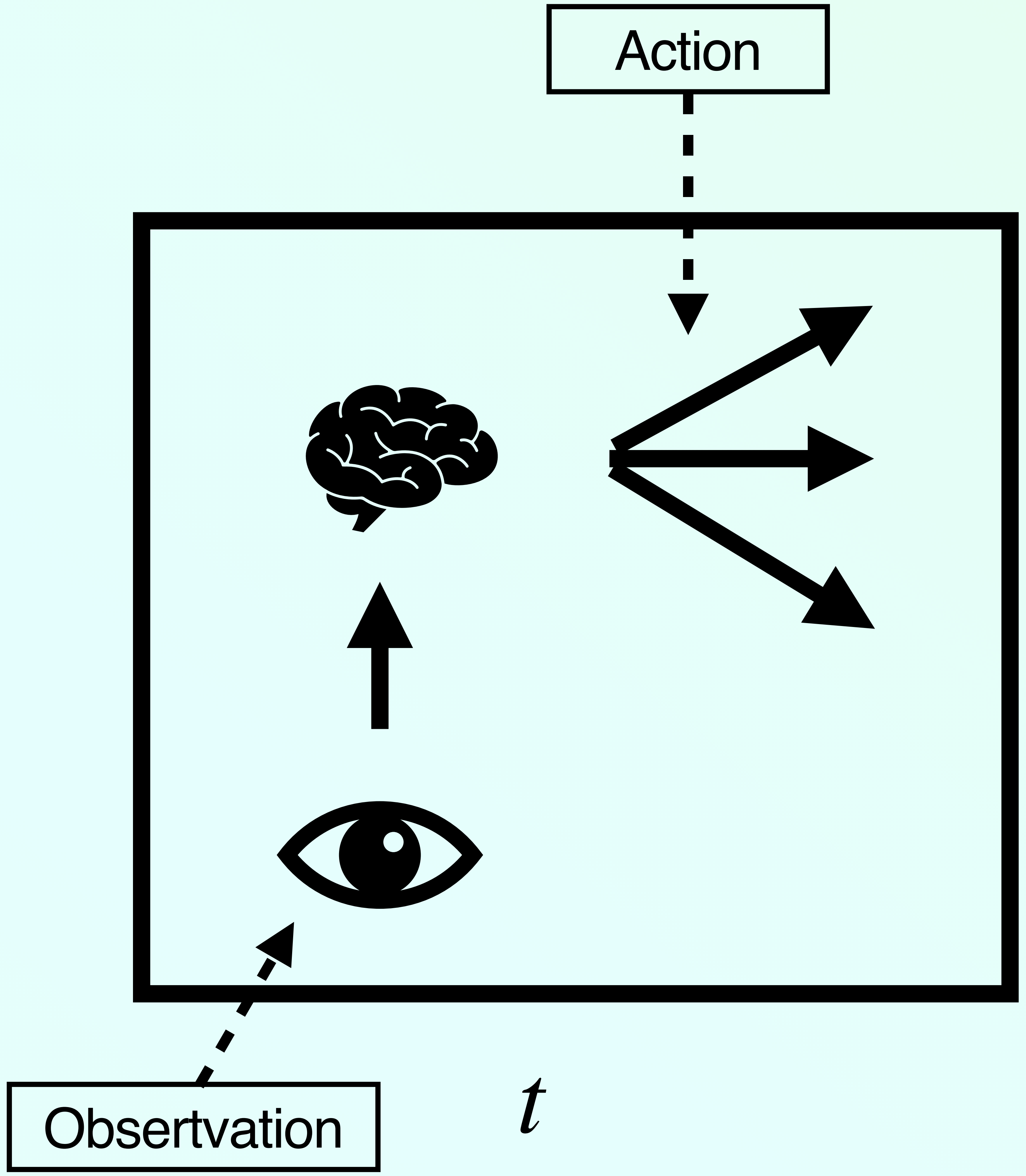
MARKOV DECISION PROCESSES

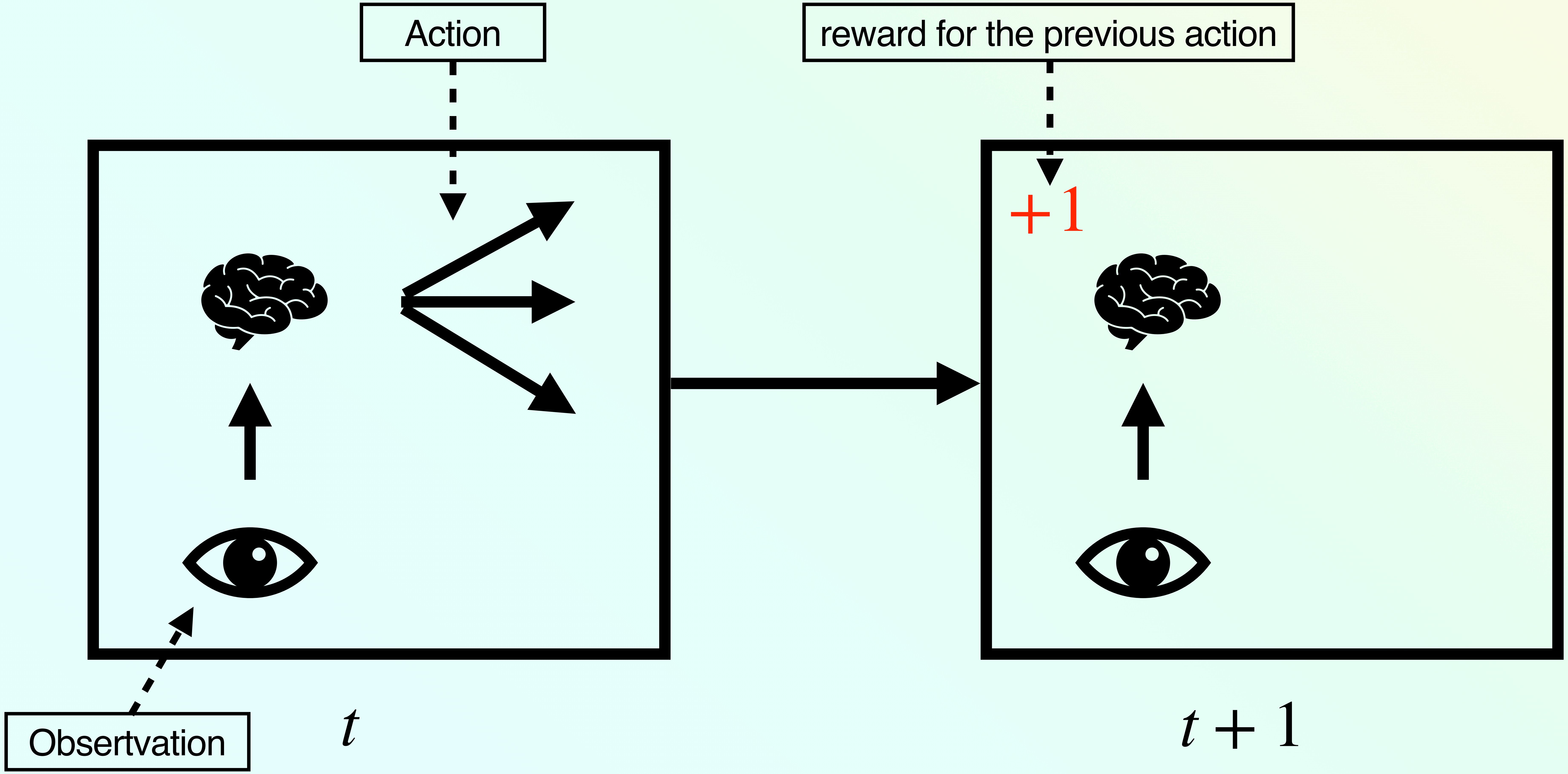
SEQUENTIAL DECISION MAKING

THE BEGING OF THE LONG JOURNEY



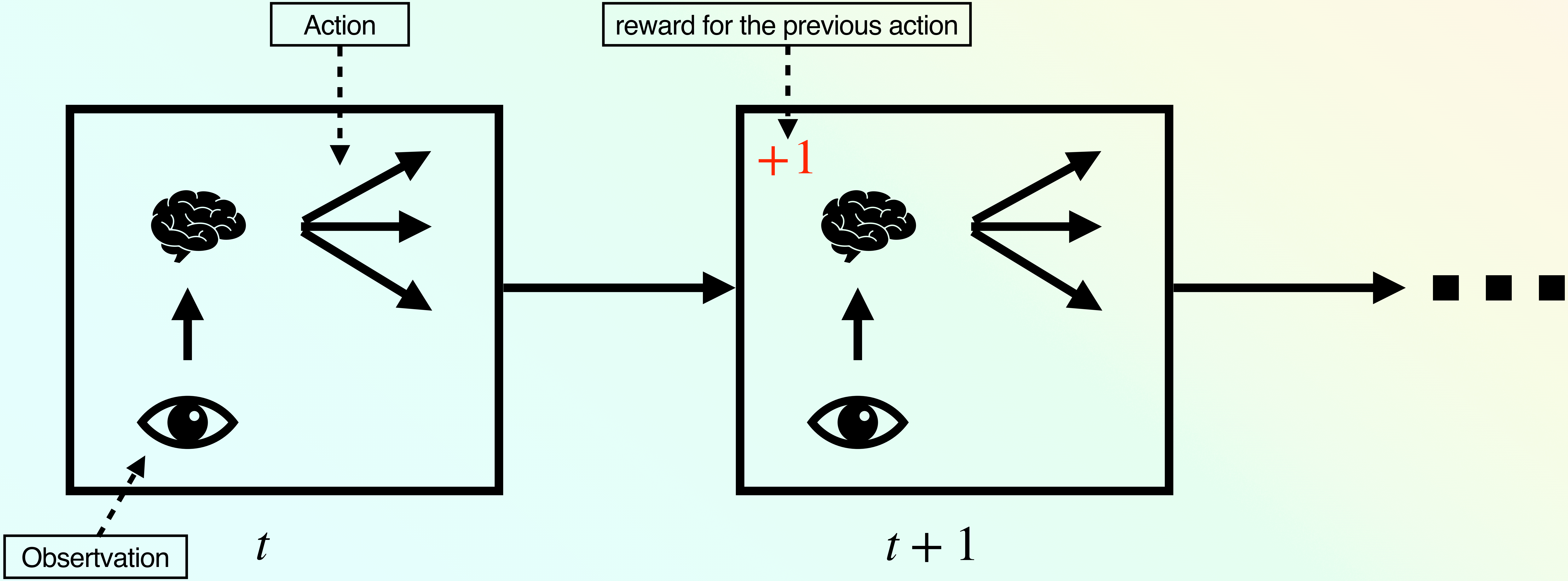
SEQUENTIAL DECISION MAKING

THE BEGING OF THE LONG JOURNEY



SEQUENTIAL DECISION MAKING

THE BEGING OF THE LONG JOURNEY



SEQUENTIAL DECISION MAKING

THE BEGING OF THE LONG JOURNEY

MARKOV DECISION PROCESSES (MDP)

DEFINITION

An MDP is a model of the agent's world and objective

An MDP is a tuple, e.g., $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p, d_0, \gamma)$

$\mathcal{S}$ set of all states — Information the agent uses to make a decision

$\mathcal{A}$ set of all actions — possible decisions

$\mathcal{R}$ set of all rewards — we use discrete notation, but it can be continuous

$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0,1]$ — a function that returns the probability of observing the next state and reward after taking an action in a particular state

$d_0 : \mathcal{S} \to [0,1]$ — function that returns the probability of starting in a given state

$\gamma \in [0,1]$ — discount factor to downweight rewards in the future

MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION

S_t the state at some time $t \in [0, \infty]$

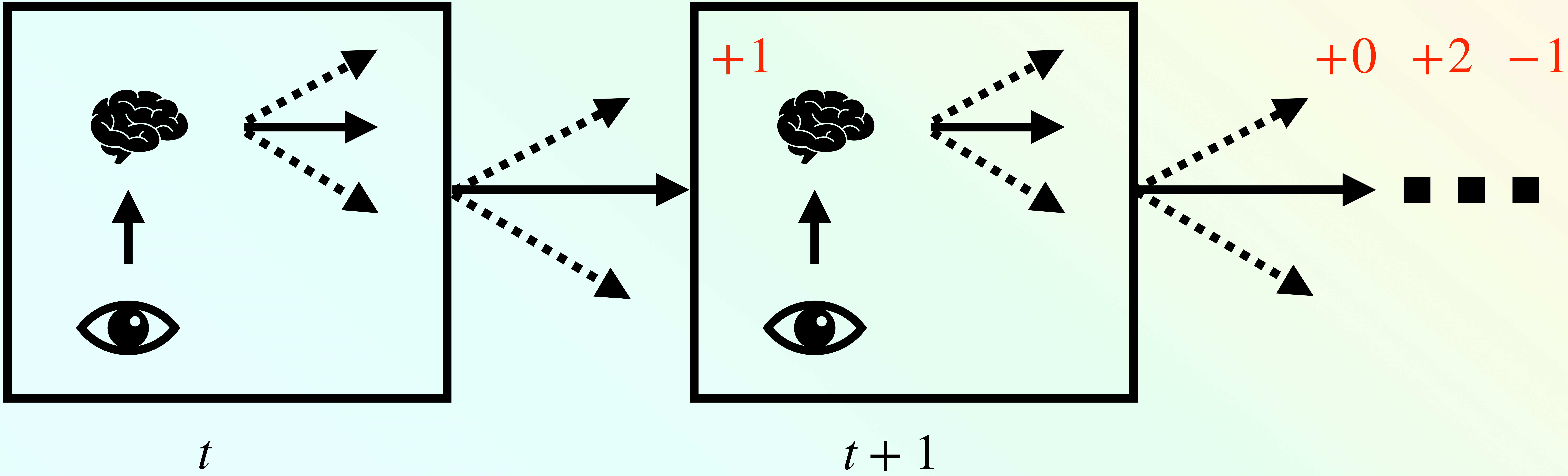
A_t the action the agent takes at time t

S_{t+1} the next state after S_t

R_{t+1} the reward received for taking action A_t in state S_t after transitioning to state S_{t+1}

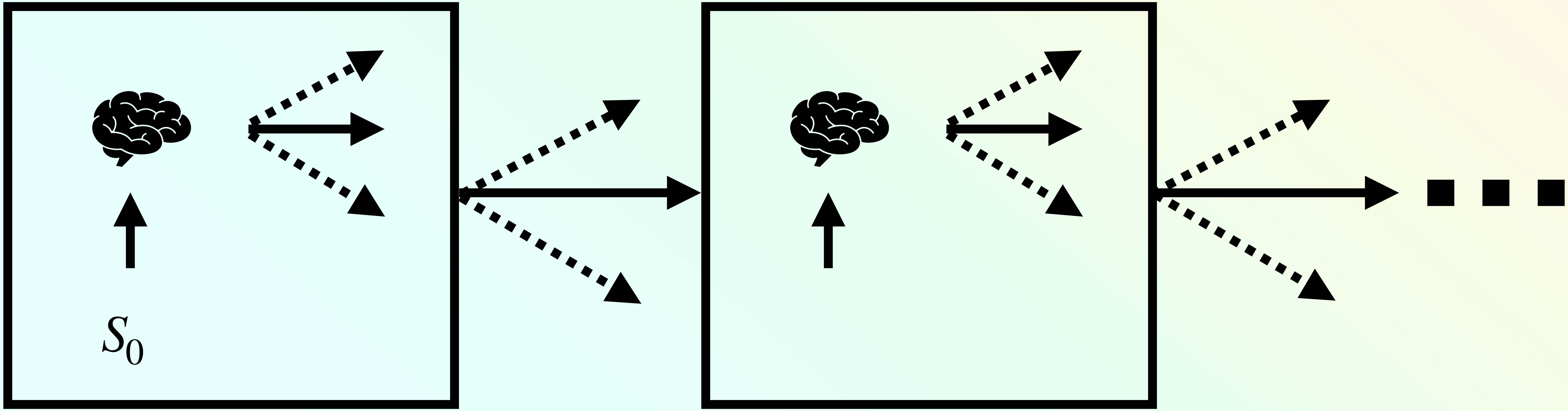
MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION



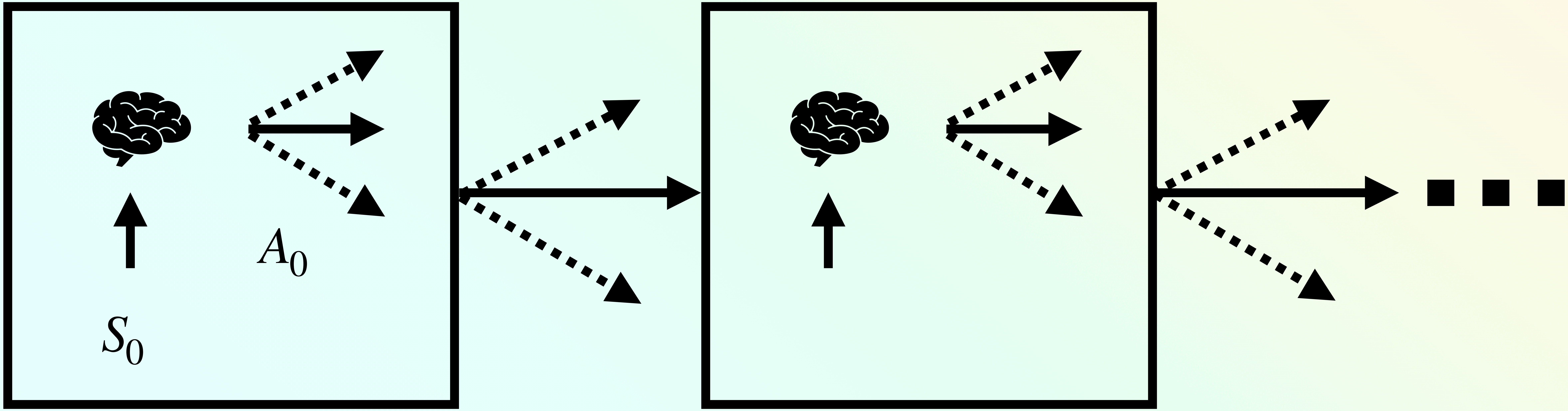
MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION



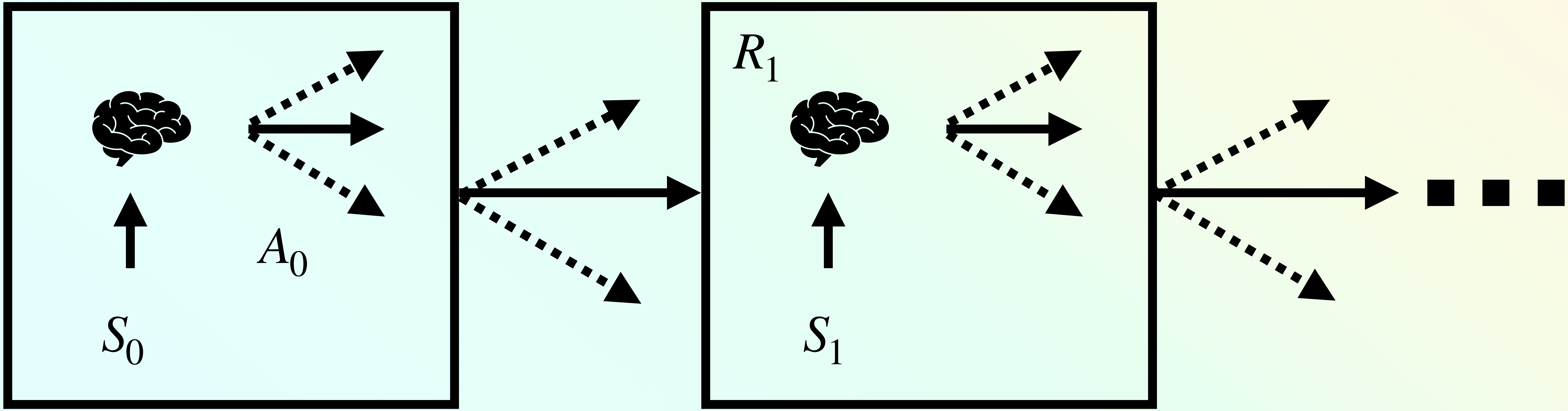
MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION



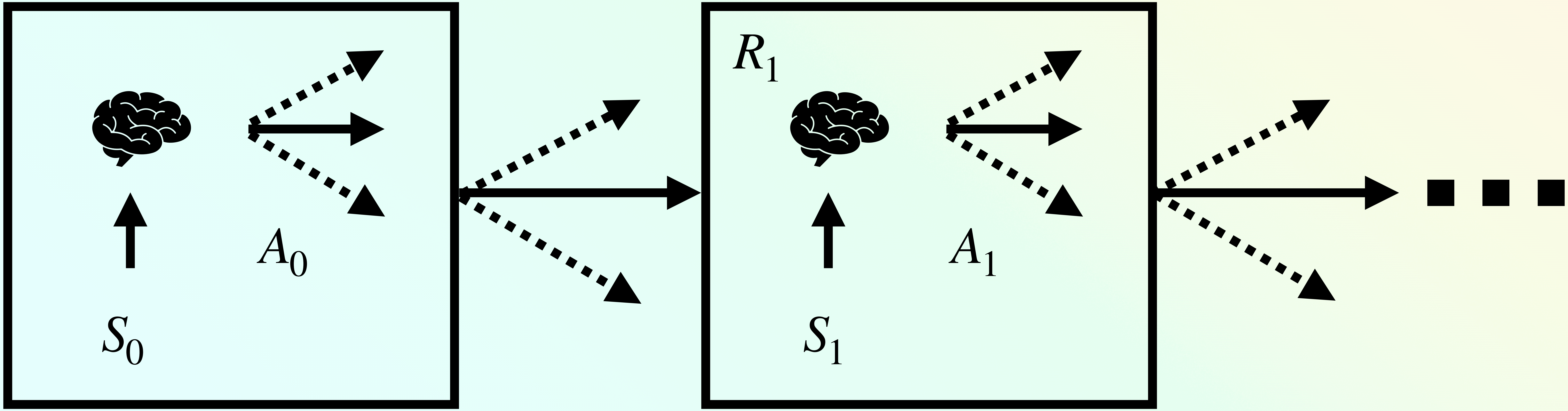
MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION



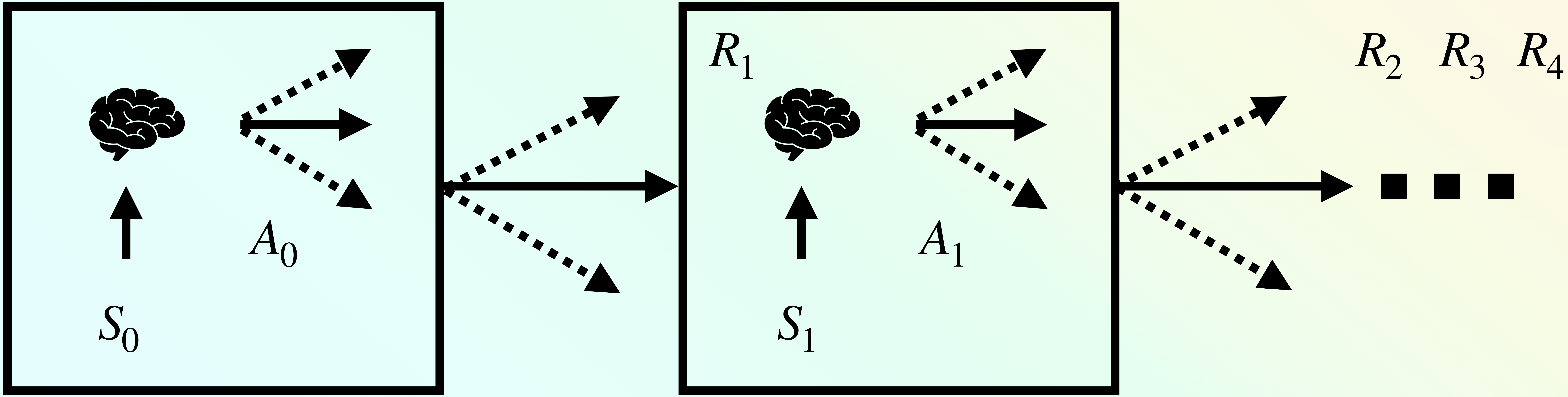
MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION



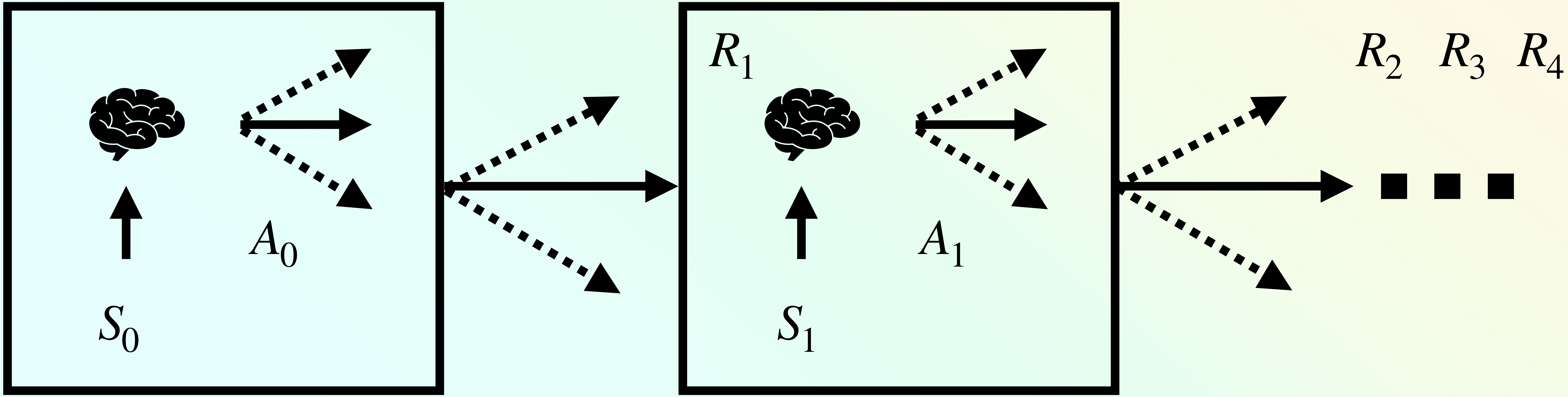
MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION



MARKOV DECISION PROCESSES (MDP)

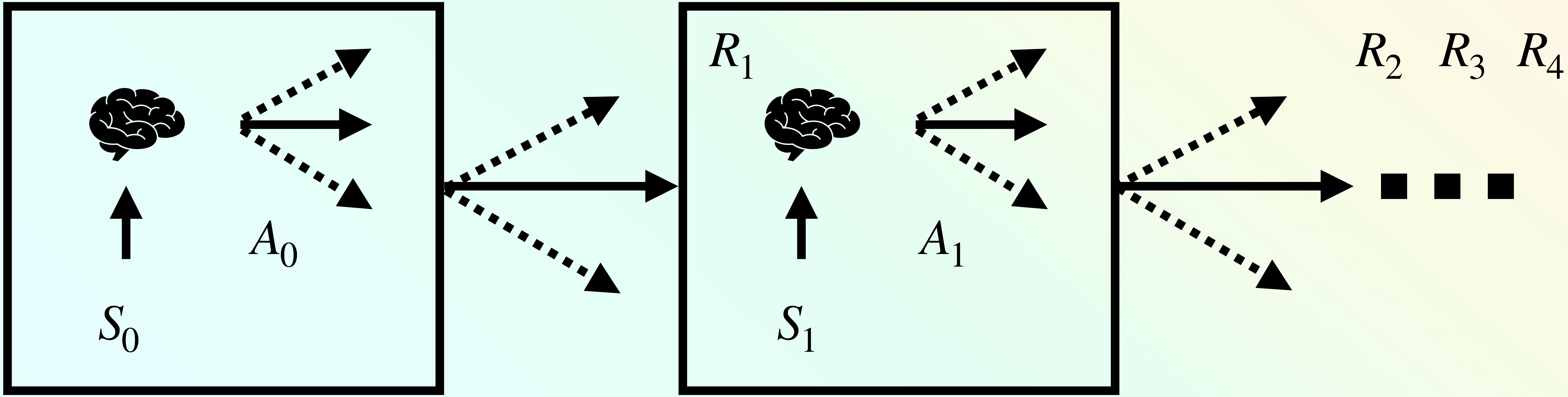
DEFINITION — INTERACTION



$$\Pr(S_0 = s) = d_0(s)$$

MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION



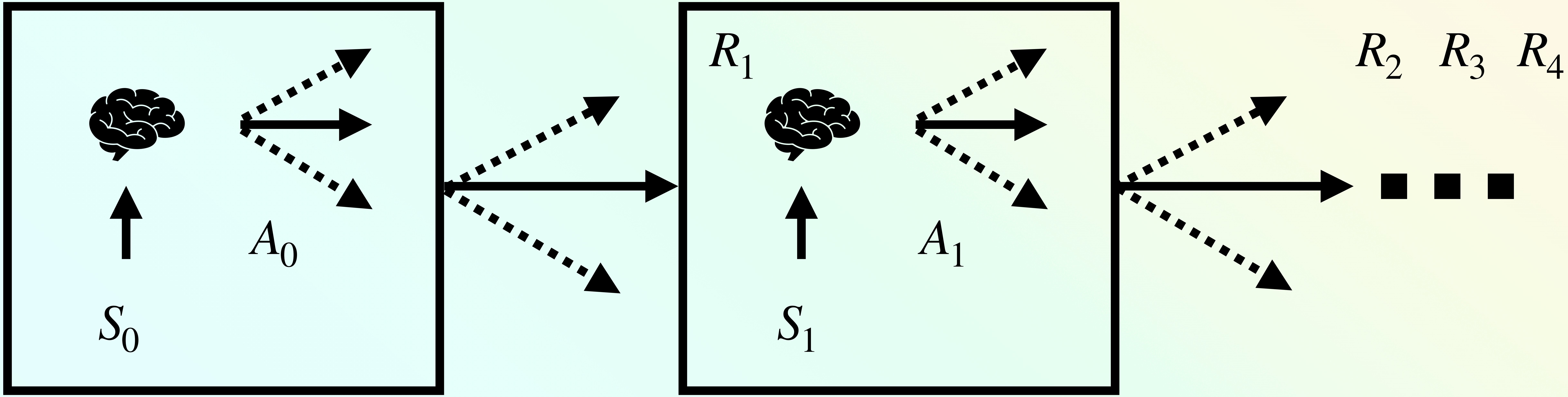
$$\Pr(S_0 = s) = d_0(s)$$

$$\Pr\left(S_1 = s, R_1 = r \,|\, S_0 = s, A_0 = a\right) = p\left(s', r \,|\, s, a\right)$$

MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION

$\Pr\left(A_0 = a \,|\, S_0 = s\right) = ?$ Covered in a future lecture



$\Pr(S_0 = s) = d_0(s)$

$\Pr\left(S_1 = s, R_1 = r \,|\, S_0 = s, A_0 = a\right) = p\left(s', r \,|\, s, a\right)$

MARKOV DECISION PROCESSES (MDP)

DEFINITION — INTERACTION

$$d_0(s) \doteq \Pr(S_0 = s)$$

$$p(s', r \mid s, a) \doteq \Pr(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1})$$

MARKOV DECISION PROCESSES (MDP)

USEFUL FUNCTIONS

Reason about the next state transition probability independent of the reward

$$p(s' \mid s, a) \doteq \Pr(S_t = s' \mid S_{t-1} = s, A_t = a)$$

MARKOV DECISION PROCESSES (MDP)

USEFUL FUNCTIONS

Reason about the next state transition probability independent of the reward

$$p(s' \mid s, a) \doteq \Pr(S_t = s' \mid S_{t-1} = s, A_t = a) \; = \sum_{r \in \mathscr{R}} p(s', r \mid s, a)$$

MARKOV DECISION PROCESSES (MDP)

USEFUL FUNCTIONS

Reason about the average reward

$$r(s, a, s') \doteq \mathbb{E}[R_t \,|\, S_{t-1} = s, A_{t-1} = a, S_t = s']$$

MARKOV DECISION PROCESSES (MDP)

USEFUL FUNCTIONS

Reason about the average reward

$$r(s, a, s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathscr{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

MARKOV DECISION PROCESSES (MDP)

USEFUL FUNCTIONS

Reason about the average reward

$$r(s, a, s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathcal{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

$$r(s, a) \doteq \mathbb{E}[R_t \mid S_t = s, A_t = a]$$

MARKOV DECISION PROCESSES (MDP)

USEFUL FUNCTIONS

Reason about the average reward

$$r(s, a, s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathscr{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

$$r(s, a) \doteq \mathbb{E}[R_t \mid S_t = s, A_t = a] = \sum_{r \in \mathscr{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)$$

MARKOV PROPERTY

DEFINITION

A state space $\mathcal{S}$ is Markov if $\forall t, s, a, s', r$

$$\Pr\left(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a, S_{t-2} = s_{t-2}, A_{t-2} = a_{t-2}, \ldots S_0 = s_0, A_0 = a_0\right) = \Pr\left(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\right)$$

S_t and R_t are conditionally independent of $S_{t'}, A_{t'}, R_{t'}$ for all $t' < t - 1$ given S_{t-1}

Memoryless: after knowing S_t, we can predict the future without considering what happened before observing S_t

MARKOV PROPERTY

EXAMPLE

$\mathcal{S} =$ set of all positions and momentums of particles and energy in the universe

p gives the transition probabilities as described by quantum theory

MARKOV PROPERTY

EXAMPLE

$\mathcal{S} =$ set of all positions and momentums of particles and energy in the universe

p gives the transition probabilities as described by quantum theory

- Valid for any MDP definition but not useful
- An MDP is a **model** for a problem

MARKOV PROPERTY

EXAMPLE

$\mathcal{S} =$ set of all positions and momentums of particles and energy in the universe

p gives the transition probabilities as described by quantum theory

- Valid for any MDP definition but not useful
- An MDP is a **model** for a problem
  - The probability that the projector correctly displays the slide on the screen
  - A meteor *could* destroy this building, but it doesn’t make sense to model this event

MARKOV PROPERTY

EXAMPLE

$\mathscr{S} =$ position and velocities of every joint on a robot in a lab

p defined by Newton equations, describe how those joints move

Is this Markov?

MARKOV PROPERTY

EXAMPLE

$\mathcal{S} =$ position and velocities of every joint on a robot in a lab

p defined by Newton equations, describe how those joints move

Is this Markov?

No

Other objects, battery level, etc

MARKOV PROPERTY

EXAMPLE

$\mathcal{S} =$ position and velocities of every joint on a robot in a **free space simulation**

p defined by Newton equations, describe how those joints move

Is this Markov?

Yes

MARKOV PROPERTY

EXAMPLE

$\mathcal{S} =$ position and velocities of every joint, battery level, camera reading for a robot in a lab

p defined by:

- Newton equations describe how those joints move
- How battery level changes as the robot moves around
- Need to model how pixels change in the camera — Can't see the whole world or even motion
  - transition dynamics would be specific to a point in time

Is this Markov?

No

MARKOV PROPERTY

EXAMPLE

Real-world problems often need a state space of the universe or many unobservable quantities

- Other modeling paradigms: nonstationary MDPs, partially observable Markov decision processes

MDPs are useful for understanding the basics of decision-making and modeling some problems

MARKOV PROPERTY

EXAMPLE — CHESS

$\mathcal{S}$ — all possible board configurations

$d_0(\text{Standard setup}) = 1.0$ if the agent is the white player

$\mathcal{A}(s)$ — all possible moves available to the agent in the current state

p — ?

MARKOV PROPERTY

EXAMPLE — CHESS

$\mathcal{S}$ — all possible board configurations

$d_0(\text{Standard setup}) = 1.0$ if the agent is the white player

$\mathcal{A}(s)$ — all possible moves available to the agent in the current state

p — probability distribution over board states after both the agent and other player make a move

Is this Markov?

MARKOV PROPERTY

EXAMPLE — CHESS

$\mathcal{S}$ — all possible board configurations

$d_0(\text{Standard setup}) = 1.0$ if the agent is the white player

$\mathcal{A}(s)$ — all possible moves available to the agent in the current state

p — probability distribution over board states after both the agent and other player make a move

Is this Markov?

Maybe

MARKOV PROPERTY

EXAMPLE — CHESS

$\mathcal{S}$ — all possible board configurations

$d_0(\text{Standard setup}) = 1.0$ if the agent is the white player

$\mathcal{A}(s)$ — all possible moves available to the agent in the current state

p — probability distribution over board states after both the agent and other player make a move

If the opponent is **human**, is this Markov?

MARKOV PROPERTY

EXAMPLE — CHESS

$\mathcal{S}$ — all possible board configurations

$d_0(\text{Standard setup}) = 1.0$ if the agent is the white player

$\mathcal{A}(s)$ — all possible moves available to the agent in the current state

p — probability distribution over board states after both the agent and other player make a move

If the opponent is **human**, is this Markov?

Probably not; humans get tired, reason about their past to inform their decisions

MARKOV PROPERTY

EXAMPLE — CHESS

$\mathscr{S}$ — all possible board configurations

$d_0(\text{Standard setup}) = 1.0$ if the agent is the white player

$\mathscr{A}(s)$ — all possible moves available to the agent in the current state

p — probability distribution over board states after both the agent and other player make a move

If the opponent is **random**, is this Markov?

MARKOV PROPERTY

EXAMPLE — CHESS

$\mathcal{S}$ — all possible board configurations

$d_0(\text{Standard setup}) = 1.0$ if the agent is the white player

$\mathcal{A}(s)$ — all possible moves available to the agent in the current state

p — probability distribution over board states after both the agent and other player make a move

If the opponent is **random**, is this Markov?

Yes

MARKOV PROPERTY

EXAMPLE — CHESS

$\mathcal{S}$ — all possible board configurations

$d_0(\text{Standard setup}) = 1.0$ if the agent is the white player

$\mathcal{A}(s)$ — all possible moves available to the agent in the current state

p — probability distribution over board states after both the agent and other player make a move

If the opponent is a **planning agent**, is this Markov?

MARKOV PROPERTY

EXAMPLE — CHESS

$\mathcal{S}$ — all possible board configurations

$d_0(\text{Standard setup}) = 1.0$ if the agent is the white player

$\mathcal{A}(s)$ — all possible moves available to the agent in the current state

p — probability distribution over board states after both the agent and other player make a move

If the opponent is a **planning agent**, is this Markov?

Yes

MARKOV PROPERTY

SUMMARY

$\mathcal{S}$ — A state needs to contain all information (except for the action) to predict the next state distribution

p — cannot change with time

p — We do not need to know a mathematical form for it, only that it exists

TERMINATION

DEFINITION

$S_0, A_0, R_1, S_1, \ldots$ can go on forever or

$S_0, A_0, R_1, S_1, \ldots S_{T-1}$ ends after $T < \infty$ time steps — Called an episode

- After an episode time starts over
- T — if constant all episodes have the same number of time step
  - Planning your schedule for 1 day (assuming you don't die)
- T — can be variable
  - Planning your route home. Some paths will require making more decisions and take different amounts of time

The episode must be guaranteed to end in a finite time.

- One way: an episode that terminates on or before some maximum time limit L, i.e., $\Pr(T \leq L) = 1$

TERMINATION

DEFINITION— INFINITE LENGTH BUT EPISODIC

Add a special state s_∞ to $\mathcal{S}$

$$\Pr(S_t = s_\infty, R_t = 0 \,|\, S_{t-1} = s_\infty) = 1.0$$

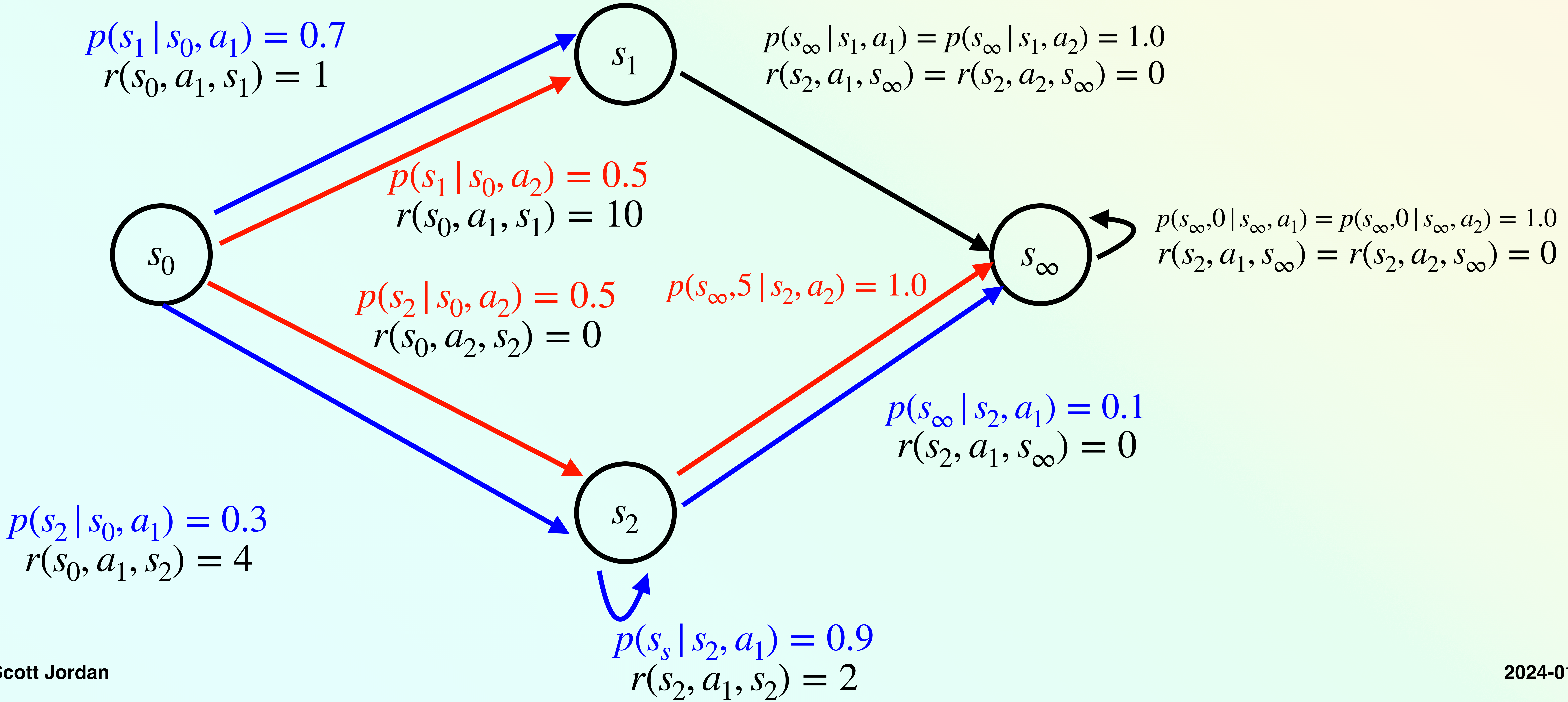
- Agent always stays in s_∞ after entering s_∞ and always receives a reward of 0 when in s_∞
- s_∞ is called a *terminal absorbing state*

A state $s \in \mathcal{S}$ is called *terminal* if $\Pr(S_{t+1} = s_\infty \,|\, S_t = s) = 1$

- S_{T-1} is a terminal state in episodic problems
- $\forall t \geq T,\ S_t = s_\infty$ in episodic problems

MDP

EXAMPLE GRAPH



MDP FORMULATIONS

EXAMPLE BANDIT

Consider a bandit problem with actions $\mathscr{A} = \{a_1, a_2, a_3\}$ and rewards $\mathscr{R} = \{1,2,3\}$. Assume we know $p(r|a) = \Pr\left(R_{t+1} = r | A_t = a\right)$. How can we model this as an MDP?

$\mathcal{S} =$?

$\mathscr{A} = ?$

$p(s', r | s, a) =$?

$d_0(s) =$?

MDP FORMULATIONS

EXAMPLE BANDIT

Consider a bandit problem with actions $\mathscr{A} = \{a_1, a_2, a_3\}$ and rewards $\mathscr{R} = \{1,2,3\}$. Assume we know $p(r \mid a) = \Pr\left(R_{t+1} = r \mid A_t = a\right)$. How can we model this as an MDP?

$\mathcal{S} = \{s_0, s_\infty\}$

1-step MDP

$\mathscr{A} = \{1,2,3\}$

$p(s_\infty, r \mid s_0, a) = \underbrace{p(s_0, a, s_\infty)}_{1} p(r \mid a) = p(r \mid a)$ and $\forall a, r,\ p(s_0, r \mid s_0, a) = 0.0$

$d_0(s_0) = 1$

MDP FORMULATIONS

EXAMPLE BANDIT

Consider a bandit problem with actions $\mathscr{A} = \{a_1, a_2, a_3\}$ and rewards $\mathscr{R} = \{1,2,3\}$. Assume we know $p(r|a) = \Pr\left(R_{t+1} = r | A_t = a\right)$. How can we model this as an MDP?

$\mathcal{S} = \{s_0\}$

1-state infinite horizon MDP

$\mathscr{A} = \{1,2,3\}$

$$p(s_0, r | s_0, a) = \underbrace{p(s_0, a, s_0)}_{1} p(r|a) = p(r|a)$$

$d_0(s_0) = 1$

$\gamma < 1$ It cannot have an infinite sum of rewards. Discuss this next class

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s) = d_0(s)$$

$$\Pr(S_1 = s_1, R_1 = r_1 \mid S_0 = s_0, A_0 = a_0) = p(s_1, r_1 \mid s_0, a_0)$$

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1) = ?$$

MDP REASONING

PROBABILITIES

Need to be able to reason about sequences of states, actions, and rewards

$$\Pr(S_0 = s) = d_0(s)$$

$$\Pr(S_1 = s_1, R_1 = r_1 \mid S_0 = s_0, A_0 = a_0) = p(s_1, r_1 \mid s_0, a_0)$$

$$\begin{aligned}
\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1) &= \Pr(R_1 = r_1, S_1 = s_1 \mid S_0 = s_0, A_0 = a_0) \Pr(S_0 = s_0, A_0 = a_0) \\
&= \Pr(R_1 = r_1, S_1 = s_1 \mid S_0 = s_0, A_0 = a_0) \Pr(A_0 = a_0 \mid S_0 = s_0) \Pr(S_0 = s_0) \\
&= p(s_1, r_1 \mid s_0, a_0) \Pr(A_0 = a_0 \mid S_0 = s_0) d_0(s_0)
\end{aligned}$$

$\Pr(A_t = a \mid S_t = s) = \,?$ — Defined by the agent's *policy.* We will discuss this in the future.

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, \ldots, R_t = r_t, S_t = s_t) = \text{ ?}$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, \ldots, R_t = r_t, S_t = s_t)$$

$$= \Pr(R_t = r_t, S_t = s_t \,|\, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= \Pr(R_t = r_t, S_t = s_t \,|\, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= p(s_t, r_t \,|\, s_{t-1}, a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

- Applied Markov property

$$= p(s_t, r_t \,|\, s_{t-1}, a_{t-1}) \Pr(A_{t-1} = a_{t-1} \,|\, S_{t-1} = s_{t-1}) \Pr(S_{t-1} = s_{t-1}, R_{t-1} = r_{t-1}, s_{t-2}, a_{t-2}, \ldots, S_0 = s_0)$$

$$= p(s_t, r_t \,|\, s_{t-1}, a_{t-1}) \Pr(A_{t-1} = a_{t-1} \,|\, S_{t-1} = s_{t-1}) p(s_{t-1}, r_{t-1} \,|\, s_{t-2}, a_{t-2}) \Pr(S_{t-2} = s_{t-2}, A_{t-2} = a_{t-2}, \ldots, S_0 = s_0)$$

$$= d_0(s_0) \prod_{k=1}^{t} p(s_k, r_k \,|\, s_{k-1}, a_{k-1}) \Pr(A_{k-1} = a_{k-1} \,|\, S_{k-1} = s_{k-1})$$

NEXT CLASS

WHAT YOU SHOULD DO

1. Quiz and programming assignment are due tonight
2. Read the book and watch the videos if you have not already

Monday: reward functions and objectives for MDPs