An introduction to Reinforcement Learning

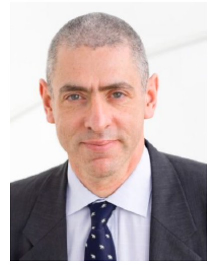
10th of May 2022

Coding: Rescorla Wagner, Blocking

https://github.com/schwartenbeckph/RL-Course/tree/main/2022_05_10

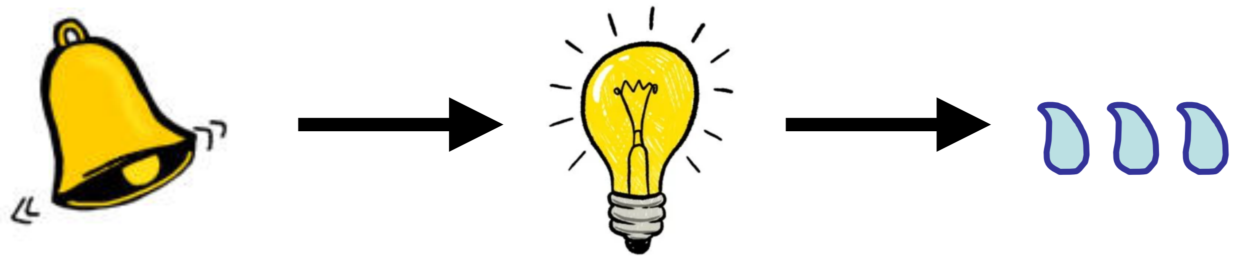
History: Learning *and* Control

- **TD learning**, Actor-critic architecture (Sutton & Barto, 1981, 1982)



- **Q learning** (Watkins 1989; Watkins & Dayan 1992):
  - Integrate dynamic programming with online learning

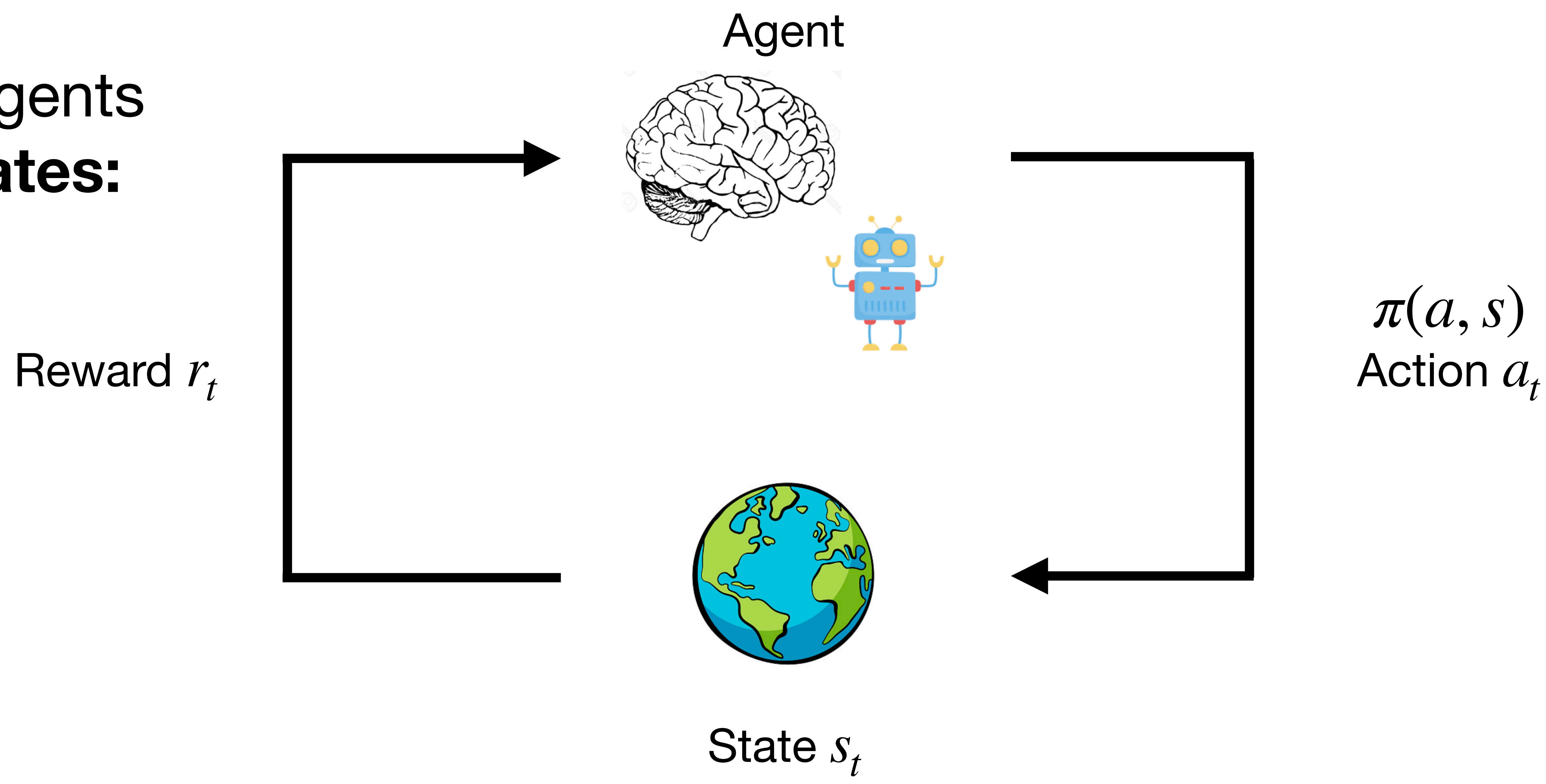
- Key idea: use **experience** and **own value estimates**!
  - Simple example: secondary reinforcement



Temporal Difference Learning

Based on a reward signal, agents learn **values of actions/states:**

$$V_\pi(s) = \mathbb{E}_\pi[R \mid s_0 = s]$$



Agent

Reward r_t

$\pi(a, s)$
Action a_t

State s_t

TD Learning:

Prediction error

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (r + \gamma \cdot V(s_{t+1}) - V(s_t))$$

Learning rate Discount rate

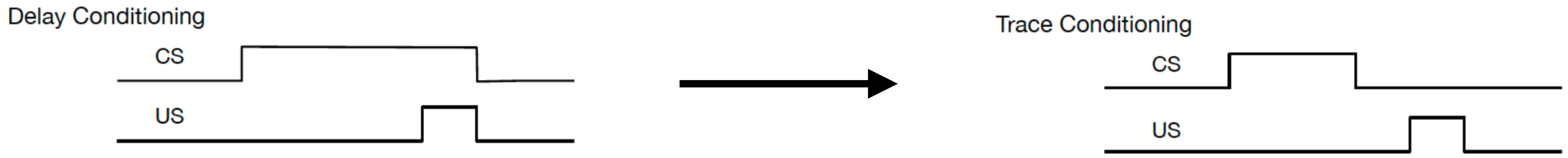
Rescorla Wagner Learning:

Prediction error

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (r - V(s_t))$$

Learning rate

Temporal Difference Learning



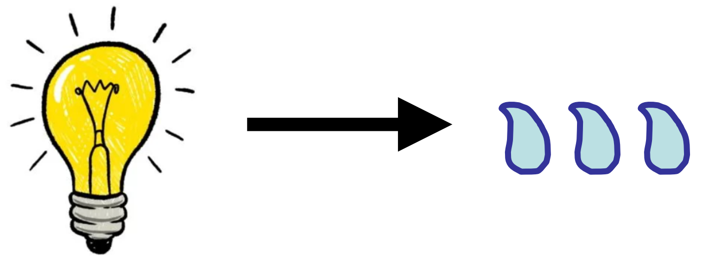
- Extends Rescorla–Wagner model
  - Learn within-trial *and* between-trial relationships
- Operates in ‘real-time’
  - t labels time steps within *or* between trials
  - Think of time between t and $t + 1$ as a small time interval (e.g. 1ms)
- Solves:
  - Higher order conditioning 🔔 → 💡 → 💧💧💧
  - NO blocking if CS_2 is moved before previously learnt CS_1 💡 → 💧💧💧 💡🔔 → 💧💧💧 🔔 → 💧💧💧
  - **Dopamine**…

Prediction error

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (r + \gamma \cdot V(s_{t+1}) - V(s_t))$$

Learning rate Discount rate

Can RL tell us anything about the brain?

- Yes, quite a lot.
- Particularly, it looks like dopamine (DA) is a key neurotransmitter for (TD) reward learning
    - Schultz, Dayan & Montague (1997): 

Dopamine neurons signal immediate reward

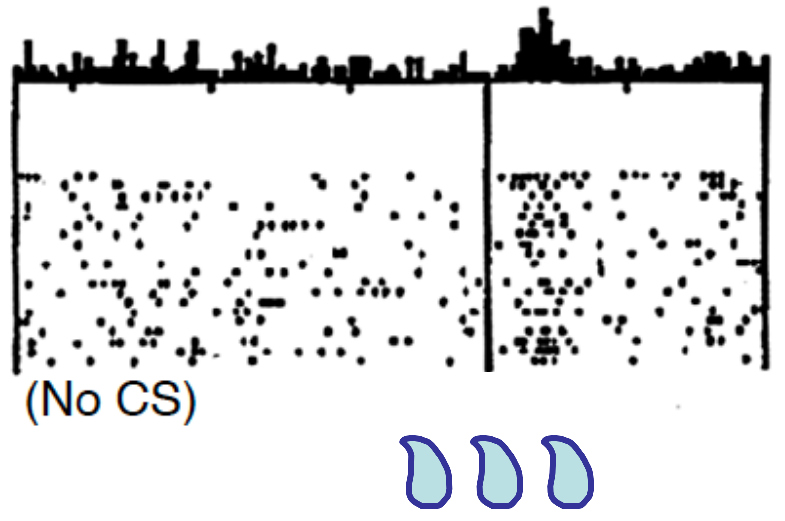
BUT: after training…

- DA signal reward prediction
- But not correctly predicted reward!

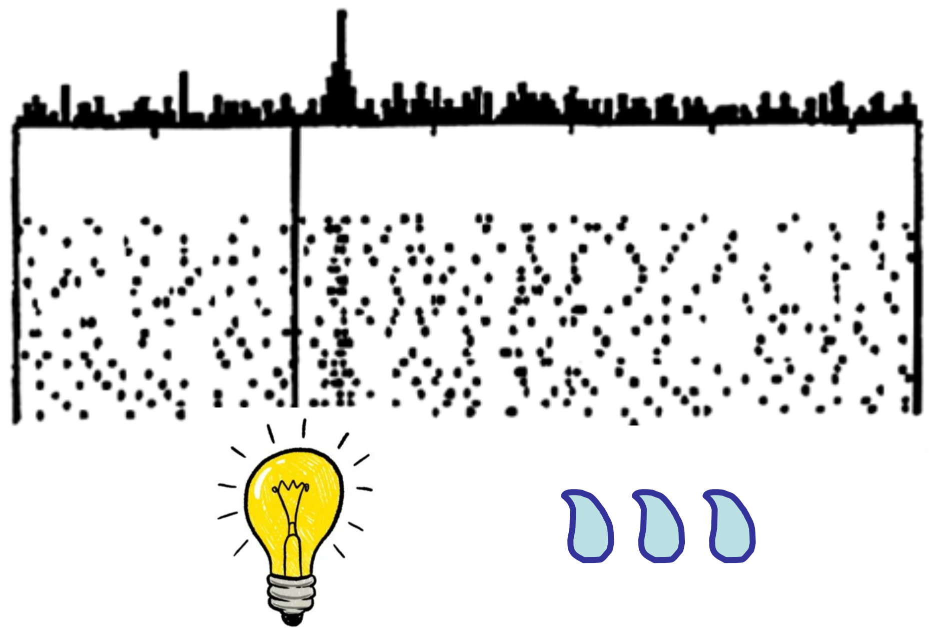
AND: it signals the unexpected omission of a reward!

Do dopamine neurons report an error in the prediction of reward?

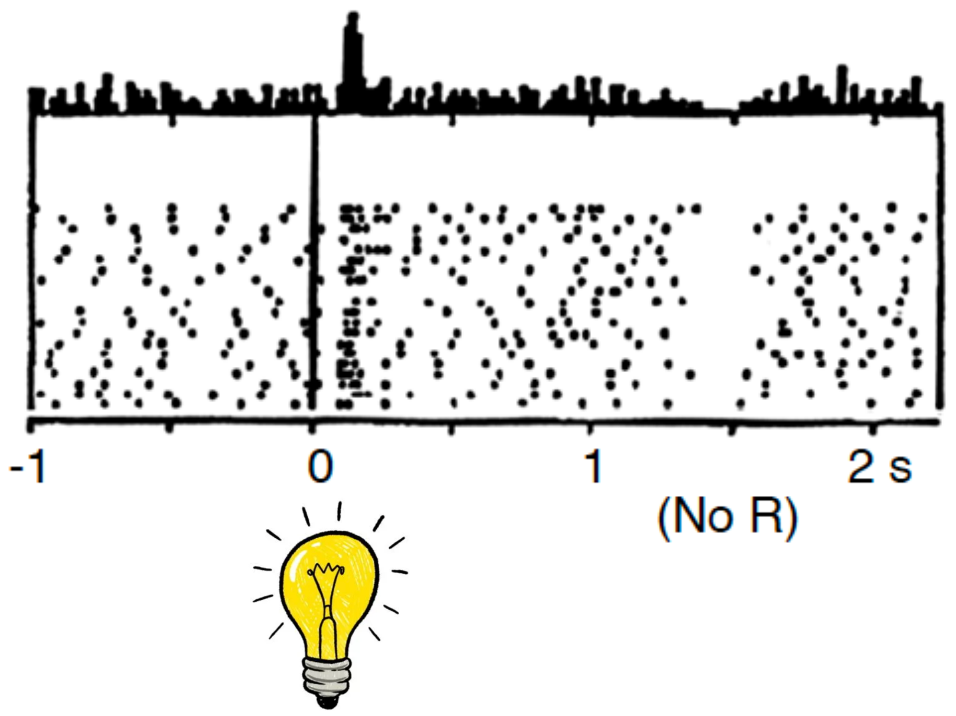
No prediction
Reward occurs



Reward predicted
Reward occurs



Reward predicted
No reward occurs



This provides strong evidence that DA signals a **reward prediction error**

Coding: TD Learning

https://github.com/schwartenbeckph/RL-Course/tree/main/2022_05_10

Recap: Basic setup: how to agents learn to act?

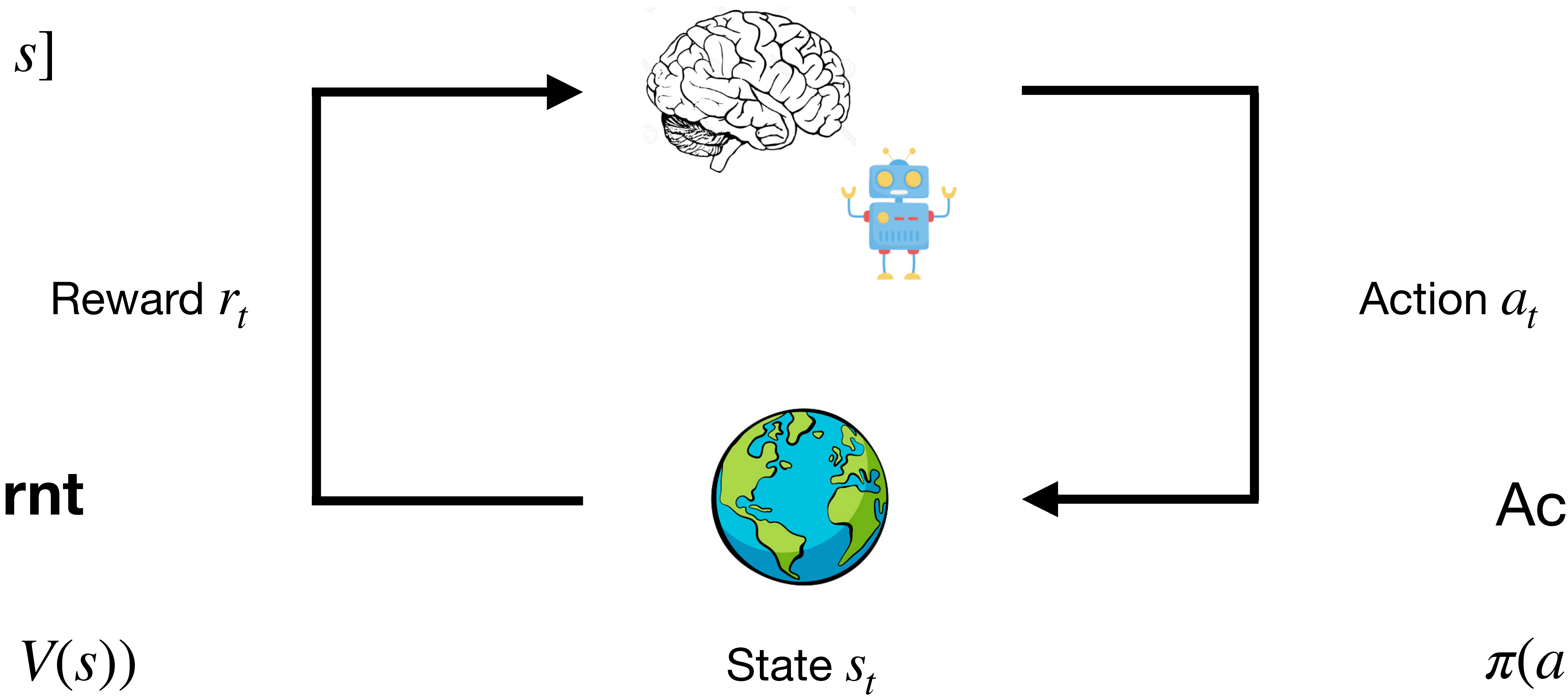
Based on a reward signal, agents learn **values of actions/states:**

$$V_\pi(s) = \mathbb{E}_\pi[R \mid s_0 = s]$$

Agents can learn a **model of the environment** to make smarter decisions, e.g.:

$$P(s_{t+1} = s \mid s_t = s, a_t = a)$$

Agent



Reward r_t

Action a_t

State s_t

Values can be **learnt** (simplified!!):

$$V(s) \leftarrow V(s) + \alpha \cdot (r - V(s))$$

Learning rate

Prediction error

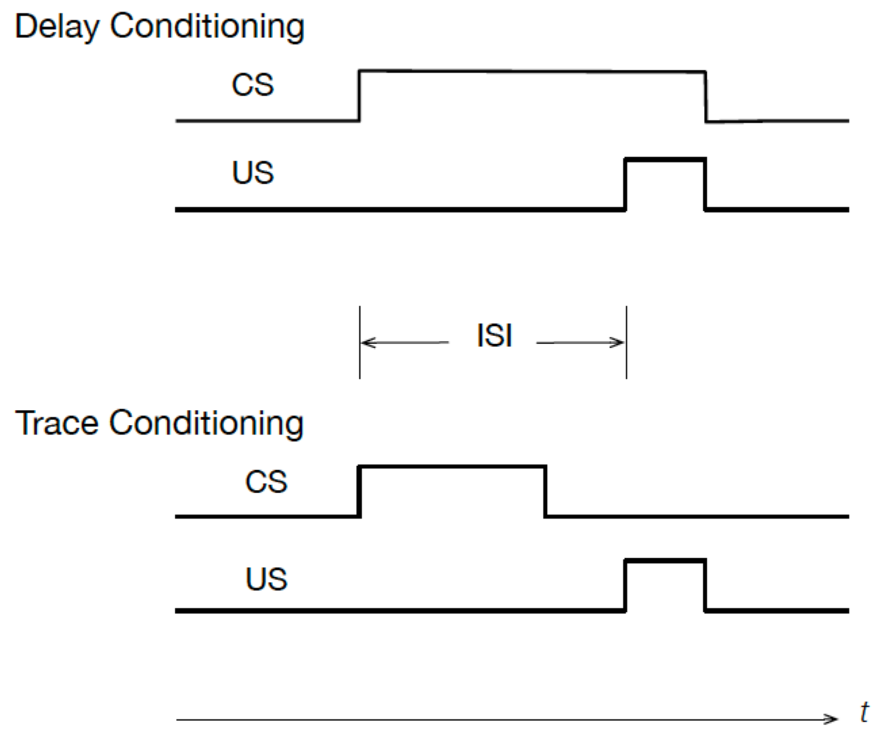
Action is governed by a **policy**:

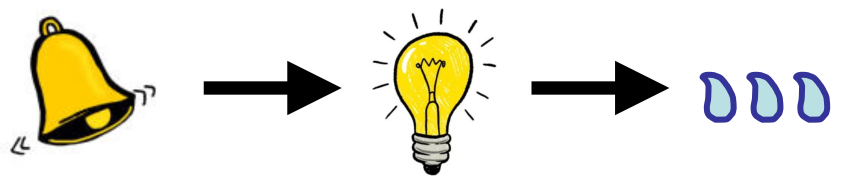
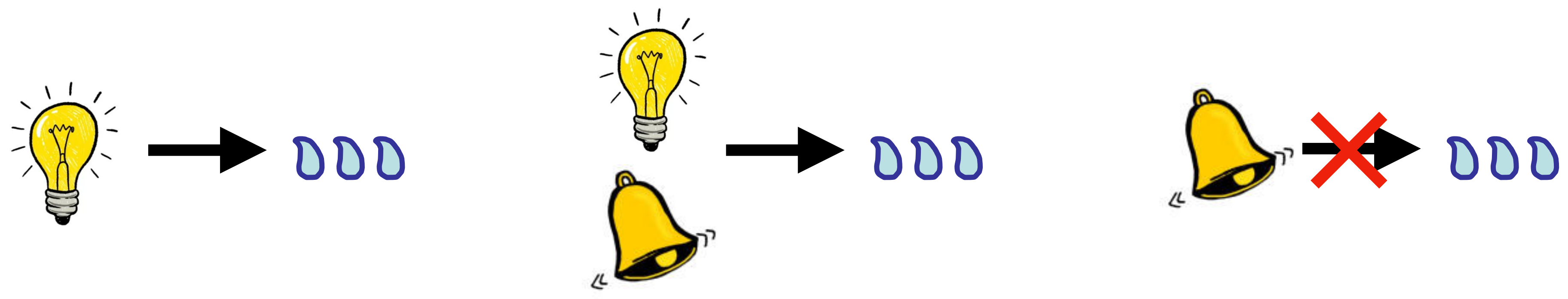
$$\pi(a, s) = P(a_t = a \mid s_t = s)$$

Recap: “Three” historical branches of RL

- Association learning, prediction (early 1900s)
- Optimal control (1950 onward)
- Learning *and* control (1980 onward)

Recap: Learning to predict reward



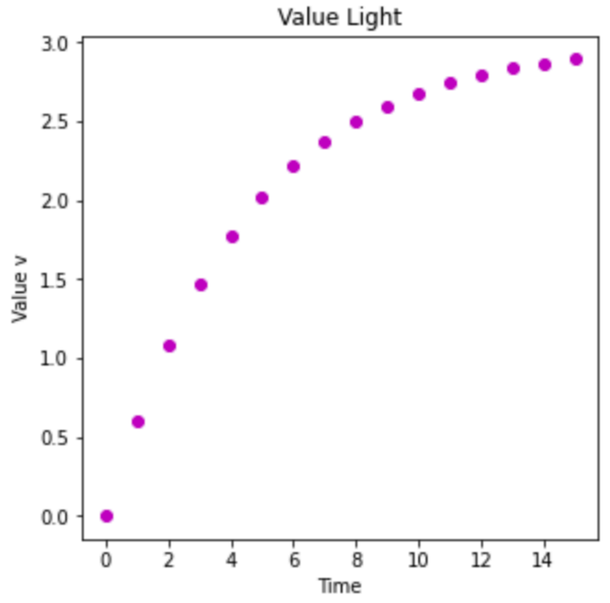
- **Classical** (Pavlovian) **conditioning** (roughly) in domain of algorithms for **prediction**
  - Algorithms for **control: instrumental** (operant) **conditioning**
- At least two interesting phenomena in classical conditioning from algorithmic perspective:
  - **Higher-order conditioning**  Temporal Difference (TD) Learning
  - **Blocking**  Rescorla-Wagner Learning

Recap: Blocking and Rescorla-Wagner Learning

Learn associative strength between a CS and US

$$V(s) \leftarrow V(s) + \alpha \cdot (r - V(s))$$

💡 ⟶ 💧💧💧



V(💡) ← V(💡) + α · V(💧💧💧 - 💡)

Introduce a second CS:

💡 🔔 ⟶ 💧💧💧

V(💡 🔔) ← V(💡 🔔) + α · V(💧💧💧 - 💡 🔔)

V(💡 🔔) = V(💡+🔔) := V(💡) + V(🔔)

V(💡+🔔) ← V(💡+🔔) + α · V(💧💧💧 - (💡 +🔔))

What does the value of the sound CS look like at different stages of learning?