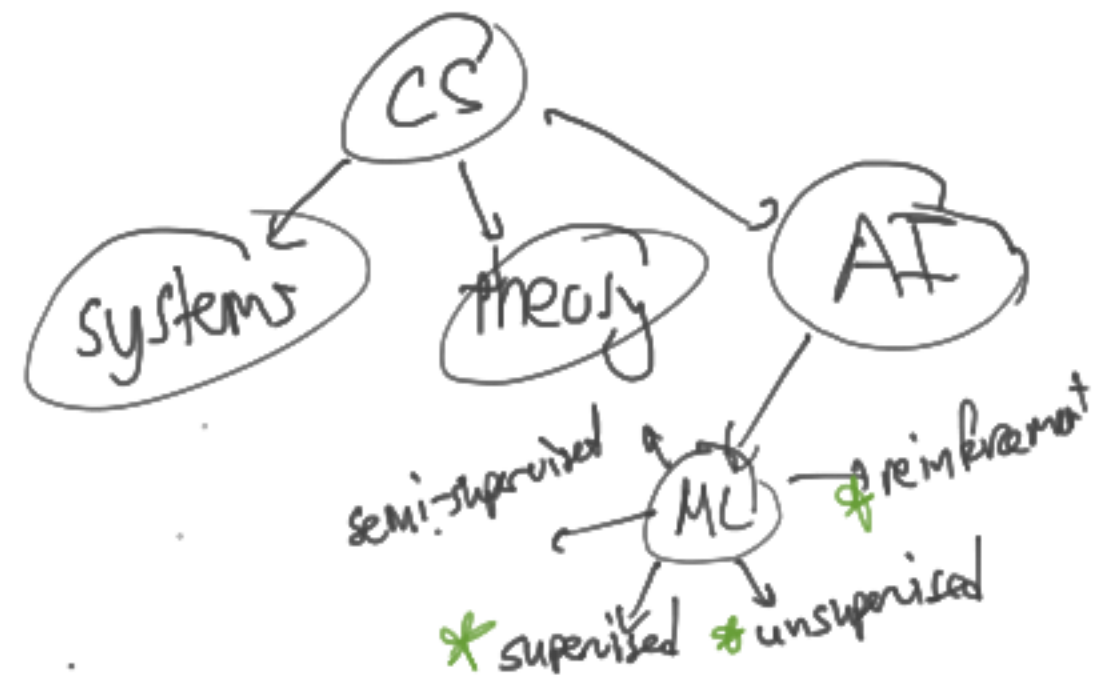
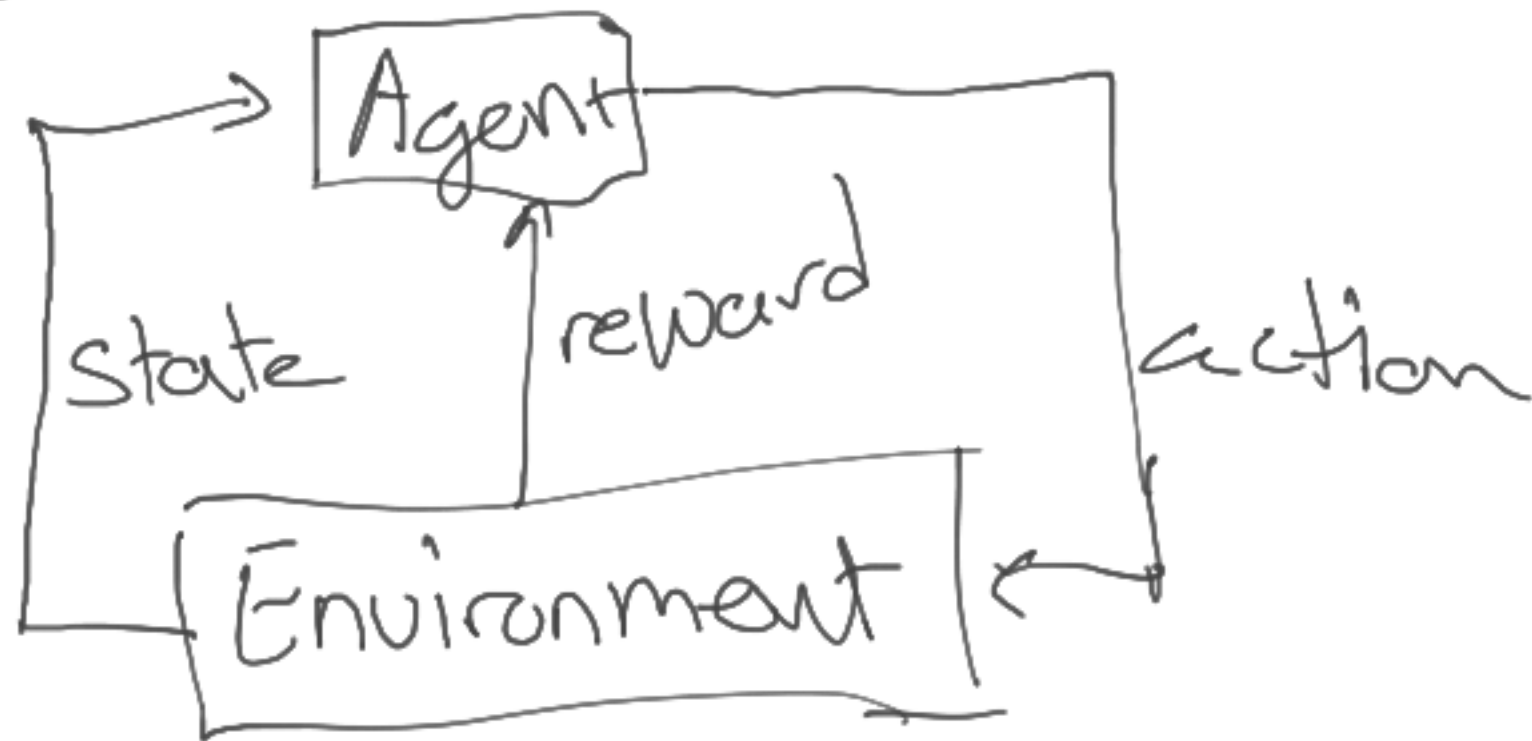


What is reinforcement learning (RL)?

"Reinforcement learning is an area of machine learning, inspired by behaviorist psychology, concerned with how an agent can learn from interactions with an environment."

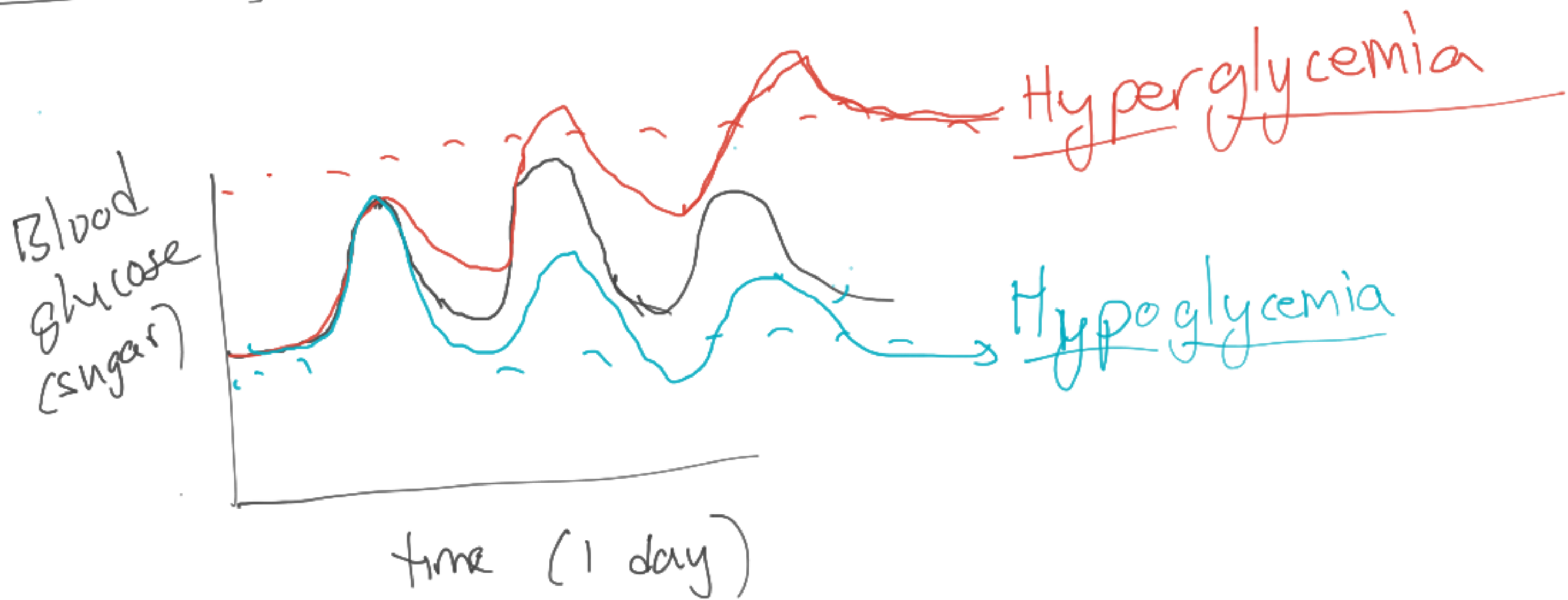
- Wikipedia / Sutton & Barto / Phil
1998

Agent - Environment Diagram



Agent: Child, dog, robot, program, ITS, diabetes treatment.

Environment: World, lab, software environment



Neuroscience:

How do animals learn?

A specific agent

or set of agents.

- The study of some examples of learning and intelligence!

RL (ML)

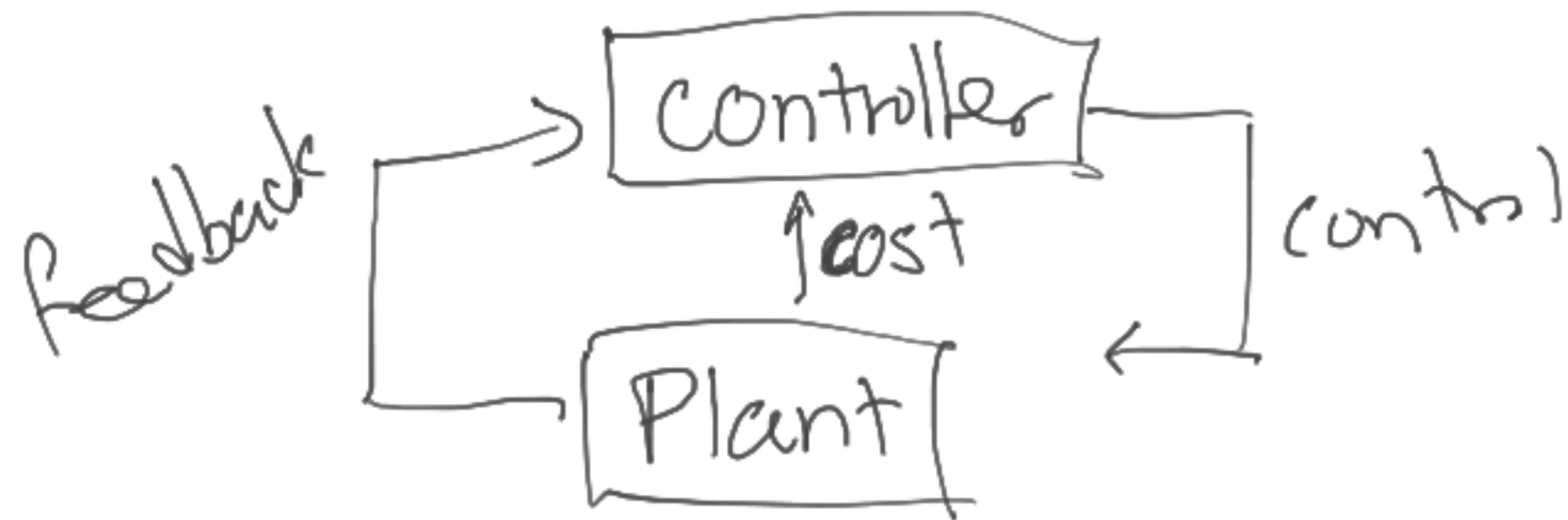
: How can we make an agent that learns?

- The study of learning & intelligence. (in general)

Dopamine \approx Temporal difference error.

Two most related fields

- Operations research
- Control (adaptive / classical)



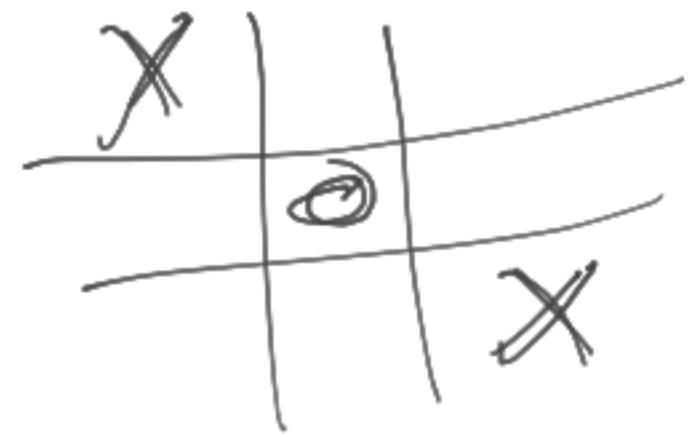
- Main difference is that these fields typically assume the environment (Plant) can and should be directly approximated.

When should you use RL?

→ As a last resort.

- Key properties:

Next time!



9
beads
9 colors.

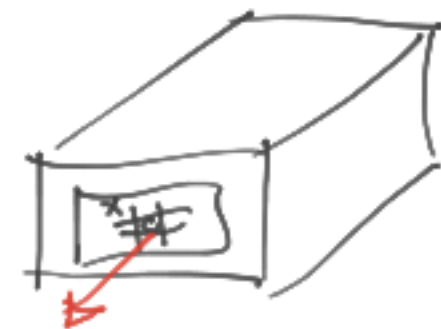
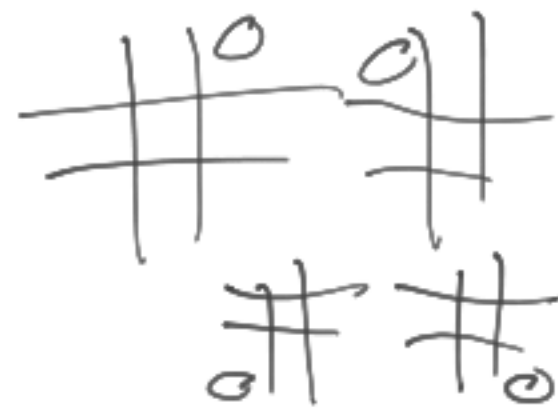
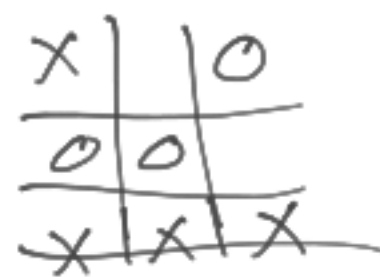




Key properties:

- 1) Evaluative feedback
↳ "This is how good the outcome was," (not instructional)
not "this is what you should have done."
- 2) Sequential.
↳ ~~no~~ no data set!

MENACE (Donald Michie, 1961)



304 matchboxes.

4 bead colors.

↳ possible moves.

Win: Put back bead, +3 more of same color.

Loss: Remove beads.

Tie: Put back bead, +1 more of same color.
(Draw)

Take:

- Act at least slightly random.
- Good outcome
↳ take chosen actions more often
- Bad outcome
↳ take chosen actions less often.

Act randomly ???

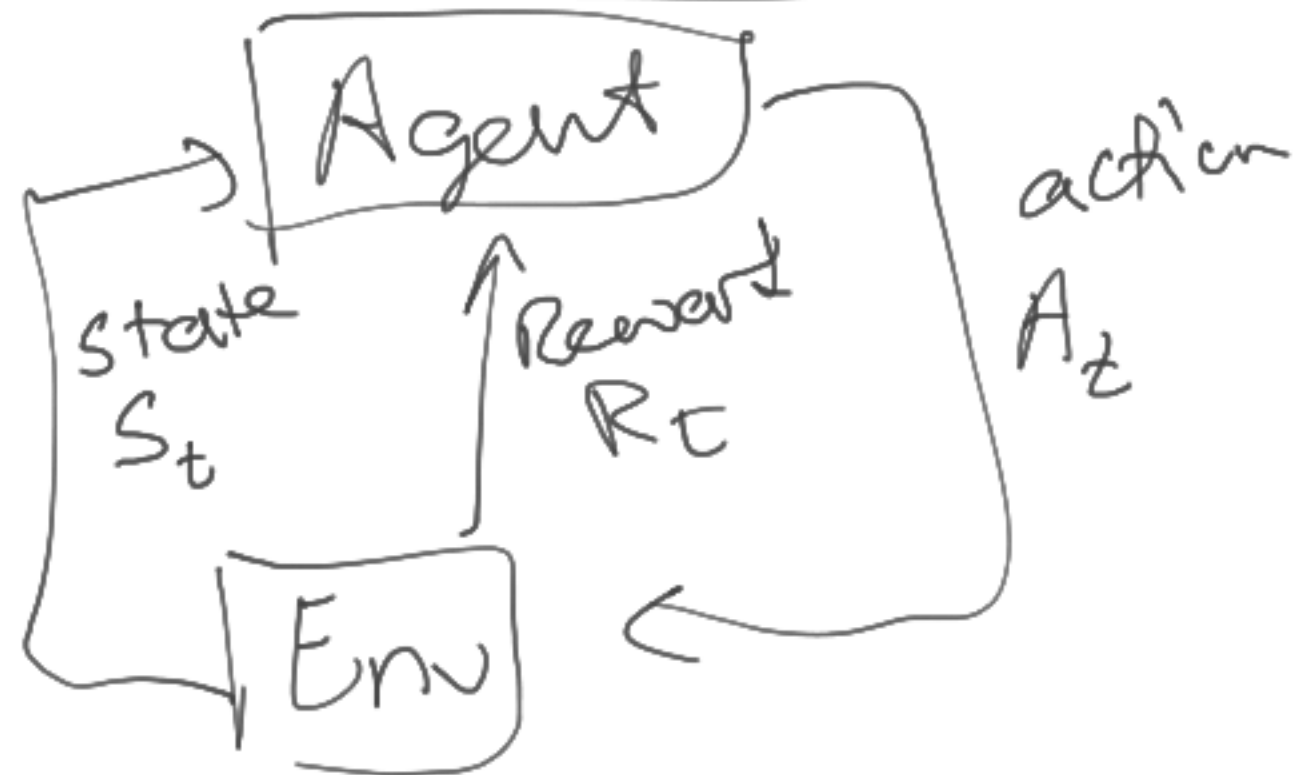
EMPSCI
687

- "Necessary" for learning!

- "exploration" vs "exploitation".

↳ taking actions
that think are
suboptimal

↓
choosing actions
that think are best.



t : time step.
 $t \in \{0, 1, 2, \dots\}$

S_t : State at time t .
★ (Agent's observation
of state)

A_t : Action at time t .

R_t : Reward at time t .

$R_t \in \mathbb{R}$

A policy π is a way of selecting actions.

$$\pi(s, a) = \Pr(A_t = a \mid S_t = s)$$

~~$A_t = \pi(S_t)$~~

- many policies - some good some bad.

Agent's goal: find a policy that maximizes the expected amount of reward the agent receives.
 ↳ not deterministic!

Objective function: $J(\pi) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R_t; \pi \right]$

~ Could be infinite! $\rightarrow J(\pi)$

- Doesn't discount based on time

Means "assuming R_t generated using π "

YATP
 ↳ yet another hyperparameters.

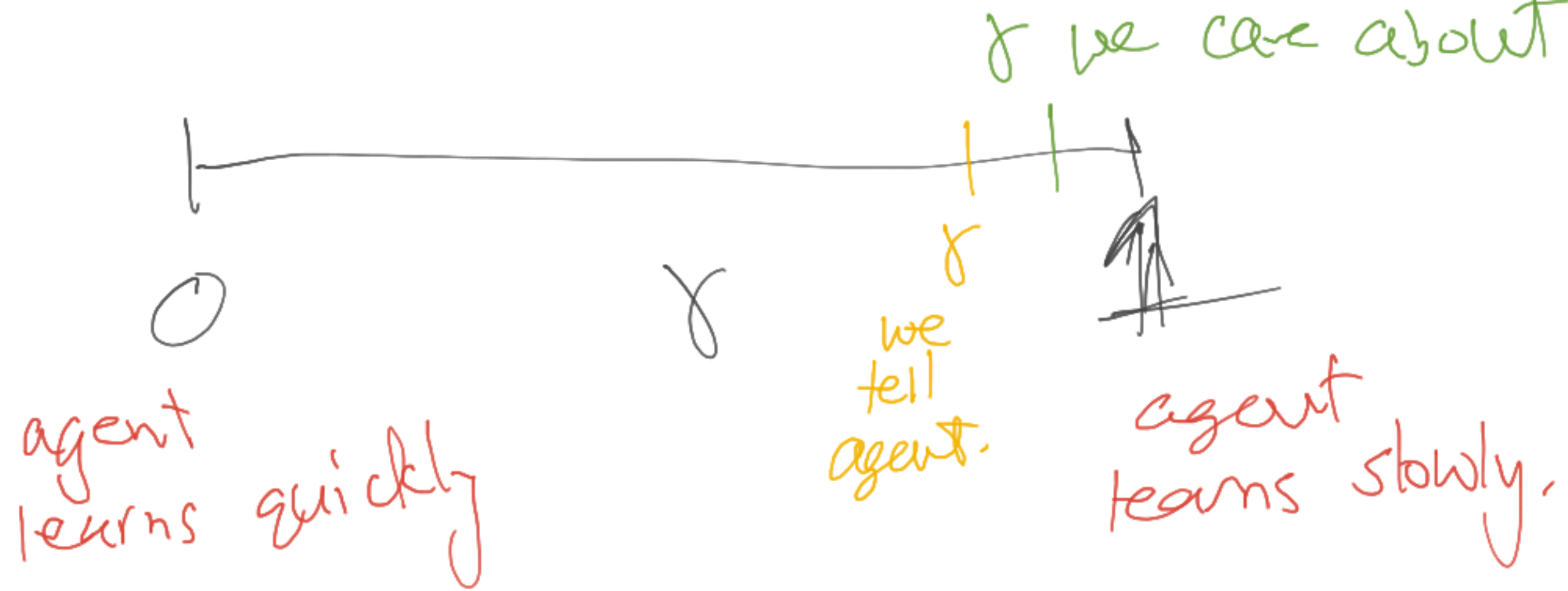
$\gamma \in [0, 1]$

care about short term

care about long term.

$\gamma^0 R_0 + \gamma^1 R_1 + \gamma^2 R_2 + \dots$
 $1 R_0 + .5 R_1 + .25 R_2 + \dots$

Typically $\gamma \approx .9 \leftrightarrow 1$



Agent's goal is to find or approximate

$$\pi^* \in \arg \max_{\pi} J(\pi)$$

π^* is called an optimal policy.