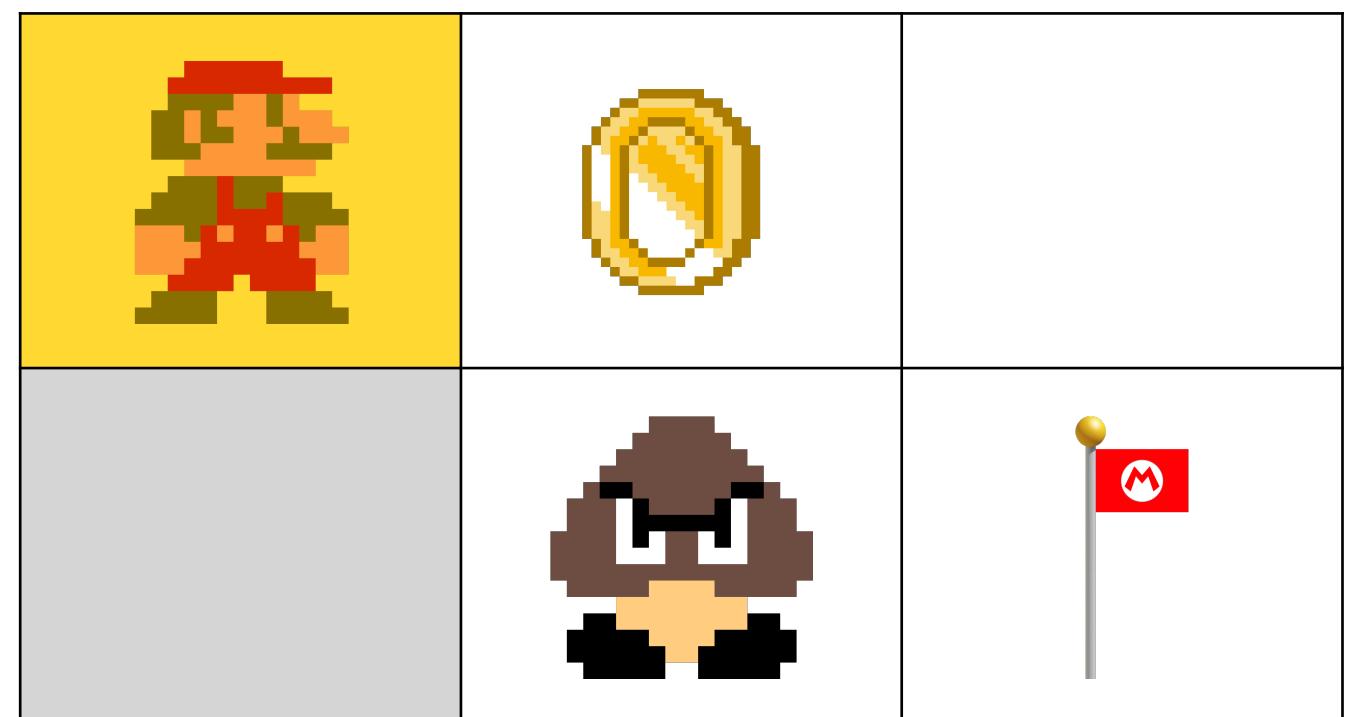
	\leftarrow	\rightarrow	\uparrow	\downarrow
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0

Q-Learning example

	←	→	↑	↓
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0

Q-Learning example

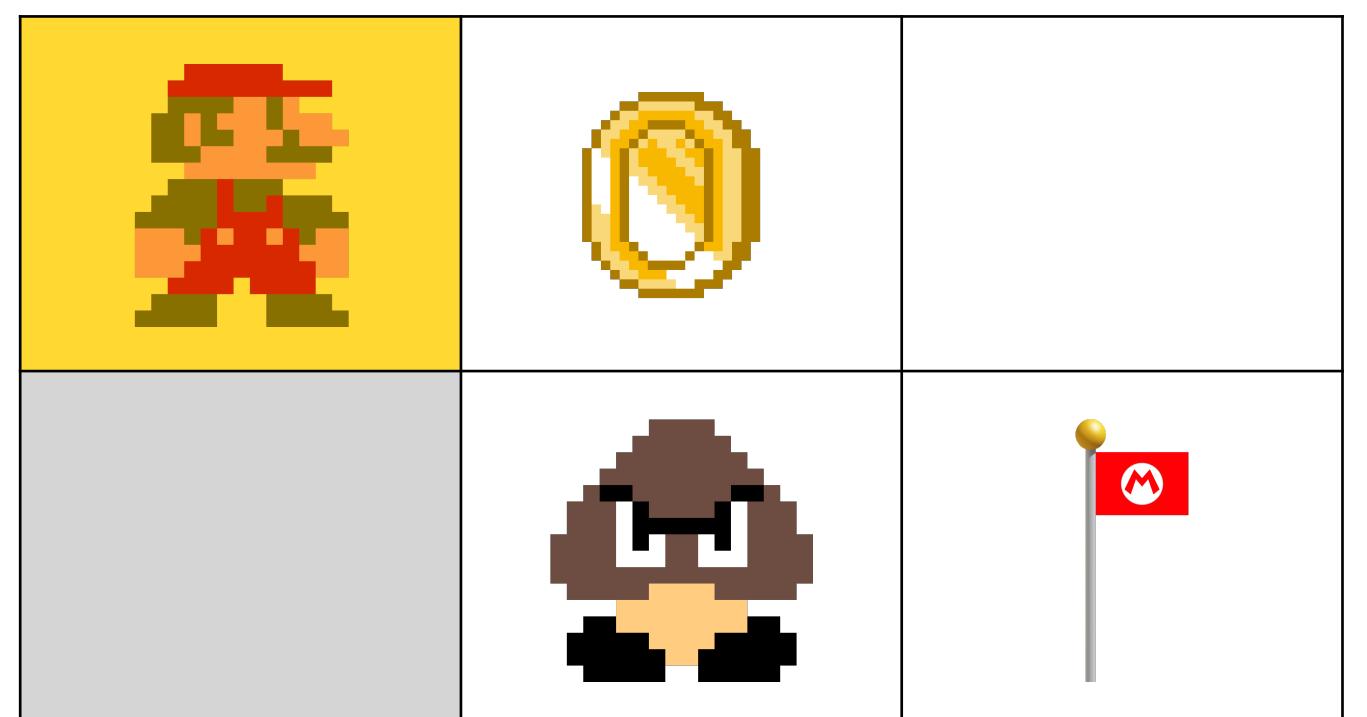
	←	→	↑	↓
Yellow	0	0	0	0
Gold Coin	0	0	0	0
Blank	0	0	0	0
Gray	0	0	0	0
Mushroom	0	0	0	0
Flagpole	0	0	0	0



Q-Learning example

	←	→	↑	↓
Gold Coin	0	0	0	0
Mushroom	0	0	0	0
Blank	0	0	0	0
Mario	0	0	0	0
Flag	0	0	0	0

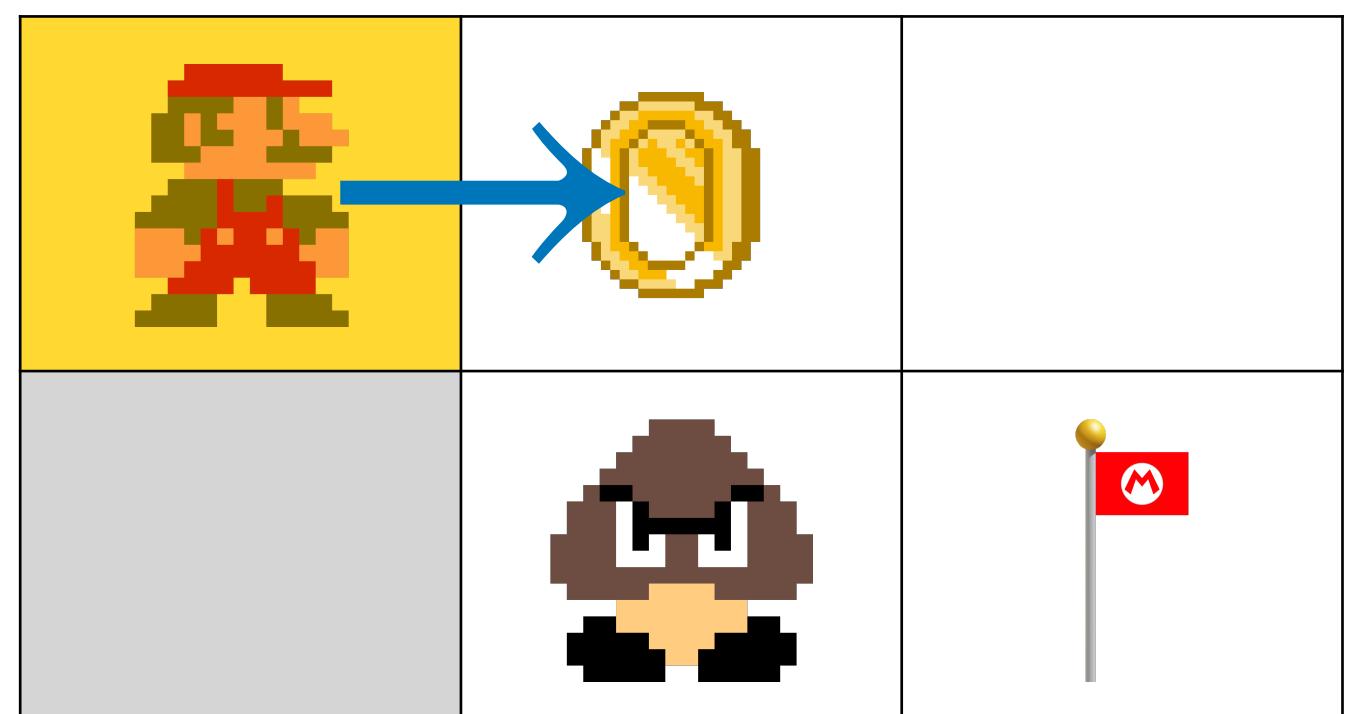
Choose A from S using policy derived from Q (e.g., ε -greedy)



Q-Learning example

	←	→	↑	↓
←	0	0	0	0
→	0	0	0	0
↑	0	0	0	0
↓	0	0	0	0
Gold Coin	0	0	0	0
Mario	0	0	0	0
Mushroom	0	0	0	0
Flag	0	0	0	0

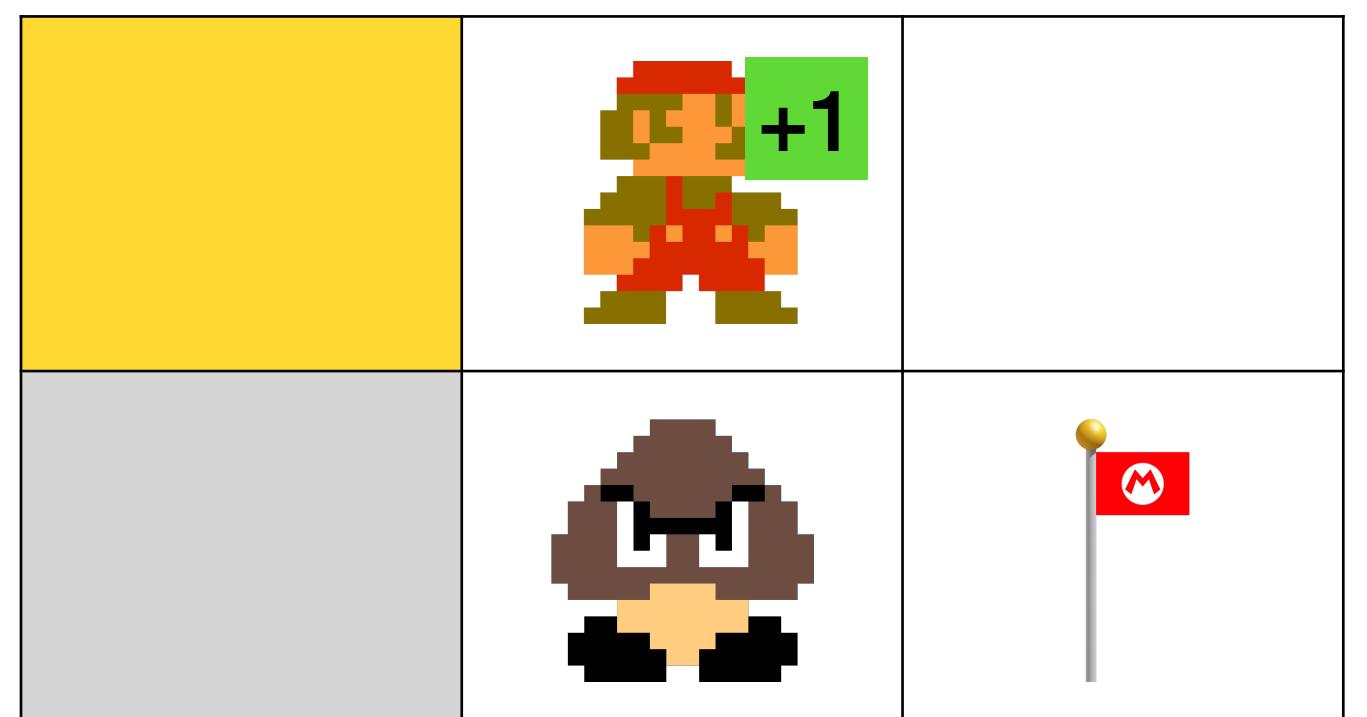
Choose A from S using policy derived from Q (e.g., ε -greedy)



Q-Learning example

	←	→	↑	↓
Gold Coin	0	0	0	0
Mushroom	0	0	0	0
Blank	0	0	0	0
Mushroom	0	0	0	0
Flag	0	0	0	0

Take action A , observe R, S'

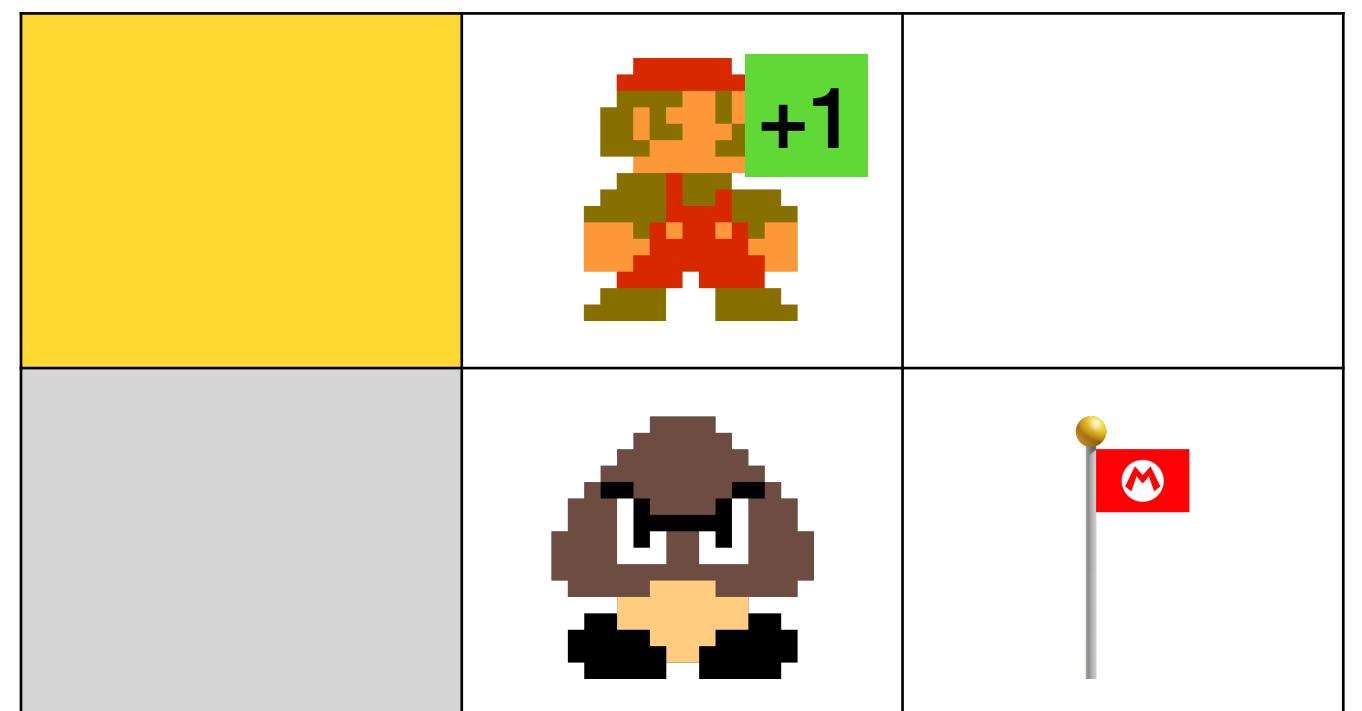


Q-Learning example

	←	→	↑	↓
←	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

Update $Q(S, A)$

Take action A , observe R, S'



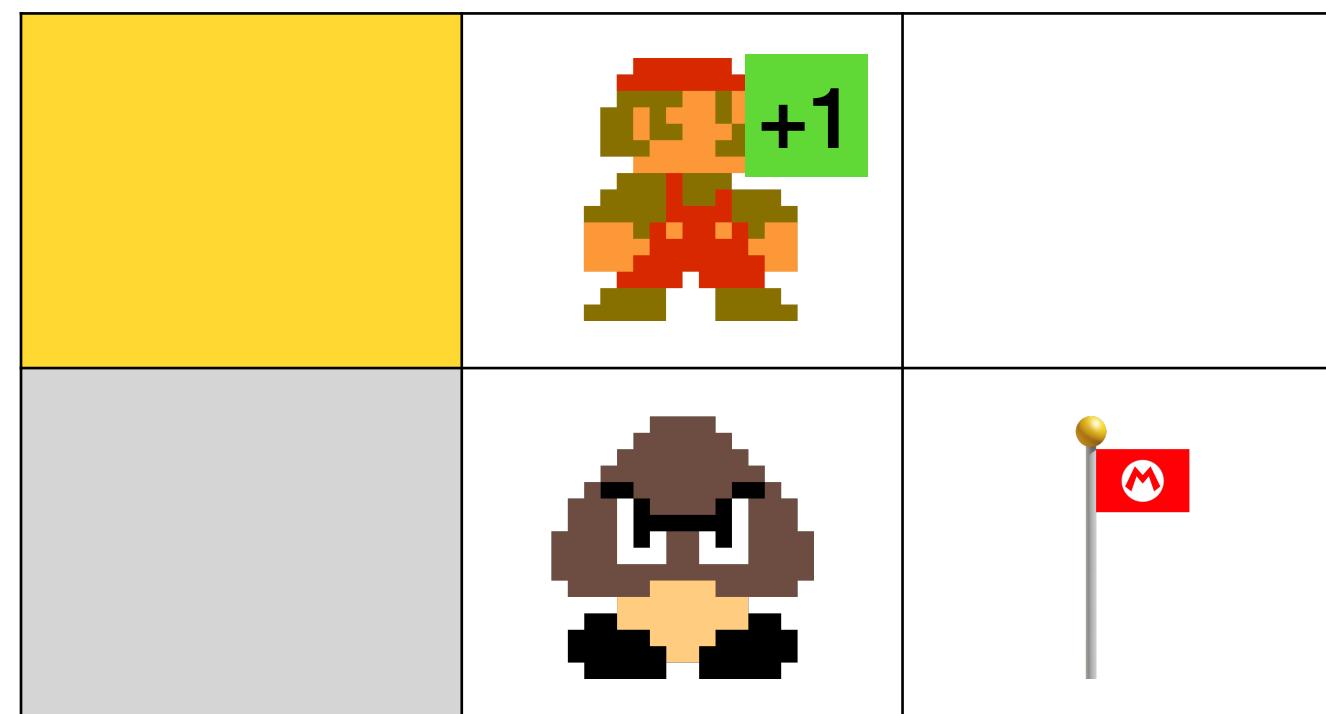
Q-Learning example

	←	→	↑	↓
←	0	0	0	0
0	0	0	0	0
0	0	0	0	0
↑	0	0	0	0
↓	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Take action A , observe R, S'



Q-Learning example

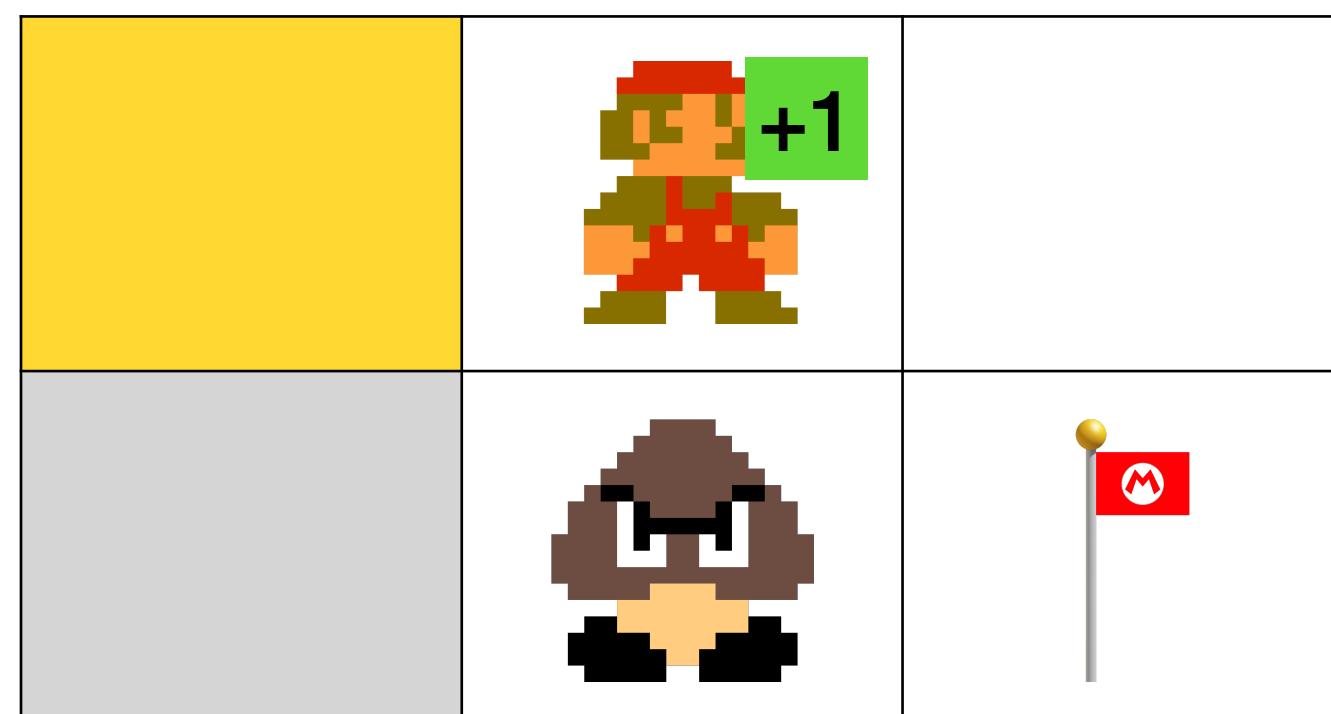
$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
←	0	0	0	0
0	0	0	0	0
↑	0	0	0	0
↓	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

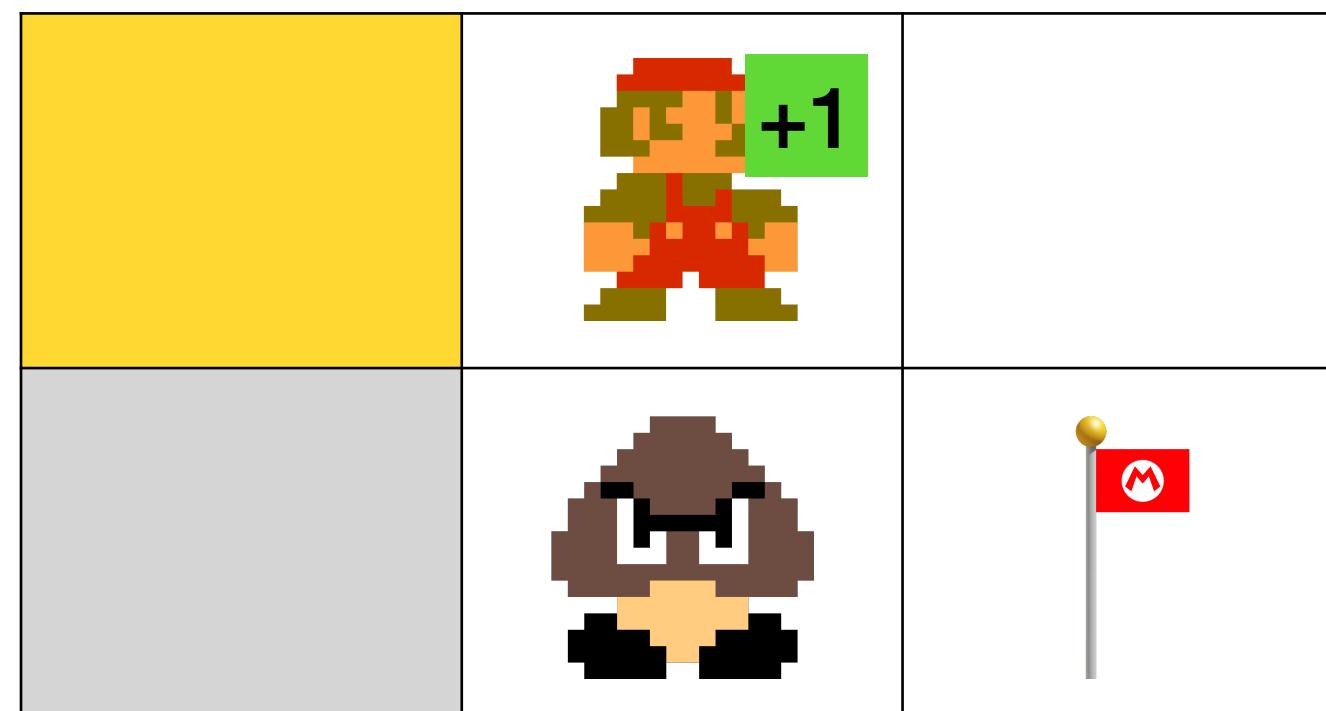
	←	→	↑	↓
←	0	0	0	0
0	0	0	0	0
↑	0	0	0	0
↓	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(\text{initial}, \rightarrow) = 0 + 0.1 * [1 + 0.99 * 0 - 0] = 0.1$$

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

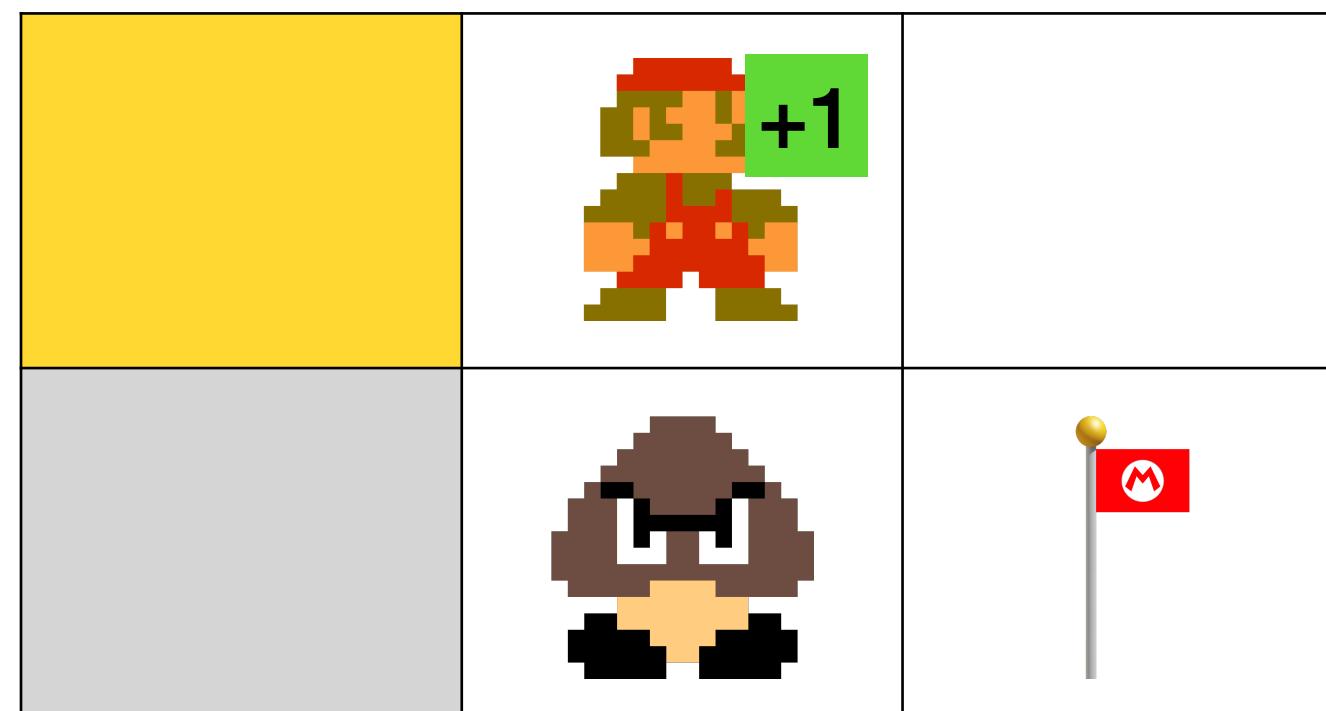
	←	→	↑	↓
←	0	0.1	0	0
0	0	0	0	0
↑	0	0	0	0
↓	0	0	0	0
Initial	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(\text{initial}, \rightarrow) = 0 + 0.1 * [1 + 0.99 * 0 - 0] = 0.1$$

Take action A , observe R, S'

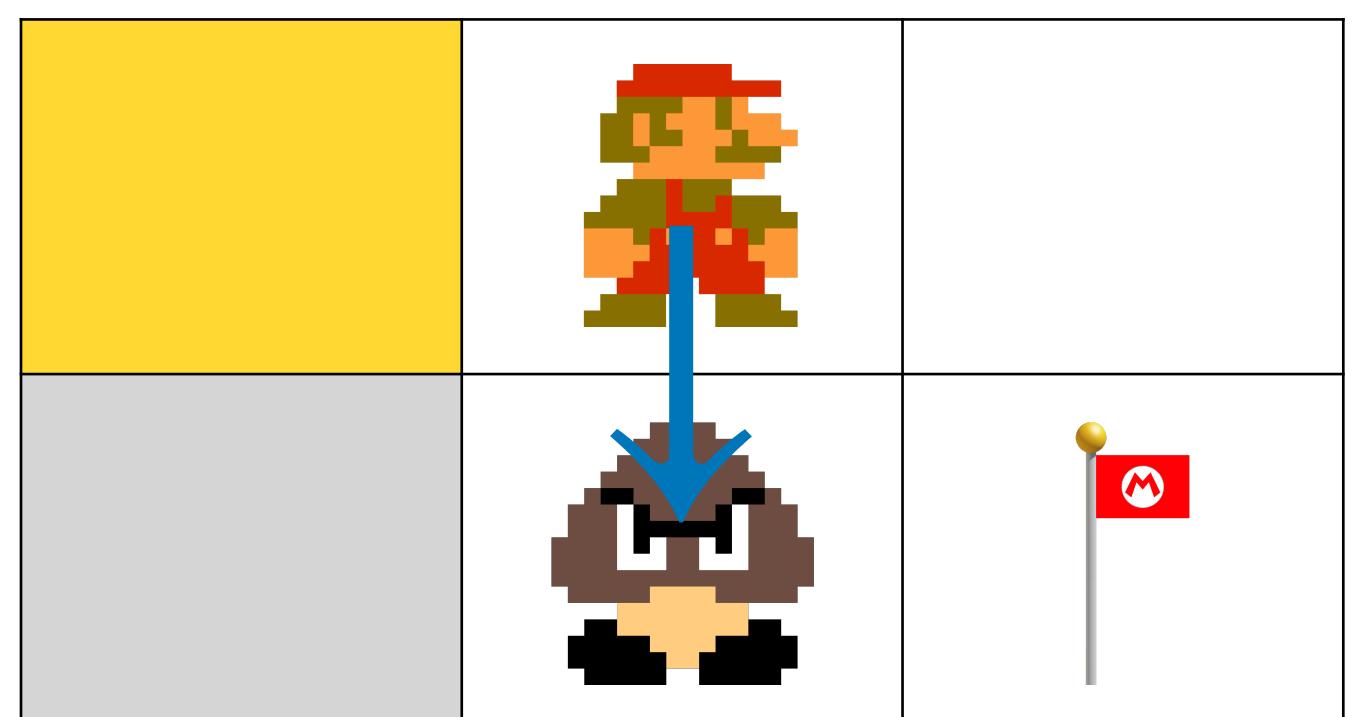


Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
←	0	0.1	0	0
↑	0	0	0	0
↓	0	0	0	0
→	0	0	0	0
Flag	0	0	0	0

Choose A from S using policy derived from Q (e.g., ε -greedy)

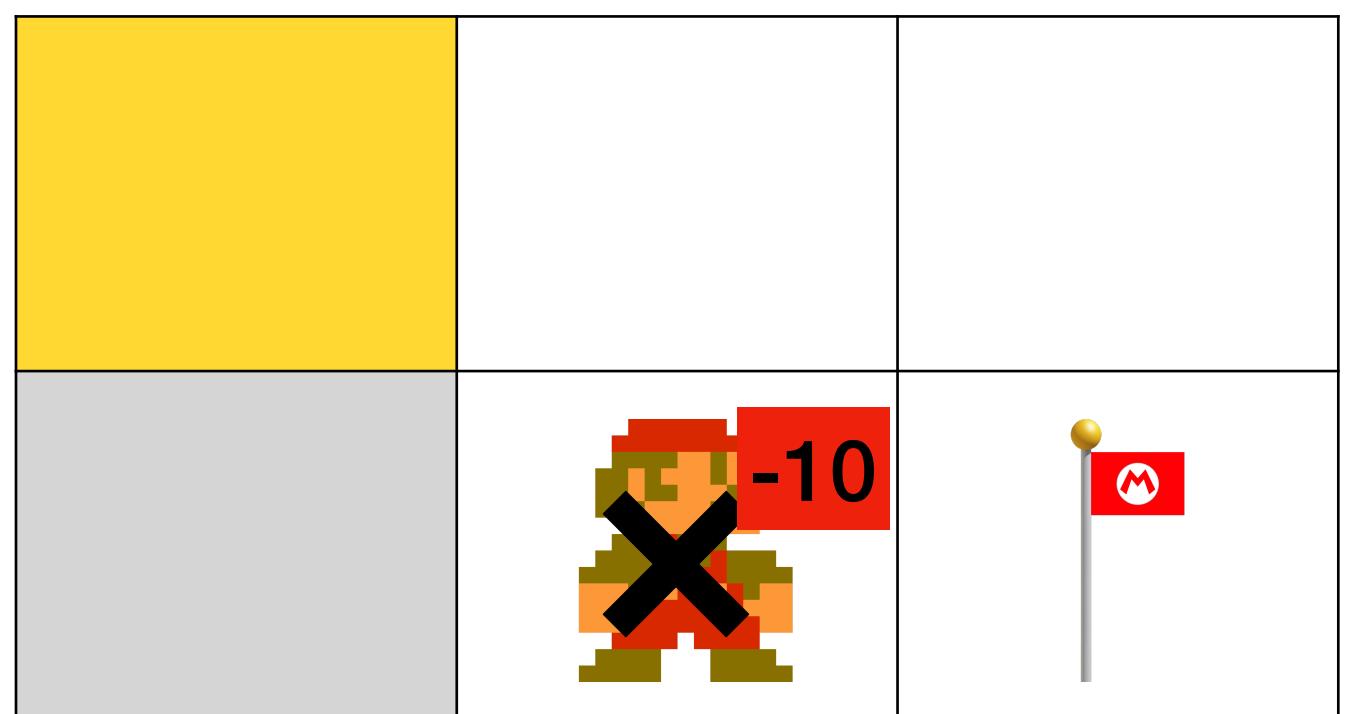


Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
←	0	0.1	0	0
↑	0	0	0	0
↓	0	0	0	0
→	0	0	0	0
Flag	0	0	0	0

Take action A , observe R, S'



Q-Learning example

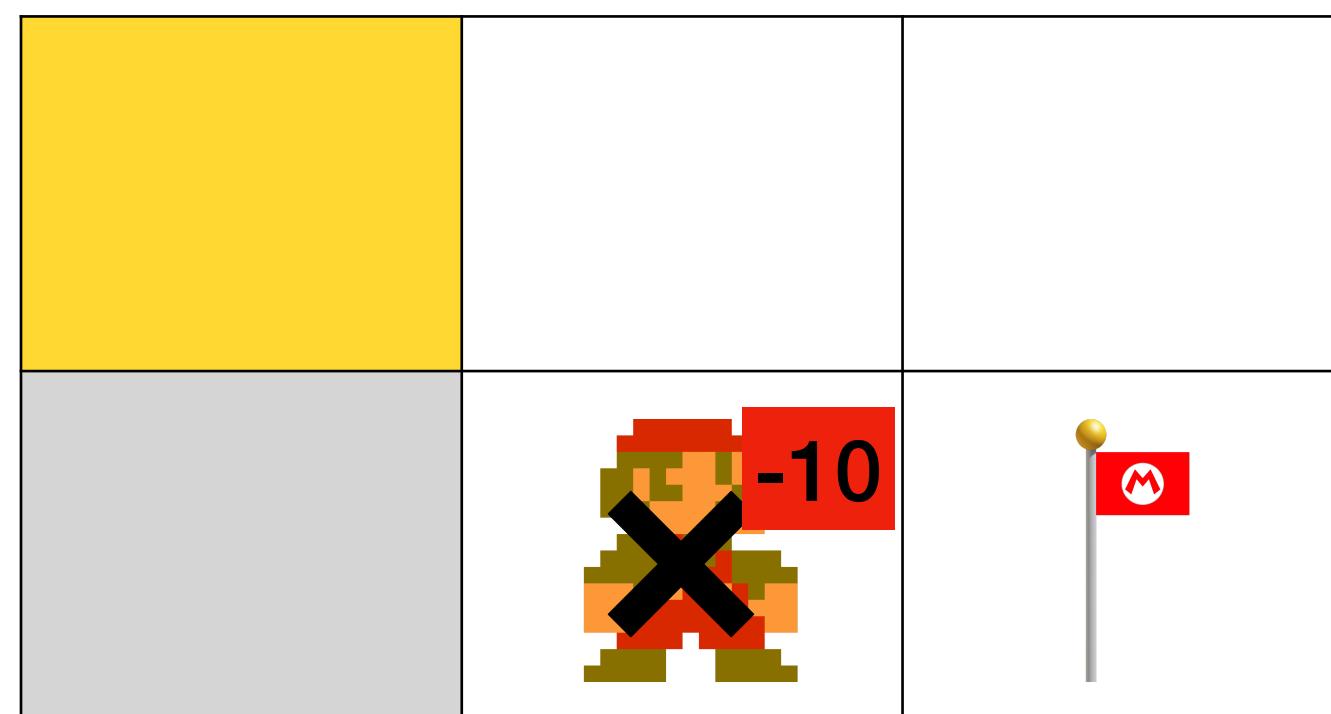
$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
←	0	0.1	0	0
0	0	0	0	0
0	0	0	0	0
↑	0	0	0	0
↓	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

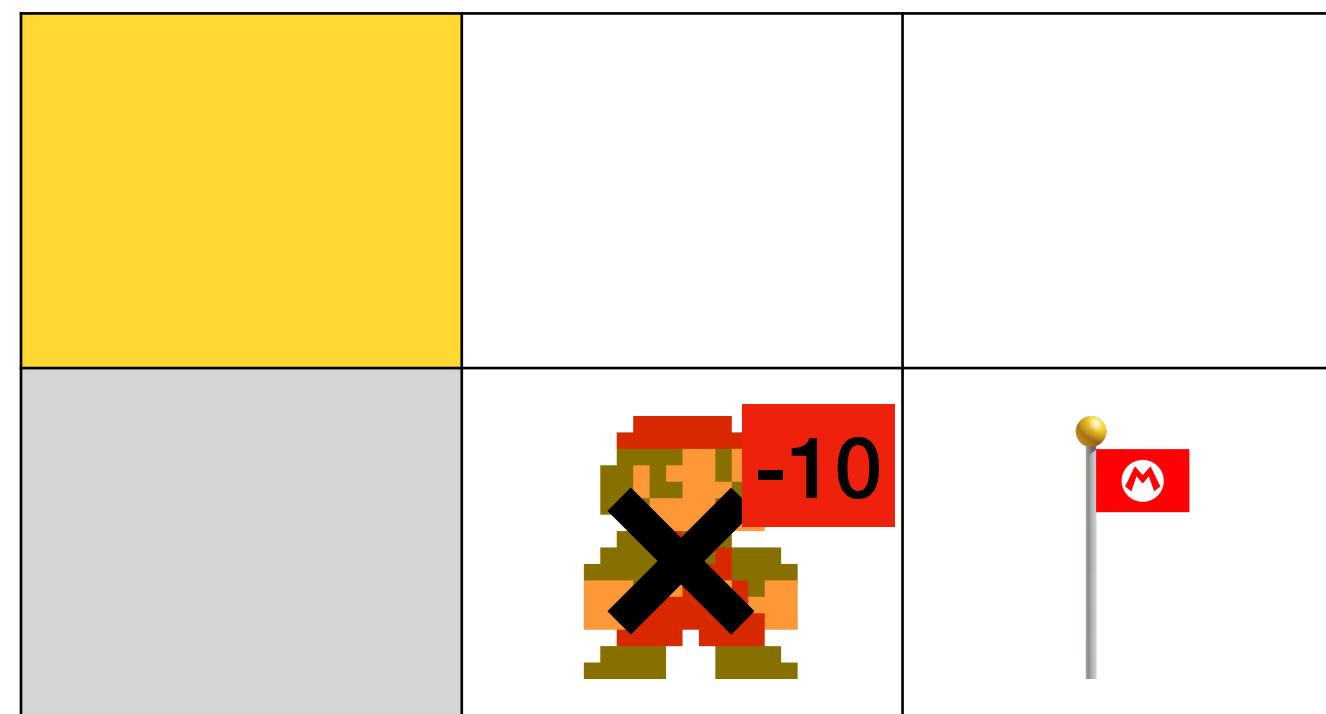
	←	→	↑	↓
←	0	0.1	0	0
0	0	0	0	0
0	0	0	0	0
↑	0	0	0	0
↓	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(\text{state_2}, \downarrow) = 0 + 0.1 * [-10 + 0.99 * 0 - 0] = -1$$

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

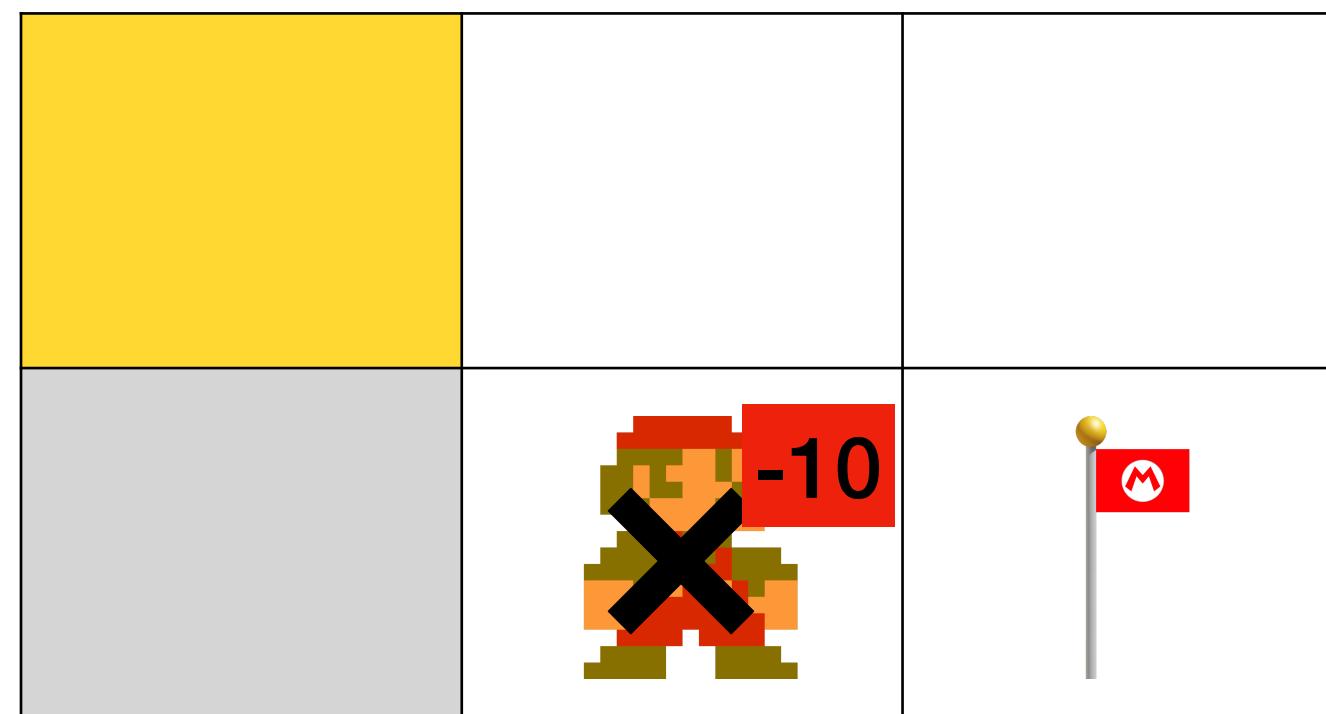
	←	→	↑	↓
←	0	0.1	0	0
0	0	0	0	-1
↑	0	0	0	0
↓	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(\text{state_2}, \downarrow) = 0 + 0.1 * [-10 + 0.99 * 0 - 0] = -1$$

Take action A , observe R, S'



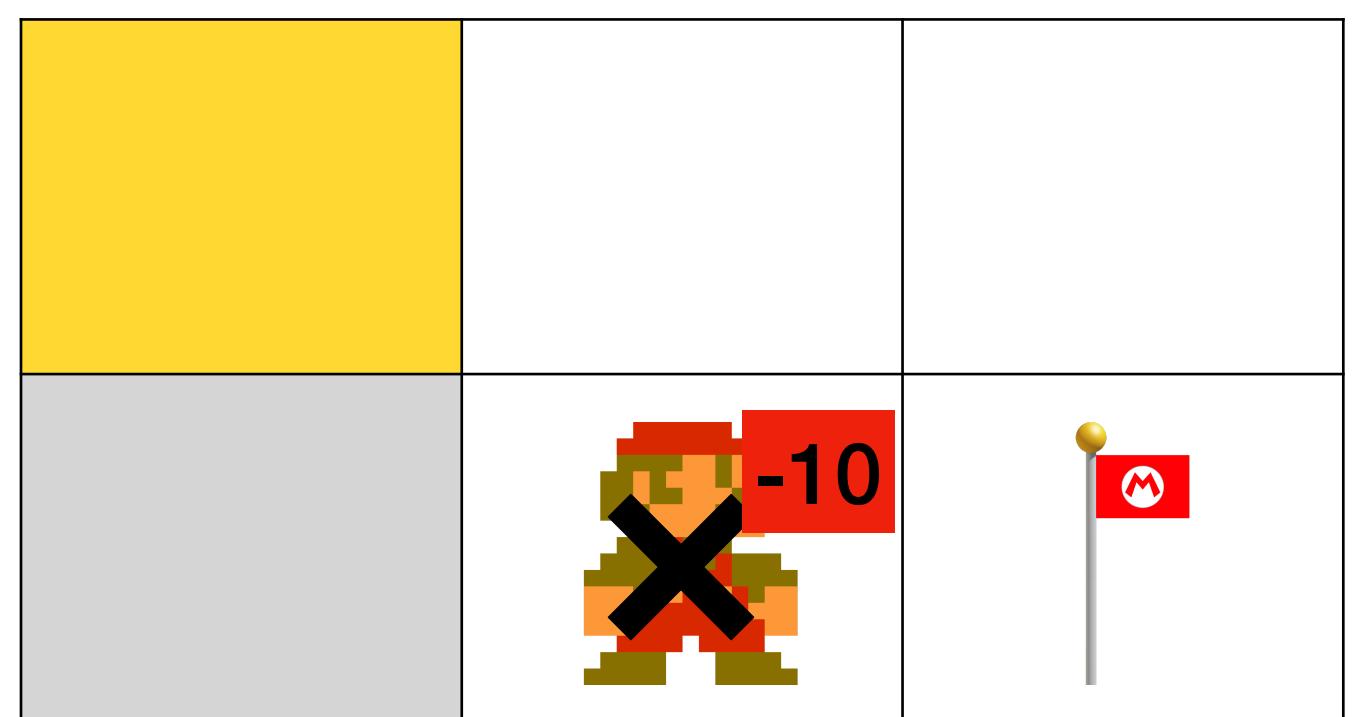
Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
←	0	0.1	0	0
↑	0	0	0	-1
↓	0	0	0	0
→	0	0	0	0
Exit	0	0	0	0

Terminal state, Episode resets

Take action A , observe R, S'



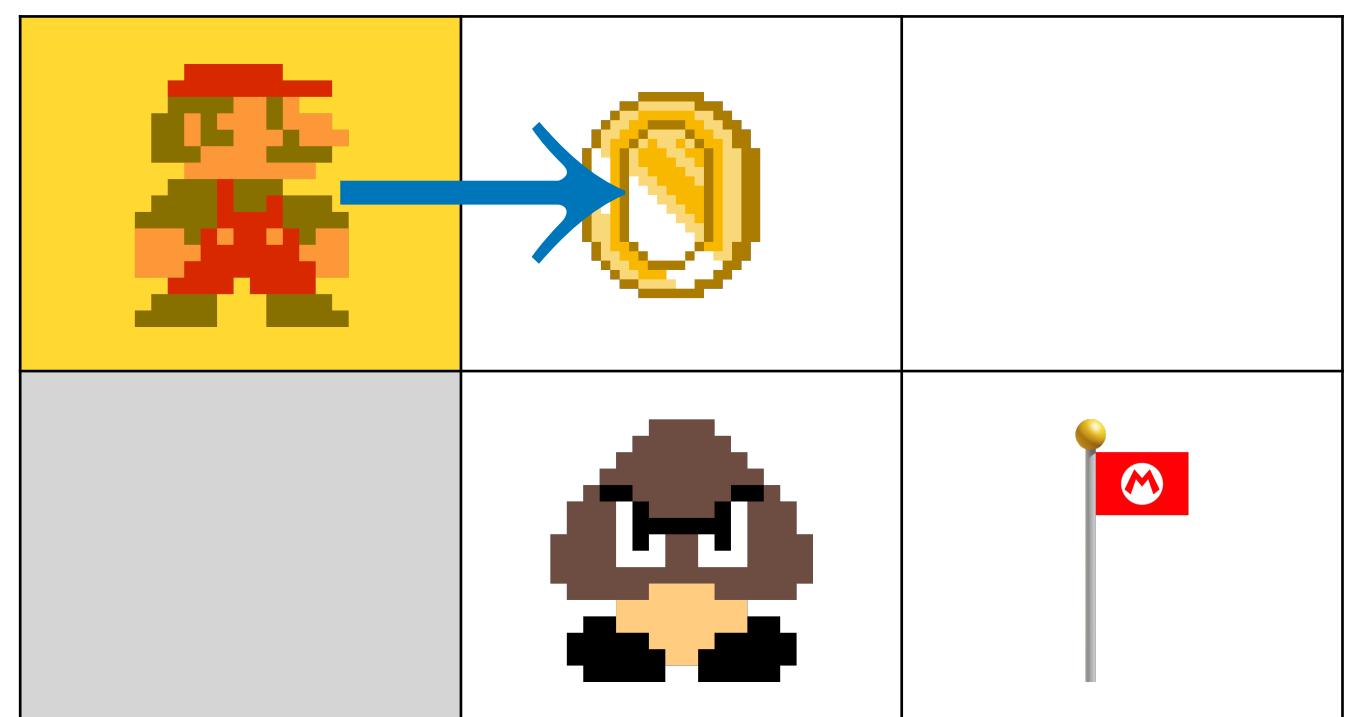
Q-Learning example

New episode

$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
←	0	0.1	0	0
→	0	0	0	-1
↑	0	0	0	0
↓	0	0	0	0
Mario	0	0	0	0
Flag	0	0	0	0

Choose A from S using policy derived from Q (e.g., ε -greedy)



Q-Learning example

$\alpha = 0.1$
 $\gamma = 0.99$

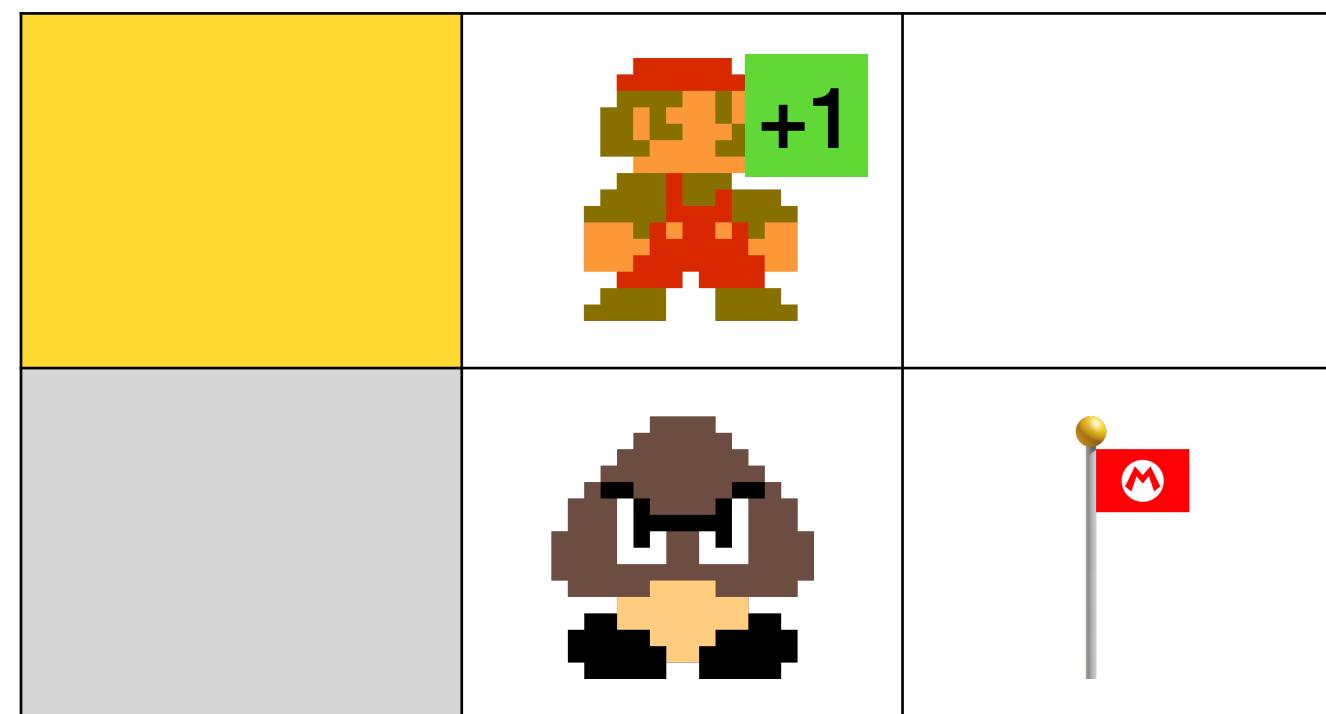
	\leftarrow	\rightarrow	\uparrow	\downarrow
Gold Coin	0	0.1	0	0
Ghost	0	0	0	-1
Blank	0	0	0	0
Mushroom	0	0	0	0
Flag	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(\text{initial}, \rightarrow) = 0.1 + 0.1 * [1 + 0.99 * 0 - 0] = 0.2$$

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

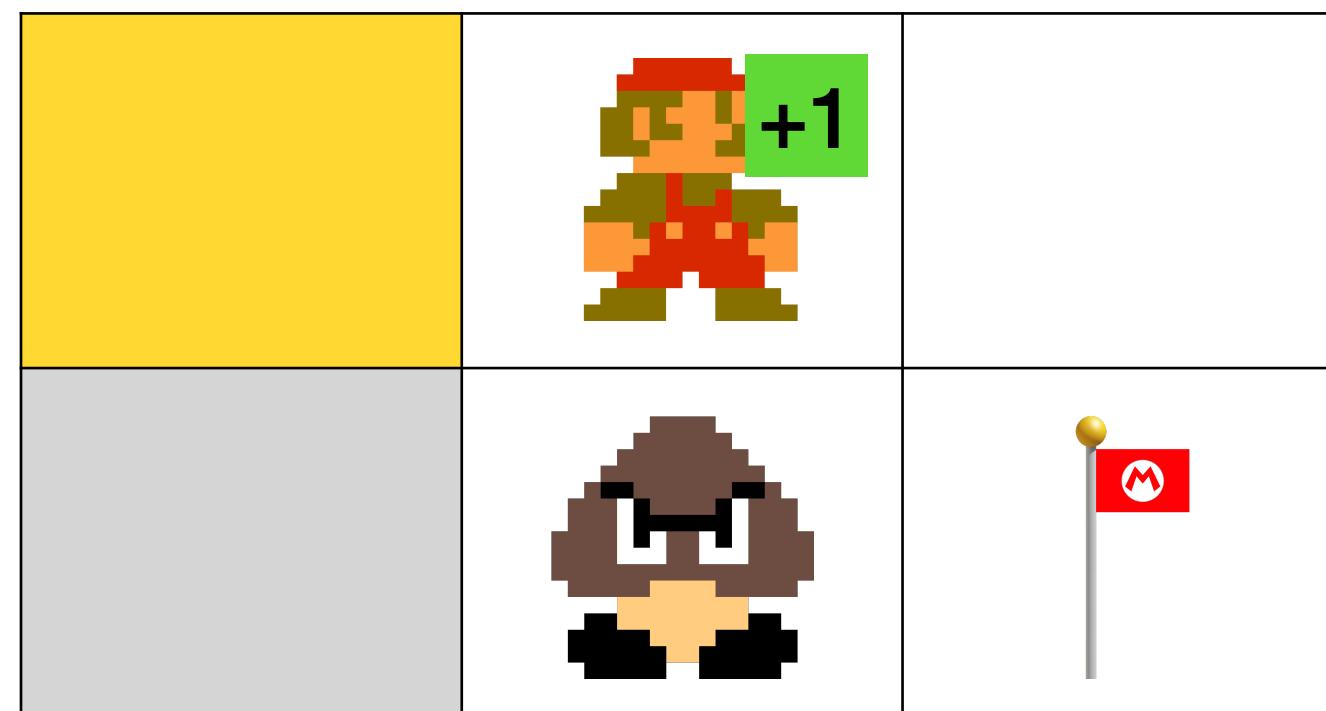
	←	→	↑	↓
←	0	0.2	0	0
0	0	0	0	-1
↑	0	0	0	0
↓	0	0	0	0
?	0	0	0	0
Flag	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(\text{initial}, \rightarrow) = 0.1 + 0.1 * [1 + 0.99 * 0 - 0] = 0.2$$

Take action A , observe R, S'

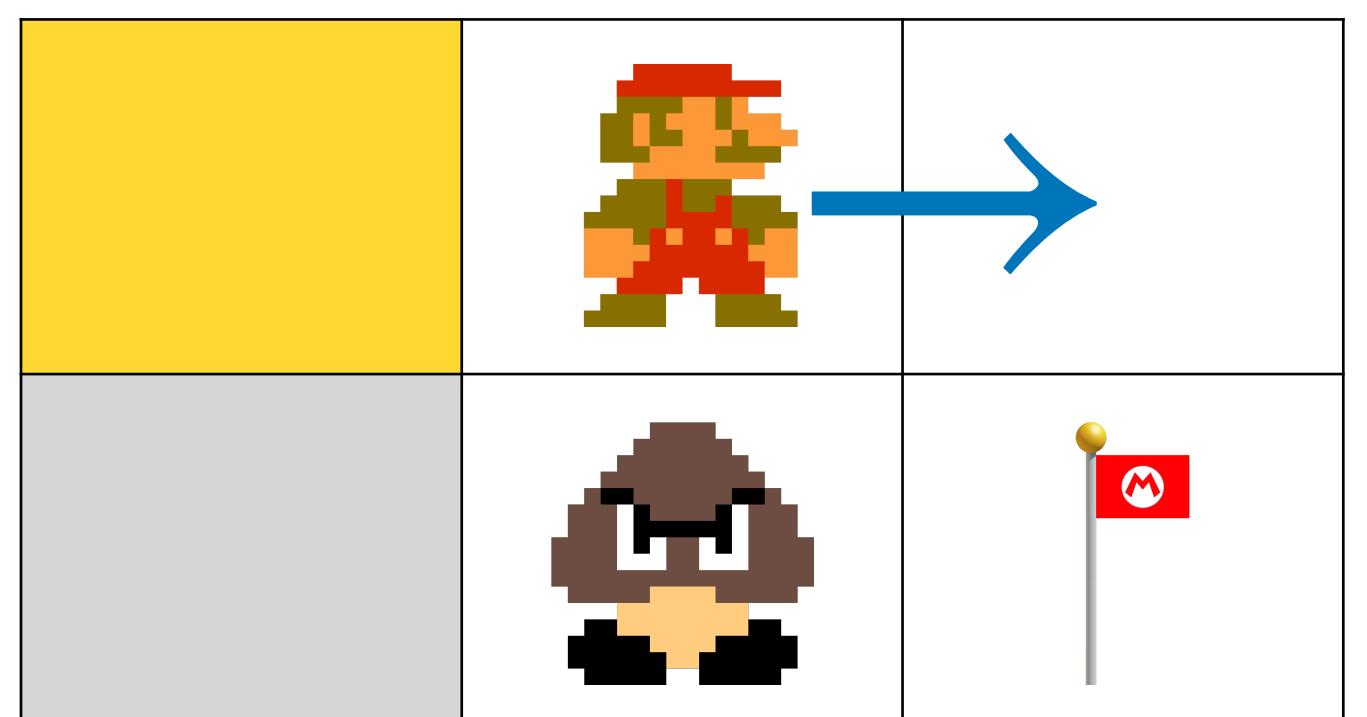


Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
↑	0	0.2	0	0
←	0	0	0	-1
↓	0	0	0	0
→	0	0	0	0
↓	0	0	0	0

Choose A from S using policy derived from Q (e.g., ε -greedy)

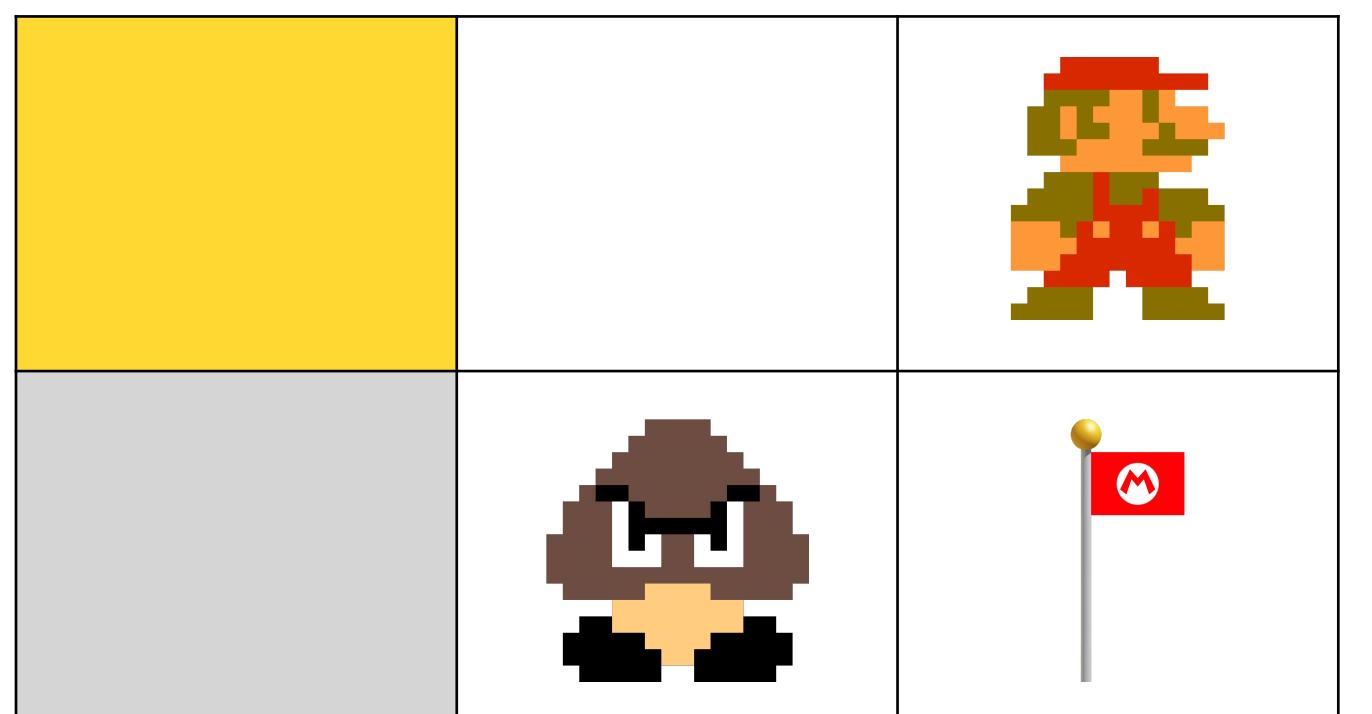


Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
←	0	0.2	0	0
→	0	0	0	-1
↑	0	0	0	0
↓	0	0	0	0
?	0	0	0	0
Flag	0	0	0	0

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

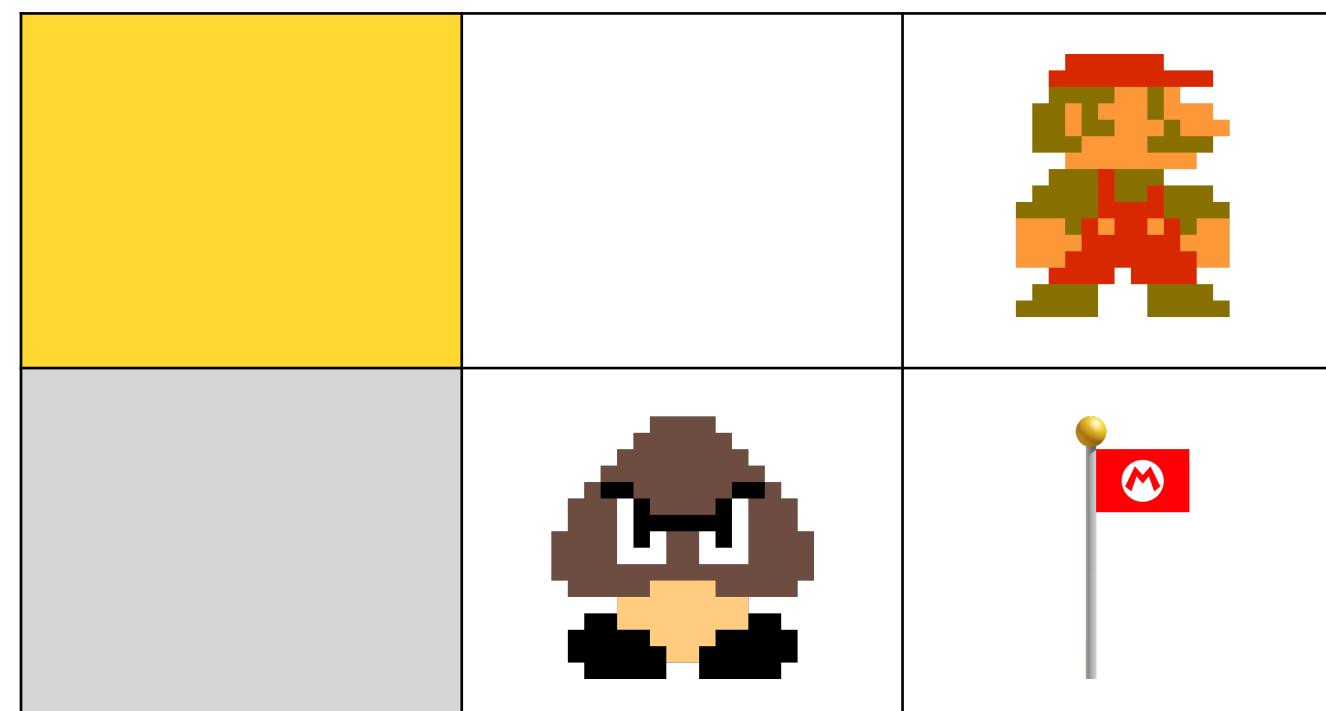
	←	→	↑	↓
←	0	0.2	0	0
↑	0	0	0	-1
↓	0	0	0	0
→	0	0	0	0
Mario	0	0	0	0
Flag	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(\text{Flag}, \rightarrow) = 0 + 0.1 * [0 + 0.99 * 0 - 0] = 0$$

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

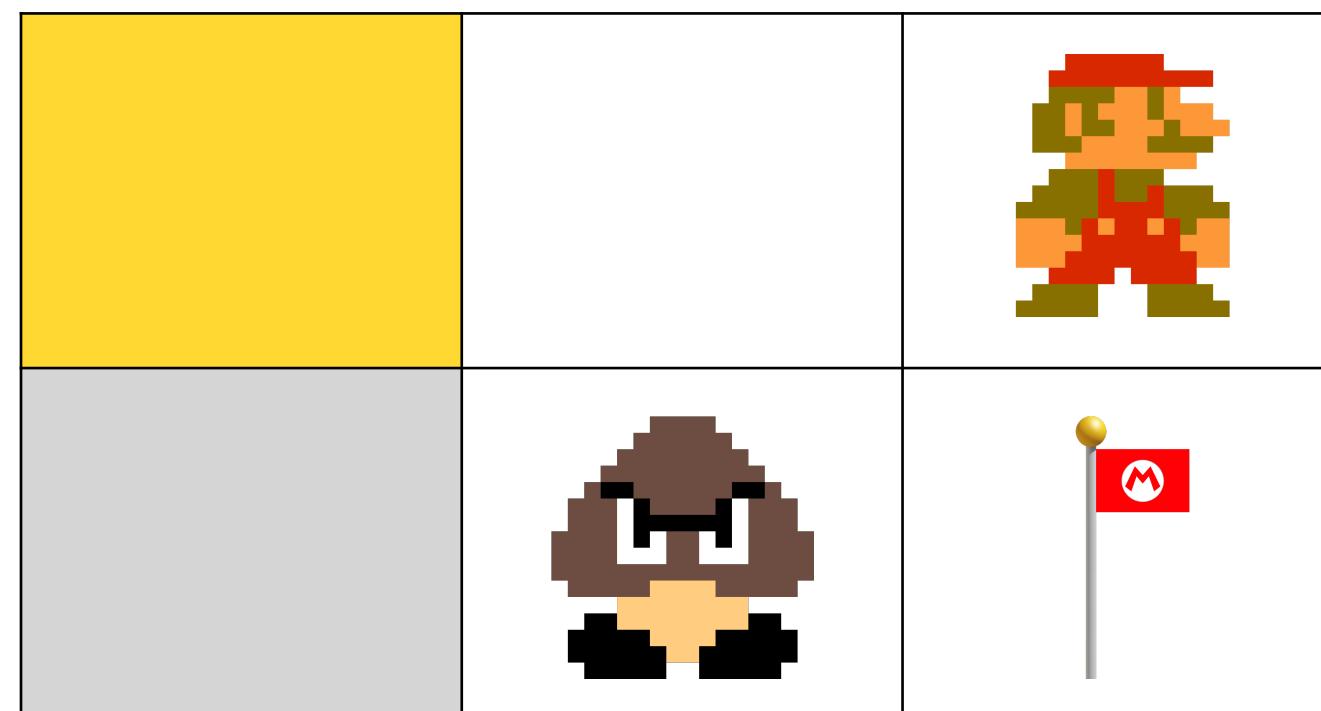
	←	→	↑	↓
←	0	0.2	0	0
0	0	0	0	-1
↑	0	0	0	0
↓	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(\text{coin}, \rightarrow) = 0 + 0.1 * [0 + 0.99 * 0 - 0] = 0$$

Take action A , observe R, S'

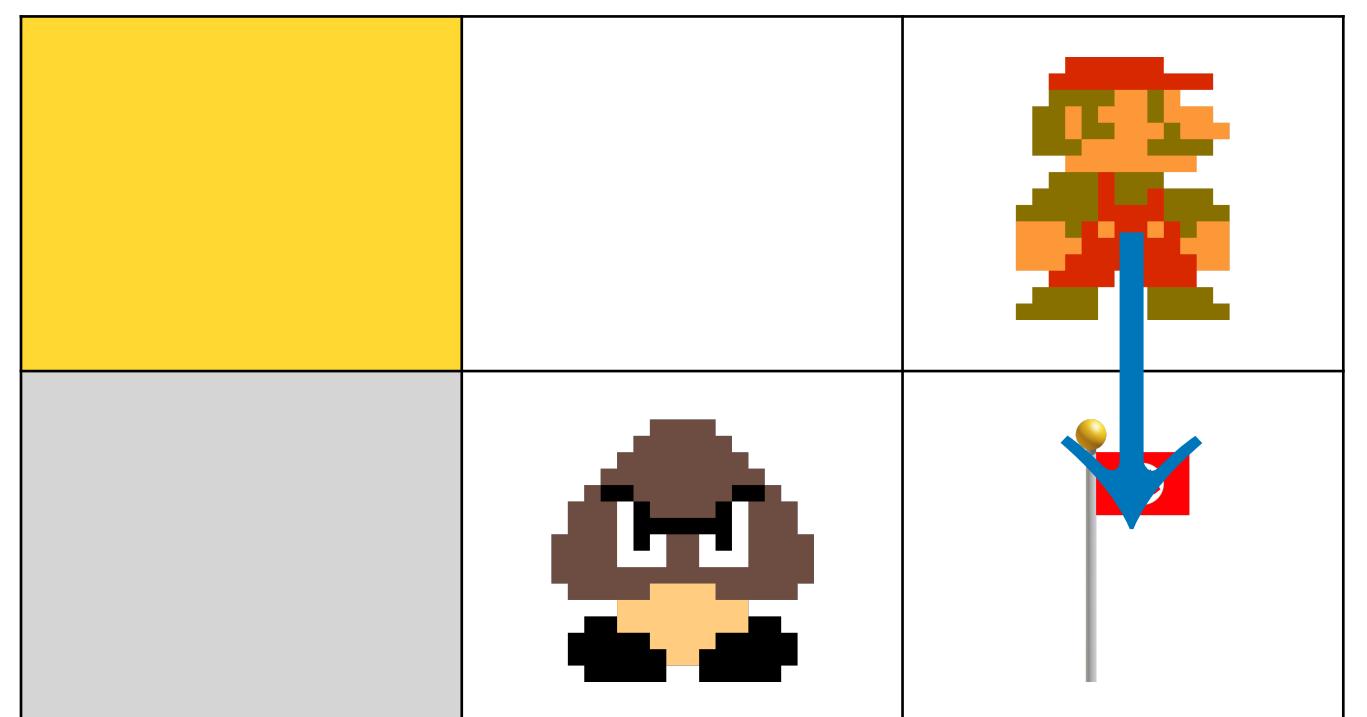


Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
Yellow	0	0.2	0	0
Gold Coin	0	0	0	-1
Blank	0	0	0	0
Blank	0	0	0	0
Mushroom	0	0	0	0
Flag	0	0	0	0

Choose A from S using policy derived from Q (e.g., ε -greedy)

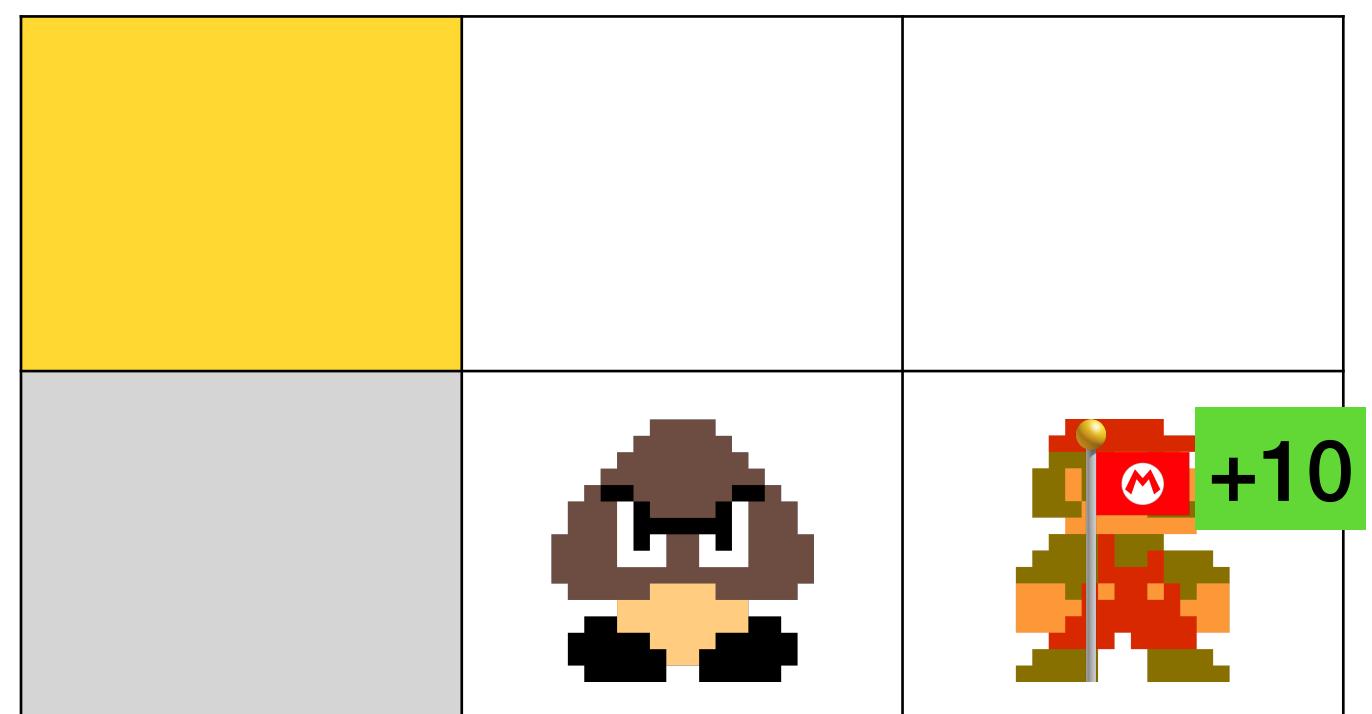


Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

	←	→	↑	↓
←	0	0.2	0	0
↑	0	0	0	-1
↓	0	0	0	0
→	0	0	0	0
Flag	0	0	0	0

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

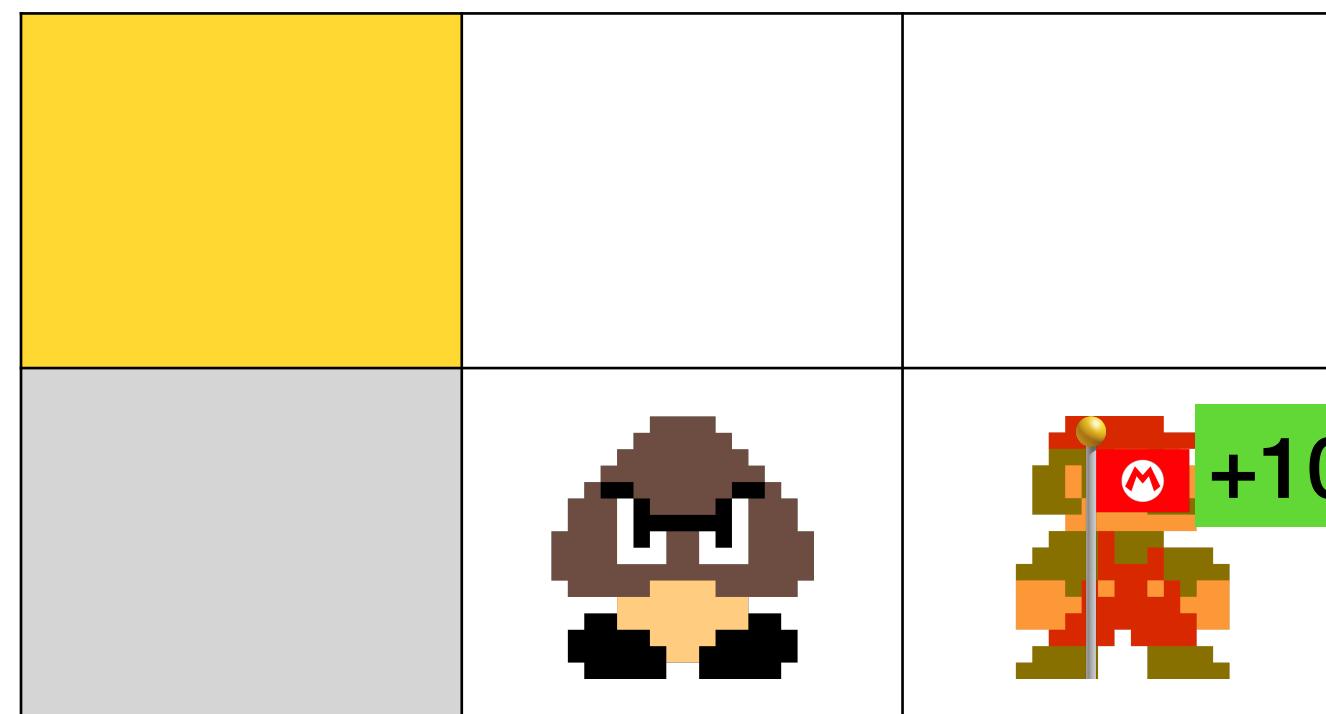
	←	→	↑	↓
←	0	0.2	0	0
0	0	0	0	-1
0	0	0	0	0
↑	0	0	0	0
↓	0	0	0	0
Flag	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(\text{ }, \downarrow) = 0 + 0.1 * [10 + 0.99 * 0 - 0] = 1$$

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

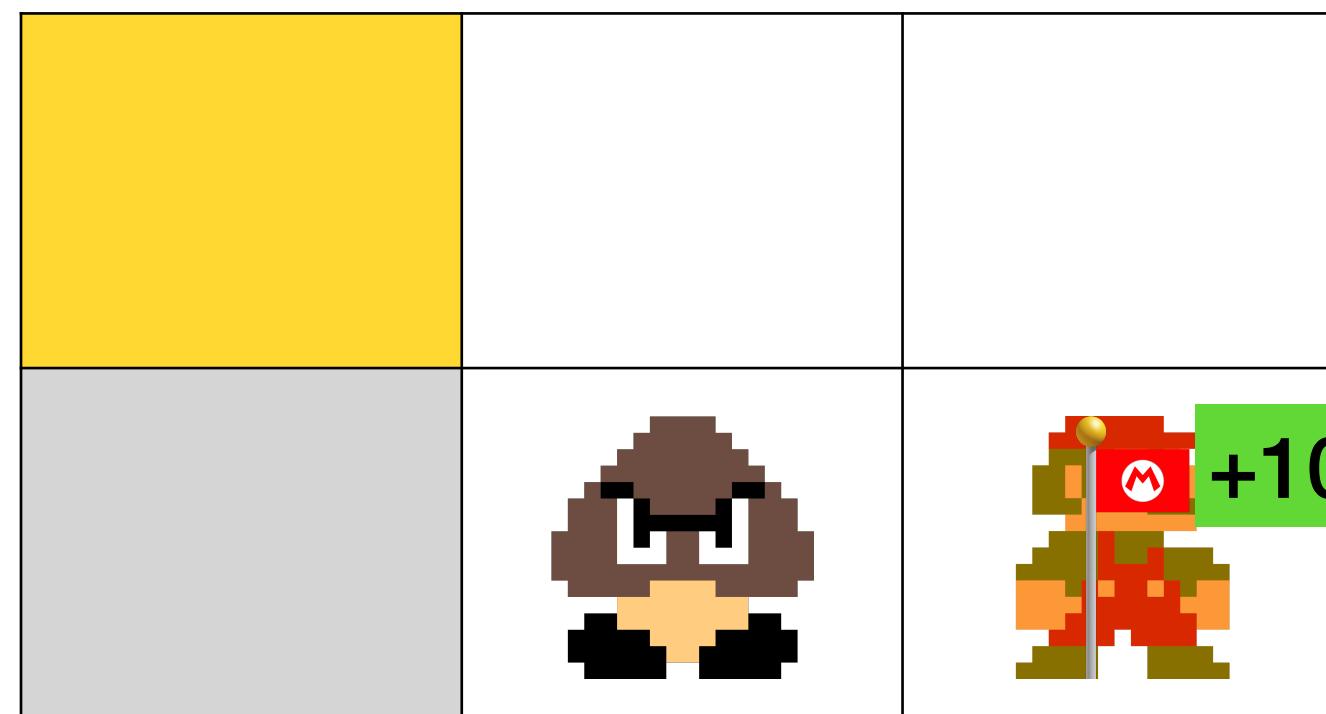
	←	→	↑	↓
←	0	0.2	0	0
0	0	0	0	-1
1	0	0	0	1
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(2, \downarrow) = 0 + 0.1 * [10 + 0.99 * 0 - 0] = 1$$

Take action A , observe R, S'



Q-Learning example

$$\alpha = 0.1$$
$$\gamma = 0.99$$

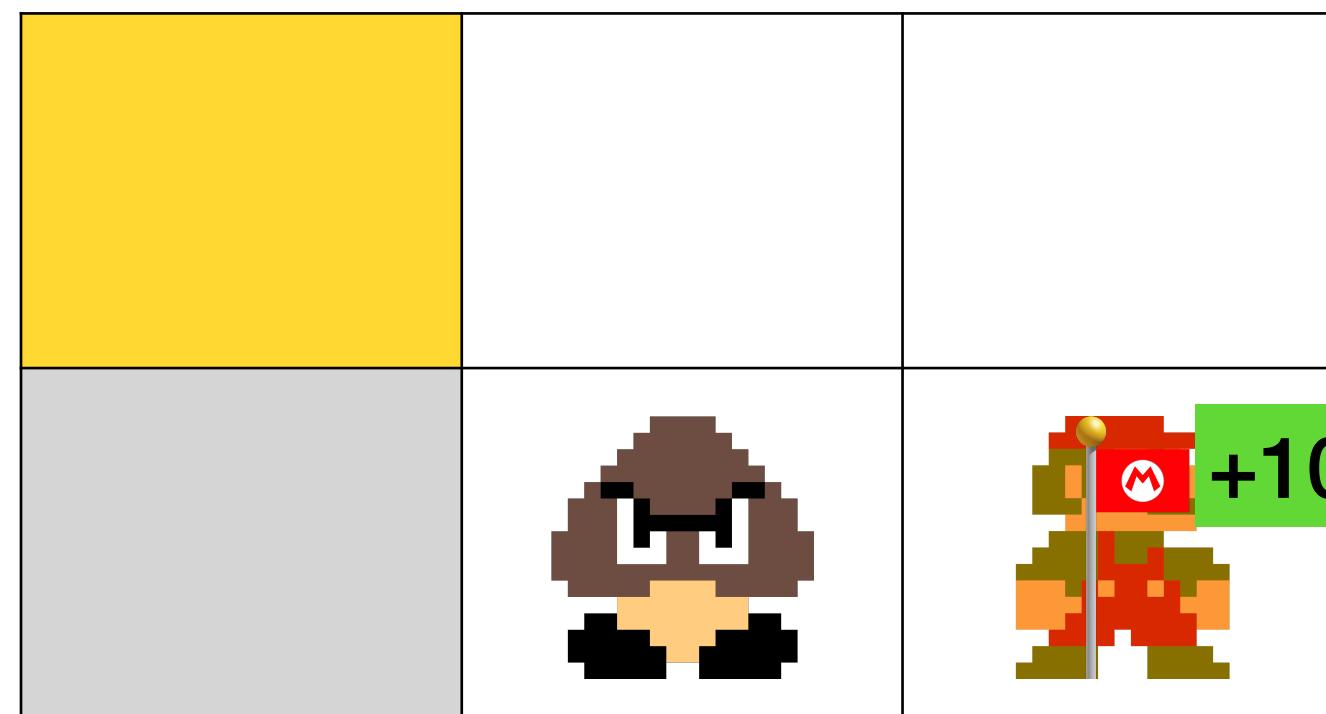
	←	→	↑	↓
←	0	0.2	0	0
0	0	0	0	-1
1	0	0	0	1
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0

Update $Q(S, A)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

$$Q(2, \downarrow) = 0 + 0.1 * [10 + 0.99 * 0 - 0] = 1$$

Take action A , observe R, S'



Episode won! Terminal state

Why ε -greedy?

Why ϵ -greedy?

Exploration-exploitation tradeoff

Why ε -greedy?

Exploration-exploitation tradeoff

Take greedy actions with probability $1 - \varepsilon$

Why ϵ -greedy?

Exploration-exploitation tradeoff

Take greedy actions with probability $1 - \epsilon$

Exploit what the agent already knows

Why ϵ -greedy?

Exploration-exploitation tradeoff

Take greedy actions with probability $1 - \epsilon$

Exploit what the agent already knows

Explore with probability ϵ

Why ϵ -greedy?

Exploration-exploitation tradeoff

Take greedy actions with probability $1 - \epsilon$

Exploit what the agent already knows

Explore with probability ϵ

Take random actions to visit new states you would not visit otherwise

Why ϵ -greedy?

Exploration-exploitation tradeoff

Take greedy actions with probability $1 - \epsilon$

Exploit what the agent already knows

Explore with probability ϵ

Take random actions to visit new states you would not visit otherwise

Decrease ϵ with time

Why ε -greedy?

Exploration-exploitation tradeoff

Take greedy actions with probability $1 - \varepsilon$

Exploit what the agent already knows

Explore with probability ε

Take random actions to visit new states you would not visit otherwise

Decrease ε with time

$$0 \leq \varepsilon \leq 1$$

Why ϵ -greedy?

Start	0	0	0	0	0	0
-1						0
-1	-10	-10				0
0						0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

Start	0	0	0	0	0	0
-1	-	-10	-	-	-	0
-1	-10	-10	-	-	-	0
0	-	-	-	-	-	0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we take only greedy actions...

Start	0	0	0	0	0	0
-1	-10	-10	-10	-10	-10	0
-1	-10	-10	-10	-10	-10	0
0	0	0	0	0	0	0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we take only greedy actions...

Start	0	0	0	0	0	0
-1	-	-10	-	-	-	0
-1	-10	-10	-	-	-	0
0	-	-	-	-	-	0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we take only greedy actions...

We end up with a suboptimal solution

Start	0	0	0	0	0	0
-1	-	-	-	-	-	0
-1	-10	-10	-	-	-	0
0	-	-	-	-	-	0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we take only greedy actions...

We end up with a suboptimal solution

Start	0	0	0	0	0	0
-1	-	-10	-	-	-	0
-1	-10	-10	-	-	-	0
0	-	-	-	-	-	0
0	10	1	1	1	1	Goal (10)

$$G = 10$$

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

Start	0	0	0	0	0	0
-1	-	-10	-	-	-	0
-1	-10	-10	-	-	-	0
0	-	-	-	-	-	0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

Start	0	0	0	0	0	0
-1			-10			0
-1	-10	-10				0
0						0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we explore...

Start	0	0	0	0	0	0
-1						0
-1	-10	-10				0
0						0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we explore...

We might discover more valuable states

Start	0	0	0	0	0	0
-1						0
-1	-10	-10				0
0						0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we explore...

We might discover more valuable states

Start	0	0	0	0	0	0
-1			-10			0
-1		-10	-10			0
0						0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we explore...

We might discover more valuable states

Start	0	0	0	0	0	0
-1			-10			0
-1		-10	-10			0
0						0
0	10	1	1	1	1	Goal (10)

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we explore...

Start	0	0	0	0	0	0
-1			-10			0
-1	-10	-10				0
0						0
0	10	1	1	1	1	Goal (10)

We might discover more valuable states

That lead to higher returns

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

If we explore...

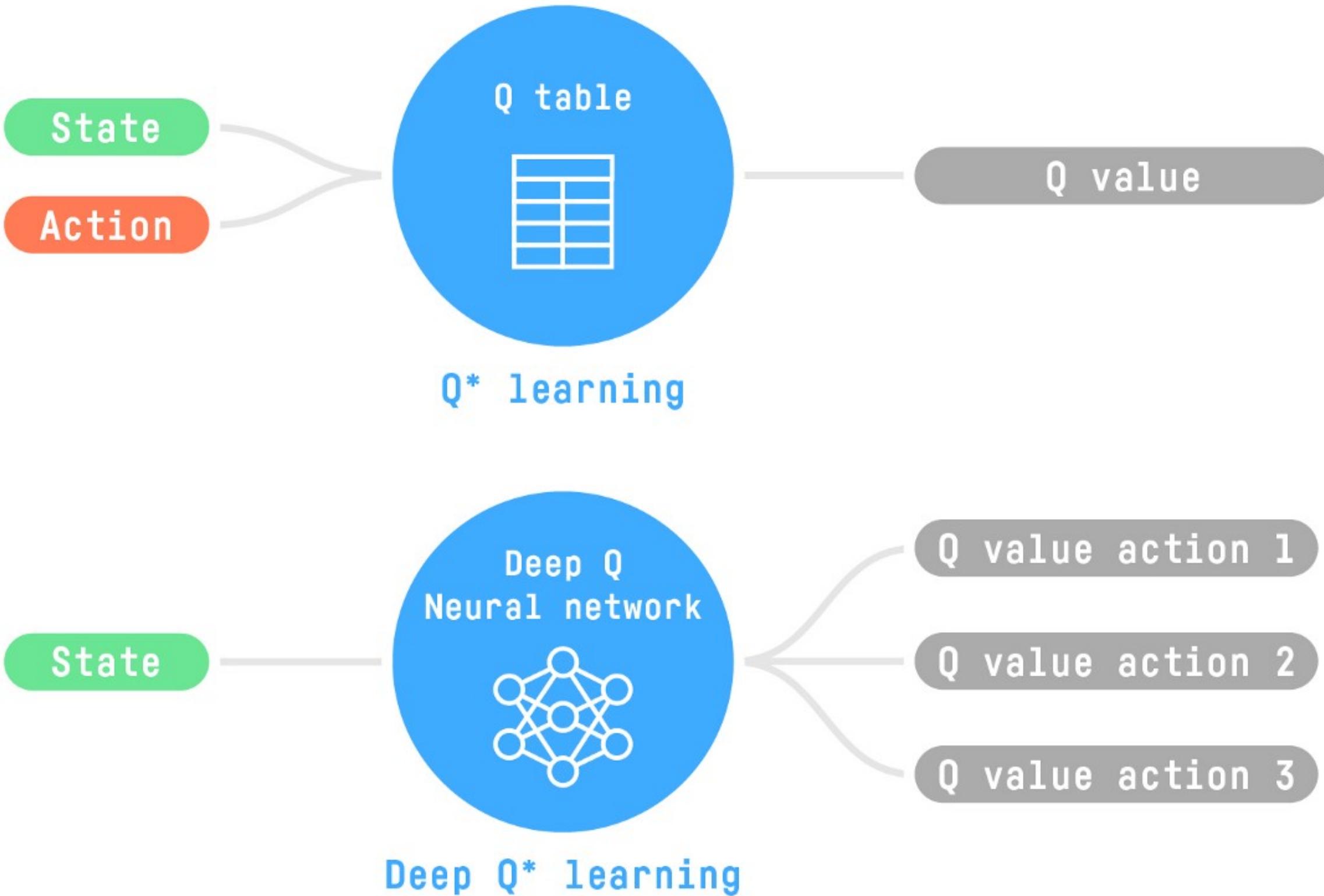
Start	0	0	0	0	0	0
-1			-10			0
-1		-10	-10			0
0						0
0	10	1	1	1	1	Goal (10)

We might discover more valuable states

That lead to higher returns

$$G = 22$$

Deep Q-Learning



Deep Q-Learning

Deep Q-Learning

Q-Learning works well for small state spaces

Deep Q-Learning

Q-Learning works well for small state spaces

Atari games, however, have observation space of shape (210, 160, 3), values in [0, 255]

Deep Q-Learning

Q-Learning works well for small state spaces

Atari games, however, have observation space of shape (210, 160, 3), values in [0, 255]

This gives us $256^{210 \times 160 \times 3} = 256^{100800}$ possible observations

Deep Q-Learning

Q-Learning works well for small state spaces

Atari games, however, have observation space of shape (210, 160, 3), values in [0, 255]

This gives us $256^{210 \times 160 \times 3} = 256^{100800}$ possible observations

Approximate Q-values using a neural network

Deep Q-Learning

Q-Learning works well for small state spaces

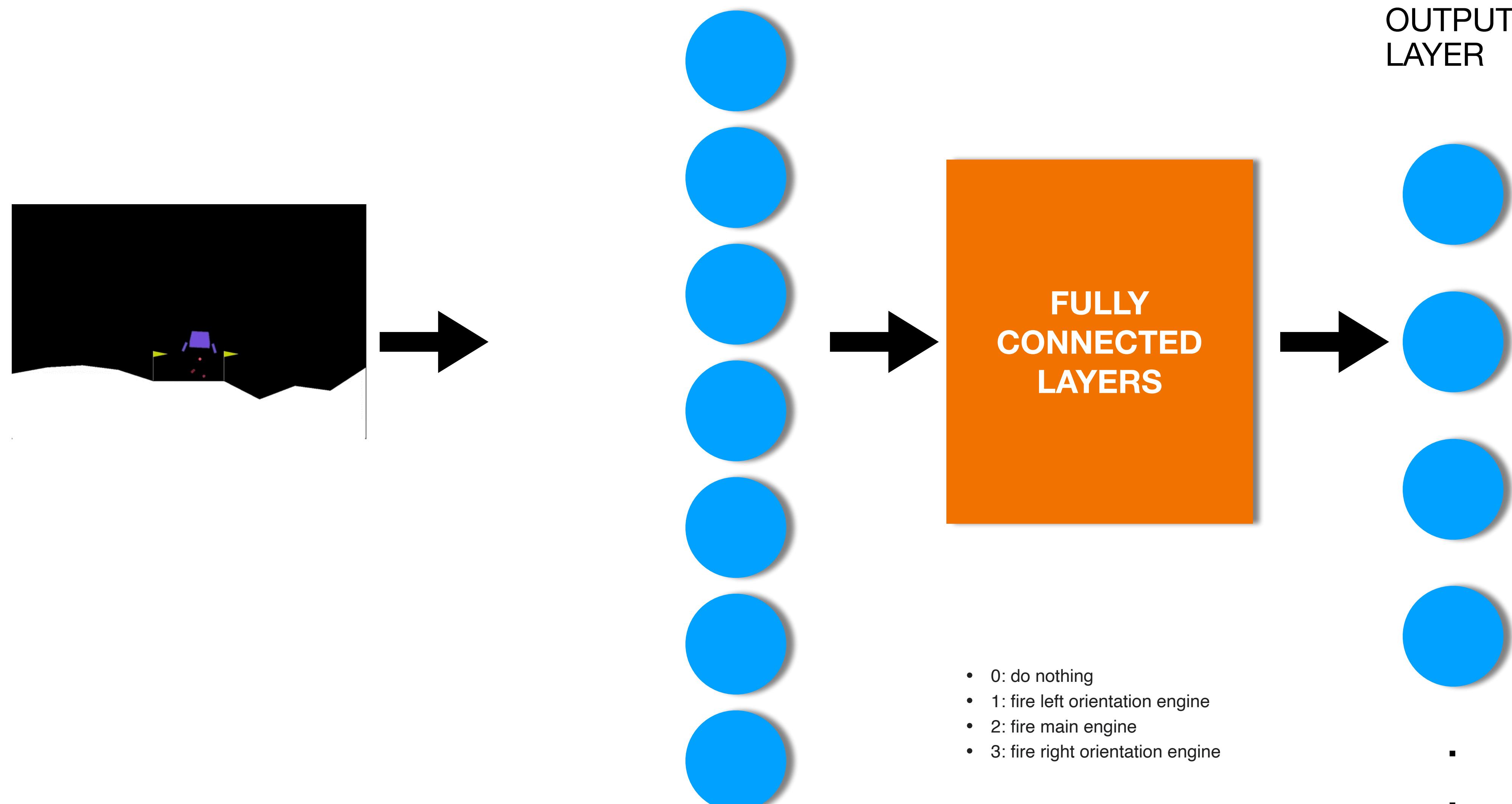
Atari games, however, have observation space of shape (210, 160, 3), values in [0, 255]

This gives us $256^{210 \times 160 \times 3} = 256^{100800}$ possible observations

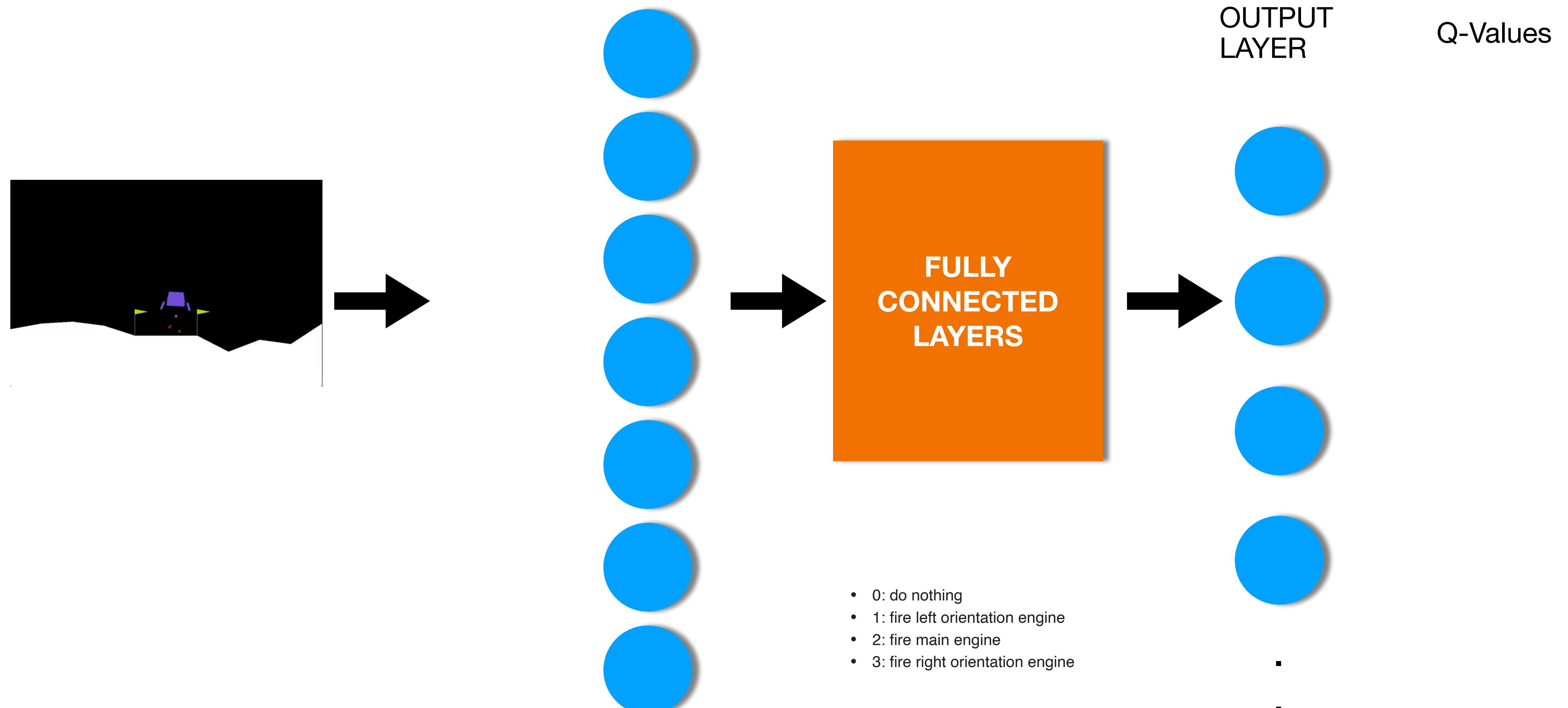
Approximate Q-values using a neural network

Parametrized Q-function $Q_\theta(s, a)$

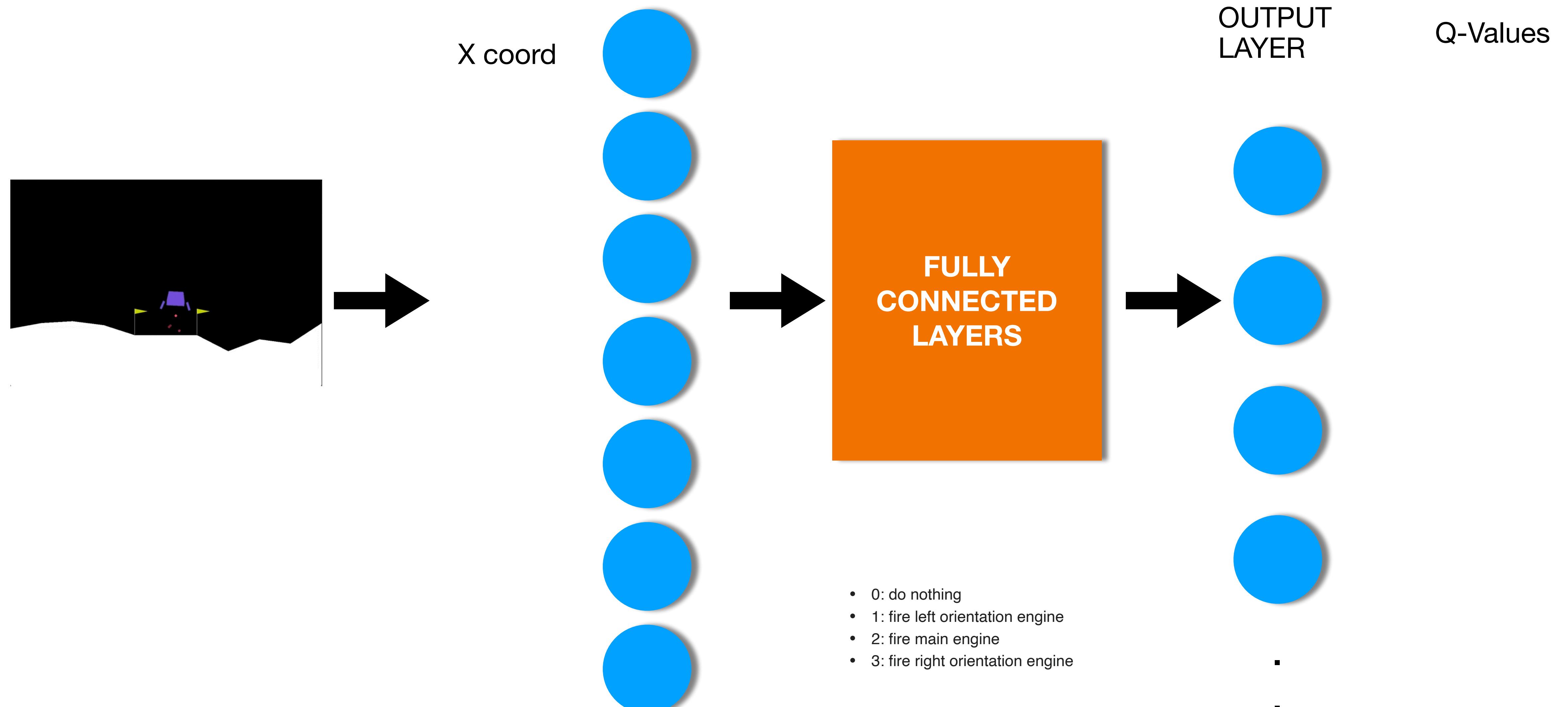
Deep Q-Learning



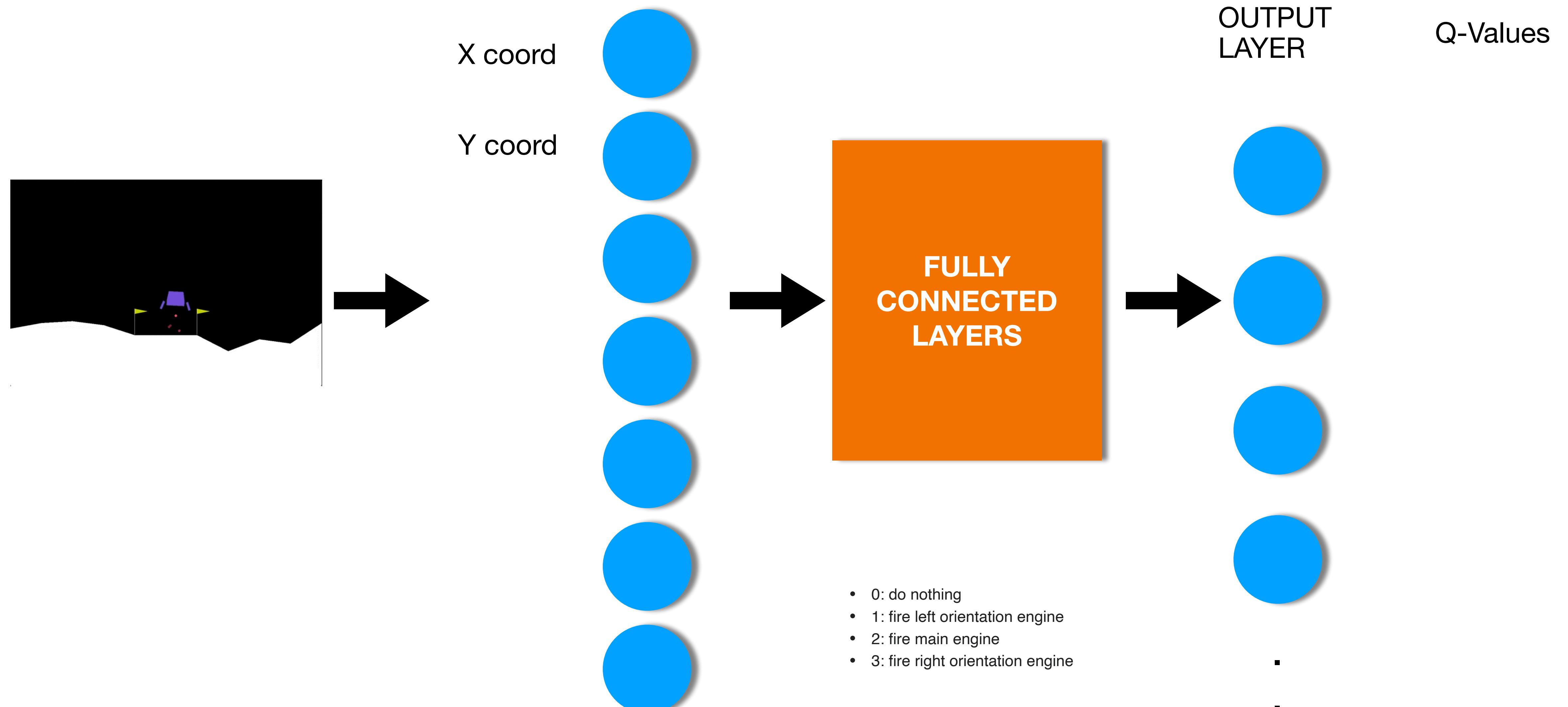
Deep Q-Learning



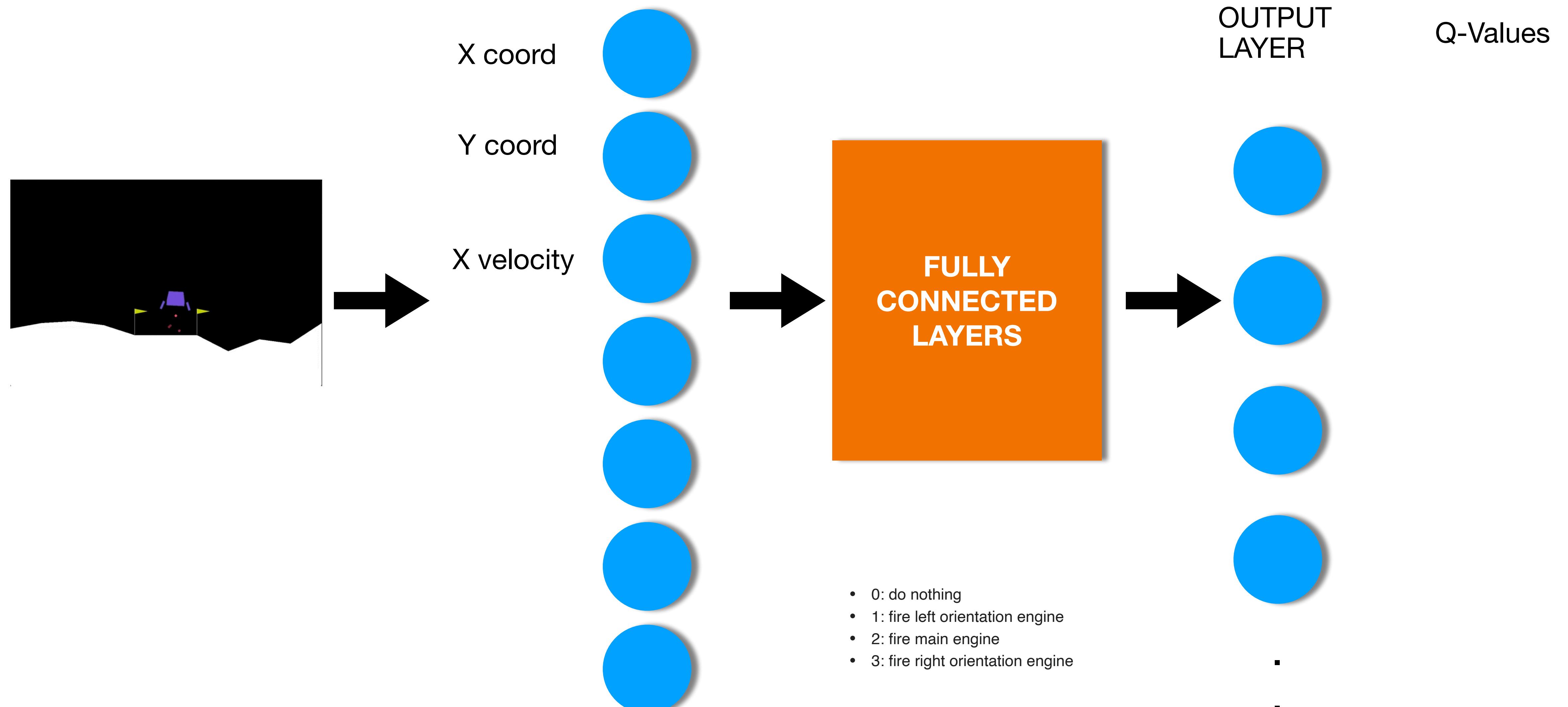
Deep Q-Learning



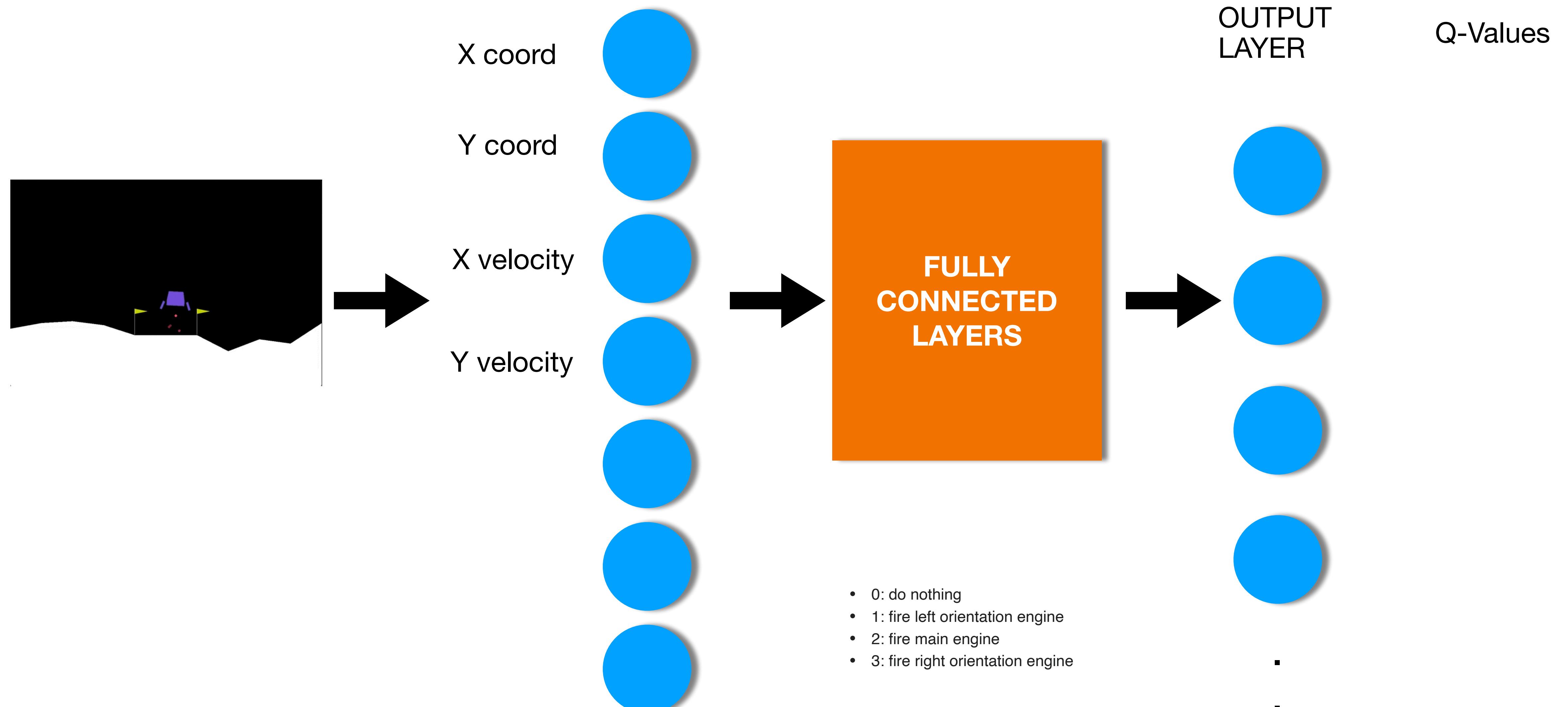
Deep Q-Learning



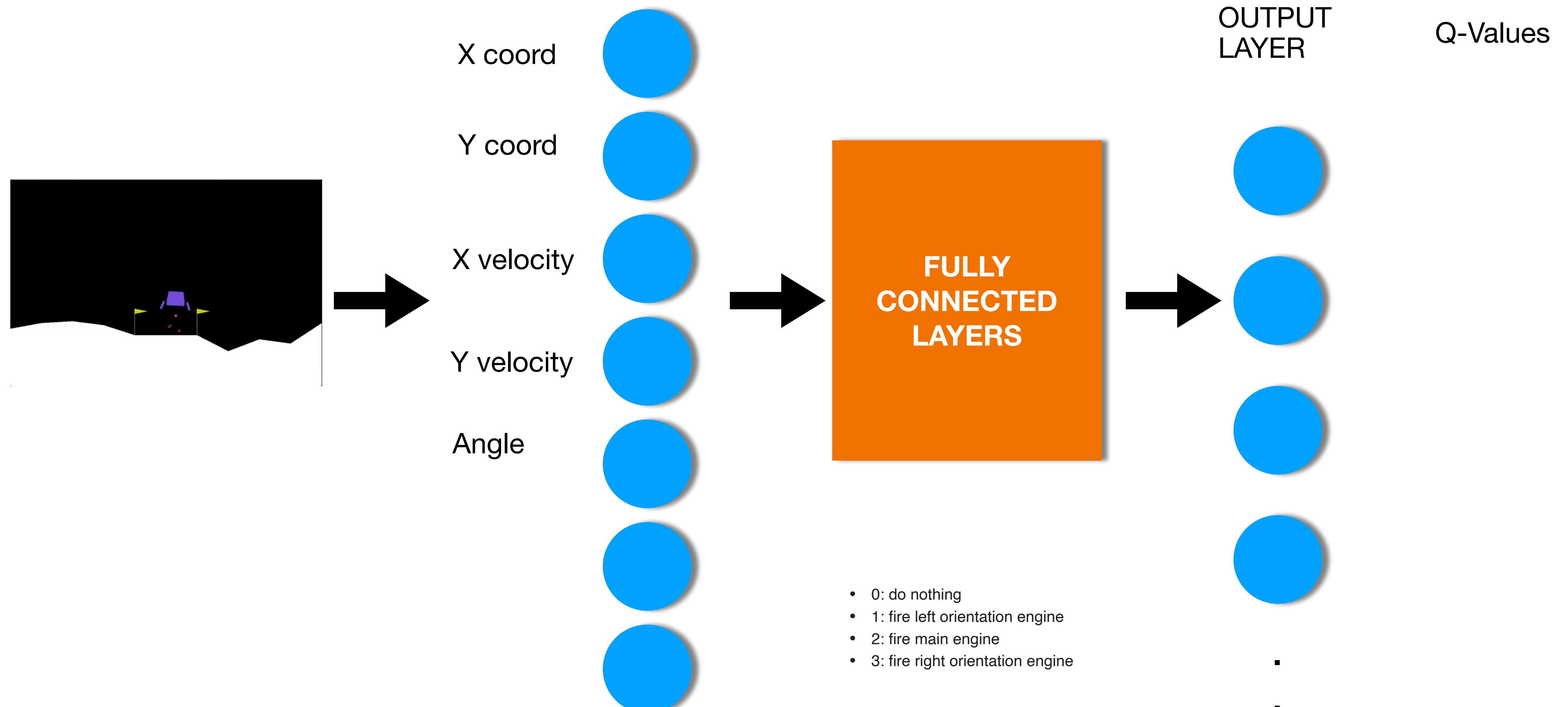
Deep Q-Learning



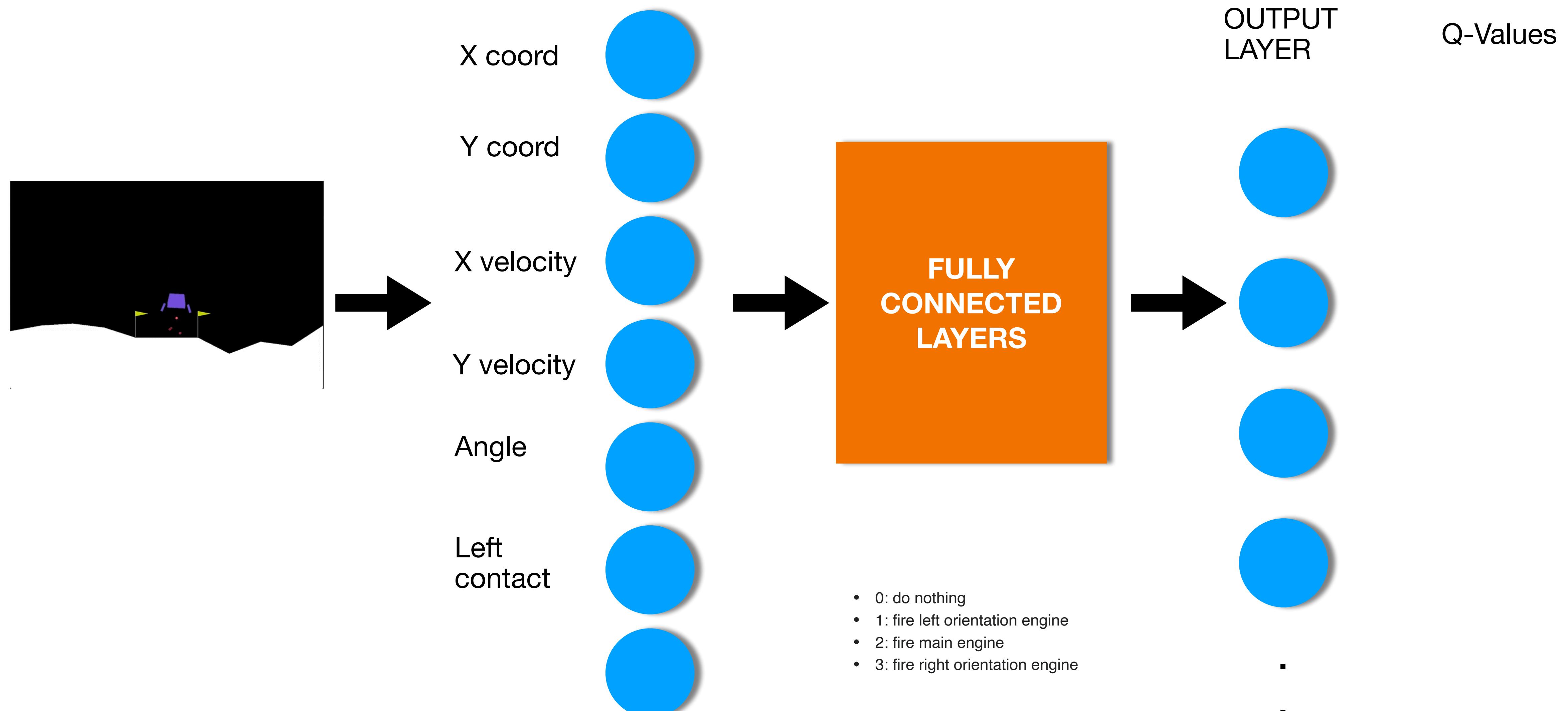
Deep Q-Learning



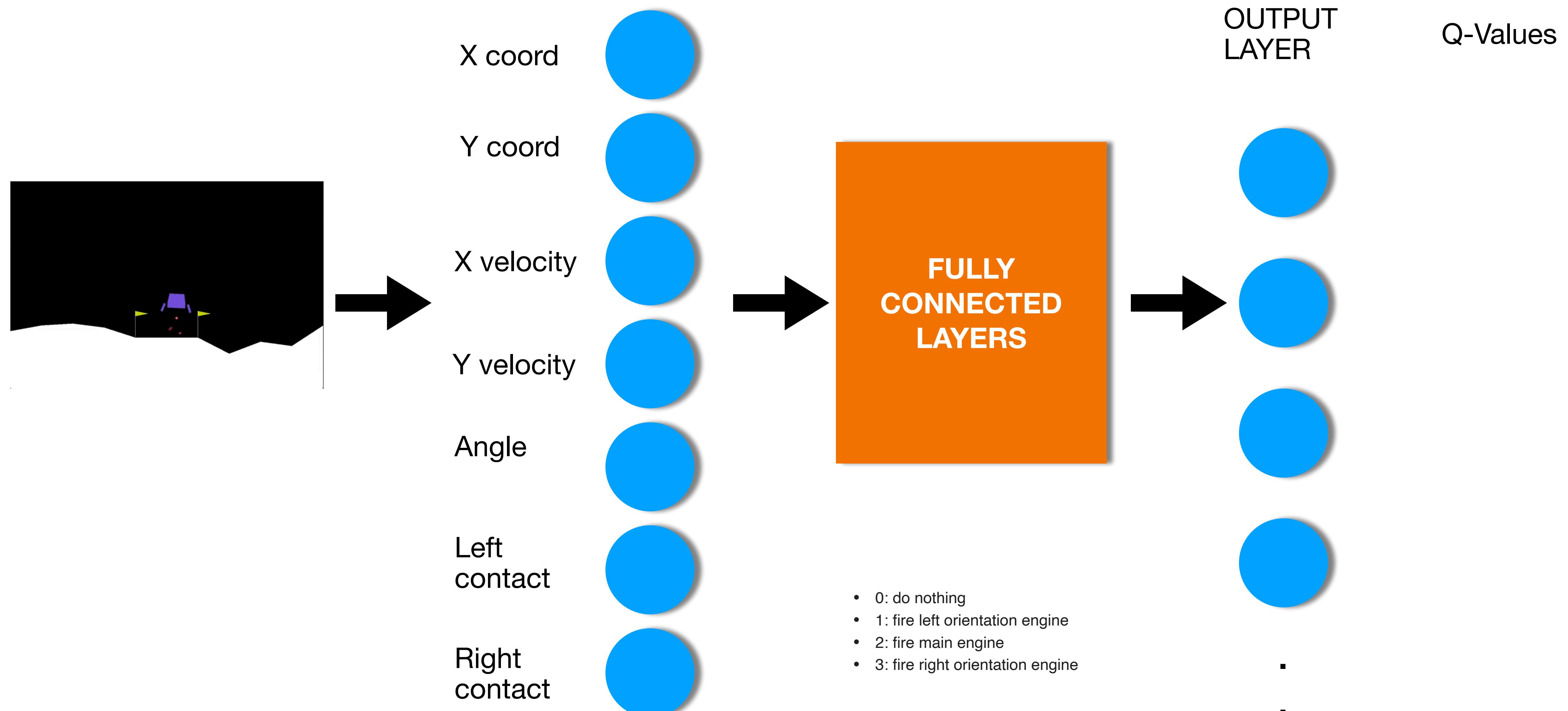
Deep Q-Learning



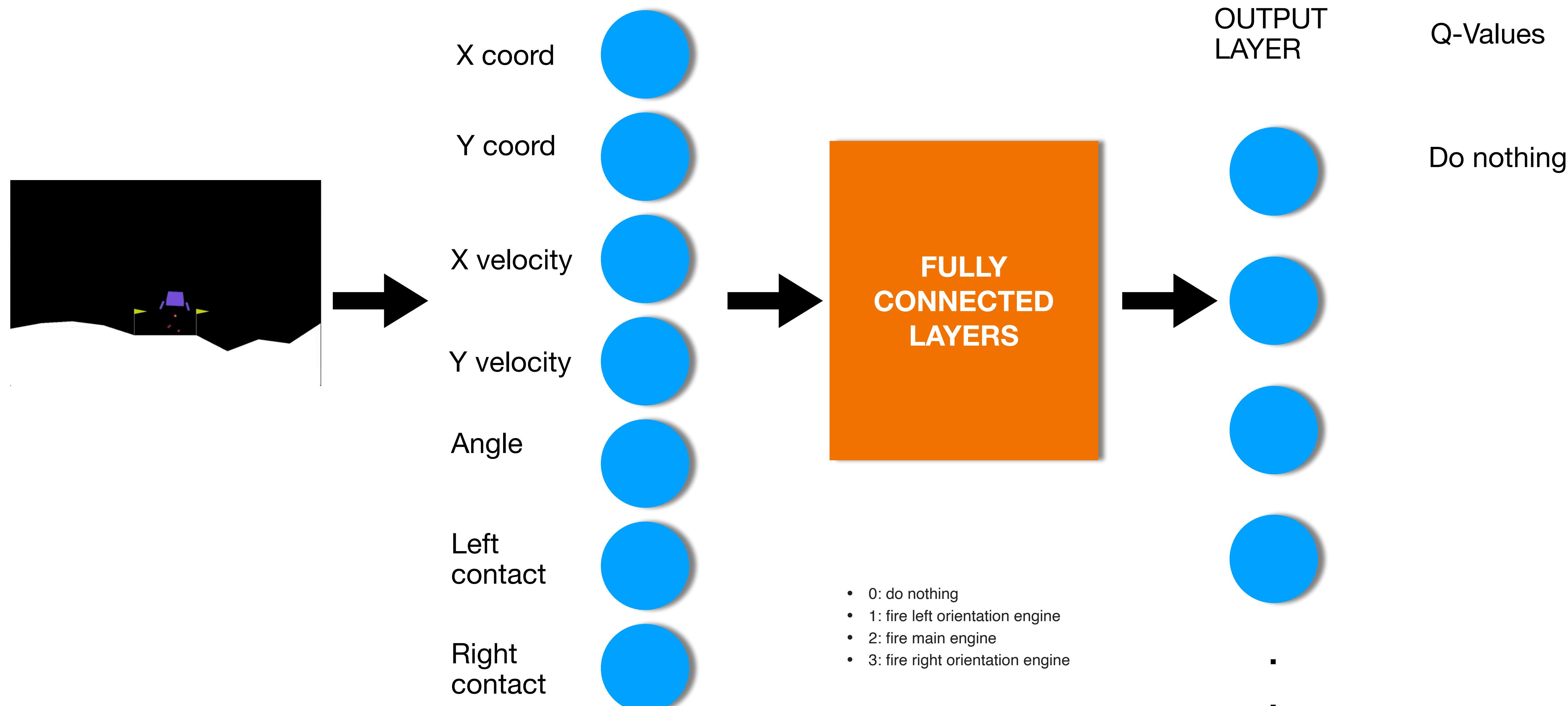
Deep Q-Learning



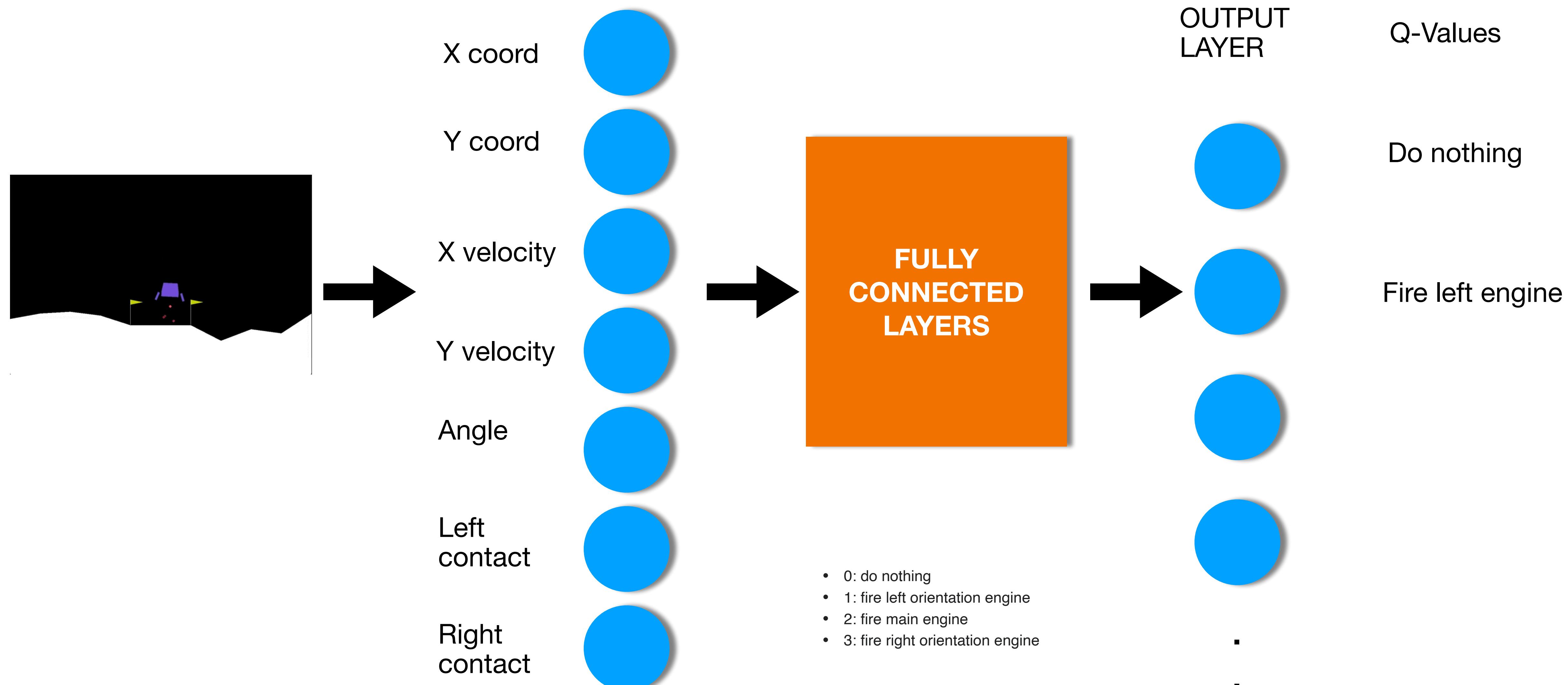
Deep Q-Learning



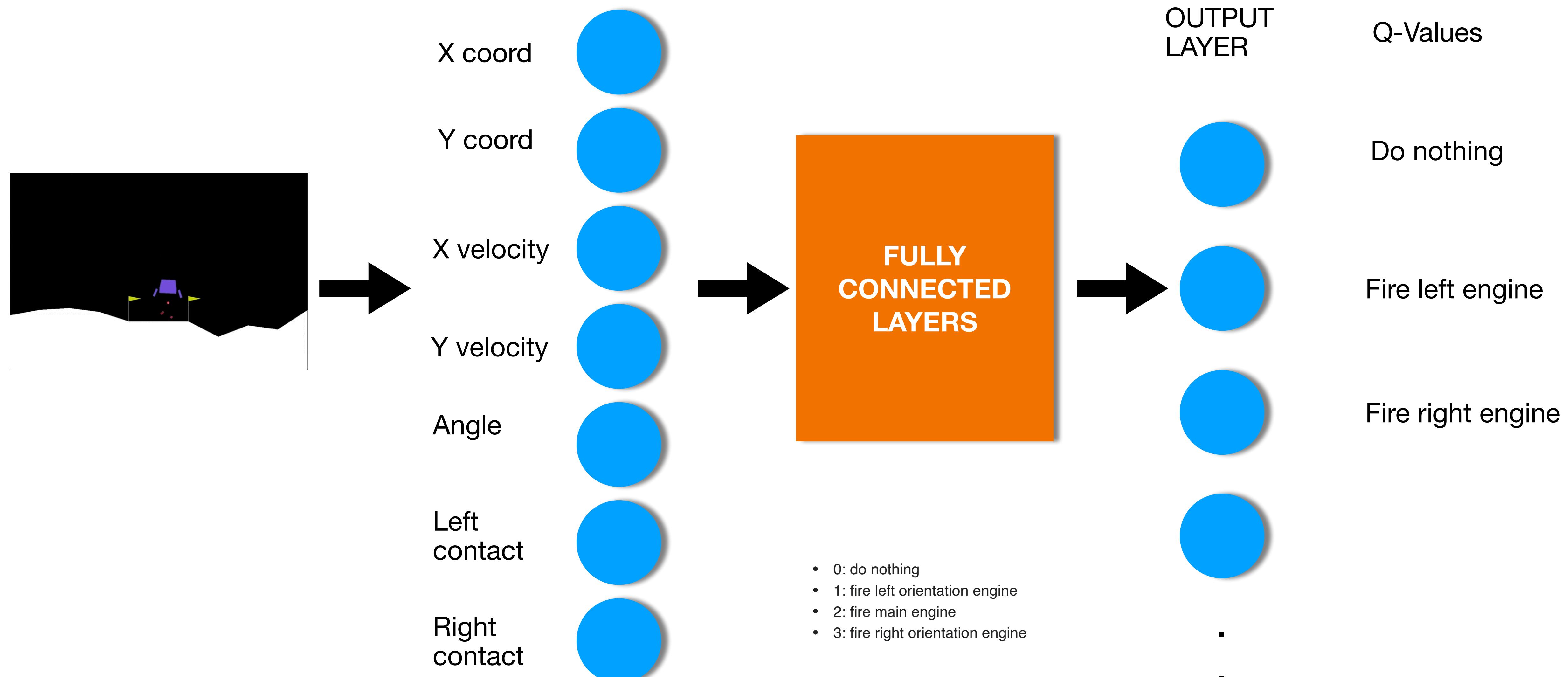
Deep Q-Learning



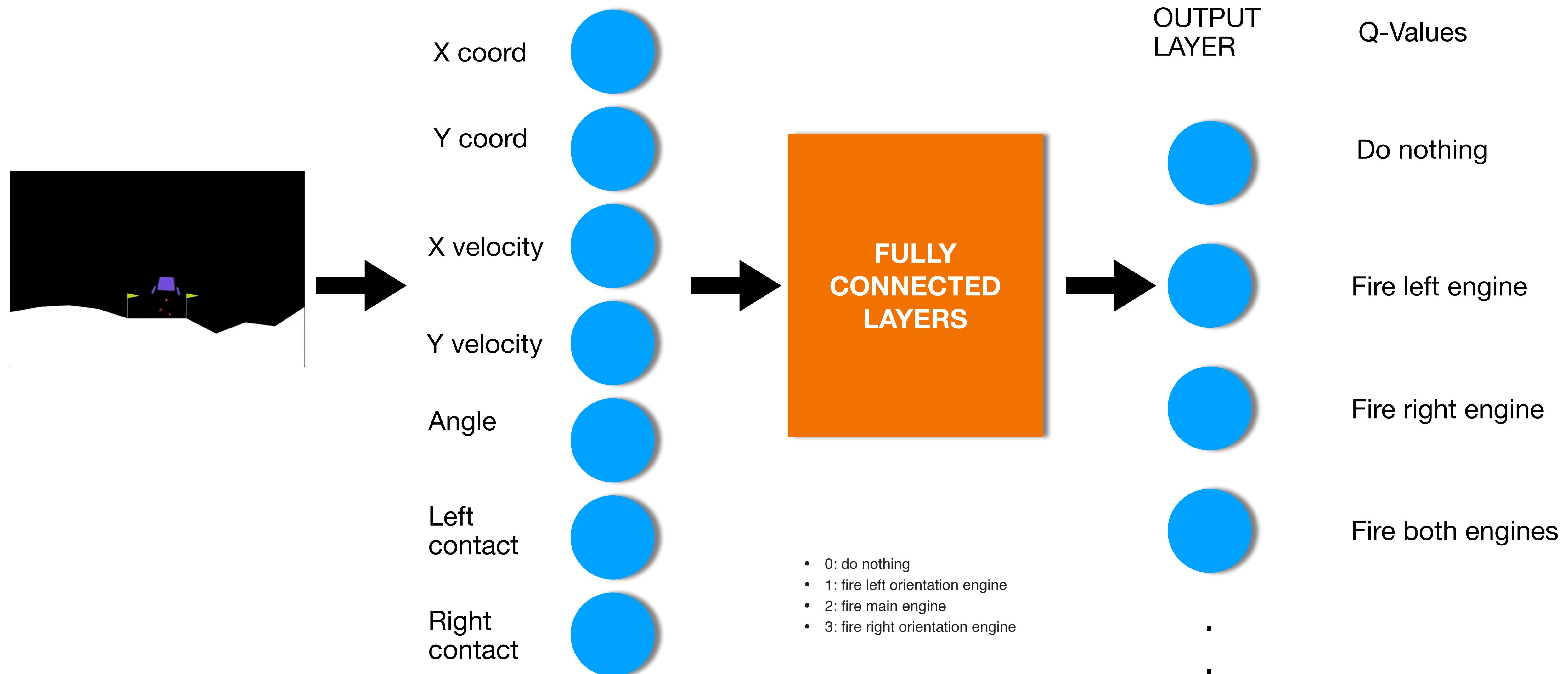
Deep Q-Learning



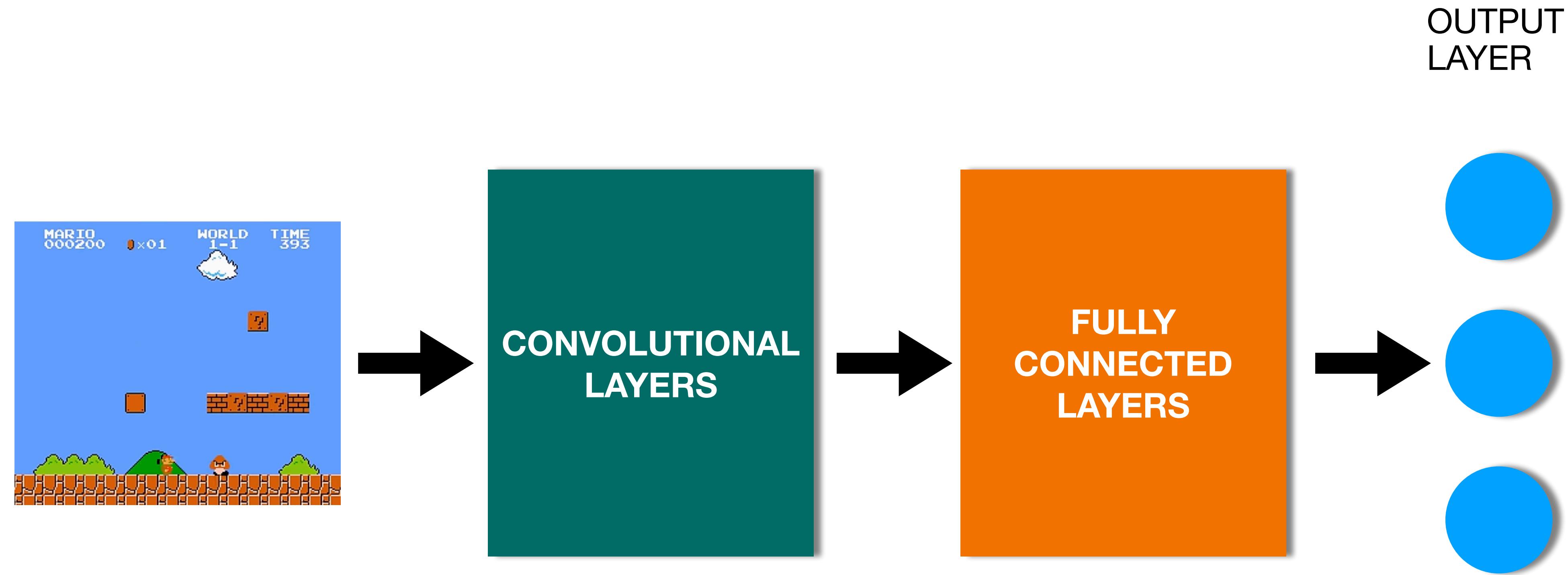
Deep Q-Learning



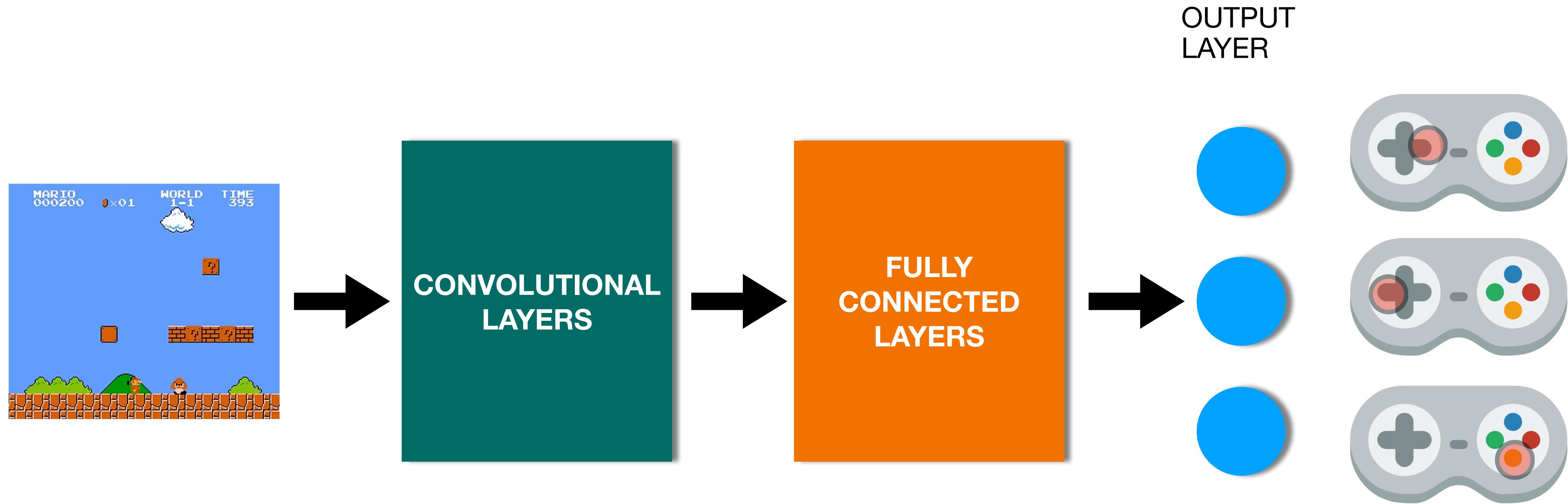
Deep Q-Learning



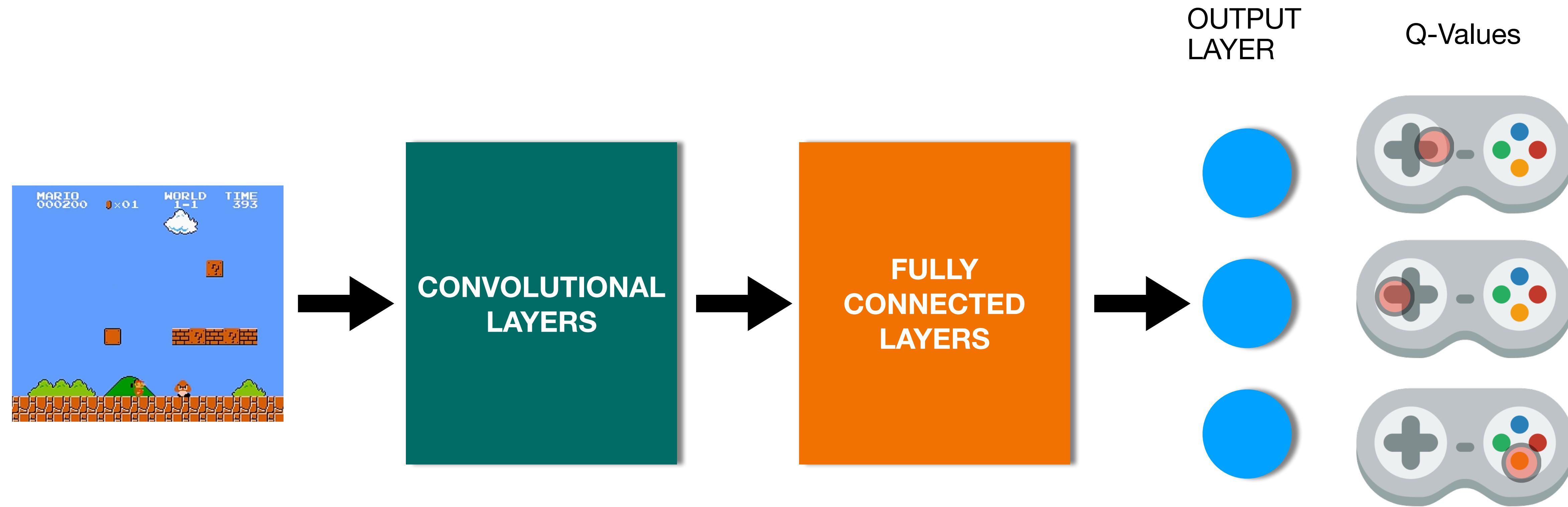
Deep Q-Learning



Deep Q-Learning



Deep Q-Learning



Deep Q-Learning algorithm

For episode = 1, M **do**

 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence $\phi_1 = \phi(s_1)$

For $t = 1, T$ **do**

 With probability ε select a random action a_t

 otherwise select $a_t = \operatorname{argmax}_a Q(\phi(s_t), a; \theta)$

 Execute action a_t in emulator and observe reward r_t and image x_{t+1}

 Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$

 Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D

 Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from D

 Set $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) & \text{otherwise} \end{cases}$

 Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to the network parameters θ

 Every C steps reset $\hat{Q} = Q$

End For

End For

Deep Q-Learning algorithm

For episode = 1, M **do**

 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence $\phi_1 = \phi(s_1)$

For $t = 1, T$ **do**

 With probability ε select a random action a_t

 otherwise select $a_t = \operatorname{argmax}_a Q(\phi(s_t), a; \theta)$

 Execute action a_t in emulator and observe reward r_t and image x_{t+1}

 Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$

 Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D

 Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from D

 Set $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) & \text{otherwise} \end{cases}$

 Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to the network parameters θ

 Every C steps reset $\hat{Q} = Q$

End For

End For

ϕ represents the NN

Deep Q-Learning algorithm

For episode = 1, M **do**

 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence $\phi_1 = \phi(s_1)$

For $t = 1, T$ **do**

 With probability ε select a random action a_t
 otherwise select $a_t = \operatorname{argmax}_a Q(\phi(s_t), a; \theta)$

 Execute action a_t in emulator and observe reward r_t and image x_{t+1}

 Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$

 Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D

 Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from D

 Set $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) & \text{otherwise} \end{cases}$

 Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to the network parameters θ

 Every C steps reset $\hat{Q} = Q$

End For

End For

ϕ represents the NN

Deep Q-Learning algorithm

For episode = 1, M **do**

 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence $\phi_1 = \phi(s_1)$

For $t = 1, T$ **do**

 With probability ε select a random action a_t
 otherwise select $a_t = \operatorname{argmax}_a Q(\phi(s_t), a; \theta)$

 Execute action a_t in emulator and observe reward r_t and image x_{t+1}

 Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$

 Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D

 Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from D

 Set $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) & \text{otherwise} \end{cases}$

 Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to the network parameters θ

 Every C steps reset $\hat{Q} = Q$

End For

End For

ϕ represents the NN

Deep Q-Learning algorithm

For episode = 1, M **do**

 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence $\phi_1 = \phi(s_1)$

For $t = 1, T$ **do**

 With probability ε select a random action a_t
 otherwise select $a_t = \operatorname{argmax}_a Q(\phi(s_t), a; \theta)$

 Execute action a_t in emulator and observe reward r_t and image x_{t+1}

 Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$

 Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D

 Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from D

 Set $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) & \text{otherwise} \end{cases}$

 Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to the network parameters θ

 Every C steps reset $\hat{Q} = Q$

End For

End For

ϕ represents the NN

Deep Q-Learning algorithm

For episode = 1, M **do**

 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence $\phi_1 = \phi(s_1)$

For $t = 1, T$ **do**

 With probability ε select a random action a_t ,
 otherwise select $a_t = \operatorname{argmax}_a Q(\phi(s_t), a; \theta)$

 Execute action a_t in emulator and observe reward r_t and image x_{t+1}

 Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$

 Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D

 Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from D

 Set $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) & \text{otherwise} \end{cases}$

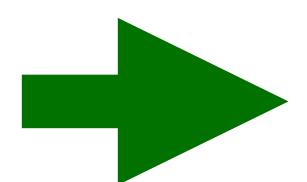
 Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to the network parameters θ

 Every C steps reset $\hat{Q} = Q$

End For

End For

ϕ represents the NN



Sampling (interaction
with the environment)

Deep Q-Learning algorithm

For episode = 1, M **do**

 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence $\phi_1 = \phi(s_1)$

For $t = 1, T$ **do**

 With probability ϵ select a random action a_t
 otherwise select $a_t = \operatorname{argmax}_a Q(\phi(s_t), a; \theta)$

 Execute action a_t in emulator and observe reward r_t and image x_{t+1}

 Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$

 Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D

 Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from D

 Set $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) & \text{otherwise} \end{cases}$

 Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to the network parameters θ

 Every C steps reset $\hat{Q} = Q$

End For

End For

ϕ represents the NN

Sampling (interaction
with the environment)

Training

Deep Q-Learning algorithm

Tabular Q-Learning

Deep Q-Learning

Deep Q-Learning algorithm

Tabular Q-Learning

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Deep Q-Learning

Deep Q-Learning algorithm

Tabular Q-Learning

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Deep Q-Learning

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) \quad \left(y_j - Q(\phi_j, a_j; \theta) \right)^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q}(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

New
Q-value

Deep Q-Learning

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) \quad \left(y_j - Q(\phi_j, a_j; \theta) \right)^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

New Q-value Former Q-value

Deep Q-Learning

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) \quad \left(y_j - Q(\phi_j, a_j; \theta) \right)^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} (R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

New Q-value Former Q-value step size

Deep Q-Learning

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) \quad \left(y_j - Q(\phi_j, a_j; \theta) \right)^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} (\underline{R_{t+1}} + \gamma \max_{a'} Q(S_{t+1}, a') - \underline{Q(S_t, A_t)})$$

New Q-value Former Q-value step size Reward

Deep Q-Learning

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) \quad \left(y_j - Q(\phi_j, a_j; \theta) \right)^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state)

Deep Q-Learning

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) \quad \left(y_j - Q(\phi_j, a_j; \theta) \right)^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state) Former Q-value

Deep Q-Learning

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) \quad \left(y_j - Q(\phi_j, a_j; \theta) \right)^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state) Former Q-value

TD-target

Deep Q-Learning

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) \quad \left(y_j - Q(\phi_j, a_j; \theta) \right)^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state) Former Q-value

TD-target

TD-error

Deep Q-Learning

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$$
$$(y_j - Q(\phi_j, a_j; \theta))^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state) Former Q-value

TD-target

TD-error

Deep Q-Learning

Q-Target

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$$
$$(y_j - Q(\phi_j, a_j; \theta))^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state) Former Q-value

TD-target

TD-error

Deep Q-Learning

Q-Target

$$y_j = r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$$

Q-Loss

$$(y_j - Q(\phi_j, a_j; \theta))^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state) Former Q-value

TD-target

TD-error

Deep Q-Learning

Q-Target

$$y_j = \underline{r_j} + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$$

Reward

Q-Loss

$$(y_j - Q(\phi_j, a_j; \theta))^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state) Former Q-value

TD-target

TD-error

Deep Q-Learning

Q-Target

$$y_j = \underline{r_j} + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$$

Reward Discounted Estimate (optimal Q-value of next state)

Q-Loss

$$(y_j - Q(\phi_j, a_j; \theta))^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state) Former Q-value

TD-target

TD-error

Deep Q-Learning

Q-Target

$$y_j = \underline{r_j} + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$$

Reward Discounted Estimate (optimal Q-value of next state)

TD-target

Q-Loss

$$(y_j - Q(\phi_j, a_j; \theta))^2$$

Deep Q-Learning algorithm

Tabular Q-Learning

$$\underline{Q(S_t, A_t)} \leftarrow \underline{Q(S_t, A_t)} + \underline{\alpha} \underline{R_{t+1}} + \underline{\gamma \max_{a'} Q(S_{t+1}, a')} - \underline{Q(S_t, A_t)}$$

New Q-value Former Q-value step size Reward Discounted Estimate (optimal Q-value of next state) Former Q-value

TD-target

TD-error

Deep Q-Learning

Q-Target

$$y_j = \underline{r_j} + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$$

Reward Discounted Estimate (optimal Q-value of next state)

TD-target

Q-Loss

$$(y_j - Q(\phi_j, a_j; \theta))^2$$

TD-error

Suggested reads

Reinforcement Learning: An Introduction (by Sutton, Barto)

<http://incompleteideas.net/book/the-book-2nd.html>

Human-level control through deep reinforcement learning (Mnih et al., 2015)

<https://www.nature.com/articles/nature14236>

Deep Reinforcement Learning with Double Q-learning

<https://arxiv.org/abs/1509.06461>

To the notebook!

Deep Q-Learning algorithm

For episode = 1, M **do**

 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence $\phi_1 = \phi(s_1)$

For $t = 1, T$ **do**

 With probability ε select a random action a_t

 otherwise select $a_t = \operatorname{argmax}_a Q(\phi(s_t), a; \theta)$

 Execute action a_t in emulator and observe reward r_t and image x_{t+1}

 Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$

 Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D

 Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from D

 Set $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-) & \text{otherwise} \end{cases}$

 Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to the network parameters θ

 Every C steps reset $\hat{Q} = Q$

End For

End For

Sample-Based RL Methods

Only learn from experience - lots of trial-error interactions

No need for a model of the world - only data collected while interacting

Temporal Difference Learning

Based on value functions and dynamic programming ideas

Monte Carlo methods for prediction and control

Often used for estimation that relies on repeated random sampling

Monte Carlo methods

Monte Carlo for prediction and control

Often used for estimation that relies on repeated random sampling

Monte Carlo in RL

Estimate values directly from experience (sequence of S, A, R)

Learning from experience

Agent can accurately estimate a value function without prior knowledge of environment dynamics

No need to know transition probability p

Monte Carlo for policy evaluation

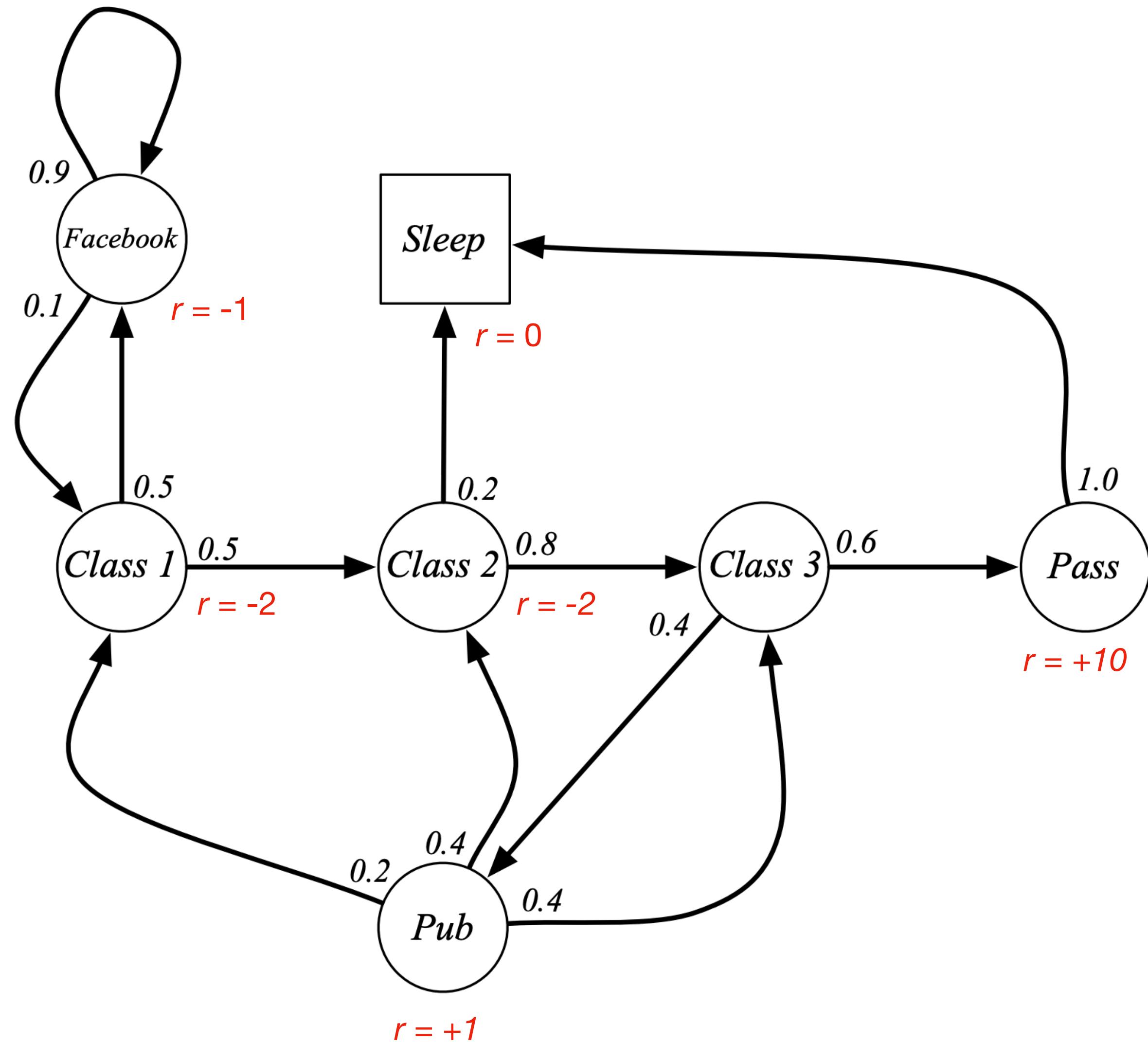
Value functions represent expected returns

Observe multiple returns from the same state

Then average those to estimate the expected return from that state

As the number of samples increases, the expected return gets closer to real avg

Example



Temporal-Difference Learning

$$V(S_t) = V(S_t) + \alpha [V(S_{t+1}) - V(S_t)]$$

Finite Markov Decision Process (MDP)

Bellman Equation

Bellman Equation

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

Off-policy:

Bellman Equation

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

Q-Learning algorithm

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+$, $a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

 Initialize S

 Loop for each step of episode:

 Choose A from S using policy derived from Q (e.g., ε -greedy)

 Take action A , observe R, S'

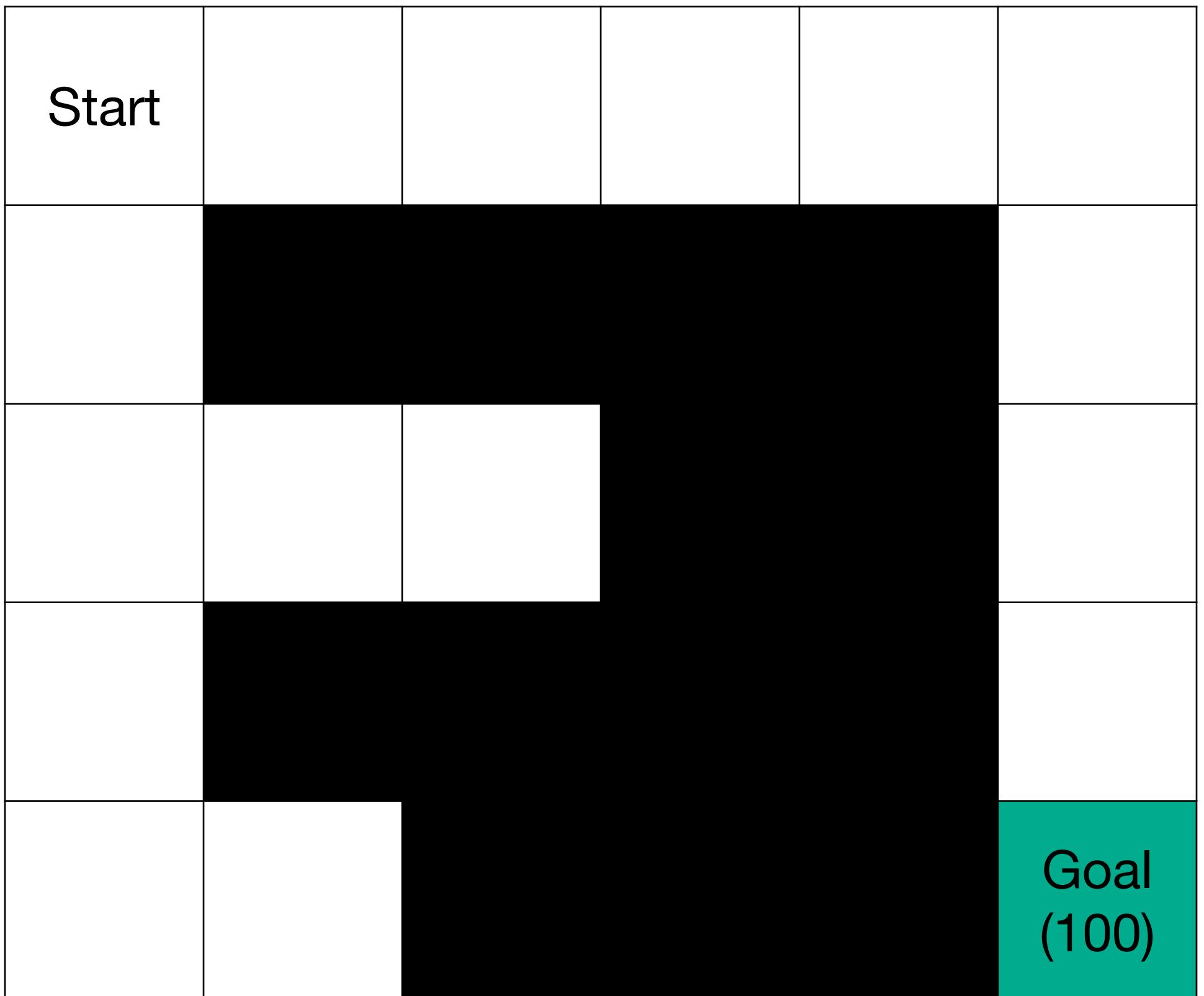
$$Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$$

$$S \leftarrow S'$$

 until S is terminal

Why ϵ -greedy?

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$



Updating the Q Table

Da controllare

Quando MDP non è a finite states.

Come si definisce nel caso di una NN (che ha immagini in input)?

Assumiamo la stessa cosa?