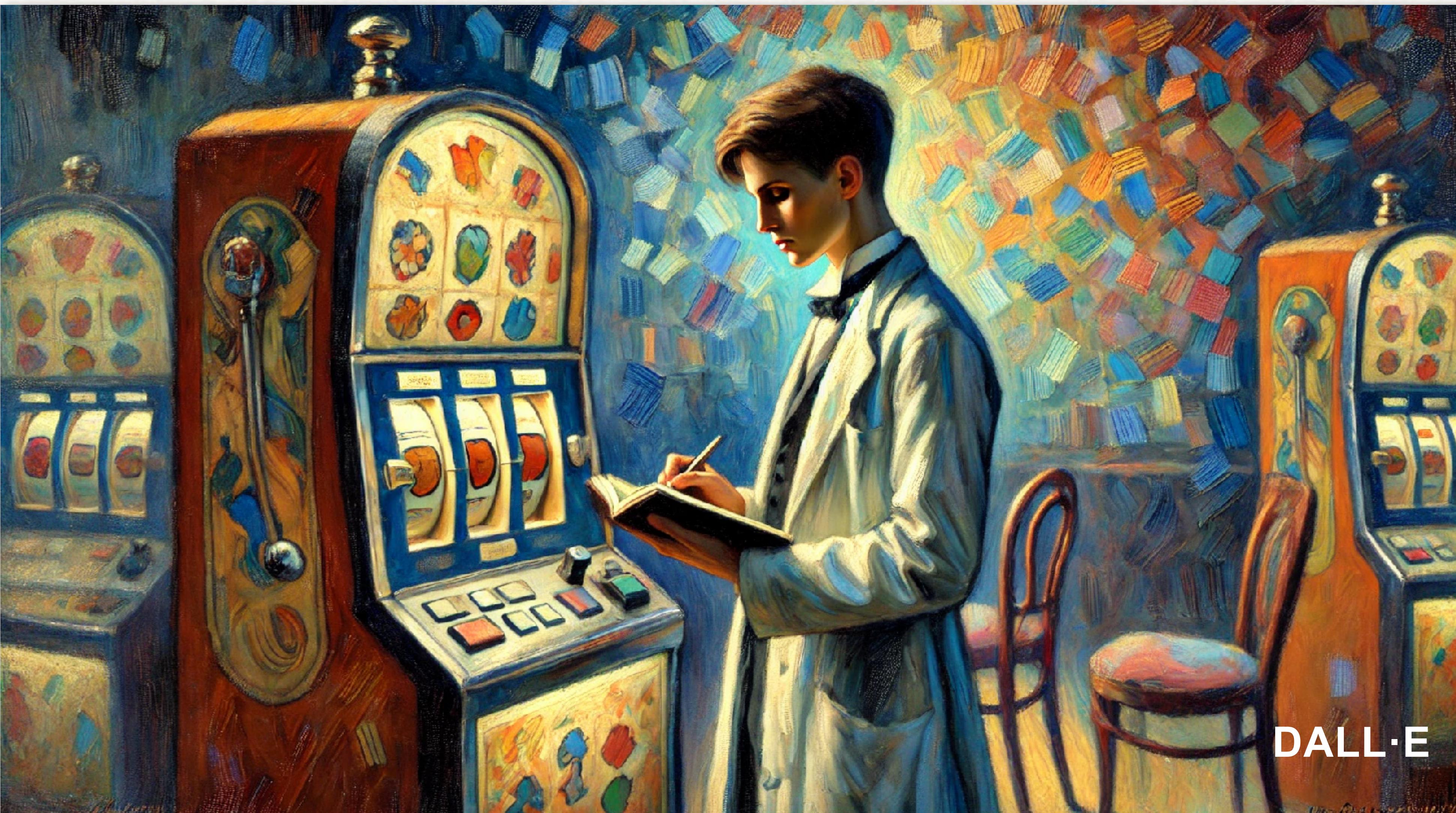


COM 3240

# REINFORCEMENT LEARNING

# IMMEDIATE REWARDS BANDITS



DALL·E

# IMMEDIATE REWARDS

## BANDITS

To know which arm to play we need the expected reward for each of them.

Why expected reward?

The setting is probabilistic.

$$q^*(a) \doteq E[R | A_t = a]$$

We cannot calculate this quantity!



DALL·E

# IMMEDIATE REWARDS BANDITS

$$q^*(a) \doteq E[R | A_t = a]$$

$$Q(a) \approx q^*(a)$$

Idea: lets “track” the reward.

$Q(a)$  tracks the rewards associated with action  $a$ .

$$R(A_t = a) \sim P(R | A_t = a) \quad \text{Sample}$$



# IMMEDIATE REWARDS BANDITS

We will do it through minimising a loss function.

Ingredients of the loss function:

$$Q(a) \quad R(A_t = a)$$

$$(Q(a) - R(A_t = a))^2$$

$$E \left[ (Q(a) - R(A_t = a))^2 \right]$$

Expectation!



DALL·E

# IMMEDIATE REWARDS BANDITS

$$L(a) = \frac{1}{2T_a} \sum_{t=1}^T (Q(a) - R)^2 \mathbf{1}_{A_t=a}$$

Replaced expectation with average.

$$\mathbf{1}_{A_t=a} = \begin{cases} 1 & \text{if } A_t = a \\ 0 & \text{otherwise.} \end{cases}$$

Indicator function

$$T(a) = \sum_{t=1}^T \mathbf{1}_{A_t=a}$$

R=R(t) is the reward at trial t



# IMMEDIATE REWARDS

## BANDITS

We can do two things (see optimisation section!).

- 1) Take the gradient and set it to 0 - stationary, convexity.
- 2) Use gradient descent.

One more thing!

$$L(a) = \frac{1}{2T_a} \sum_{t=1}^T (Q(a) - R)^2 \mathbf{1}_{A_t=a}$$

Use gradient descent, works for non-stationary environments!

# POLICIES

## THE EXPLORATION EXPLOITATION DILEMMA



DALL·E

# **POLICIES**

## **THE EXPLORATION EXPLOITATION DILEMMA**

**While we exploit we are missing out on potentially better rewards.**

**While we explore we may not collect reward, or collect less reward.**

# POLICIES

## THE EXPLORATION EXPLOITATION DILEMMA

**Greedy**  $a_t = \operatorname{argmax}_a Q_t(a)$

**Optimistic Greedy:** initialise Q-values unrealistically high

**Epsilon-Greedy :** explore with probability epsilon, greedy otherwise

**Softmax:**  $P(a) = \frac{e^{Q_t(a)/\tau}}{\sum_b e^{Q_t(b)/\tau}}$

What happens if  $\tau$  grows very large or tends to 0?

# IMMEDIATE REWARDS

## BANDITS

- Initialise Q-values, and count  $T(a)$  for all actions  $a$
- For each time step
  - Select an action  $a$  using policy
  - Increase the count  $T(a)$
  - Take action  $a$  and observe reward  $R$
  - Update  $Q(a)$ 
    - For a stationary environment:  $Q(a) = Q(a) + (R - Q(a))/T(a)$
    - For non-stationary environment:  $Q(a) = Q(a) + \eta(R - Q(a))$

# FUTURE REWARDS

## CHESS, BACKGAMMON, MAZES



Stochastic



Deterministic

DALL·E

# RL THROUGH THE LENS OF OPTIMISATION

## REWARD MAXIMISATION

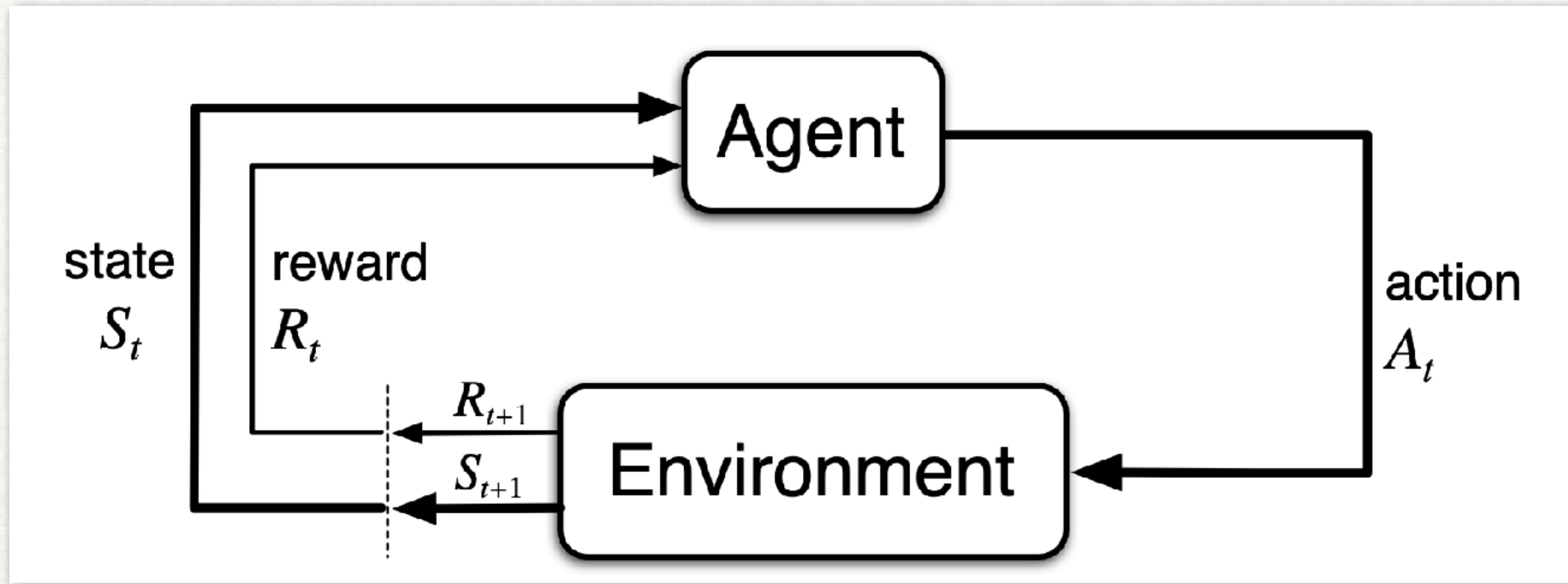
$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \dots + R_{t+N}$$

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots$$

$$0 \leq \gamma < 1$$

# FUTURE REWARDS

## MARKOV DECISION PROCESS

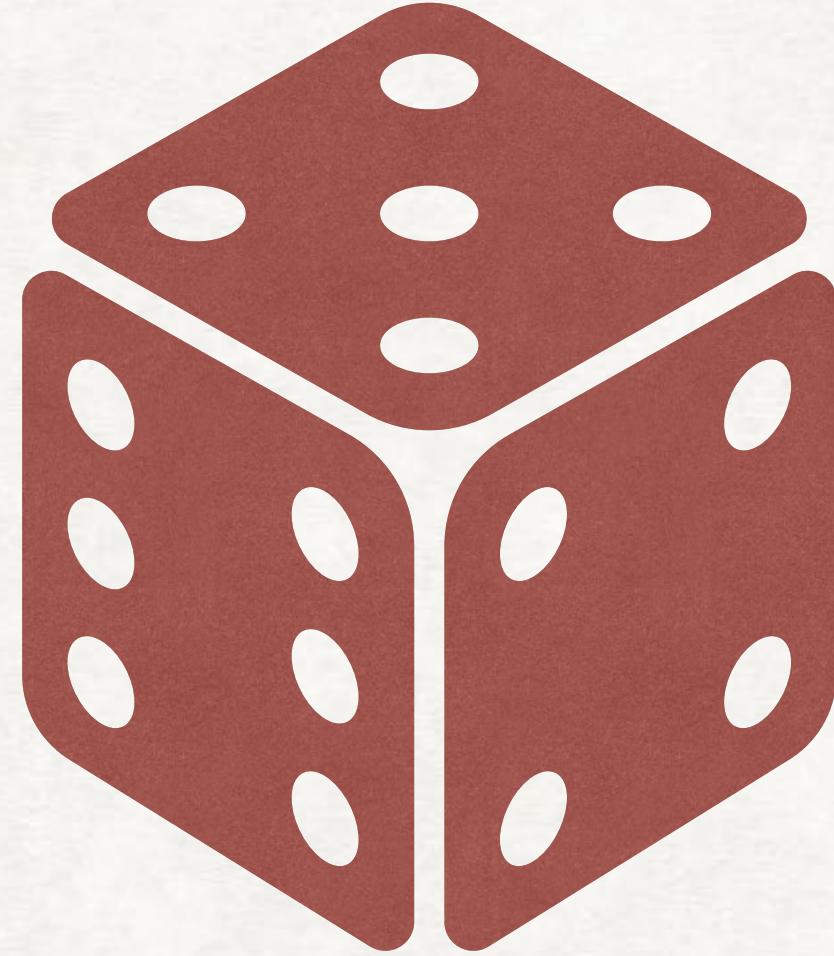


The agent-environment interaction from Sutton & Barto, 2020 - with permission

# FUTURE REWARDS

## MARKOVIAN PROPERTY

The future depends only on the present state and action taken, not on the past history!



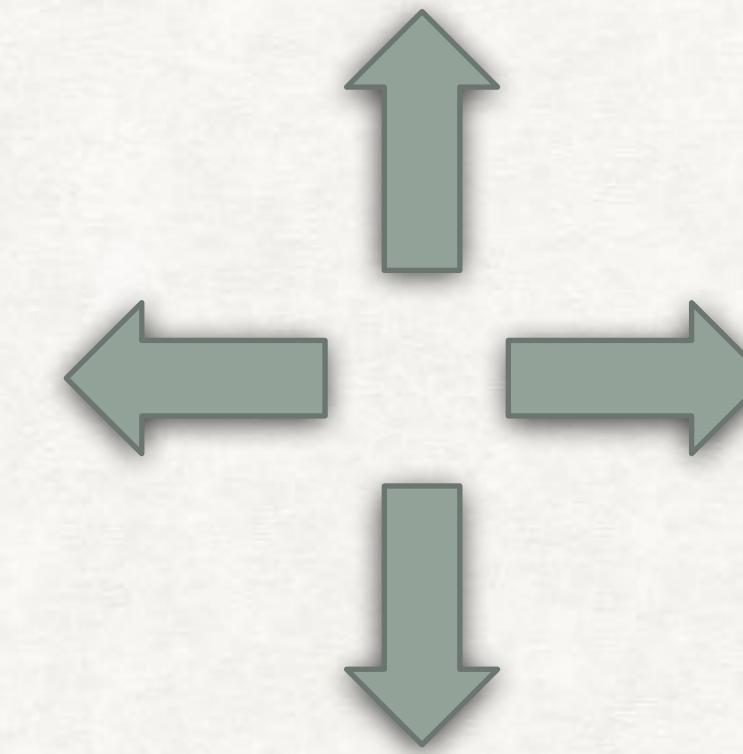
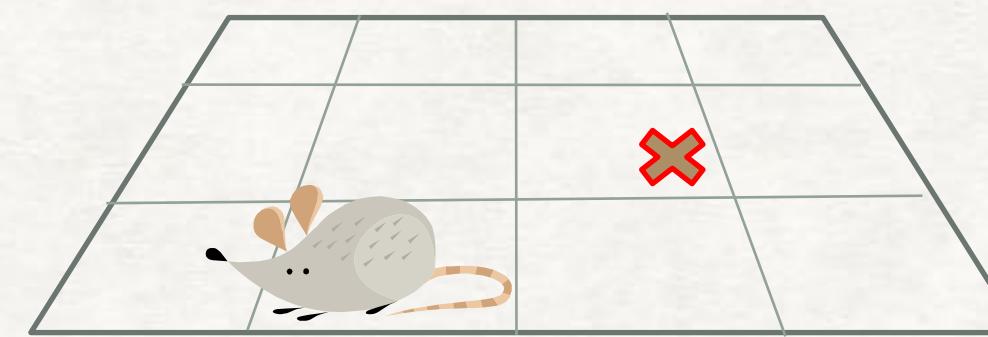
D6 dice: the probability of rolling e.g. 6, is  $1/6$  and it does NOT depend on the previous rolls.

Not true in real life and in many practical scenarios, but there is a remedy.

# FUTURE REWARDS

## MARKOVIAN PROPERTY

The future depends only on the present state and action taken, not on the past history!



$$P(S_{t+1} | S_t, A_t, S_{t-1}, A_{t-1}, \dots, S_0, A_0) = P(S_{t+1} | S_t, A_t)$$

# FUTURE REWARDS APPROACH

- We will see:
  - Scenarios of known environment, i.e. known transition probabilities and rewards.
  - We will be able to calculate the total expected rewards.
  - Next week we will study an algorithm that is able to estimate the rewards, so we can use it in “unknown” environments.
  - It will be again via a sensible loss function and gradient learning.

# FUTURE REWARDS

## MARKOV DECISION PROCESS



2 dials {0,1}

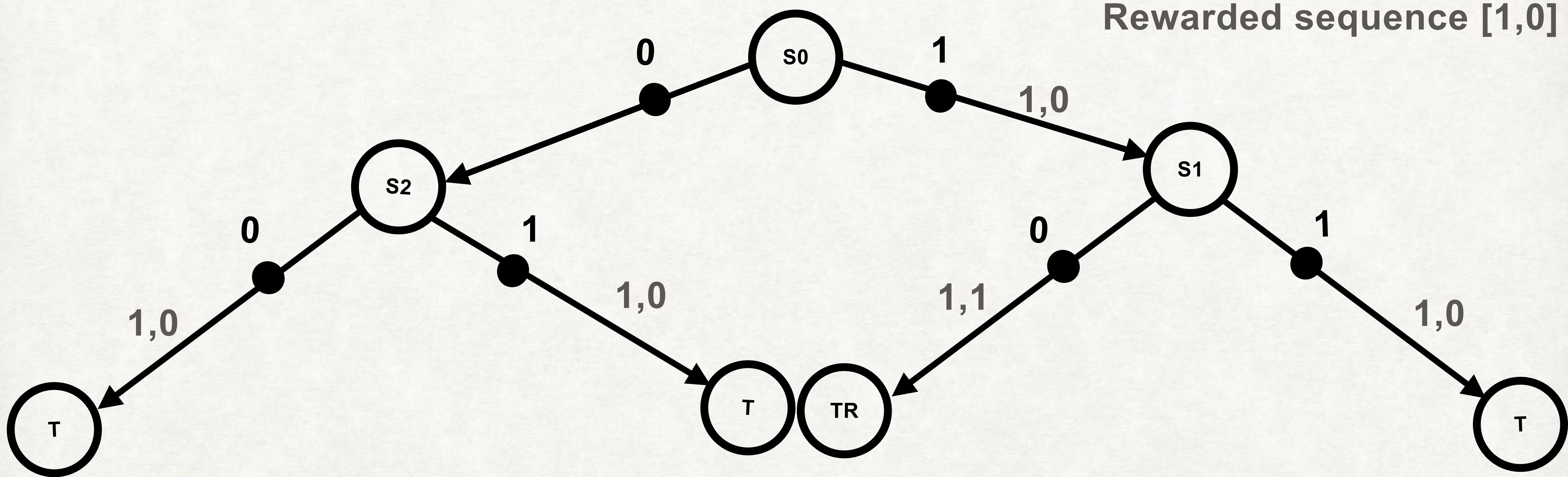
2 digits code

Example of code [1,0]

DALL·E

# FUTURE REWARDS

## DETERMINISTIC EXAMPLE



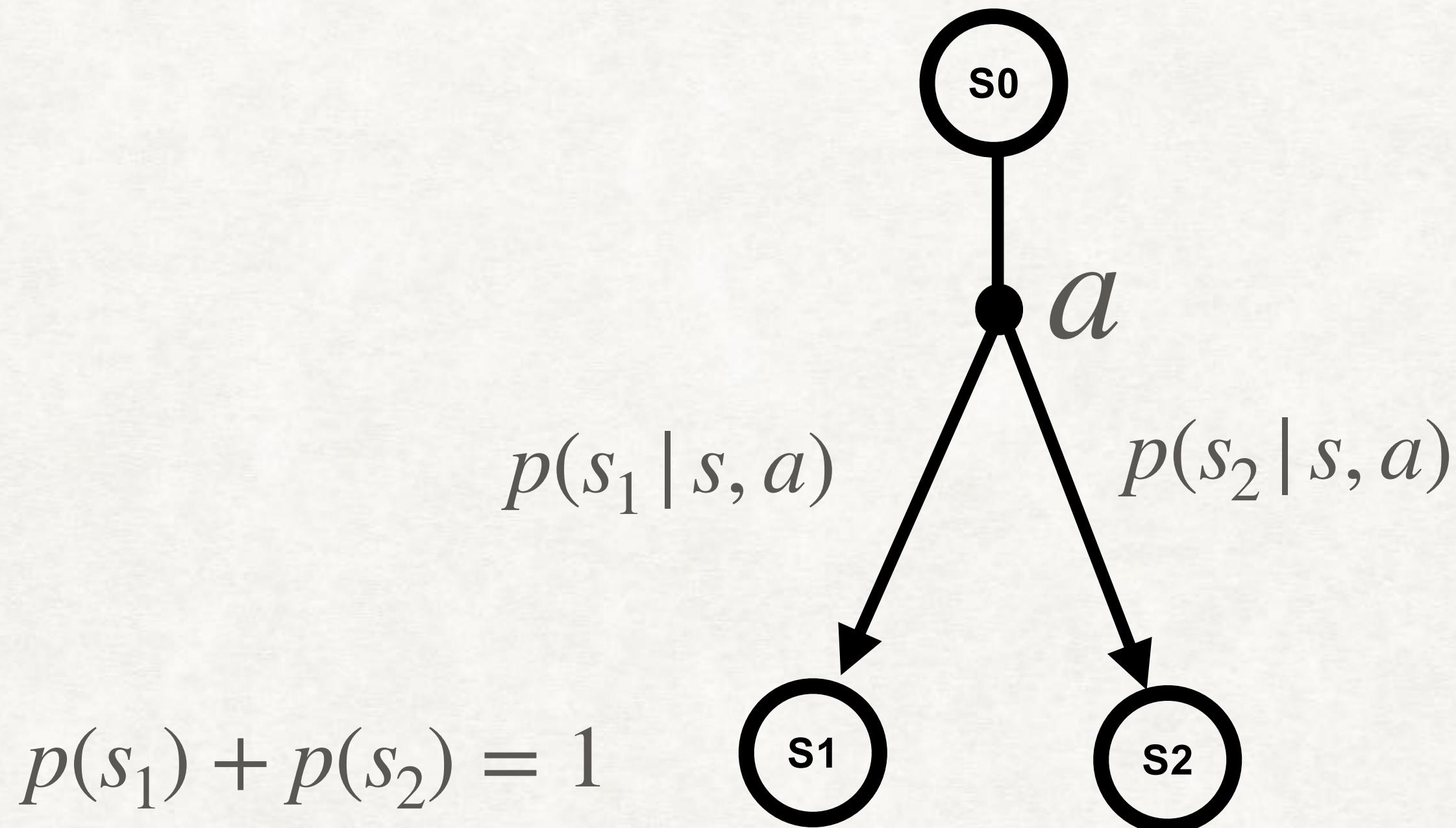
Dots correspond to actions (turning the dial to 0 or 1).

Number pairs correspond to transition probability, reward (here 0 or 1).

Deterministic setup, all probabilities are 1.

# FUTURE REWARDS

## STOCHASTIC EXAMPLE

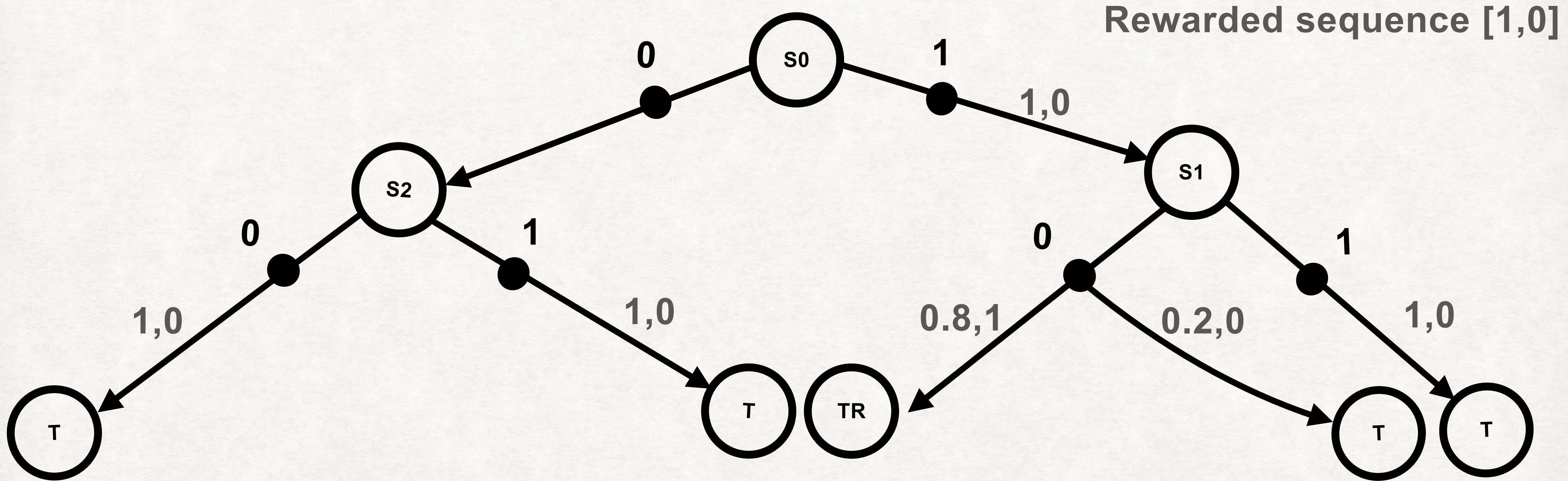


$$\sum_{s'} p(s' | S_t = s, A_t = a) = 1 \quad s' \in \{s_1, s_2, \dots\}$$

DALL·E

# FUTURE REWARDS

## STOCHASTIC EXAMPLE



Dots correspond to actions (turning the dial to 0 or 1).

Number pairs correspond to transition probability, reward (here 0 or 1).

Probabilistic setup, at state  $S_1$  the safe sometimes jams instead of opening.

# FUTURE REWARDS

## STATE AUGMENTATION



DALL·E

# FUTURE REWARD

EXPECTED RETURN

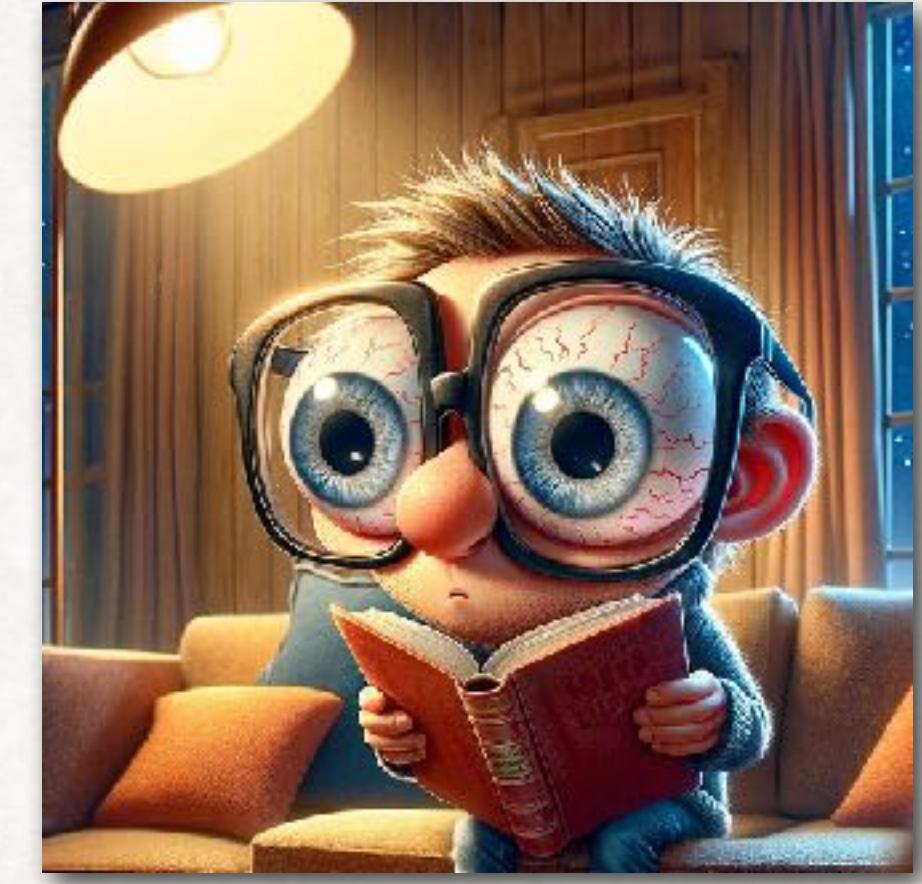
$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \dots + R_T$$

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

# FUTURE REWARD

## EXPECTED RETURN

$$0 \leq \gamma < 1$$



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

Agent behaviour will be different depending on gamma values.

DALL·E



# FUTURE REWARDS

## EXPECTED RETURN

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

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

$$G_{t+1} = R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \dots$$

$$G_t = R_{t+1} + \gamma G_{t+1}$$

**Recursive!**

## FUTURE REWARDS ACTION-VALUE FUNCTIONS

$$G_t = R_{t+1} + \gamma G_{t+1}$$

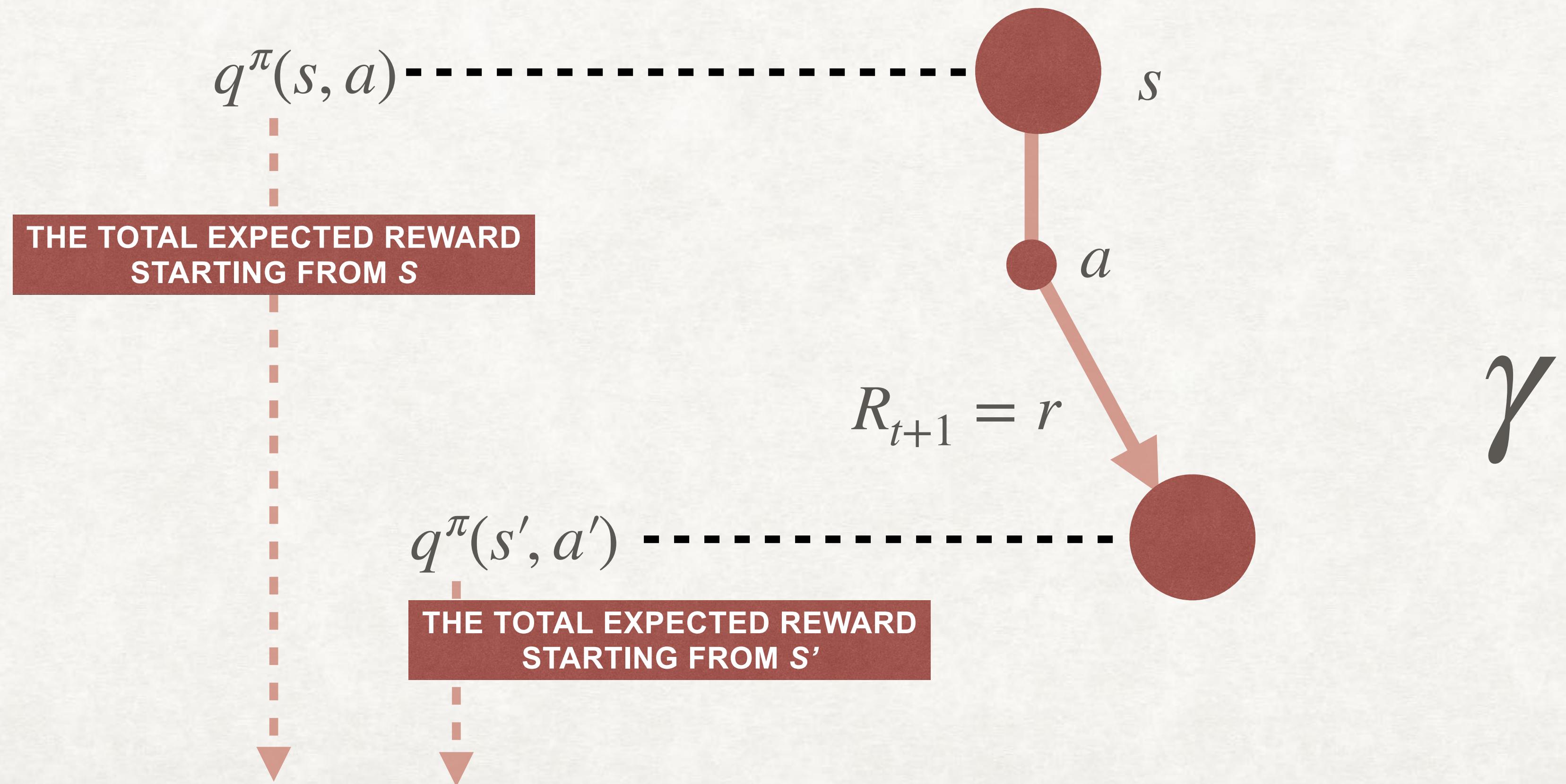
$$q^\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$

$$q^\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma q^\pi(s', a') | S_t = s, A_t = a]$$

**The recursive form of total returns is also followed by the expected reward.**

# FUTURE REWARDS ACTION-VALUE FUNCTIONS

$$q^\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma q^\pi(s', a') \mid S_t = s, A_t = a]$$



# FUTURE REWARDS

## OPTIMALITY & BELLMAN EQUATION FOR ACTION-VALUES

$$q^\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma q^\pi(s', a') \mid S_t = s, A_t = a]$$

$$q^*(s, a) = \max_\pi q^\pi(s, a)$$

$$q^*(s, a) = \mathbb{E}[R_{t+1} + \gamma \max_{a'} q^*(s', a') \mid S_t = s, A_t = a]$$

# FUTURE RETURNS STATE VALUE FUNCTIONS

$$v(s) = \mathbb{E}[G_t | S_t = s]$$

$$v^*(s) = \max_a q^{\pi^*}(s, a)$$

# SUMMARY

## FUTURE REWARDS

- Total return and its recurrent form.
- Expected reward: action-value functions  $q(s,a)$
- Optimal policy
- Optimal value function  $q^*$  under optimal policy  $\pi^*$
- “Greedy” selection of  $q$  - but here it is different from the bandit scenario, how?

**THANK YOU!**